Consistency in Performance in the Group Oral Discussion Test:

An Interactional Competence Perspective

By

David A. Leaper

B.A., University of Auckland, New Zealand, 1993

M. App. Ling. (TESOL), Macquarie University, Australia 1999

Submitted August 29, 2014 to the Department of Linguistics, Faculty of Human Sciences, Macquarie University.

This thesis is presented for the degree of Doctor of Philosophy in Applied Linguistics.

To my family:
Sohyun, Celine, Liam
- for you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Language teaching institutions may require a dependable test of conversational ability to assess student development and encourage communication in the curriculum. The group oral test (GOT) is a format in which three or four test-takers discuss a prompt for up to ten minutes that has the potential to do this. Using Interactional Competence (IC) as its foundation and a mixed-methods research (MMR) framework, this dissertation traces 53 Japanese university students' performance over three administrations in two years to investigate the test's appropriacy as a tool of assessment for their conversational ability. The quantitative phase of the study measured the participants' performance using indices of complexity, accuracy, fluency, vocabulary and interactive functions and found varying developmental patterns: most indices showed significant improvement only in the second administration, while others showed gains in just the third administration, and some improved in each successive administration. Overall, gains made in the second administration had larger effect sizes than in the third administration. However, the scores awarded showed that students only improved significantly in the second administration. The qualitative phase of the study adapted Young's Interactional Competence (IC) framework (2009, 2011) to investigate a subsample of eight test-takers to investigate the extent that students' performances were represented by the quantitative indices and their scores. The qualitative analysis illuminated the consistent and variable elements in the students' performances, and revealed the difficulties the raters had of scoring individuals independently of their IC displayed relative to the group's performance: higher level test-takers could be scored over-generously for being the best in their group, and lower level students rewarded for being tested with supportive group members. Among the implications for the GOT are that students of similar abilities should be grouped together, rating scales need to reflect the higher order conversational skills identified by this study, and the raters need to be aware of situations in which their interpretation of the rating scales may be compromised.

STATEMENT OF CANDIDATE

I certify that the research described in this thesis entitled “Consistency in Performance in the Group Oral Discussion Test over Time: An Interactional Competence Perspective” has not already been submitted for any other degree nor has it been submitted as part of the requirements for a degree to any other institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that to the best of my knowledge all sources and literature used are indicated in the thesis.

The research presented in this thesis was approved by the Macquarie University Ethics Review Committee, reference number HE23SEP2005-M04311 on 23rd Sept 2005, amended 2007.

David Leaper

David A. Leaper (31340482)

August 29, 2014

ACKNOWLEDGEMENTS

Without a supportive home life it simply would not have been possible to complete this dissertation. The first to thank is my wife, Sohyun, for putting up with with my obsessive drive to "get it finished". I also have to thank the little ones, Celine and Liam, though they don't understand why Dad has to spend so much time in front of the computer. Thanks for inspiring me to persevere with this work. Also, thanks to my parents who supported my efforts from afar.

I need to acknowledge the guidance and support of my supervisor, Dr. Mehdi Riazi. Thanks for your patience, understanding and knowledge. I am deeply appreciative of your encouragement and considerate advice, not only for the decisive final years of the dissertation, but also for co-authoring the *Language Testing* article. I also would like to express my gratitude to the professors at Macquarie who counselled me along the way. Thanks to Dr. Geoff Brindley for taking me on in the first place, and helping me through the early stages. Dr. Stephen Moore guided me through the middle stages, in particular a discussion I had with Dr. Moore a few years ago helped to sustain my motivation.

Amongst current and past colleagues, I would like to acknowledge the help of James Brawn, Kyle Devlin, Dr. Shaun Manning and Jeff Durand for their help with various tasks and questions that I threw their way. Gavin Farrell's dedication in proof-reading this work was extraordinary – any remaining mistakes are mine. I also employed people to help at various times, and I would like to acknowledge the role they played in the finalization of this dissertation: Shawnna Barrett, Austin Brick, Jinyoung Lim, Hyemin Lee, Ji Young Lim, Jiyoung Seo, Soohyun Kwan and Sangmi Park. Finally I would like to thank the students of Kanda University of International Studies whose utterances in the speaking tests comprise the data of this study.

Other distinguished scholars helped by answering questions I emailed to them, and I would like to thank them for their illuminating answers about various aspects of their work. They took time out of their busy schedules to answer questions from somebody who most of them had never met before, and some exchanges went to a considerable length. Here I acknowledge Lindsay Brooks, Dr. Tom Cobb, Dr. Glenn Fulcher, Dr. Zhengdong Gan, Dr. Philip MacCarthy, Dr. Fumiyo Nakatsuhara, Dr. Gary Ockey and Dr. Michael O'Donnell. Finally, I must acknowledge my three external reviewers whose comments, advice and suggestions helped make this a stronger dissertation.

To all of the above, thank you. This dissertation could not have been attempted or completed without you.

CHAPTER ONE

Introduction to the Study

1.0 Introduction

This chapter is a brief introduction to the background and objectives of this study. It begins by outlining the basic issue for this dissertation in the statement of the problem, before going on to describe the context in which it takes place. It then provides a conceptual framework, before explaining the objectives of the study and the research questions it is aiming to investigate, and finishes with an outline of the remaining chapters of this dissertation.

1.1 Statement of the problem

Over the last two decades, in the field of the direct assessment of spoken language there has been a shift away from tests in which an examiner interviews a test-taker towards those in which the test-takers' speaking ability is assessed as they interact with each other. Previously, the dominant form of oral test was one in which the test-taker's interlocutor was an interviewer who asked questions and set tasks to which the test-takers were obligated to respond. Research into this format of oral test, particularly in its most influential form, the Oral Proficiency Interview (OPI) (ACTFL, 1986, 1999), made it clear that not only were there problems with the OPI itself, but it was the status and authority inherent in the role of 'interviewer' that was skewing the language elicited by this method (see Bachman 1988; Bachman & Savignon, 1986; Johnson, 2001; Johnson & Tyler, 1998; Lazaraton, 1992; Van Lier, 1989). Given the rise of communicative language teaching at about the time of these criticisms of the OPI and the growing recognition of the importance of interaction in conversation in language learning (Mackey, 2012, Van Lier, 1996), more interactive methods of assessing speaking were sought. If it is the presence of the interviewer that exerts an unavoidable influence over the language elicited in a speaking test, then one solution would be to remove the interviewer altogether and have the test-takers interact with each other. The group oral test (GOT), in which three or four test-takers discuss a prompt in a group discussion, is one such format which emerged from this context.

To date research has been conducted on various aspects of test-taker performance in the group oral discussion. Among the positive findings for this format of test are that it has been found to be efficient for testing on a large scale (Bonk & Ockey, 2003, Folland & Robertson, 1976; Hilsdon, 1991), capable of eliciting a wide range of more 'natural' or conversational language (Fulcher, 1996a, Gan, 2010; Gan, Davidson & Hamp-Lyons 2006), not as nerve-inducing for test-takers (Fulcher, 1996a; He & Dai, 2006) and capable of providing positive washback on teaching and learning (Folland & Robertson, 1976; Fulcher, 1996a, Van Moere, 2006). On the negative side, threats to the validity of this form of testing have also been found. The personality traits of shyness and assertiveness of the participants can affect their scores (Bonk & Van Moere, 2004, Ochey, 2009, 2011), the number of participants in a group can exacerbate the influence of extraverts on the interaction (Nakatsuhara, 2011), participants may opt not to contribute (Nakatsuhara, 2011), the prompt may affect the discourse and language elicited (Leaper & Riazi, 2014), high scoring reliability may be difficult to achieve (Van Moere, 2006), and there are concerns over the range of language functions it elicits (He & Dai, 2006; Van Moere, 2007).

To summarize the literature on the GOT, it is fair to say that although there are positive reasons to favour this format for assessing oral ability, there are concerns about its validity. The range and extent of the threats to the GOT's validity raise doubts about its utility as a method of testing at an institutional level, despite its potential to elicit authentic conversation-like interactions. An alternative perspective on the GOT may yet give test-makers reason to value this format. Instead of examining the GOT for yet more factors that affect it negatively, the question may be turned around to ask "Just what does this test measure?" The need to identify not merely the variable but also the stable elements of performance in interactive oral assessment tasks has been pointed out in the literature, particularly related to the notion of interactional competence (Kramsch, 1986; Young, 2000, 2009, 2011, 2013) and the need to explain "situated language use" of "abilities – in language users – in context" (Chalhoub-Deville, 2003, p. 378).

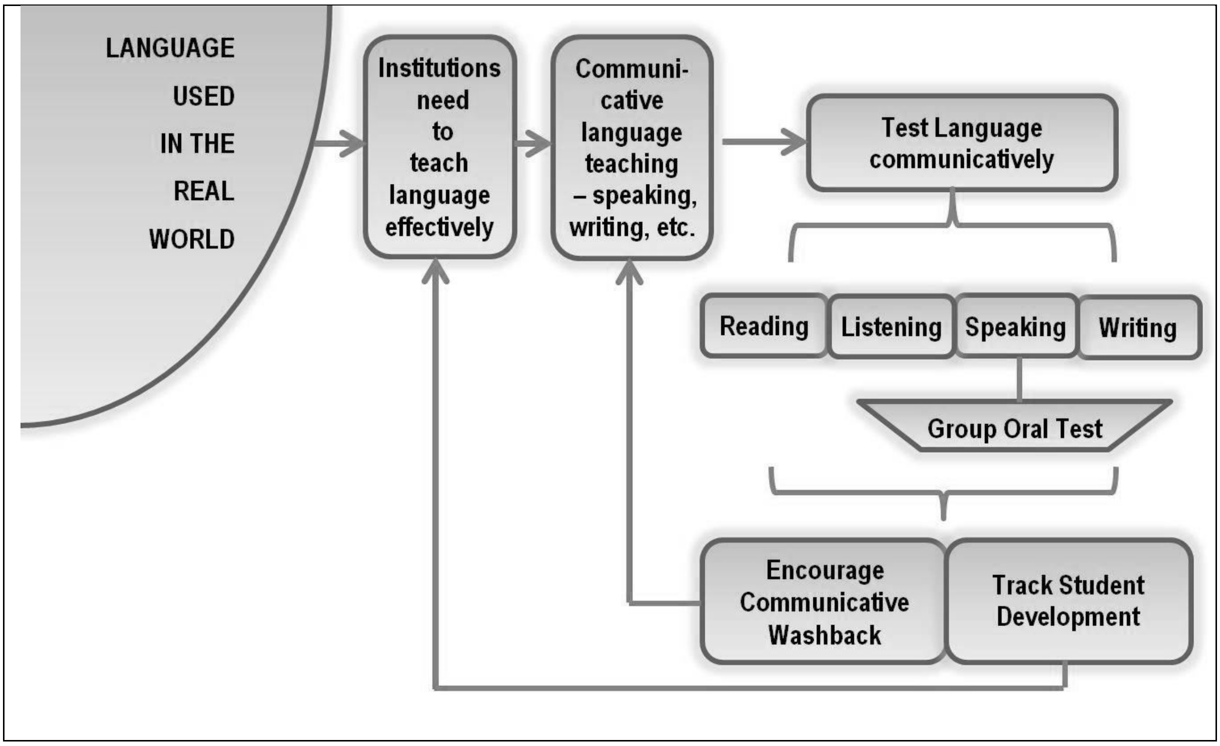
To answer this question, this study will use as its primary data videos taken of the same candidates from three GOTs administered over two years to investigate the consistency with which the test-takers perform. Research using longitudinal data of this nature is something that no other research on the GOT has attempted to do, to the author's knowledge. If this investigation reveals that the test is capable of consistently eliciting a set of core language abilities, then this would allow administrators more certainty about what this format can be used for. On the other hand, if little or no consistency of performance is found to exist, then it can be considered that the negative factors raised by critics of this format are dominant and the GOT, as implemented in the context of this study, is less likely to be viewed as a viable method of assessment for medium to high stakes tests.

1.2 Context of the study

The GOT data is collected from the speaking section of a four skills language test administered at a private languages university in Japan. This university emphasizes the ability to communicate with the world as a primary mission, (Our Philosophy, n.d.) and its curriculum aims to achieve this through a curriculum that focuses on fostering communicative skills. The test as a whole was designed to relate closely to this curriculum, which is compulsory for first and second year English major students, and is thus expected to have a positive washback effect on the teaching and learning at the university (Van Moere & Johnson, 2002). As well as the need for positive washback, the scores of the test are used for two practical purposes: the placement of students into classes at the beginning of their first and second years, and as a final proficiency test in which the sections of the test contribute 20% towards a student's grade in certain subjects. Beyond this the university administration uses it to monitor student language development and for research purposes. Research conducted by Bonk and Ockey (2004), Ockey (2006, 2009, 2011), Leaper and Riazi (2014) and Van Moere (2006, 2007) are notable examples of the work that has gone into the GOT at this institution. The role of the GOT in the administration is summarised in Figure 1.

At this private languages university, the entire body of first and second year as well as many third and fourth year students take the test every year, bringing the total number of participants to well over 2000 for each administration. The students take the first test before classes begin as incoming students, then at end of each academic year. They sit the reading, listening, grammar and writing sections in the morning, and the GOT in the afternoon. For the group discussion, out of consideration of fairness, students are assigned to groups at random, subject to two conditions: they are not assigned group members who were their classmates and they are assigned raters who were not their teachers.

Figure 1: The role of the group oral test at an institutional level


LANGUAGE USED IN THE REAL WORLD
Institutions need to teach language effectively
Communi-cative language teaching – speaking, writing, etc.
Test Language communicatively
Reading
Listening
Speaking
Writing
Group Oral Test
Encourage Communicative Washback
Track Student Development

For security reasons, the prompts created for the administration are confidential, and one is selected at random by the raters immediately before each test. The test-takers are given one minute to read the prompt and consider their response before they start talking, and the raters do not intervene until they have collected a rateable sample, up to ten minutes later. Some students are placed into rooms equipped with video cameras where they are recorded for training and research purposes. For this study, the same participants were placed in these video rooms the three times they did the test in their first two years of study, and these videos along with the test-takers' scores comprise the source data for this study.

For each administration three or four prompts are created by the test committee. Each GOT is rated by two of the native English speaking M.A. qualified teaching staff, whose scores are adjusted by Rasch before being reported to the students. The raters are not specialized language testing staff; scoring this test is one of their duties as laid down in their contract. They are trained in one 2-3 hour session, which is run by the testing committee of the university, which is composed of teachers with an interest in assessment.

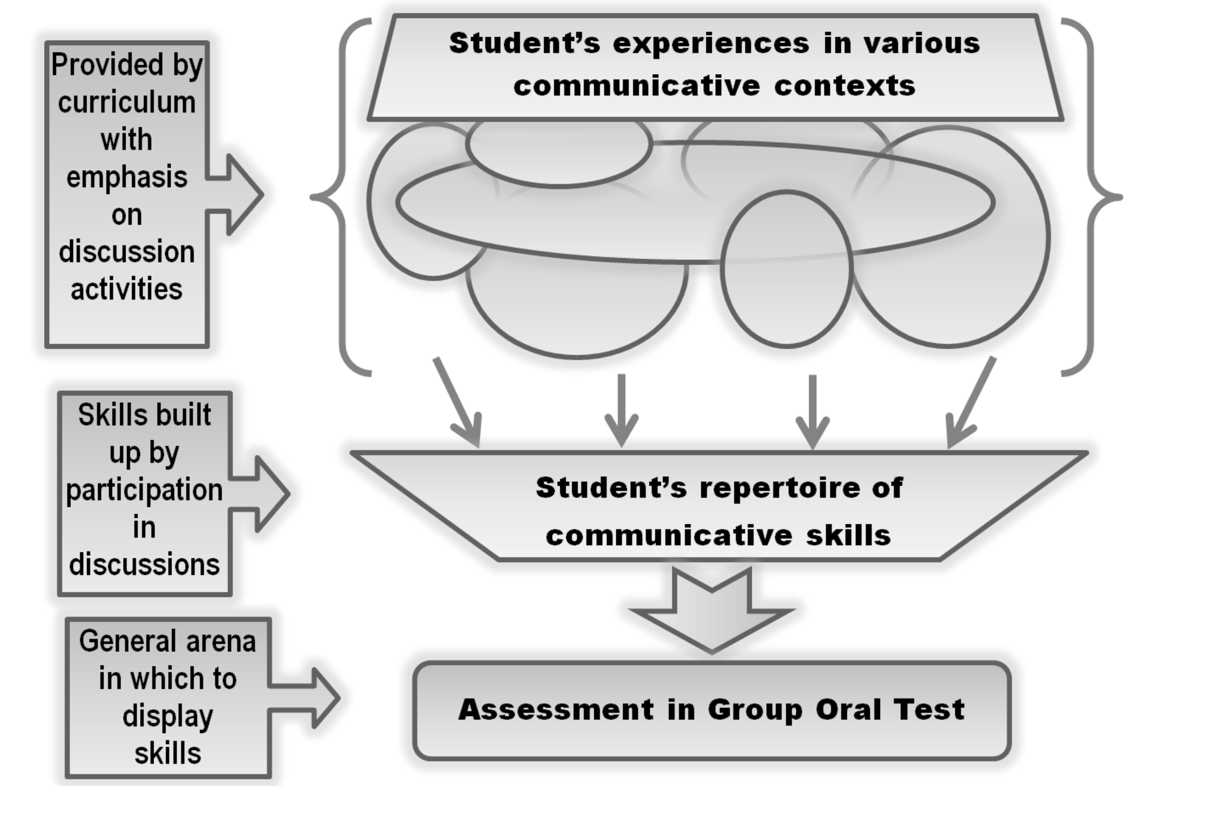
The students who take this test are a relatively uniform group in terms of their background in studying English. They were almost all 18 years old at the time of the first test and had all studied English for at least six years at junior and senior high schools. Many of them had not focussed on studying or practicing spoken English at high school since the exams they take to graduate from high school only assess reading or listening by means of multiple-choice questions. Many of them had never previously been exposed to English to such an extent as they were at this university. Over their first two years of study at this university, they attend four 14 week semesters in which they are exposed to 15 hours per week of English classes taught by the native speaking staff of the university, making about 840 hours of English over the two years that is the span of this dissertation. Additionally, they are encouraged to use the university's Ministry of Education award winning Self-Access Learning Centre, where teaching staff are assigned to be available to talk to students on any topic and they can study at their own volition.

1.3 Conceptual framework

In language testing, a primary consideration is that the assessment should be consonant with its purpose (Hughes, 2003, p. 3). This is especially important for language education institutions with large scale testing needs in which considerable resources must be expended to achieve a workable solution for assessing students' language ability and learning progress. As can be seen in Figure 1 above, an institution may have a need for a test to track its students' language development and encourage positive washback in accordance to its goals; in this case that communicative language is being taught and learned. For the administration, a test that shows the scores of its students as a rising slope would be ideal, since it could be taken to indicate that its students' abilities are improving and that their program is working efficiently. The underlying assumptions to this reflect the cognitive perspective which holds that the test score represents a trait that is present in the language learner: for *trait theorists* the score indicates the test-takers ability (a trait) that is relatively stable over time, (Messick, 1988, p. 15) and could be expected to improve with education. Moreover, since the score represents a generalizable underlying ability, the assumption is that it has meaning in other contexts. Although a convenient perspective to take, shortcomings are revealed when confronted with the nature of communication, which has been found to be a collaboration of the participants, who co-create it according to the shared understanding of the context in which it occurs (Hall, 1993). Speaking conversationally is inherently unpredictable in that what is said and how it should be said depends on such various factors as the participants, the situation and the subject they are talking about, among other factors. In this situation, a *behaviourist* would interpret the score as merely reflecting what that test-taker happened to do on a particular occasion, and is only generalizable when the context is duplicated (Chapelle, 1998). While the GOT as a test may replicate the unpredictability of conversation and thus fulfil its role of generating positive washback by encouraging the teaching and learning of language as communication, it creates a tension with the needs of a language education institution for a test that can consistently track the development of the individual students' communicative ability over time.

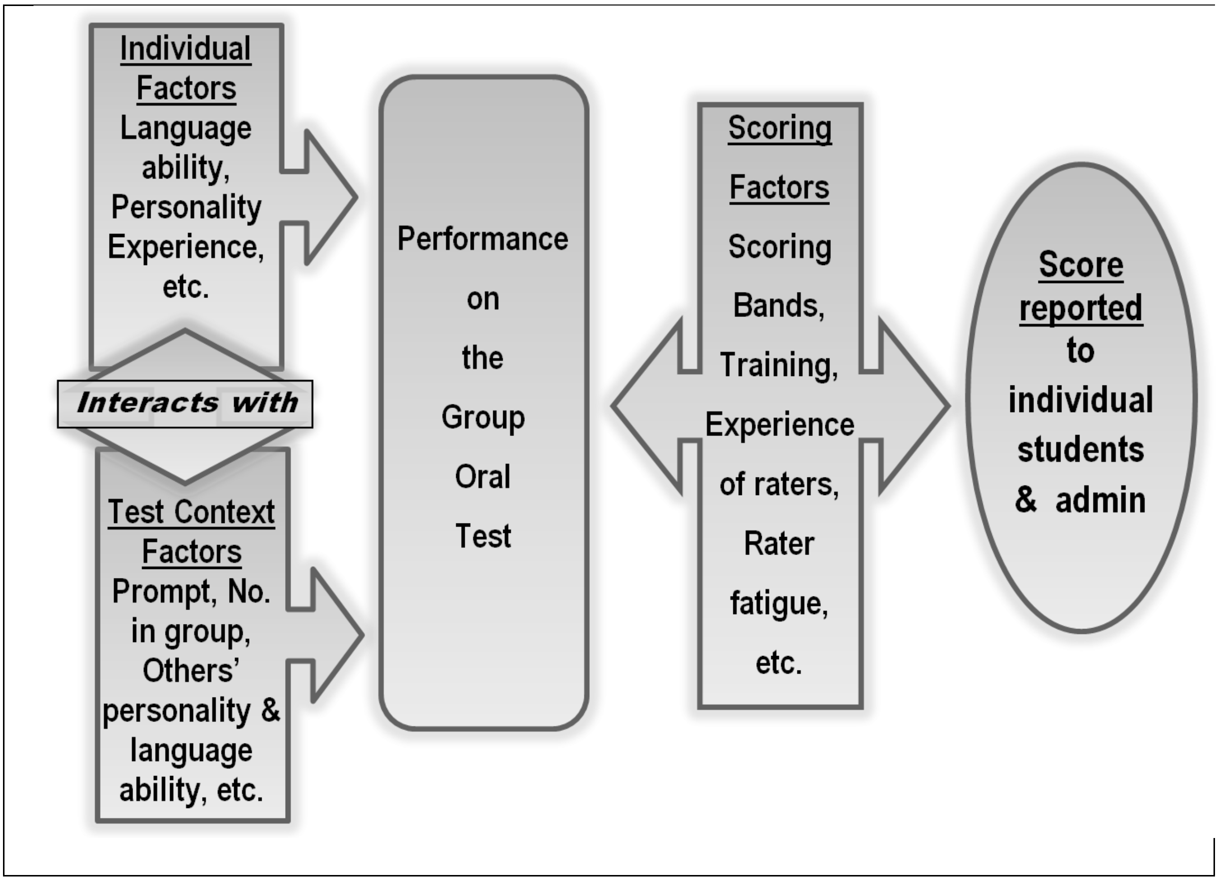
The positions of the trait and behaviourist theorists are extremes; the notion of *Interactional Competence* (IC) finds a middle way between them. By describing a spoken assessment as an interaction between the participants and the context, this perspective acknowledges the value of describing not only the trait and the context of the assessment event, but also the relationship between them (Young, 2000). Subscribing to an IC framework for assessment brings challenges that need to be explicated. Among these is claim that there is no general language proficiency, only local competency (Young, 2011). Although this may not seem an advance on the behaviourists' stance, generalizability is still possible and can be accounted for according to IC. Through stages of observation, reflection, active construction and response individuals acquire interactional competence in the 'oral practice' of the specific context (Hall, 1993, 1995), and by building up the resources necessary to operate in the local context, the learner is acquiring and developing a repertoire of skills that can be transferred to other, similar contexts (Johnson, 2001), as displayed in Figure 2. As shown in this diagram, the institution provides opportunities through its curriculum and facilities for students to build their experience of communicating in the target language, and this gives students skills they can transfer to other situations. The institution monitors the success of the program by evaluating how well the students transfer the resources they have developed to an appropriate assessment event. At this institution, the GOT is the form of assessment that provides a general arena in which students can deploy the language skills they have acquired from specific contexts provided through the university's curriculum and facilities. While the literature suggests that the GOT is capable of eliciting conversation-like language (Gan, 2010; Leaper & Riazi, 2014), it also has made it clear that the scores awarded are subject to influence from a wide variety of factors (as briefly summarized in the Statement of Problem above). While acknowledging their complex interaction with each other, these various influences may be categorized into those that are related to the format of the GOT, those related to the individual, and finally, those involved with the scoring of the GOT. This is displayed graphically in Figure 3. This figure shows that individual factors include language ability, personality, experience as well as other factors such as whether the individual test-taker slept or ate well before the test, his or her comfort level and how well he or she knows the other participants or not. These, combined with test context factors such as the prompt, the number of other participants, the other

Figure 2: The GOT in an interactionalist framework


Provided by curriculum with emphasis on discussion activities
Student's experiences in various communicative contexts
Skills built up by participation in discussions
Student's repertoire of communicative skills
General arena in which to display skills
Assessment in Group Oral Test

participants' ability and so on, affect their performance in the GOT. Their performance is observed by the raters, who are influenced by their understanding of the scoring bands, training, experience, fatigue

Figure 3: Factors affecting performance on the GOT



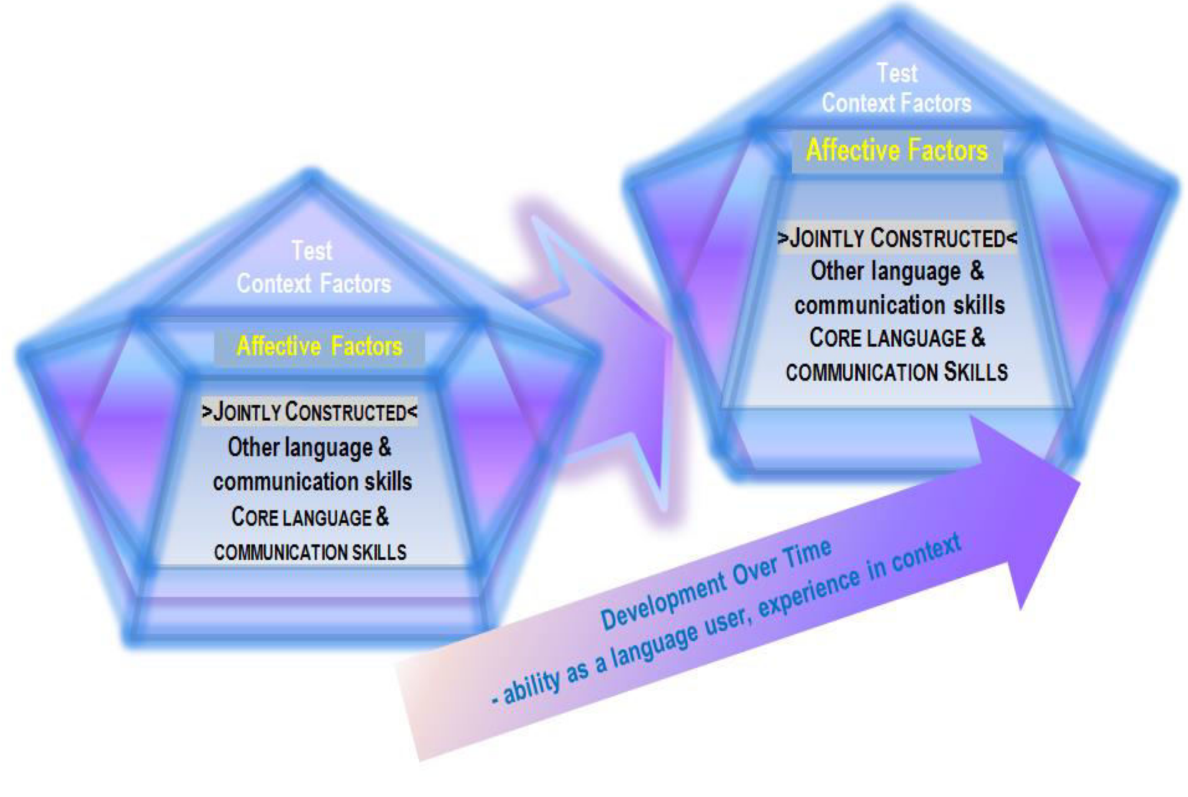
Individual Factors
Language ability, Personality Experience, etc.
Interacts with
Test Context Factors
Prompt, No. in group, Others' personality & language ability, etc.
Performance on the Group Oral Test
Scoring Factors
Scoring Bands, Training, Experience of raters, Rater fatigue, etc.
Score reported to individual students & admin

and so on when giving their score. The score, after being adjusted by a Rasch analysis based on Item Response Theory (IRT), is reported to the students and used for administrative purposes.

According to IC, a student's performance on the GOT is a collaboration with the other participants, and mediated by a combination of both individual and test context factors. No matter how these various elements interact, the final factor that plays a potentially decisive role on the overall impact of this test is the scoring. This dissertation will use the scores awarded as an important source of data. Although the factors involved with rating are not the main focus, this dissertation will most certainly illuminate issues and offer suggestions as to how it can be improved.

The dimension of time brings the students' development of language skills into consideration when investigating the impact of elements on performances in the GOT. At the time of writing this dissertation, no other study in the literature that has examined performances in peer interaction speaking assessment has used data from the same participants over a period of time, making this dissertation unique in this respect. In Figure 4 the various factors that may be involved in performances on the GOT are shown in diagram form. The layered jewels represent the various elements that are bound together in an individual's performance in the GOT. At the peak of the jewel are the 'test context factors' that affect performance on the test and change the most between different test administrations. It is the factors in this category that the test-takers have the least control over, like the number of people in their group (Nakatsuhara, 2011), the personality of the other members of the group (Ockey, 2009), and the prompt (Leaper & Riazi, 2014). The facets in the middle layer consist of the affective factors of the individual's performance: their personality, test-nerves, shyness, attitude, mood, condition of health on the day of the test and so on. Some of these factors may change slowly between administrations, but perhaps have their biggest impact the first time the individual takes the test, both in relative and absolute terms, as that is when test-takers are most unfamiliar with the test and their surroundings, and will also be the least adept at speaking. At the base is the test-taker's display of language and communication skills available to be assessed by the raters.

Figure 4: Relationship of factors impacting on performance on the GOT over time



This can be divided into context specific and core skills portions. The context specific language appears as a result of the local conditions that pertain to that instance of the unique interaction between individuals and setting. The ability to deal with such situations will develop over time with the test-takers continued exposure to various ‘oral practices’ (Hall, 1993) they encounter in their classes and facilities at the institution. Sharing space with these exceptional language uses will be the core language and communication skills that this format of test regularly elicits, and it may be hypothesized that these skills will show more consistent development over time, as they are more frequently called upon in the specific contexts that the individual is exposed to. It is these skills that this dissertation seeks to identify.

1.4 Objectives of the study and research questions

An issue that concerns this dissertation is related to the core communicative skills that the test-takers can demonstrate in the GOT. To do this, students taking the GOT three times in their first two years of study at the institution have been videoed, and the language they produce transcribed and analysed. The tools used for measuring the core language skills they display in their transcripts are the indices of syntactical complexity, grammatical accuracy, fluency (CAF), and range of lexis that the test-takers use over the times they take the test. CAF and lexis have become established as indices of L2 performance (see Housen & Kuiken, 2009) and so were employed in this study to capture the ability its participants demonstrate in the GOT. It is assumed that the students would show improvement in these indices over the time of the three administrations. However, given that the first time they take the test they are incoming students who usually have not on improving their their speaking skills and unused to the oral practices of this context, they will perform the least well in the first administration. Due to this, they are expected to gain the most in the second administration given that in their first academic year they have more to gain, as well as having become more familiar with the test context. It can be expected that improvements they show in the third from the second administration of the test will not be as marked as this first gain, perhaps due to ceiling effects and the fact that there is less to improve. This would be consistent with patterns of development noted in longitudinal research on study abroad programs (Serrano, Tragant, & Llanes, 2012). The first research question will confirm whether this supposition is in fact the case:

RQ1. How does the language elicited by the GOT show the development of syntactic complexity, accuracy, fluency, and range of lexis in test takers' performance in the three GOTs taken over the first two years of study at university?

Other key indicators of the pattern of discourse have been found to be words and frequency of turns (O'Sullivan & Nakatsuhara, 2011; Van Moere, 2007). In a study that involved one of the administrations in this dissertation's data set, significant differences were found according to the prompt that the test-takers took (Leaper & Riazi, 2014). Another study that investigated different levels of ability that was based on a cross-sectional design suggested that more advanced learners would tend to use more shorter turns (Galaczi, 2013). Using data that consists of an individual's performance over three administrations would cast light on the extent to which the length of turn is subject to varying factors, and hence its use in research in this area as formulated in the following research question.

RQ2. To what extent do the test-takers' number of words, frequency and length of turns show a consistent pattern of development in their performances in the three GOTs they take over two years?

The next objective of this study is to investigate the specific context for interaction that the GOT provides. This context needs to be researched because it is at the centre of the role that the GOT plays in the institution itself. As we saw in Figure 1, one of the primary reasons for choosing this format is the link that it has to the communicative curriculum at the university; particularly the positive washback of encouraging students to practice communicating in English and teachers to use communicative activities in class. The underlying assumption is that conversational interaction is beneficial to language learning, since it is on this basis that communication is part of the curriculum and the GOT used to assess it. This assumption will be investigated by reviewing the literature to identify the features of interaction in conversation that are beneficial for the acquisition of language, and then investigate the GOTs in this administration to see if such features are used in this test.

Although previous research suggests that this format of test is capable of eliciting such features (see Gan, 2010 for example), it has yet to be investigated as a function of the development of ability among test-takers who take the test on multiple occasions over time leading to the third research question.

RQ3. To what extent is the GOT able to elicit features of conversation that have been found to be beneficial to language learning?

The final research questions are related to the role the GOT plays in the university's language learning administration. As pointed out above, a major role of the GOT is to generate positive washback within the learning system, and the principle means of doing so is by the scores assigned to the students (see the rating bands in Appendix I, p. 426). Since it is likely that the students' perceptions of the test will be strongly influenced by the scores they are assigned for their performances, it is necessary to investigate this issue. For the university administrators too, the scores are used for monitoring the performance of the students and the curriculum. For both students and the administration then, the justifiability of the scores based on actual student performance over consecutive administrations of the test is a crucial issue.

Aside from institutional uses, there is a wider interest in the outcome of this research question. If it can be shown that regular teachers with minimal training can assign scores that are justifiably based on the performance of the test-takers, then this would encourage wider use of this test and the administrative procedures described in this dissertation. Conversely, if there is little relationship between the performances and the scores, then it would seem that the variable factors picked out in the literature dominate this format of test, and would make extensive changes to the format or administration necessary before its further use could be recommended. In this case, recommendations can be made for improving the test. This question can be answered statistically by using the speaking performance indices from the quantitative findings, and results in the following research question:

RQ4. How well do the speaking performance indices of test-takers who take the GOT over the three administrations in two years predict the scores awarded to them?

Since the indices may capture a narrow view of the progress it is important to consider this question qualitatively as well, using a complementary mixed methods research (MMR) design (Riazi & Candlin, 2014). This will be done by choosing a smaller sample of the participants, and analysing their performance qualitatively, using framework informed by the theory of Interactional Competence (IC) (Young, 2009, 2011). This analysis will be synthesized with the scores they were awarded according to the rating bands used at this institution (Appendix I, p. 426) and their performance as measured by indices. The final research question is thus:

RQ5. How well do the test-takers speaking performance indices and scores awarded to test-takers of the GOT over the three administrations of the test taken in two years represent their performance compared to a qualitative analysis of their performance?

1.5 Organization of the dissertation

This dissertation is composed of six Chapters. Chapter 1 is the Introduction, in which I have outlined the statement of the problem, the context and conceptual framework, and finished with the objectives and research questions. In Chapter 2 the literature will be reviewed. As this dissertation is embedded in the construct of interactional competence (IC), I am bound to start the literature review with a discussion of its major features. This will allow the next section, an overview of the history of the direct assessment of speaking, to be seen in a wider context as the field moved from an unquestioning acceptance of interviewer-led tests, to the realization that formats in which peers interact with each other without an interviewer are more likely to elicit conversational ability. This section will also be of value since the previously predominant format, the interview, was criticized for its inability to measure conversational interaction, and thus, it includes a description of important features of conversation. The following section will move onto conversation and its role in language acquisition, which will inform the analysis of interactional features in Chapter 3. Chapter 3 will present the quantitative phase of the MMR study; providing both the methodology and results which answer research questions 1-4. Chapter 4 presents the qualitative phase of the MMR study, which is an examination of a sub-sample of participants' progress, and does not itself answer a research question, but provides the basis for research question 5 to be answered in Chapter 5, which synthesizes the findings of the quantitative phase with those from the qualitative phase. Chapter 6 will conclude the dissertation by summarising the findings and discussing its implications and limitations.

CHAPTER TWO

Literature Review

2.0 Introduction

The first section of this literature review starts by explaining a framework of interactional competence (IC) which can take into account the context of communication while allowing meaning to be inferred from the scores. The second section reviews the literature on speaking tests, starting with interview-led tests. Although this format has been subject to criticism from many different quarters, the literature review will focus on studies that investigated the language that interview-led tests elicit. The review simultaneously presents the history of speaking tests since it was widely accepted that interviews were *the* method of directly assessing speaking. Reviewing the shortcomings of interviewer-led tests naturally leads to the alternative: interviewer-less formats in which the test-takers interact with each other. For the purposes of this dissertation, these will be divided into paired and group formats of three or more test-takers, though this should not be seen as implying that one format is superior to the other – they are simply alternatives, each with their own advantages and disadvantages. Finally, since one of the main reasons for using a group format is the positive washback of encouraging test-takers to use the target language with each other and teachers to provide practice for such activities (Bonk & Ockey, 2003), the literature on conversation as a mode of language learning will be reviewed by way of a justification for peer-interaction testing. The review of the role of conversation forms the background for this dissertation by not only describing how NNS-NNS interaction benefits language learning, but also by providing some of the terminology and tools that have been used in research in this field.

2.1 Interactional Competence

A fundamental question for language testers is, ‘What do the scores of a test mean?’ If the scores that result from an assessment vary widely every time the test is taken, the question of what the test measures may be confounded with questions about the quality of the test itself. Yet even if a test produces consistent scores, the question still has to be answered. As Messick pointed out, it becomes necessary "to account for the consistency in behaviours or item responses, which frequently reflects distinguishable determinants" (Messick, 1989, p. 14). In the case of this dissertation, the question is just what are the distinguishable determinates for the GOT?

The two opposing perspectives on possible meanings of the scores in relation to the quality that is being measured can be characterized as that of the *trait theorists* and *behaviourists.* For trait theorists, consistent scores on a test reflect underlying abilities and knowledge that exist in the test-taker. The trait is relatively stable, though change can be expected over time as the test-taker's ability either improves as they learn and practice more, or declines if they fail to maintain their skills or knowledge. A consequence of the notion that there is such a relationship between the test-scores and the trait is that scores are meaningful indicators of the test-takers' ability in other domains, and such characteristics should be specified independently of the context in which they were assessed (Chapelle, 1998). Since the test measures underlying abilities, the scores that represent them are generalizable. For example, to a trait theorist, consistent scores in a GOT would generalize to the test-takers' ability in general conversation. For language testers this is somewhat natural territory, since over much of its relatively short history as a discipline, since Lado's (1961) textbook on testing, the focus has been on cognitive models of underlying abilities that permit generalizability (Chalhoub-Deville & Deville, 2005).

For behaviourists, on the other hand, the scores derived from test performances indicate the test-takers ability and skills in the specific context of the test, and as such the scores only show how well they did on that particular occasion. The particular circumstances of the test elicit a *response class;* a set of behaviours that emerge consistently in the context of this particular kind of activity and environment (Messick, 1988, p. 15), and, according to behaviourists, it is a mistake to think of them as traits. The test, in this view, elicits a sample that is only generalizable to other environments that resemble in detail the context in which it was collected. To determine if the scores of a test are relevant then, it is necessary to provide specifications of the context in which the scores were awarded (Chappelle, 1998).

These two perspectives are extremes, and both contain elements useful for assessing the validity of a test: while we would not wish to abandon the generalizability of scores, neither would we wish to deny the impact of the context. When it comes to the assessment of spoken language, the failure to take into account the contextual element leaves the test developer open to criticism of ignoring the co-constructed nature of communication (Jacoby & Ochs, 1995), while specifying the behaviour leaves it open to question what the construct actually is (Read & Chapelle, 2001). It is no accident that the *interactional competence* (IC) that informs this dissertation was borne from the perceived shortcomings of the proficiency movement that was inspired by the Oral Proficiency Interview's (OPI) *ACTFL/ETS Proficiency Guidelines* (Kramsch, 1986). Moreover, the body of work in the literature that is built on the *communicative competence* of Canale and Swain (1980) is beholden to this criticism, since it focuses on what skills and knowledge the individual needs to have and know in order to communicate within a social context, without considering the jointly constructed nature that encompasses all participants in the act of communicating (Bachman, 2007; Young, 2003).

The concept of IC seeks a middle way between behaviourists and trait theorists by attempting to incorporate a degree of generalizability, while at the same time recognizing that the performance in the test cannot be separated from the circumstances in which it was elicited (Chapelle, 1998). This approach is not an additive process of appending the effect of the context to the trait; rather it is a case of describing how they *interact* with and through each other. To understand the meaning of test scores it is necessary to not only collect information on the trait that the test is supposed to measure, but also the context it is collected in, and investigate the relationship between them.

In the years following Kramsch's article, IC has come to encompass a variety of positions depending on the extent to which the trait and the context can be distinguished. Bachman (2007) identified minimalist, moderate and strong versions of IC. The minimalist position is one that frames IC as a characteristic of the interaction in which an individual participates, and for Bachman, Chapelle (1998) is representative of this position. Chapelle (1998) points out the necessity of a bridge between trait and context. This bridge is conceptualised in Bachman's (1990) framework, considered a "general interactionalist construct definition of communicative language ability" (Chapelle, 1998, p.44), by *strategic competence*, which has the role of mediating between trait and context.

The difficulty for the minimalist position is that it is barely distinguishable from that of a trait theorist's. This can be seen when Chapalle states that "an interactionalist construct definition comprises more than trait plus context; it includes the metacognitive strategies responsible for putting person characteristics to use in context" (Chapelle, 1998, p.44). As Chalhoub-Deville (2003) and others (for example, see Bachman, 2007; Johnson, 2001; McNamara, 1997) have pointed out, Bachman's (1990) framework is still a cognitive model in which the individual is solely responsible for an interaction, and is static with regard to the social context (Johnson, 2001), making it ill-suited as a framework for IC.

By contrast, a moderate position on IC proposes that the ability the participant brings to a communication event dynamically interacts with the situational facets. When a participant attends to certain aspects of the communication, it activates certain abilities in the participant's repertoire and in this way the context is inextricably linked to the communicative performance (Chalhoub-Deville, 2003). Chalhoub-Deville represents this as "ability – in language user – in context", where there is interaction between each of these elements. By contrast, Bachman's (2002) stance is characterised as "'ability – in language user' based on 'language user – in context'" (Chalhoub-Deville, 2003, p. 372), and this would also apply to the minimalist conceptualization of IC.

For Bachman, Young (2000) represents the strong interactionalist position. For Young, an oral interaction is locally produced, contingent on its participants, and bound to the context. A spoken interaction consists of a series of *discursive practices*, which are similar to Hall's *interactive practices* (1995). These are episodes of communication that regularly occur, allowing participants to recognize and respond according to what is considered acceptable in the circumstances. The extent an individual can use such knowledge is constrained by the extent that it is shared by the participants in the interaction, thus the *competence* that is referred to is one that is both shared and jointly constructed in the interaction. Young also points out the necessity to place the interaction in the wider context of practices so that the extent to which it shares resources can be understood. Usefully, Young provides a list of interactional resources that participants of a conversation may draw on that can serve to define the particular interaction that is taking place. These form an "interactional architecture of that practice" (Young, 2000, p. 5) that allows comparison with other contexts to identify which features are shared and which are unique, and to allow judgments about the generalizability of the scores. They were later expanded to seven resources that can be organized into three categories (Young, 2009, 2011) as can be seen in Table 1 below. The question about the meaning of the test scores is answered by investigating the configuration of the identity, linguistic and interactional resources of the assessment event, and comparing these to the discursive practices in non-test events.

Table 1: Resources through which participants' interaction may be analysed (Young, 2013)

1. Identity resources	Participation framework	Identities of all participants, whether officially there or not, their role in the interaction, the identity they construct as participants
2. Linguistic resources	Register	Use of pronunciation, vocabulary, grammar that characterizes the practice
	Modes of meaning	Interpersonal, experiential and textual meanings participants construct for the practice
3. Interactional resources	Speech acts	The sequential organization of acts selected for the practice
	Turn-taking	The allocation of turns in an interaction, including who has a right to claim them, how they know a turn has ended
	Repair	How interactional difficulties are dealt with once they arise
	Boundaries	The means by which the various parts of the discourse are marked by the participants.

Bachman (2007) finds a number of points to critique the earlier formulations of interactional competence as discussed in He and Young (1998) and Young (2000). Firstly, he points out the contradiction between speaking being termed a 'subset' of language ability, which is consistent with the trait theorist's stance, and then later on their suggestion that the "abilities, actions, and activities do not belong to the individual but are *jointly* constructed by *all* participants" (He & Young, 1998, p. 5, italics from the original). For Bachman, confusion persists with He and Youngs' alignment of interactional competence with the co-construction of Jacoby and Ochs (1995), which he points out is a process. Bachman criticizes He and Young (1998) and Young (2000) for not adequately explaining how a locally developed practice tailored to a particular context can then be utilized as a general resource that can be applied to other contexts, a point that Chalhoub-Deville (2003) concurs with. At the same time, Bachman argues that Young's (2000) list of resources that participants bring to a communicative situation is "essentially the same as that of language ability as an attribute of individual participants" (Bachman, 2007, p.60), and is consistent with his "ability – in language user" based on "language user – in context" position (Chalhoub-Deville, 2003).

These criticisms were answered by Young (2011) who pointed out that although IC may be seen as adding another component to those already found in formulations of communicative competence such as Bachman (1990), the difference is that they can only be deployed in a communicative context, and how they are used is contingent on the other participants. In this sense IC is “distributed across participants” (Young, 2011, p.430), and will naturally vary according to who the participants are and their context. Young (2011) answers other criticisms by grounding IC in the notion of intersubjectivity. This is the necessity in communication that not only both the sender and receiver of the message be aware of each other, but both have awareness that the other is aware that the message is being attended to. Moreover, these layers of awareness do not exist in a vacuum, but a “triangular relationship between the sender, the receiver, and the context of the situation” (Wells, 1981, p. 47). Within the context the participants employ their linguistic resources, but what and how they say depends on who they are talking to and the context, and this integrates with the identity, linguistic and interactional resources they draw on to construct their participation for the context.

For IC to be of use to assessment, and indeed in a more general sense, it is necessary to account for the participants’ ability to transfer their conversational skills. A strict interpretation of IC as locally produced goes against the common sense notion that a speaker should be able to communicate appropriately in some novel situations never before encountered. Young (2000, 2003, 2011) argues that generalizability occurs because elements of the ‘architecture’ are similar in different contexts, and to this extent participants can employ practices developed elsewhere. As McNamara (1997) pointed out, this requires a specification of interactive practices that the assessment is focussed on “to reveal the standards that apply in reality” (p. 457).

Evidence of the transferability of interactive practices were discussed in Young (2003, 2011) in which he compared the interaction of teacher assistants in the Maths department with an interaction by teacher assistances in the Italian department. He found that there were sufficient similarities to characterize it as a genre, and so presumably these similarities have the potential to be used to construct an assessment task for this sphere of activity. He also gives negative examples where the difference between contexts was the issue. Here he points to the literature that critiqued interview formats as tests of conversational ability (see Section 2.2.2). These papers showed that the "interactional architecture" (Young, 2011, p. 440) of interviews differed substantially from what might be expected in conversation, and so the assumption that a high (or low) score in such an assessment inferred a certain level of conversational ability would be unfounded.

In summary, the theory of IC as described by Young (2011) offers a viable portrayal of how an individual participates in a communicative event, and provides a framework for analysing it (see Table 1). According to this perspective, language learners build up a repertoire of resources in the contexts they are exposed to, which then can be deployed in other contexts to the extent of their similarity to familiar contexts. Test-takers need to develop not only the more static abilities to speak fluently, and use the vocabulary and grammar at their disposal, but also to build their experience in communicative events so that they know when it is appropriate to speak, how to take turns, and so on. When they communicate with other interlocutors, what, how and when they communicate is contingent on the skills and abilities of the other participant, and in this way their competence is shared. At the same time, through their participation they construct an identity as a language user through which these abilities are deployed.

To apply IC to the assessment context in this dissertation means that there will inevitably be conflict between the test-takers performance in the conversation as jointly constructed among the participants and the institutional necessity to award scores to individuals. This is not surprising given the field of assessment's tradition of measuring individual abilities (Hughes, 2003) and the basic need to assess individuals. It must be recognized that when perusing the curriculum vitae of a prospective employee, an employer is likely to prefer information about individual characteristics, rather than conditional information about what the candidate might be able to do in certain conditions that may not be relevant in the workplace. This has repercussions for the contribution this dissertation can make. Firstly, since this dissertation tracks student performance of the same participants over three administrations, it can make some progress in the issue of appraising what skills an individual can be expected to consistently display in this assessment context. Secondly, by investigating both the scores awarded and the performance, it may be able to contribute by describing conditions in which the context is most likely to impact on supposedly individual scores.

Having examined IC as the theoretical underpinning of this study, the next section will review the literature on the direct assessment of speaking.

2.2 Assessing speaking

In this section the literature on the assessment of speaking will be reviewed starting with the interview format, which also comprises much of the early history of testing speaking. Over this time there were papers critical of the particular format that was the most widespread at the time, ACTFL's Oral Proficiency Interview (OPI) (1982, 1986) as well as of interview-led tests in general, on a wide range of issues. The central role that is of interest is the ability of the test to engender the washback of encouraging students to use the target language with each other, and so this dissertation will focus on the literature that critiqued the interview format for the language it elicits. This review will thus explain one of the most important reasons why a language learning institute might choose a peer interaction test over the longer established interview format tests. After the sections on interview-led tests, attention will switch to paired and group format speaking tests.

2.2.1 The development of the interview format for oral assessment

In language assessment, the testing of speaking ranks as the most recent of the skills to become the focus of development (Fulcher, 2003), and is inextricably linked to development of interview-led tests since this is how the early tests were conducted.

The first consistent use of a speaking test seems to have been the compulsory oral component of the *Certificate of Proficiency in English* (CPE) in Great Britain in 1913, which involved reading aloud, a conversation with an examiner and dictation (Taylor, 2003). However, it was new technology such as the short-wave radio and especially the advent of the Second World War that thrust oral and aural foreign language skills to the forefront (Kaulfers, 1944), and the testing of speaking along with them. In the United States it was realized that the military would need personnel who could talk and understand the language of the various nations that they would be dealing with, and so such language programs as the Army Specialized Training Program (ASTD) were set up to focus on training military personnel in speaking and listening skills (Fulcher, 2003). With the creation of such programs, valid and reliable assessment became of particular importance, especially given the consequences of making mistakes in a time of war. Kaulfers (1944) laid out some important principles for the testing of speaking, notably a focus on "the examinee's *ability to perform in very specific real-life situations*" (p.140, italics from the original). The tasks were short directions for the test-taker to perform a function in the foreign language. For example; "How would you tell a Spanish speaking native to take you to a military hospital?" (p. 145). Yet it was an alternative approach described by Agard and Dunkel (1948) that proved to be more influential. Rather than asking for a direct performance, they elicited a sample by means of tasks in which test-takers described pictures or spoke on a certain topic. As Fulcher (2003) points out, the approach by Agard and Dunkel (1948) can be seen to be a direct forerunner of the Foreign Service Institute's (FSI) Oral Proficiency Interview (OPI).

The FSI itself had been set up before the war to teach language to diplomats, but it was in 1952 that the measurement of speaking skills came into focus when they were required to set up a register of those with language skills who might be useful to the United States government. As a result of this, the first rating scale was constructed and published in 1958. This six band holistic scale was developed and used to grade performance on an interview which would last 10-40 minutes. The interview opened with a series of simple questions and proceeded with prepared questions designed to elicit various grammatical structures, along with a role play (Chalhoub & Fulcher, 2003).

The need for a language test was being felt by other governmental organizations in the 1960s, particularly the Central Intelligence Agency, Defence Language Institute and the Peace Corps, who each adapted the scales for their own needs. These organizations came together in 1968 as the Interagency Language Roundtable to construct a standardized version (Malone & Brooks, 2013). Perhaps the most important of these users was the Peace Corp, since it was through this organisation that academics were introduced to the OPI, and by the 1970s the rating bands were being used in universities in the US for bilingual teacher certification. It was soon recognized that the scales would need to be revised so that it could better discriminate between lower level speakers, and the Educational Testing Service (ETS) and American Council on the Teaching of Foreign Languages (ACTFL) became involved, eventually resulting in the *ACTFL Proficiency Guidelines* in 1986, which were revised again in 1999.

The Guidelines were an impressive development at the time and came to be used as a basis for curriculum design, with the OPI being used to measure learner progress. The 1986 revision of the Guidelines gave impetus to this approach, which became known as the *Proficiency Movement*. The Guidelines describe language learner proficiency on a holistic scale that ranged from Novice, with little functional ability in the language, to Superior, representing a well-educated native speaker. The OPI was based on this scale, and with funding from the US government for workshops to train raters, became the most widespread method of assessing speaking proficiency at universities across the USA.

In the field of language education, the use of these scales and method of testing was taken up with some enthusiasm. As it was described at the time:

> The boldness and purposefulness of the ACTFL Proficiency Guidelines… have generated enthusiasm, rekindled the faith of many foreign language teachers, made our profession highly visible, and won us generous financial support from private and public funding. (Kramsch, 1986)

Along with its popularity, there was a rising tide of criticism of it, and some of the most notable papers critical of the OPI and other interview speaking tests are summarized in Table 2.

The next section focuses on the papers in the literature that investigated the language elicited by interview format speaking tests.

2.2.2 The interview as an assessment of conversation

At the time it was simply assumed that the language elicited by the OPI was evidence of conversational ability of the test taker. Indeed in the training manual for TOEIC's Language Proficiency Interview (LPI), which is, according to Johnson (2001) the same as ACTFL's OPI, they describe the interview as "a face-to-face conversation with one or two trained interviewers" (ETS, 1996, p. 1), and in the ETS training manual of 1982 they pronounce that the OPI "tests speaking ability in the real life context – a conversation. It is almost by definition a valid measure of speaking ability" (ETS, 1982, as cited in Johnson, 2001, p. 2). Such statements became increasingly questionable as researchers investigated the nature of the language used in interviewer-led tests.

Table 2: Summary of studies critical of interview format tests

Study	Test	Participants	Main findings
Bachman 1988, 1990	OPI	-	Holistic scale not supportable, lack of generalizability, confounds abilities with elicitation procedures
Brown 2003	IELTS	1 T-T[1], 2 interviewers	Raters showed substantial variation in organizing topic, questioning technique and feedback
Brown & Hill 1998	IELTS	32 T-T, 6 interviewers	Characteristics of easy vs difficult interviewers, may affect a test-takers score
Fulcher, 1996b	OPI	-	Evidence the empirical support for the ACTFL scales is flawed.
Johnson & Tyler, 1998	OPI	1	Turn-taking, topic nomination dominated by interviewer, unequal obligations, lack of conversational involvement
Johnson, 2001	OPI	35	OPI resembles sociolinguistic & survey research interview, asymmetric power distorts the interaction
Lantolf & Frawley 1985, 1988	OPI	-	Scales not empirically based, use of well-educated native speaker as ultimate criterion, circular definition of proficiency
Lazaraton, 1992	Entrance interviews	20 T-T	Only opening and closings of interviews resemble conversation
Raffaldini, 1988	OPI	60 T-T	OPI correlates strongly with grammar & communicating message, no relationship with sociolinguistic competence
Ross & Berwick 1992	OPI	60 T-T, 12 Interviewers	Interviewers accommodated differently to test-takers at different levels
Ross 2007	OPI	1 T-T, 2 interviewers	Divergent levels of interviewer support influenced rating assigned to the same candidate
Young & Milanovic 1992	Cambridge First Certificate	30 T-T	Interviewers goal oriented, T-T reactive oriented; T-T affected by significant variation due to variability in tasks, gender, theme.
Van Lier, 1989	OPI	-	Substantial differences exist between conversation and interviews

[1]T-T = test-takers

An early indication of how interviewer-led tests fail to encompass the range of communicative skills involved in conversation was a study by Raffaldini (1988). She correlated the results of students' OPI ratings with their performance on two tests designed to investigate their knowledge of sociolinguistically appropriate language. The results showed that the OPI correlated most strongly with *grammar* and *communicating the message* – the two parts which require the heaviest use of grammar – but not at all with sociolinguistic elements. This was largely due to the dominant interaction in the OPI being question-and-answer, leaving little opportunity for other functions to be elicited. She goes on to point out that even the role-plays in OPIs seem designed to elicit language that has already been produced in the interview, instead of sampling different language functions. She concludes that the OPI elicits but a limited range of interactional functions.

While Raffaldini examined the sociolinguistic range of language elicited by the OPI, van Lier (1989) compared interviews with the characteristics of conversation. He pointed out that conversations are unplanned, participants have equality of status and there is an unpredictable sequence and outcome, whereas interviews are "almost inescapably asymmetrical" (p. 501). Indeed, he declared that in the OPI "there is no official requirement to produce conversation… The emphasis throughout is on successful *elicitation of language*, not on successful *conversation*" (p. 501, italics in the original).

The first to investigate the OPI from the "inside" was van Lier (1989) in what he admitted was an "exploratory attempt" (p. 490). Clearly a more rigorous, principled investigation was required. In response to van Lier's paper, Lazaraton's (1992) conversation analysis (CA) study of 20 video and audio taped course placement interviews answered this call. She found that only the opening and closing moves of the interviews in her study had features in common with conversation. In the OPI, these sections are not considered in grading. In all other parts of the interaction, the interviewer dominated, as was particularly evident in the way the interviewer finished one section and moved to the next part of the interview. Although on some occasions the interviewees initiated the closing sequence itself, they could only do this because the interviewer introduced a pre-closing sequence, and as Lazaraton notes, "'Thanks' as a closer is not typical of conversation" (p.383). It is precisely these features which mark these exchanges as one of an interview, not a conversation.

Lazaraton's (1992) study was not on the ACTFL OPI but her findings were confirmed in a CA analysis of an OPI test that was used as an example of a Level Two speaker in the workshop for training OPI examiners. Like Lazaraton, Johnson and Tyler (1998) based their analysis on procedures of Schegloff and Sacks (1973), Sacks, Schegloff, and Jefferson (1974) and Schegloff, Jefferson and Sacks (1977) amongst others, to examine the same features of conversation as Lazaraton. In addition, Lazaraton (1992) looked into the role of topics in the conversation as spontaneously created and collaboratively constructed by conversationally involved participants. The researchers expected that participation in a naturally occurring conversation would be approximately balanced between the participants, and that there is a need to maintain face of those who take part. However, in the analysis they found that turn-taking, length of turns and their distribution were determined by the two interviewers who took part in this OPI. Their data showed how dominant the interviewer is in the OPI, and by extension, the interview as a format for assessing language. When the turn-taking sequences were examined it showed a clear sequence of interviewer asking questions and the interviewee being obliged to respond. The obligation is not reciprocal because the interviewers ignored two of the five information questions addressed to them, breaking another conversational rule that the next speaker is chosen either through the current speaker choosing, self-selection, or the current speaker continuing (Sacks et al., 1974). Furthermore, in this example there were two interviewers who did not interact with each other, instead taking turns to ask questions that the single interviewee was obliged to reply. The next speaker was automatically one of the interviewers, and even if the test-taker tried to choose the next speaker she was ignored, as noted above.

When length of turns was looked at, they found that there was a heavy imbalance between the interviewers and the interviewee, with the average length of the interviewers' turn being 4.3 words against the average length of the interviewee being 19.5 words. What is more, the interviewers seemed to be monitoring the length of response. If they perceived it as being too long, or they needed to move on to the next task, they would interrupt to cut the response short and if it was thought to be insufficient, prompted the interviewee for more details. This can lead to exchanges that violate conversational norms, as can be seen in the following exchange:

Int: What does he look like… originally looked?

NNS: Well, he just looks like any other oriental man.. Korea man. (Silence) he is big though, like, uh, he weighs about 180 pounds and... he is about six foot.

Int: What about his eyes? What are they?

NNS: Well, uh I think he has . . dark brown. Uhm . . all my family are big you know, my father is big except my mother she is kina . . little . . and my brother is big… etc

(Johnson & Tyler, 1998, p.39)

As Johnson and Tyler point out, in this exchange the interviewer's prompting results in an incoherent conversational exchange. The object of this part of the interview is to elicit a description, and when the reply is insufficient the interviewer prompts the test-taker. The interviewer is ignoring the conversationally pertinent information about the man being big, breaking the 'relevance' rule of conversation (Grice, 1975).

When topic nomination was examined, in interviews the topic is often decided in advance and contrived to suit the interview's purpose. This is shown clearly at the point in the transcript when the next stage of interview starts and is particularly evident in the transition to the role play task in the OPI. The interviewer baldly states that a role play will be the next activity in the interview, and the role play ends when the interviewer thanks the interviewee for her participation. Such explicit bracketing also took place in the beginning and end of seven of the eight major topics of the interview. A later study that used data from 35 LPI interviews conducted over the telephone, confirmed interviewer dominance of topic (Johnson, 2001). In one typical example the interviewer changed the topic from a description of a room to a question designed to elicit a response about current events in South-East Asia, two topics that obviously bear little relationship to each other. Only in terms of repair did Johnson find anything in the OPI that comes close to what would be expected in conversation, typically being self-initiated self-repair. Johnson states that this may be surprising when it is considered that the participants of these dialogues are a native and non-native speaker, and it might be expected that the native speaker would help the non-native on occasion. It should be pointed out that this too is an artefact of the OPI, as interviewers are directed not to give corrective feedback during the interview (ETS, 1996).

It is clear then, that while there are a few features that OPIs and conversation have in common, in terms of the range of socio-linguistic knowledge, turn-taking, topic nomination, turn sequencing, and type of turn the OPIs depart from what might be expected. The crucial difference is that the role of the interviewer is assigned prior to the interaction and introduces a power difference that is not conducive to producing conversational language (Savignon, 1985; Young, 1995). Other studies have investigated how this power difference manifests itself in various ways (see Brown 2002; Brown & Hill 98; Lazaraton & Saville 1994; Reed & Halleck 1996; Ross 1992; Ross & Berwick 1992 for example, as summarized in Table 2).

If it is the construct of 'conversation' that is the objective of the assessment, then a solution would be to remove the interviewer from the interaction, and this is what happens in a peer interaction test: a number of participants are given a task that elicits a sample of them talking to each other without outside intervention. The next section examines alternative formats that use fellow test-takers rather than interviewers as interlocutors.

2.2.3 Peer interaction speaking tests

At this point I should distinguish paired tests from group oral discussion tests (GOTs) of three or more participants. While they are both formats in which the test-takers interact with each other, they differ in some ways. Firstly, participation is obligatory in paired interactions since if one person is talking, the next turn is necessarily the listener's. In GOTs the next turn is potentially any of the listeners'. The participants of a GOT take part as much as they are capable of or willing to. Also, the extra participants make the interaction in GOTs more complex than that found in paired interaction. Pairs can only interact with each other, but having (typically) three or four participants multiplies the possible combinations of speaker and listener, and also creates an audience of participants if the speaker chooses to direct a question at another individual. Despite this added complexity, this study makes the assumption that interactions in paired and grouped discussions are more similar to each other than they are to interviewer-led tests, and while the focus of this dissertation is on group discussions of three or four students, the literature on paired formats is also relevant.

2.2.4 Testing in pairs

2.2.4.1 The early literature on the paired format

Although the last decade has seen a plethora of articles on the paired format (see Table 3 for a summary), one of the complaints in the earlier literature is about the lack of research on it (Csepes, 2002; Foot 1999a, 1999b). This lack is quite surprising when it is considered that paired speaking tests have been used since the 1990s by what is now called Cambridge Assessment, one of the largest testing organizations in the world (Saville & Hargreaves, 1999). Yet, one of the earliest reports in the literature is an enthusiastic report of a paired format being used with Italian university students, in which the author noted how "there seemed to be more experimentation, greater spontaneity… [without] the visible cringing that some students exhibit when the teacher-examiner is the one asking the questions" (Lombardo, 1984 p. 4). Lombardo's paper considers issues of validity and reliability, but it remains an anecdotal report of classroom assessment.

Table 3: Summary of selected studies investigating paired formats

Study	Test Context	Test-takers (T-T)	Main findings
Bennett 2012	Language Institute	12 Italian	Differing proficiency level had no effect for 9 of 12 test-takers; no pattern of better or worse with 3
Berry 2004	University	54 Hong Kong	Extraverts perform better in paired tasks, introverts perform better on individual tasks; extraverts do better on high interactive tasks, introverts on low interactive tasks
Brooks 2009	University	16 varying	10 interactive functions were produced in interview test, 17 in paired interaction, scores were higher in paired test
Davis 2009	University	20 Chinese	Proficiency level had no significant impact on scores received; low proficiency learner produced more language when paired with higher level.
Ducasse & Brown 2009	Spanish as a 2nd Language	34 Aus. T-T 11 Span. raters +1	Effective interaction divided into three categories: Non-verbal interpersonal communication, Interactive listening, interactional management
Egyud & Glover 2001	High School	14 Hungarian	Participants in favour of the paired format, compare language produced in interview and paired
Csepes 2002	High school	120 Hungarian	Proficiency level had no significant impact on scores received
Iwashita 1996	JSL	17 Australian	High level students seem to get high scores with similar level partners, low partners did better with high partners
Fulcher 1996a	EFL institute	47 Greek	Comparison of T-T on different formats of speaking test
Galaczi 2008	Cambridge FCE	60 varying	Collaborating dyads with low dominance tended to score higher than those with blended or asymmetric interaction
Galaczi 2010, 2013	Cambridge Main Suite	52 varying	Development of IA skills of test-takers at CEFR levels B1, B2, C1, C2
Lazaraton & Davis 2008	Cambridge PET, FCE	8 varying	Identity as test-takers described as doing ‘being proficient’, ‘being interactive’, ‘being supportive’, ‘being assertive’
May 2009	University	2 Chinese	Raters difficulties dealing with dysfunctional pairing
May 2011	University	12 Chinese	Raters expand on the wording of the rating bands
Nitta & Nakatsuhara 2014	University	32 Japanese	Test-takers with 3 min. planning time rated significantly higher in fluency, accuracy and complexity; indices showed a significant difference in fluency and length of turn.
Norton, 2005	Cambridge FCE	27 Japanese	Test-takers influence each other’s use of vocab. & grammar, positively affected by familiarity, negatively affected by gender of partner
Orr, 2002	Cambridge FCE	4 T-T 32 raters	Raters paid varying attention to the rating scales, sometimes arriving at the same grade for different reasons
O’Sullivan 2002	University	32 Japanese	Japanese females performed better with a friend, but accuracy improved with non-friend.
Plough, MacMillan, & O’Connell 2011.	University	39 Greek	Greater range and number of interactive functions produced in paired tests (34) than interviews (21); functions found only in the paired more challenging
Sandlund & Sundqvist 2011	High School	20 Swedish	Management of task related trouble compared between higher and lower scoring students and raters

The first major testing organization to use the paired format was the Royal Society of Arts Certificate in the Use of English as a Foreign Language test that was developed in the 1980s. After it was transferred to the University of Cambridge Local Examination Syndicate (UCLES) the system of paired testing was retained. This organization gradually introduced the paired format into their main tests from 1991 with the Certificate in Advanced English (CAE), the Key English Test (KET) in 1993 and First Certificate in English (FCE) in 1996, and other tests as they were revised (Taylor, 2000). Currently the speaking sections of the Cambridge Main Suit exams, as well as more specialized ones (for international finance and law), are conducted in a paired format[1] in which an interlocutor leads the test-takers through three or four sections including one which is a 3-4 minute peer interaction task in which the rater-interlocutor does not intervene.

Given this increasingly widespread use of the paired format by a major testing organization, the lack of published information must have been of concern to those involved in testing at the time. These concerns were given voice by Foot (1999a) who raised several questions about the validity of the format. He pointed out that little is known about how stress interacts with performance, and just because test-takers may feel less anxious interacting with a peer than with the interviewer, it does not mean they will perform better. Also, paired candidates with different L1s might be unfairly penalized due to the difficulty of understanding each other than if they shared language backgrounds. Foot continued by raising questions about the effect of different proficiency levels, age and personality factors such as assertiveness on the interaction that results from being tested as a pair. In particular, Foot felt that lower level test-takers would be particularly disadvantaged without the presence of an interviewer, and that “there is a risk the assessment will reflect the paucity of data rather than the candidates’ abilities” (Foot, 1999a, p. 40).

These early questions were not answered satisfactorily by Cambridge’s representatives, Saville and Hargreaves (1999), who countered the contention that talking in pairs is not fair due to the impact of the interlocutor, by merely claiming that “these concerns cannot be addressed with definitive answers” (Saville & Hargreaves, 1999, p. 44), before going on to justify the introduction of the format

[1] If there are an odd number of candidates, the final test will be conducted in a group of three.

and describing administrative procedures. Perhaps not surprisingly, as it did not address the particular concerns raised in the original article, Foot (1999b) was unsatisfied with this response.

2.2.4.2 Comparisons between paired format and interviews

Saville and Hargreaves (1999) could have referred to a study by Fulcher (1996a) to bolster their defence of the paired format. Fulcher compared the relative performances of 47 students on a 15 minute paired discussion of a reading article, with two interview based tasks, one a picture based FCE task, the other based on an IELTS interview task. The survey of the test-takers found that the students "overwhelming" perception of the paired discussion was that it elicited more "natural" language, allowing them to "say what they wanted" as opposed to what an interviewer may ask (Fulcher,1996a, p. 33). Of the three tasks, overall the students preferred the paired one. The fluency of the students was rated by five raters, and these scores were used in a G-study and Rasch partial credit analyses. These statistical analyses showed that while there were differences between the tasks, they were not sufficient to rule out generalizability from one to the other. Since the results show that at least in terms of the particular tasks used in this study, being assessed in a pair will result in scores not significantly different to scores from an interview test, this study can be seen to be limiting the impact of criticisms by Foot (1999a, 1999b) and Norton (2005) (see Section 2.2.4.3). Nonetheless, such a specific operationalization of the paired task as used in this study may constrain the generalizability of the results.

A more robust response to the issues raised by Foot (1999a, 1999b) was provided by Egyud and Glover (2001), who found strong support from their Hungarian high school test-takers when they were surveyed about the format. Against Foot's (1999a, 1999b) claim that test-takers might not be able to show their ability when working in pairs, they provide examples of language produced in response to the same task done in an interview and in a pair. In their view, the language proficiency interview often turns out to be more like an "interrogation" (Egyud & Glover, 2001, p. 74) than an assessment of language. Although the performance in the paired format supports their case, it would have been stronger if the same candidate had been involved, and if they had been more systematic in their analysis. They also point out that in their practical experience, the paired format works just as well for low level test-takers as it does for those who are at a higher level. While they are passionate in their defence of the paired format, the evidence they use to support it is little more than anecdotal. Since they only asked 14 students and it appears the survey was conducted informally, it was still not convincing evidence.

A more systematic analysis of the language produced in paired assessment compared with the interview would have to wait until Brooks (2009). Brooks compared scores and analysed the discourse of 16 students who took both paired and interview tests and found that overall they scored significantly higher in the paired than the interview test. Nine of the 16 test-takers achieved a higher score on the holistic scale in the paired test than the interview, three had the same score and four did better in the interview than the paired test. When Brooks examined the features of interaction she found that of the 17 interactive functions that occurred in the two tests, seven only occurred in the paired interaction, demonstrating that students use a strikingly wider range of language when their interlocutor is another test-taker rather than an interviewer. These functions included such co-constructed features of conversation as prompting elaboration, finishing sentences, referring to partner's ideas, paraphrasing and managing the topic. Moreover, the negotiation for meaning that occurs in a paired format coerced the test-takers into producing more complex language than was required in the interview. Even students who scored lower in the paired test than the interview used a greater range of interactive functions when talking to their peers. That the richness of their interactions was not reflected in the holistic scoring band used in the study prompted Brooks to call for rating scales that include such features, pointing out that the ability to co-construct is a real-world skill that should be seen as a challenge to be included in an assessment, rather than a threat to the construct being measured. Indeed, given that the OPI was also criticized for its holistic scoring scale (Bachman, 1990), using analytic scales might seem an obvious and necessary part of an effective speaking test.

Interview tests and paired discussion tests were also compared in research by Plough, MacMillan, and O'Connell (2011) in which 39 Greek first language students of English took part. The purpose of this research was to compare a newly developed paired test with an established interview test. The researchers investigated the scores the students got from taking the paired test as a 'practice test' before doing the interview test, to answer research questions about the complexity, range of vocabulary and linguistic functions the different tests elicited. No significant differences were found for complexity as measured by clauses or sub-clauses per AS-unit[2]. Lexical diversity (LD) was measured by a standardized type/token ratio procedure, and although the paired test had a higher ratio, it was not significantly different. However, consistent with Brooks (2009), the paired test elicited both a greater range and number of linguistic functions than the interview test, with 34 different functions for the paired test compared with the interview's 21. Not only this, but the authors argue that some of the functions that were only used in the paired test, for example, negotiation, presenting and summarizing, are more challenging than the typical functions in the interview, such as 'question answering'. This is all the more impressive when it is considered that any 'warm-up' effect would have benefitted the performance on the interview test.

Such comparative research is useful for illuminating how the paired format differs to other speaking tests and providing impressions of them from a test-takers point of view, but it is necessary to understand how other factors mentioned by Foot (1999a, 1999b) affect the language and scores of those taking part, and this is what the next section will look into.

2.2.4.3 Impact of other factors on paired interactions

One of the first articles that outlined potential negative effects of using a paired format was an article by Norton (2005). The data was collected for an investigation into why Japanese test-takers seemed to underperform in Cambridge's paired speaking assessments. In her study, 27 Japanese candidates were either paired with each other or with one of six European test-takers at either Cambridge FCE or CAE level. Based on this data Norton notes several ways one candidate can affect another's performance. She found that candidates who use particular grammatical forms or vocabulary may influence their partners to use the same structure, as happened in one of her tests where there were a particularly high proportion of conditional statements used to link what they say. She also finds examples of pairs who have a high degree of familiarity with each other who also did well in certain bands in the scoring, and notes that three of the four top scoring Japanese in her sample were with partners they knew well, although Norton does point out that her sample size is too small to make firm conclusions. Finally, Norton notes some cultural phenomenon that may affect candidate performance. When Japanese

[2] AS-unit is an abbreviation of Analysis of Speech Unit, and defined as which is defined as "an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn & Wigglesworth, 2000, p. 365)

females were paired with male candidates she found that they tended to adopt a supportive role by providing back-channelling tokens and letting the male speak first. In support she points to three of five Japanese women speaking less than their male partners.

While Norton's report does point out some potential problems, the support she provides for them is often deficient. For example, she gives examples of candidates influencing each other's grammar and lexis, and she infers that the highest scoring Japanese candidate benefitted from being paired with her high scoring partner, but she cannot show that this Japanese test-taker would not have achieved a high score anyway. The data that she presents in her report cannot provide support for claims about the effect on the test in terms of the construct being measured or the way it is scored. Also, her evidence to support her claim that the scores of Japanese females may be affected adversely by their cultural predisposition to give males precedence is that three out of five Japanese candidates in her study spoke less than their male partners. This falls far short of being persuasive, firstly because three out of five cases is not a substantial number on which to make such a claim, and secondly because she fails to provide qualitative support for what else was happening in these conversations. As Lazaraton (2006) puts it in her acerbic critique of this article, Norton (2005) leaves herself open to the accusation of making "unsubstantiated quantitative claims" (p. 288).

The effect of the proficiency of a test-taker's partner has not surprisingly been a common concern of those involved in paired assessments. Iwashita's (1996) pilot study used 17 Japanese as a foreign language students who were paired alternately with other similar and different level students. Due to the small number of students, no tests of significance were used, but the general results were that high level students got higher scores with similar proficiency partners, and low proficiency students scored better with a higher level partner, and Iwashita related this largely to the quantity of talk, which followed the same pattern, though with more variation. However, due to the small numbers and the variation noted, firm conclusions cannot be drawn from this study.

A thoroughly conducted study by Csepes (2002) did not suffer from such problems. Csepes used 120 Hungarian high school students who were split into four groups to ascertain the effect of partner proficiency. Here, 30 core students were tested with 30 who were higher, 30 who were lower and 30 the same level as themselves, but no significant difference was found in their scores. This study provides strong evidence that when conducting paired tests with the same cultural background and well trained raters, the scores the test-takers receive will not be affected by the proficiency of their interlocutors.

Studies conducted since then have supported Csepes's findings. In research conducted with 20 Chinese students by Davis (2009) the test-takers did the test once with an interlocutor of the same proficiency level, and once with one at a higher or lower level. Davis found that there were significant differences in the language produced, with lower proficiency candidates producing 35% more words when being tested with a higher proficiency partner, a similar pattern to Iwashita (1996). Nonetheless, there was little difference between the scores awarded no matter who they did the test with.

Further confirmation of different proficiency levels not affecting paired test scores can be found in Bennett (2012), using Italian participants, who were, unlike the other studies, mostly unfamiliar with each other. The first part of her study was a survey of 43 Italian EFL students who were practicing for a Cambridge paired examination. Of those surveyed, 40 believed that the proficiency of their partner could affect their score in the paired examination, but in practice, of the twelve students who did tests with similar level and different level students, for nine of them there was no difference of greater than 0.5 points over the two tests (Bennett, 2012, p. 342). Of the other three candidates, Bennett could find no consistent trend in the differences, though from the evidence garnered from responses in the questionnaire and the scoring she concludes that one of the candidates was particularly shy and lacking in confidence, and this may have caused her to produce a weaker performance with a higher scoring partner. Although this study had a low number of students, it has consistent results with previous studies that examined proficiency of partner as a possibly confounding factor in the scoring of paired speaking tests.

That Bennett ascribed shyness as a factor that had an impact on her candidate's performance has support in the literature. Berry (2004) investigated this issue using 32 male and 22 females who had been classified as either extreme extravert or extreme introvert on the Eysenck Personality Questionnaire. They were put into 27 pairs, of which nine were homogenous introvert or extravert pairs and nine were heterogeneous pairs, and they also were scored on individual tasks. The students had been in the same classes for almost a year, so they were at least familiar with each other, but they were not allowed to self-select their partners. To avoid effects of gender, there were no mixed sex groups. From these groupings it was predicted that both extraverts and introverts would do better in homogenous pairs. However, no significant effect for this was found and the hypothesis was rejected. Evidence was found that extraverts performed significantly better when in homogenous and heterogeneous pairs than when they did an individual task.

Berry's (2004) study included the impact of interactivity on shyness as the three tasks they were tested on were rated for requiring low, medium and high levels of interaction to complete. It was hypothesized that introverts would do best on the least interactive tasks and vice versa for extraverts, and significant differences were found, confirming the hypothesis. The discourse features that Berry examined were Lazaraton's (1996) list of eight accommodations that native speaker interviewers make. Berry found that the candidates who used the greatest range were female introverts, and the least were male introverts. It appears the raters responded to this range of features by awarding this group with higher scores. This study has made an important contribution to the validity of peer interaction speaking tests by informing the possibility of this kind of bias. Still, it is not clear from this study whether the raters were affected by the actual language elicited by the test, or if it was the contrast between the introvert and extravert pairings that influenced them.

Familiarity with one's partner is another factor that has been thought to make a difference to the language elicited in a paired speaking test. This was investigated by O'Sullivan (2002), who had 24 female and 8 male Japanese university students performing in two paired activities; one with a friend and one with somebody they had not met before. O'Sullivan also included the variable of whether the stranger was a male or female, but the low number of males meant that all possible male-female-stranger combinations could not be included. The participants did the first test with a friend of the same gender, but only eight could do the second test with a previously unknown male, and sixteen with a previously unknown female. The results showed that when grouped with a friend, the performance of Japanese females was significantly superior to their performance with a stranger, although the actual difference in the score was small. There was a trend towards worse performance for females if grouped with a male stranger than a female stranger; however, the statistical evidence was not strongly significant, and the overall impact in terms of the score was small. When grammatical complexity and accuracy were examined, it was found that the complexity of their utterances did not differ between the two groups, but accuracy showed a significant improvement if working with a female unknown person, and a non-statistically significant decrease with a male. Although the results of this study may have been affected by socio-cultural background the Japanese participants shared, and the small numbers involved have to be considered a limiting factor, it does demonstrate that scores and language in paired interactions can be affected by the familiarity of the participants.

2.2.4.4 Analyses of interaction in paired formats

The interaction of the paired test and its relationship to the scores in the interactive band of the analytic scoring scale used in Cambridge FCE tests has also been investigated in a ground-breaking study by Galaczi (2008). Her study comprised of 30 paired tests, using participants of a variety of nationalities on their performance in the interactive task of the test. She analysed the conversations using the dimensions of mutuality, the extent to which the conversation was co-created; equality, the balance of topics that were initiated and developed; and dominance in terms of the count of words spoken, interruptions and questions. Using these categories she identified three different patterns of interaction: collaborative, parallel, asymmetric; as well as blends in which different patterns combined. When she investigated the dyads in her study, 30% were collaborative, 30% were parallel, 23% a blend of collaborative/parallel, 7% a blend of collaborative/asymmetric, and the remainder were asymmetric (10%).

The collaborative dyads were characterized by high mutuality and equality as the participants smoothly switched roles from speaker to listener. The conversational move that most characterized their talk was the ‘topic extension’ move, in which they could signal their involvement with the other speaker and then develop the conversation. Within this category, Galaczi (2008) found two subcategories: those with high conversational dominance, and those with low. The high conversational dominance dyads were differentiated most clearly by their use of interruptions, competition for the floor and questions, which limit the contribution of their partner by imposing on them the necessity to give a relevant answer. This contrasted with low conversational dominance pairs in which the focus was more on the interlocutor and questions were much fewer and used to extend the topic.

Parallel interaction was characterized by a lack of engagement with their interlocutor's ideas, the test-takers sole concern being to develop their own ideas. Although there was high equality this was accompanied with low mutuality. Galaczi noted two subtypes: low dominance interactions in which there were long gaps between turns, and high dominance ones in which there was competition to speak, and interruptions were frequent. Asymmetric interactions are those which are marked by an imbalance in the quantity of talk as one of the pair dominates the discussion. This can be a result of either 'expert/novice' or 'dominant/passive' relationships between the test-takers. The 'expert/novice' relationship is one of low conversational dominance, as the dominant partner coaxes the partner to take part by asking questions or by allowing openings for his partner to participate at points of speaker transition. In dominant/passive relationships, the partner who does most of the talking and dominates the interaction does not give opportunities to their partner to take part.

Although Galaczi (2008) looked at the scoring of these, firm conclusions could not be drawn because the raters based their scores on the entire test, not just on the interactive section of the test that was the subject of this study. Nonetheless, she found that the dyads in her study that had a collaborative interaction with low conversational dominance had the highest average scores in the relevant scale, and these were followed by collaborative interactions with high conversational dominance. Next highest equal scoring were those with blended or asymmetric interaction, with the dominant partner doing better than the passive one. The lowest scoring were those with parallel interaction patterns. Although the relationship between scores and the scoring scales is not direct, this does seem to show that the raters are influenced by these different kinds of interaction.

As Galaczi (2008) points out, there are also implications for the acquisition of conversational skills, implying that the final stage of development is the low dominance, collaborative interaction that was rewarded mostly highly by the raters. However, this point of view may take a rather idealized view of what 'conversation' is. The possibility needs to be considered that some people simply have a certain conversational style that would not garner the top grades awarded, and indeed conversational styles have been found to differ from culture to culture (see Clancy, Thompson, Suzuki & Tao, 1996), raising questions of fairness. This point has been addressed somewhat in later research by Galaczi (2013) that is covered in Section 2.2.4.6.

Another limiting factor that recent research has revealed may be the time available to the participants. Research by Nitta and Nakatsuhara (2014) demonstrates that collaboration among the pairs in their study was built over a five minute period. A test time of 3-minutes may not have been sufficient for some of Galaczi's test-takers to establish a collaborative pattern.

Conversational patterns were also investigated by Sandlund and Sundqvist (2011) who researched Task Related Trouble (TRT) among five low ranking dyads and five high ranking dyads from 40 that took a paired interaction test in Sweden. The researchers investigated three sub-skill scales that were most closely related to topic management: overcoming difficulties in communication, interactional ability, and treatment of topic. The task consisted of questions or statements on cards which the dyads should discuss until the topic is exhausted, in which case they draw another card. The high scoring students in the sample confirmed the procedure of the test with the teacher, and managed the interaction smoothly by introducing the topic and seemed to reach a mutual understanding of when conversation on the topic was exhausted. When TRTs occurred in the transcripts they found that ambiguities in the task itself could be the source for some students. When it came to rating the participants, those who oriented to the task excessively were not valued so highly. In one example, a dyad seemed to focus on the task as one in which reaching agreement or disagreement acted as a signal for the closing of the topic and that a new card should be drawn. Scoring lower than these students were those who abandoned the task due to non-understanding of a word in the prompt, and finally those who 'resisted' the task by using their L1s in the face of uncertainties over the task. To Sandlund and Sundqvist (2011), the teachers' understanding of what constitutes good and bad task management also played a part in the distribution of grades. Sandlund and Sundqvists' (2011) study shows how important it is to have clear procedures for the test and practice materials for the test-takers so that the final results represent their abilities more fairly.

In an interesting conceptualization of the problem of assessing a co-constructed endeavour such as conversation, Lazaraton and Davis (2008) posited the existence of a Language Proficiency Identity (LPID) through which test-takers position themselves as "proficient and competent" speakers of English (p. 318). To do this they examined four examples of test-takers displaying their proficiency in various ways, first looking at the scores they received then looking for evidence by Conversation

Analysis (CA). The first example shows how a dyad which had the highest scores represented ‘being proficient’. The two speak with remarkable coherence, using turns that are lengthy, well-constructed, with clear discourse markers, and that build on each other’s ideas collaboratively. As Lazaraton and Davis (2008) put it “doing ‘being proficient’ positions them as competent English users, thus deserving of high marks on the speaking test” (p. 321). Their next example shows ‘doing interactive’ by a pair that received their highest scores in the interactive band. The conversation is notable for its short rapid turns in which the test-takers display their ability and confidence by their overlapping speech, laughter and by one inviting the other to travel with her. Through this interactivity, their proficiency could be demonstrated, and thus rewarded by the raters. Lazaraton and Davis (2008) also show how a test-taker can present proficiency when there is a mismatch in abilities between the pairs. The higher of the pair does ‘being supportive’ and displays her proficiency by supporting enthusiastically what her lower proficiency partner says, even though the researchers admit that what he said was “incomprehensible” (p. 325) to them. She affirms her partner’s contributions and uses them to build in her own opinion, showing herself to be a sympathetic, supportive partner, who should be, and was, rewarded with high scores by the raters. The final example is of a test-taker doing ‘being assertive’, as she commands the conversation by taking the opportunities presented by her partner’s inability to contribute to collaboratively complete what he was trying to say, and move the conversation along. In the end, both participants were awarded the same score, and while Lazaraton and Davis can understand why the test taker doing ‘being assertive’ got a high score in this category, seem mystified as to why the passive interlocutor was awarded the same score.

When considering the potentially useful construct that is the LPID, it is important to be aware that in these paired tests the extent that an individual’s LPID is allowed to flourish may depend on the interlocutor. For example, the dominant partner through the agency of proficiency was considered to be ‘doing supportive’ because of her partner’s low ability, but Lazaraton and Davis (2008) do not comment on what LPID her partner was trying to display. It seems the extent to which an LPID can be displayed would be firstly limited by an individual’s proficiency, and then by the proficiency and willingness of the interlocutor to co-construct the conversation. As Lazaraton and Davis (2008) note “that proficiency identity can also be influenced by a partner through his or her appraisal of the displayed proficiency and that appraisal can be ratified, maintained or modified – in and through the discourse" (p. 333).

Such studies as the above are leading the way in uncovering how pairs interact in assessment contexts. Other research has focussed on how the task facet may influence a paired discussion. In a recent study by Nitta and Nakatsuhara (2014), the effect of planning was investigated quantitatively and qualitatively. They put 32 English major students at a Japanese university under the conditions of either 3-minute or no planning time. In their procedure they allowed the test-takers, who were paired with the same gender partner of their choice, to warm-up by introducing themselves, and then gave them two FCE interactive stage tasks under the two conditions. The quantitative analysis revealed a small but significant advantage for those in the planned condition in rated scores for complexity, accuracy and fluency. In the indices used however, there was no significant difference in the measures for complexity (clauses per AS-unit), accuracy (errors per 100 words) or lexical diversity as measured by Measure of Textual Lexical Diversity (MTLD). Significant differences were found for fluency, with the unplanned condition having significantly higher words per second, but with longer pauses per second, as well as fewer words per turn.

The most interesting findings come from the CA analysis of two test-takers. In the unplanned condition their turns built momentum as they initially used short turns to improve their understanding of each other and the topic, before using longer turns to explain their ideas towards the end of the test time. By contrast, the test-takers under the planning condition had longer turns at the beginning of their discussion, but the exchange resulted in the parallel development of topic as each test-taker reproduced the ideas they had generated in their planning time, rather than develop the conversation collaboratively. Following this was a "stagnant middle" (Nitta & Nakatsuhara, 2014, p. 17) in which their original ideas had become exhausted and they searched for ways to continue speaking. This explains the seemingly paradoxical result from the questionnaire the test-takers took, that those in the planned condition reported that generating ideas for their talk was more difficult than for those in the unplanned condition.

In Nitta and Nakatsuhara's (2014) study there seems to be a contradiction between the raters scoring fluency higher in the planned condition despite the index of speed fluency in the unplanned condition being significantly higher in the unplanned condition. Indeed, the raters also scored complexity and accuracy significantly higher even though the indices showed no significant difference. A possible explanation can be found in the significantly longer turns of the planned condition. This may reflect two relevant factors: firstly raters may find it easier to score fluency in longer turns since extended lengths of discourse may be considered more salient to the construct of fluency; and secondly, in shorter turns speakers may be able to deploy a greater proportion of memorized chunks of language, allowing them to be spoken more quickly. For example, in the quoted examples one speaker in the unplanned condition had single turns that consisted of the pattern "I think X is not important for me" twice within a short period of time (Nitta & Nakatsuhara, 2014, p. 14 and p. 15) with evidently little hesitation, and such a memorised chunk would contribute to a higher fluency index. However, in a longer turn a speaker has to connect memorised chunks together into a coherent reply, which will inevitably introduce a greater proportion of pauses into the speech. It would have made the comparison between fluency in unplanned and planned conditions fairer if they had only counted turns that were longer than a certain period of time, such as 10 seconds in length. Nonetheless, the finding that planning time may be detrimental to a candidate's ability to show their ability to collaboratively build a conversation as it causes them to rely overly much on the speech they have prepared in their planning time is a valuable contribution to the field.

In this dissertation, the test-takers of the GOT have a planning time of 1 minute as opposed to the 3 minutes in Nitta and Nakatsuhara (2014), and yet even so the initial turns of GOTs often have features similar to the parallel development noted here. For example, Van Moere (2007), whose data was collected at the same institution as that in this dissertation, noted that this pattern "usually occurred towards the beginning of discussions, when the understanding among candidates was that each of them had something to contribute because they had had one minute to read and think about the prompt before the discussion began" (Van Moere, 2007, p. 325). Leaper and Riazi (2014) noted that in GOTs it depended to a certain extent on the prompt. The test-takers who produced the longest turns did so to a prompt that elicited their family history, and these GOTs were almost invariably started by an exchange of lengthy turns; the prompt with the highest number of turns was on the more factual subject of mobile phones, and were often started by a test-taker posing a question to the other group members.

Given that the research reviewed has revealed aspects related to the co-constructed nature of the paired format, it becomes necessary to investigate how this affects the rater. The next section reviews literature on this topic.

2.2.4.5 Raters and the paired format

This section is on the literature dealing with the issue that was noted in the first section of this dissertation: How can a rater give a score to the individual when it is recognized that interaction is jointly constructed? This question is highlighted in May's (2009) study of a dysfunctional pair in which one of the test-takers dominates long sections of the exchange while ignoring her partner's contribution, even to the extent of sitting in a position oriented away from him. In this test, the dominating test-taker exhibits such brusque behaviour as cutting off her partner and seeming to ask for his opinion, but then not waiting for his response. In the final turns of the test she summarizes their conversation and ends it peremptorily with the use of the first person pronoun without regard to her partner: "I think it's more benefit for human society… that's all" (May, 2009, p. 411). One of the raters in the study describes this interaction as an 'interview', which May (2009) seems to agree with (p. 413). The 'situationally located' nature of the scoring of this interaction is displayed by the dominated test-taker being awarded a failing score of 2 out 5 on the 'effectiveness' band which is related to interactive skills in this test. Yet, in a parallel discussion task conducted at the same time in which the same test-taker collaborates effectively, the same rater awards the test-taker a passing mark of 4 out of 5.

When the raters discussed the grades their struggle became evident, with debate over which test-taker was to blame for the asymmetric interaction. One suggestion to resolve this situation that May (2009) makes is to include a joint score that characterizes the interaction. While this may be a step in the right direction, a further solution may be to make the objectives of the peer interaction clear to the students so that they know what is expected. In any event, as May points out, there was little guidance as to what the raters should do in such a situation, and so at the very least administrators of any test that involves peer interaction should set clear policies on how to deal with asymmetric interactions.

Dealing with the co-constructed nature of conversation was explored further in May (2011). In this article May used rating notes, verbal recalls and discussions to investigate how four trained raters scored paired format tests by 12 Chinese university students in the band that referred to their interactive skills. Given that the rating scales reduced ‘interaction’ to three aspects related to *understanding* the speaker, *responding* appropriately and using *communicative strategies* (that were not specified), May thought it not surprising that the raters were forced to incorporate aspects of performance not mentioned in the bands. With *understanding*, the raters found it problematic when they did not understand what the test-taker said, but the interlocutor apparently did. They had to interpret whether the lack of understanding was due to the speaker not explaining clearly enough or the listener for not comprehending. The raters reacted positively to use of backchannel to indicate interest in what the other was saying. In the *responding* category, the raters were positive when the speaker developed a partner’s idea since it showed the ability to engage with their interlocutor’s ideas. On the other hand, minimal and irrelevant responses counted against the test-taker, and from such characteristics emerged the profile of a “passive speaker” (May, 2011, p. 135). The ability to challenge their interlocutor was another positive feature of a test-takers ability to *respond*. Although ‘authenticity’ was not mentioned in the bands, raters drew upon this notion to describe interactions that flowed with inclusive and cooperative participants, as opposed to those that were not, and these were often with parallel development of topic in which the test-takers used monologues and talked past each other. Non-verbal behaviour was often taken into account, with ‘open’ stances being positively viewed, along with eye contact and gestures to emphasize points. The term ‘communicative strategies’ was referred to as the ability to ask clarification questions, ask for the opinion of their partner and facilitate the interaction through the use of functional language. Finally, May briefly notes some aspects that tended to be referred to as mutual achievements rather than as individual accomplishments. These include understanding and responding to the interlocutor’s message, working co-operatively and contributing to the authenticity or quality of the interaction. It is disappointing that May does not discuss these to the same length as the other rater comments.

The finding that such elements were drawn upon by raters echoes an earlier study by Orr (2002) who collected verbal reports from 32 raters of two pairs of FCE test-takers, and found that the scores assigned varied widely. Orr gives three reasons for this: firstly it was evident from the comments that the raters had different standards in severity and applied the scales incorrectly; secondly, even when they applied the scales, they did not focus on it in the same way as each other; finally, they varied in the extent to which they heeded criteria that was not mentioned in the scales (p. 149). When Orr tallied the number of non-criterion aspects referred to in the comments, he found that they ranged from a mere 2.78% for some who obviously paid careful attention to the scales, to 78.95% who seemed to barely use the provided scoring bands at all. These comments were further categorized: Orr found that the three most often used non-criterion information were references to the test-takers physical presentation, followed by their global impression, and then comparison with another candidate. As Orr points out, improving this problem involves give and take: on one hand, it is an issue that rater training has to improve; on the other, it also showed the difficulty of applying these particular rating scales, suggesting that reform to what raters actually pay attention to may be a necessary consideration for consistent scoring.[3]

The above mentioned studies found that raters often draw on aspects outside the rating bands that they are supposed to apply. This may be due to the rating scales being “based on theory with little empirical justification” (Fulcher, 1987, p.291) as has all too commonly occurred in the past. One source of empirical data would be to investigate what raters pay attention to when assessing language (Upshur & Turner, 1995) and this was the purpose of the paper by Ducasse and Brown (2009). They used videos of 17 pairs of Spanish as a second language test-takers, three of which were sent to 12 raters who were told to comment on the interactive success or otherwise of the speakers, purposely without defining what ‘interactive’ meant. From the verbal report data Ducasse and Brown (2009) derived three main categories, each with subcategories. These are summarized in Table 4.

[3] The FCE rating scales have since been updated (Ffrench, 2003).

Table 4: Empirically derived rating criteria from Spanish as a foreign language pairs

Major Categories	Sub-elements	Definition
Interpersonal, non-verbal communication	gaze, body language	An interlocutor in a pair demonstrates communication through non-verbal communication using gaze and body language
Interactive listening	supportive listening, comprehension	An interlocutor… demonstrates communication by indicating comprehension to the speaker or through supportive audible feedback
Interactional management	horizontal cohesion (turn-taking), vertical cohesion (topic management)	An interlocutor in a pair demonstrates communication in the current turn or over different topics

Adapted from Ducasse and Brown (2009)

As suggested by the previously reviewed studies on raters of paired format speaking tests (May, 2009, 2011; Orr, 2002), body language and eye contact play an important role in determining the success of the interaction, and yet it is rarely included in the scoring scales, as indeed it is not present in the rating scales used for the GOT in this dissertation. The raters in Ducasse and Brown's study pointed out positive aspects such as "they look at each other and never lose thread of the conversation" and negative: "they look at the paper and not at each other" (2009, p.434). Hand gestures too could be seen in a positive light, as noted by one rater who commented that they help "her interaction to be more positive and fluent" (ibid); or negatively: "there are too many gestures and that gives me the impression that they lack verbal resources" (ibid).

In the listening section, *comprehension* included demonstrating verbal support for the interlocutor by asking questions for clarifying or understanding what was said, and use of backchannel falls into the supportive listening subcategory. The use of backchannel is another aspect that often does not appear in rating scales, but was an important part of what these raters mentioned. In the interactional management category, *horizontal management* was about allowing the turns to flow smoothly and transition between topics, and includes such aspects as the speed of response, turn length and domination. *Vertical management*, by contrast, is the ability to extend a topic by relating it to what was previously talked about so that the conversation appears coherent.

The descriptions in Ducasse and Brown (2009) provide a useful methodology for providing empirical data for rating scale construction. For this dissertation they provide a useful insight into the various aspects that a rater may be thinking about when assigning grades, whether or not it is actually described in the rating scales. The results must also be treated with some caution since they are from a Spanish as a foreign language context rather than with learners of English, and from paired interactions not group interactions.

2.2.4.6 Development in the paired format

Less well researched has been how language learners develop their speaking skills over time, particularly their interactive skills. Galaczi (2010, 2013) has investigated this aspect on the interactional section of Cambridge ESOL tests, using 26 pairs of students between the Common European Framework of Reference (CEFR) B1 and C2 levels. The pairs were chosen because their scores fell in the middle of Cambridge's scoring bands, as it was thought that high scoring students would display characteristics from the next highest band. The overall findings are summarized in Table 5. Galaczi found that at the lowest level in her study, B1, a typical engagement consists of each test-taker suggesting new topics which they extend themselves, without much engagement in each other's ideas other than a brief token of agreement. There is little listener support in the form of backchannels here as it seems the speakers deal with the high cognitive demand of producing their own turn of speech and they do not have spare processing capacity to actively engage as a listener. Also indicative of this need for processing capacity are the gaps between the speakers. In the example provided, Galaczi considers gaps of 0.5 seconds between turns being "indicative of weak alignment between speakers" (Galaczi, 2013, p.9). Topic shifts for B1 speakers tend to be abrupt as they fail to transition effectively in a "stepwise" manner (Sacks, 1992). The next level, B2, many of these features are still present, but these learners also have some capacity to develop other-initiated topics, and their conversations have some collaborative features, such as the occasional jointly constructed turn. The role of 'listener' has advanced from B1, but it is mostly limited to the use of simple backchannels, such as 'yeah' to show support.

At C1, the speakers have demonstrably higher level of mutuality and reciprocity and can now jointly construct meaning in conversation. They can confidently develop both self- and other-initiated topics, and these can be extended over several turns. Where before active listenership was limited to backchannel, now they are capable of confirming the other's opinion by using short statements such as 'yes, I see' and 'indeed'.

Table 5: Development of interactive features mapped onto CEFR levels

Feature	B1	B2	C1	C2
Topics **Degree of development:** **Topic extension of own vs other:**	- mostly self-initiated, self-extended, soon exhausted, and with abrupt topic shifts	- evidence of other initiated topics, other's topic extension possible - longer life of topics	- confidence in developing other & self- initiated topics	- development of self-initiated and other initiated topics mastered
Listener support **Backchannel:** **Confirmation of comprehension:**	- brief acknowledgement of others ideas - little backchannel, active in role of speaker	- backchanneling present, more active in roles of listener weak but present	- backchannel & confirmation of comprehension, switch listener-speaker roles	- effortless switching of speaker/listener roles
Turn-taking management **Gaps & overlaps:** **Following overlap/ latch gap/pause:**	- gaps between turns indicative of longer processing time - overlaps or latches rare	- no gap-no overlap between turns more usual, - overlaps & latches possible, - fewer pauses between turns, pauses within turns remain	- no gap-no overlap - orient to latching, over-lapping turns, flow of conversation	- short turns, rapid speaker changes, supportive overlaps latches

Adapted from Galaczi (2013)

At the next level, C2, the test-takers are capable of all this but in a smoother and proficient manner. Galaczi admits that the features of interaction of speakers at the C2 level is “very similar” (2010, p. 101) to that seen at C1 level but points to the subtle differences in the wording of the CEFR descriptor bands to justify the distinction between these levels. For C1 the learners can “relate… contributions to those of other speakers” compared with C2 in which they “interweave… contributions into joint discourse” (2010, p. 28).

These findings provide a useful benchmark for comparing the developmental patterns of interactional skills that has been little researched. However, this research also raises further issues that need to be addressed about how interactional ability progresses. Notably as it is a cross-sectional study it leaves open the question as to how learners actually improve in the development of skills over time, and how much time it takes to move from one level to the next. In the development of other language abilities, it has not gone unnoticed that a picture gained from cross-sectional studies may give a somewhat distorted picture of development compared with longitudinal studies (see Pine & Lieven, 1990) and there is no reason to think that interactional skills are any different. Also, the skills and abilities shown by the participants at the described developmental steps are questionable with respect to the lack of smooth progression. An examination of the abilities shown at each step seems to show a large jump in interactional abilities between B2 and C1, and yet the differences between C1 and C2 are described as ‘subtle’. Naturally this comes from the difficulty of taking snapshots of a moving target, and further evidence of development of interactive skills over time will confirm the value of Galaczi’s findings.

2.2.4.7 Summary of research on paired interactions

The literature on paired assessments has made a substantial contribution to our understanding of peer-interaction tasks which are relevant to the GOT. This literature shows that paired peer interaction format tests:

- produce a greater range of language than interviews (Brooks, 2009; Plough, MacMillan & O’Connell, 2011)

- composed of participants at different proficiency levels seem to have little or no effect on the scores (Bennett, 2012; Csepes, 2002; Davis, 2009).
- may be affected by the participants' shyness, familiarity and gender (Berry, 2004; O'Sullivan, 2002) and planning time (Nitta & Nakatsuhara, 2014)
- may produce interactions that can be categorized into certain patterns, of which co-construction tends to be rewarded more highly by raters (Galaczi, 2008, Sandlund & Sundqvist, 2011).
- provide a context in which test-takers can assert a certain identity as a test-taker as they take a stance (Lazaraton & Davis, 2008).
- allow the test-takers to show their ability to interact at different levels (Galaczi, 2010, 2013)

The studies that examined the scoring of paired interaction showed

- the difficulty of rating dysfunctional pairs (May 2009)
- what elements of performance raters may pay attention to (Ducasse & Brown, 2009; May, 2011; Orr, 2002)
- the importance of body language and eye contact to raters (Ducasse & Brown, 2009)

It seems likely that many of the aspects of paired interaction will be similar to groups larger than two, and thus the question may well be posed 'Why test groups with more than two members?' The most obvious answer to this question is a practical one: When you have many test-takers and time is a limited resource it is simply more expedient to assess more than two test-takers at a time, and expediency is of no small concern in language testing. Once extra test-takers are added it becomes necessary to investigate how the dynamic of the interaction changes. Research into paired formats provides a useful reference point for the next section, which reviews the literature of the GOT.

2.2.5 Testing in groups of three or more

2.2.5.1 The early literature on the GOT

The GOT has appeared intermittently in the literature since Folland and Robertson (1976) before the recent upsurge began in the mid-2000s (as summarized in Table 6). The first centre that published on the extensive use of a group oral format was at the University of Tampere in Finland, where Folland and Robertson (1976) recommended it as a solution to large scale oral testing. In the form described by Folland and Robertson (1976), groups could be composed of up to seven test-takers and the time allowed was calculated as five minutes per candidate. They found that an advantage of using this form of speaking test stemmed from the lack of rater involvement in the discussion. This allows the rater to focus on the test alone rather than be distracted by having to ask questions as in an interview. In addition, it removes the interviewer as a source of variability that they had found to negatively affect reliability. They also point out that the lack of an examiner gives the test-takers maximum talking time and seemed to make students more inclined to participate.

Without providing evidence, Folland and Robertson (1976) stated that their students felt less stressed compared to interview tests. Surveys in different contexts have since supported this claim: Hilsdon's (1991) survey found the GOT had high face validity amongst both teachers and students in Zambia, despite concerns about factors such as shyness and the fairness of the rating system. In a high stakes context in China, He and Dai (2006) survey found that 74.5% of 144 candidates were either not nervous or relaxed for their GOTs – a surprising finding considering the importance of the exam. A more medium-stakes GOT conducted at a private languages university in Japan similarly found most of the 1088 surveyed had positive opinions about their experience of testing in the GOT (Van Moere, 2006). The marking system Folland and Robertson (1976) describe counted the number of mistakes and set these against 'plusses' under categories of pronunciation, lexis, grammatical structure and use, as well as the number of 10-word plus turns contributed by each test-taker.

Table 6: Summary of the literature on group oral tests with empirical findings

Study	Test	Test-takers	Main findings
Berkoff, 1985	High school	24 Israel	Students mostly self and peer assessed each other lower than raters, except for 3 lowest level test-takers rated higher
Berry, 2004	University	447 Hong Kong	Extravert test-takers score higher in low extravert groups; Introverts score higher in high extravert groups
Bonk and Ockey, 2003[1]	University	1103 & 1324 Japan	Rasch analysis showed test could discriminate 2-3 levels, prompt plays a minimal role, scores allocated efficiently
Bonk & Van Moere, 2004[1]	University	1055 Japan	Different vs same proficiency levels in a group: no significant difference found, shyness has a small negative impact
Gan, 2010	High school	8 Hong Kong	Higher level group collaboratively engage in the task, lower level group approach as 'question answering task'

Gan, 2011	High school	39 Hong Kong	No significant correlations between extraversion and scores or complexity, accuracy, fluency indices.
Gan, 2013	High school	30 Hong Kong	GOT performance less complex, accurate and fluent than performance in a presentation
Gan & Davidson, 2011	High school	8 Hong Kong	Gestures of higher level students are synchronized to their words, lower level groups they are not.
Greer & Potter, 2008	University	40 Japan	Various uses of "How about you", including its use to give reticent speakers an opening
He & Dai, 2006	CET-SET	196 China	Failed to elicit negotiation & other advanced interactional feature, most common function is question and answer
Hilsdon, 1991	High school	3000 Zambia	No relationship found between paper test score and GOT score; high face validity despite reliability concerns
Kobayashi & Van Moere, 2003[1]	University	60 Japan	Number of words in a group has no effect, number of turns has no effect, words an individual speaks related to score, mostly in Comm. Skills band
Leaper & Riazi, 2014[1]	University	141 Japan	Prompt influences turn-taking patterns, fluency, complexity of responses
Liski & Puntanen, 1983	University	698 Finland	Most mistakes were in grammar, those who talked more scored higher & made less mistakes, females scored higher
Liski & Puntanen, 1985	University	639 Finland	Statistical model for error frequency was developed, with the dependent variable following a compound Poisson distribution
Luk, 2010	High School	43 Hong Kong	Test-takers managed interactions, gave the most content in initial stages, few disagreeing moves, reluctance to negotiate
Nakatsuhara, 2011	High School	269 Japan	Extraverts play a bigger role in groups of 4 than 3, groups of 3 more collaborative, supportive of low proficiencies
Negishi, 2010	University Junior & High school	135 Japan	Lower level students use 'asking for information or opinion' more than higher levels, modifying or developing moves show clear development sequence from low to high
Ockey, 2009[1]	University	228 Japan	Assertive speaker score less when grouped with other assertive speakers, higher with non-assertive speakers
Ockey, 2011[1]	University	360 Japan	Assertive students have a small overall advantage, self-consciousness has no effect.
O'Sullivan & Nakatsuhara, 2011	High school	269 Japan	Use of measures of quantitative dominance, goal orientation, interactional contingency as distinguishing features of GOT
Pavlou, 1997	High school	60 Cyprus	GOT scores correlated the least with scores on other formats
Shohamy, et al 1986	High school	103 Israel	GOT scores correlated the least with scores on other formats, showing it measures different quality
Van Moere, 2006[1]	University	113 Japan	Inter-rater reliability was .74; variation due to test-taker performance or group dynamics
Van Moere, 2007[1]	University	63 Japan	Group discussion elicits fewer long turns, mostly asking and providing information compared to picture difference or consensus tasks

[1] Data collected at the same institution as this dissertation

Data from this test were subject to a detailed statistical analysis, as described in a series of articles (Liski & Puntanen, 1983, 1985). In these group discussions males spoke significantly more and achieved more plus marks than females, who made significantly fewer mistakes per 10 utterances. Overall, those who spoke more tended to get higher grades (Liski & Puntanen, 1983). A sophisticated statistical model for the scoring data was outlined by Liski and Puntanen (1985) in which a student's overall ability could be described in terms of a weighted sum of Poisson variables. Whatever is thought about using the number of errors and "plusses" to determine speaking proficiency, it was a feature of this university for many years, with at least 5000 students being tested between 1973 and 1983 (Liski & Puntanen, 1983).

Another major centre in these early years was in Israel, where the GOT was part of an experimental battery of oral tests with the intention of reforming the high school examination system (Berkoff, 1985; Shohamy, Reves & Bejarano, 1986). Berkoff (1985) describes a system in which four high school students select one of 20 cards that have two or three questions on a subject that students are expected to have opinions on, and had been discussed during the students' school year - similar to that described by Sandlund and Sundqvist (2011), as described in Section 2.2.4.4. They are allowed to firstly discuss whether they want this topic or to choose another one before the 15 minute discussion begins. Berkoff then got 24 students who had been tested in six groups to self- and peer-assess and found that all but three of them rated themselves lower than the grades given by the raters. Those who had a higher peer evaluation than their final grade also happened to be the three weakest candidates, which Berkoff notes may have been done "with some element of 'pity'" (1985, p.98). This interesting finding shows how being part of a group may give rise to feelings of solidarity for their peers (and foreshadows a study by Luk [2010], as described in Section 2.2.5.4).

An early paper by Hilsdon (1991) described how the GOT was used to test large numbers of students in Zambia's high school examination system. As well as reaffirming many of the advantages Folland and Robertson (1976) mentioned, Hilsdon (1991) pointed out the positive washback the GOT had in encouraging teachers to have their students practice by talking in groups. Problems were also reported, particularly with respect to the language it elicited which transcribed tests revealed to be mostly "imparting and seeking factual information" (p.191), a harbinger of later studies' findings (He & Dai, 2006; Van Moere, 2007). Hilsdon noted that the interactions were often conducted by the test-takers taking their turns in hierarchical order. She pointed to the Zambian culture of not encouraging females or learners to speak out of turn as a factor that may have inhibiting more natural discourse, but similar turn-taking practices have been noted in GOTs in more recent literature (Luk, 2010; Van Moere, 2007). Hilsdon reports that a study that correlated the oral scores of 3000 test-takers with their written test scores found no relationship, which was interpreted as meaning that the test was indeed "assessing the interaction rather than reflecting other test scores" (p.193). Given the difficulties she reports in her article, it is questionable the extent to which this conclusion was justified.

Bonk and Ockey's (2003) paper added quantitative support to the noted ability of the GOT to asses a large number of students' oral proficiency per administration (see Berkoff, 1985; Folland & Robertson, 1976), and can be seen as the forerunner of the current research interest on GOTs. This was a large scale many-faceted Rasch analysis of GOTs conducted at a private languages university in Japan. This study used data from two separate administrations taken in consecutive years of 1103 and 1324 students respectively. The analysis took into account the difficulty of the prompt and rater severity and found that the test could separate the candidates by ability, but only into 2 or 3 levels of the rating bands. One possible reason Bonk and Ockey gave for this was the uniformity of the student population being studied, although it could also have been a result of other factors, such as unclear wording in the rating bands, for example. The effect of the prompt was found to be negligible and although the raters varied considerably in terms of their severity, this could be ameliorated using Rasch. These findings suggest that the GOT is a reliable and efficient method of judging the oral ability of large numbers of students in a relatively short time. Though of course, Bonk and Ockey (2003) cannot take into account what individuals in the test are doing at a discourse level.

2.2.5.2 Comparisons with other formats

Not surprisingly, a theme of the early research was how the GOT compares to other oral tests. The first to do this was a study that reported the results of a variety of oral tests taken by 103 Israeli students and found that although their version of the GOT was not quite as well liked as interview and role play formats, it was still rated favourably (Shohamy, Reves & Bejarano, 1986). The differences between the formats were small though tests for statistical significance were not employed. This study also found that the GOT had the lowest mutual variance compared to the other tests (interview, role-play, monologue reporting), showing that it captures a different quality to them. As such, the authors recommended the four tests, including the GOT, be used as a battery of tests.

Pavlou's (1997) study compared 60 high school students who took the same set of formats as Shohamy, Reves and Bejarano (1986) -- an oral interview, group discussion, role play and an oral report. He found that correlations between their scores on the different tasks correlated very highly – between 0.876 and 0.979. Similar to Shohamy, Reves and Bejarano (1986), Pavlou notes that the correlations for the group discussion were consistently the lowest amongst the tasks, lending support to the conclusion that it tests different skills. The lack of significant differences between scores on the different tasks is also consistent with Shohamy, Reves and Bejarano (1986). Pavlou attributes this to the use of the same scale, which masked differences found in the language produced in these tests. It should be pointed out though, that this finding is not unusual; other studies have found that test scores to be insensitive to changes in prompt or format (Brooks, 2009; Fulcher & Reiter, 2003; Leaper & Riazi, 2014).

The most recent article to compare the GOT with another format is Gan (2013), whose study included thirty 14-16 year old Hong Kong high school students whose speaking test includes not only a GOT but also a presentation, and for both of these tasks they had about 10 minutes of planning time. They found that among the fluency indices there were mixed results, with test-takers being less fluent in the GOT than the presentation according to a significantly higher number of filled and silent pauses per minute with medium effect sizes; but with a significantly higher speech rate, albeit with a small effect size. The presentation task showed more complexity in its significantly higher verb phrase ratio, longer T-units[4] and utterances than the GOT. Interestingly, although the presentation utterances were significantly longer in the presentation, contrary to Gan's expectations, the ratio of clauses to T-units and dependent clauses were actually higher in the GOT, though not significantly so. This phenomenon

[4] "a nuclear sentence with its embedded or related adjunct" (Harrington., 1986, as cited in Gan, 2013, p. 6)

was explained by a high rate of usage of utterances prefaced with ‘I think’ in the discussion task. Finally, in an index for accuracy as indicated by error free clauses, the presentation task again had significantly less errors than in the GOT. Gan (2013) contends that an important reason for the less complex, accurate and in some cases less fluent performances on the GOT was that it demands more attentional resources, meaning these test-takers had less to spend on producing fluent, accurate and complex stretches of talk as they could in the presentation task.

As well as comparisons to other formats, different tasks within the GOT format have also been compared. In illuminating research conducted by Van Moere (2007), 63 test-takers performed two or three different tasks: a group discussion of a prompt, a consensus-reaching task, and a picture difference task. The group orals were analysed by interactional function, words and turns, and then by Conversation Analysis (CA). Van Moere found that the discussion task tended to elicit fewer longer turns; the picture task a greater number of shorter turns, with the consensus task in between. The interactional functions also varied, with the discussion requiring the most “asking for and providing information”, the picture the most “negotiation of meaning”, and the consensus task a wider variety of the functions (Van Moere, 2007, p.15). However, only one discussion prompt was used, and as research by Leaper & Riazi (2014) found, different prompts can elicit significantly different patterns of interaction. If Van Moere (2007) had used a different prompt it is possible that a different overall pattern of long and short turns would have been found.

Different tasks within the group format were also investigated to a certain extent by O’Sullivan and Nakatsuhara (2011), whose 269 Japanese high school students undertook information gap, ranking and free discussion tasks in groups of three or four, with minimum discussion times of 4-5 minutes or 6-7 minutes respectively. Each group did at least two of the three tasks, and as there was no maximum time, it means direct comparisons between the groups should be made with caution. Their purpose was to test three quantifiable measures of group conversation styles: goal-orientation (as measured by number of topics initiated), interactional contingency (the number of topics ratified by the other speakers) and quantitative dominance (the number of words spoken). In their quantitative results they found a significant but small difference in goal orientation between the three tasks, and a significant and large difference in quantitative dominance, but no significant difference for topic initiation. The post-hoc tests showed that in the information gap task, significantly more topics were initiated than in the GOT, and significantly more words than both the GOT and ranking task.

It seems that these limited results fail to convincingly show that different group discussion tasks can be characterised by the three dimensions proffered in this article. Since the indices they were based on were not normalized for time and person, it renders interpretations of the significant differences problematic, at best. Moreover, since the measure of interactional contingency was not significantly different between the three tasks, it does not seem to be suitable for the role of characterising patterns of conversaiton. Perhaps the explanation for this finding of non-significance is related to developmental issues that are not considered in O'Sullivan and Nakatsuhara (2011). In Galaczi (2013), the ability to confidently ratify other's topics is a feature of speakers at the relatively high CEFR C1 level, and though this ability does appear in the next level down (B2), Galaczi (2013) distinguishes the speakers at the lowest level in her study, B1, as only being capable of limited topic development, especially if the other speaker had initiated it. Since the Japanese high school students in this study had a level of proficiency judged to be between A2 and B1 (Nakatsuhara, personal communication, February 11, 2014), the ability to ratify other initiated topics was likely to have been developmentally beyond them, making it an inappropriate vehicle for distinguishing tasks along the dimension of conversational style among lower level learners.

Research comparing the GOT to other formats reveals that it seems to measure different abilities to other types of speaking test (presentations, interviews, role-plays, oral reports). It also suggests that the GOTs reliance on the willingness of its test-takers to participate allows some in the group to take part to a minimal extent. Finally, concerns have been raised over the variety of functions that it elicits.

2.2.5.3 Impact of other factors on GOTs

From an administrator's point of view, the ability of the GOT to assess large numbers of students in an efficient way may make it seem ideal. However, to be a useful test it also needs to not unfairly disadvantage the test-takers who comprise the group. That is, affective and group composition factors need to be investigated. Many of these findings come from a series of research projects at a private languages university in Japan (where the data for this dissertation was also collected). Bonk and Van Moere (2004) investigated the effect of different levels of ability within a group among 1055 test-takers at a Japanese university. The researchers asked the home room teachers to estimate the speaking ability of students so that they could manipulate the groups to be homogenous or heterogeneous. For Bonk and Van Moere, a group was homogenous if it had all of the students rated at the same level, or all but one, for example LLLM, MMMM (where L is Low and M = Medium level); heterogeneous groups had no more than two students at the same level, for example LMMH, LLHH (where H is High level). This permitted two variables for the regression model that was used: one for group mean proficiency level in relation to the individual's level; and one for the homogeneity of proficiency levels in each group. No significant effect was found on the expected individual's scores.

The quantity of talk produced in terms of the number of turns, words and incoming proficiency level of the candidates was investigated by Kobayashi and Van Moere (2003). Their data came from 60 test-takers in 25 GOTs that were analysed for the number of turns the speakers took, the amount of talk produced and the ability of the participants. They found that the number of turns taken had no impact as a main effect, and neither did the interaction between number of words spoken and incoming proficiency level of the student. This shows that high level students do not improve their scores by talking more and that low level students are not penalized for speaking less. An interesting finding was a positive correlation between the number of words spoken and the score in the 'communicative skills' band (Appendix I, p. 426 for the rating bands used in this test). This band recognizes a candidate's ability to manage and interact in the discussion, but the finding leaves open the question as to whether those who speak more are using their words to carry out these functions.

Concerns have been voiced about the impact of the personal characteristics of the participants on the performance in the GOT. Some fear that shyer students may be at a disadvantage in this test (Hilsdon, 1991), and it is possible that gender might have an impact on either the raters or the participants of the study, as has been found in paired formats (O'Sullivan, 2002). These aspects were investigated by Bonk and Van Moere (2004) who administered a seven question shyness survey immediately after the test to 1,055 students in 300 GOTs. Although their study found no significant impact of gender, there was a small but significant relationship between teachers' pre-test predictions about their students' scores and shyness. At the most extreme ends of the scale, they calculated that the difference between the most outgoing and the shy would result in a difference of 2.5 points out 20.

The importance of these findings depends on the stance taken towards performance testing. Those that believe that speaking tests should test only the linguistic aspects of language might deplore the fact that any test-taker is disadvantaged due to an aspect of personality like shyness. They would claim that the ability to produce language should be separated and assessed independently of the context in which it is sampled (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008). The opposing camp points out that language by its nature is a co-constructed context dependant event (Chun, 2006), and that a factor such as personality is an inevitable part of the performance of the construct of speaking that we are attempting to assess. In this view even though such aspects of personality as shyness might impair a student's score, the test is a fair reflection of the student's ability in a speaking situation. In this case a maximum difference of only 2.5 points out of 20 might be considered acceptable considering it is a performance test.

Not all research into the effect of shyness on the group oral has returned significant results. A study that investigated the slightly different construct of self-consciousness, a subset of the NEO-PI-R neuroticism scale, found no impact on scores in a study of 360 Japanese university students (Ockey, 2011). These seemingly contradictory results can be explained by the two constructs being slightly different, as Ockey points out "the extent to which the concept of self-consciousness as defined by the NEO-PI-R is different from that which most practicing teachers view as shyness" (2011, p. 981). In the same study, Ockey also looked into assertiveness, which is one of the six elements of the 'extraversion' scale of the NEO-PI-R. This element is related to an individual's inclination to take the lead, speak without hesitation and lead a group. Ockey found that assertive students had a small but significant advantage on all scales, with slightly stronger relationships with the communication skills and fluency bands than grammar, vocabulary and pronunciation. However, as Ockey points out:

> when the effect is quite small as it was with the effect of assertiveness in the study, the instrument and the context will have an effect on the results that may mask the true relationship among the variables. These effects are really small, and may or may not manifest themselves in a particular study. (Ockey, February 6, 2014, personal communication)

This applies equally to the shyness findings for Bonk and Ockey (2002).

A more serious threat to the validity of the group oral was found by Ockey (2009) whose soundly constructed investigation found that it is not necessarily the test-taker's own personality, but also the personality of the test-taker's interlocutors that may affect a rater's scoring. Ockey investigated how assertiveness affected the scores of 228 students. Ockey points out that researchers have most often focussed on the wider domain of extraversion with L2 proficiency and with a few exceptions have been inconclusive. Assertiveness, on the other hand, is one of six facets on the NEO-PI-R extraversion scale, and specifically refers to an individual's inclination to be a leader and put forward opinions without hesitation. Ockey found that assertive test-takers got lower scores than expected when grouped with other assertive test-takers, but higher than expected when grouped with non-assertive ones. Ockey suggests that the raters were influenced by the context of the group by rewarding assertive test-takers when their fellow groupmates were more passive participants, but had no such context when they were grouped with equally assertive groupmates. This finding has important implications for rater training and the need for raters to pay attention to the interaction as a whole rather than just each test-taker individually.

Ockey's (2009) results are in some parts consistent with an earlier study which investigated extraversion and introversion among 447 Hong Kong university students by Berry (2004). In her study, the test-taker groups consisted of between four and six test-takers who were given short texts to read and then discuss for a span of time calculated as five minutes per participant plus an optional extra five minutes at the raters discretion. Unlike Ockey (2009), she found that when introverts were in groups with a low average level of extraversion, their scores decreased, but are elevated when grouped with more extravert participants, whereas Ockey found no difference in this case. A plausible reason Ockey gives for this difference was that in the longer span of Berry's (2004) GOTs, introverts may have been unable to sustain the effort necessary to stay engaged in the discussion.

Consistent with Ockey (2009), Berry (2004) found that extraverts in groups with a low level of extraversion were rated significantly higher, and being grouped with similarly extravert test-takers had the effect of depressing their scores. To explain this, she states that "when trying to award a rating to five students at the same time, [raters] reward the extravert for simply being more noticeable when in a low mean extraversion group, whereas in such a group, introverts do not stand out" (2004, p.115). Neither Berry (2004) nor Ockey (2009) investigated what was happening at the level of discourse by reference of their scores, and requires support that a qualitative examination of GOTs would accord, such as in this dissertation.

More recent research on this matter has produced inconsistent results. Supporting the findings of Ockey (2009) and Berry (2004) is research conducted by Nakatsuhara (2011) who investigated extraversion in GOTs, comparing groups consisting of three participants with groups of four participants. Participating in her study were 269 Japanese high school students who took the extraversion scale of the Japanese Eysenck Personality Questionnaire (EPQ), and were allowed to put themselves either into a group of three or a group of four. During their discussions the conversational styles were quantified by taking measures of each group's goal orientation and quantitative dominance. Goal orientation was the proportion of topics an individual initiated of his or her groups total, and the number of words spoken relative to his or her groupmates was used as a measure of quantitative dominance. She found that in groups of four, extraverts were more influential than in groups of three, and while both groups of three and four were influenced by the participant's proficiency, the effect size for groups of three was larger.

While Nakatsuhara's (2011) make a substantial contribution for investigating this subject from a discourse level, it is important to point out that it was conducted on experimental groups and so was not an authentic assessment context. This was not the case with Gan's (2011) investigation into the impact of extraversion in the GOT in a school-based assessment context in which 39 Hong Kong high school students took part. Significantly, this study examined not only the impact of extraversion on scoring, but also the complexity, accuracy and fluency with which they performed. Gan found no significant difference correlation between extraversion as measured by the EPQ and their scores, or the indices.

Although Gan (2011) shows that extraversion need not affect the scoring of GOT, there are several issues that should be raised that may at least partially account for his findings. Firstly, differences in the personality scales should be considered. Gan's instrument was the concise 12 question version of the extraversion scale, whereas Ockey used an eight question assertiveness subsection, which may have been more specifically targeted, and the possibility has to be considered that this impacted on different findings. A second point to be raised is the very different methods of forming the groups in these studies. In Gan (2011) the groups were self-selected, whereas Ockey (2009) and Berry (2004) manipulated group membership in order to produce groups that were asymmetric with regard to extraversion. It is possible that by using self-selection, Gan's groupings did not include the same widely varying mixes that appeared in Ockey (2009). Indeed, Nakatsuhara (2011) also used self-selection, but the large numbers in her study make it seem more likely to produce asymmetric groups, which may not have been the case with the much smaller scale of Gan's study. As such, the findings of Ockey (2009) and Berry (2004) would be better interpreted as illustrating a potential threat to validity that may arise should that particular mix of assertiveness occur in the test. It would need to be empirically demonstrated how often such asymmetric groups occur, as well as the extent that rater training could overcome it.

The final point that arises is one that raises further questions for research: It is apparent from the discourse analysis presented in Gan (2011) that the level of ability displayed by his participants is qualitatively superior to the subjects of Ockey (2009) or almost certainly Nakatsuhara (2011). There is the possibility then, that with increasing proficiency in the target language, differences between assertive and non-assertive personalities becomes less important in their performances. That different aspects of personality may affect performance at different levels of proficiency is something that has yet to be investigated in the assessment literature of speaking tests, to this author's knowledge.

All the above research into factors that may affect scores usefully illustrates some of the issues that need to be addressed if the GOT is to be a useful vehicle for assessing spoken language. Other research has focussed on gaining an “insider’s view” (Van Lier, 1989, p. 489) of the performance of test-takers in the group oral. The next section reviews literature that has taken this approach.

2.2.5.4 Analyses of interaction in GOTs

The first study in the literature to examine the language produced in the GOT was He and Dai (2006), who examined the group discussion task that is part of the College English Test – Spoken English Test (CET-SET), a high stakes assessment that is a requirement for a Bachelor’s degree at many universities in China. The GOT in this test lasts for four and a half minutes and has three or four participants who are expected “to engage in communicative interaction while arguing with each other, asking each other to clarify a point and trying to reach an agreement” (He & Dai, 2006, p. 376). The study analysed 48 transcripts of 197 candidates for eight interactional language functions that included: agreeing/disagreeing, asking for opinions or information, challenging, supporting and modifying developing opinions or points, persuading and negotiating meaning. They found that nearly half of the exchanges were agreeing and disagreeing, and nearly a quarter were asking for opinions or information, followed by challenging (7.4%), and the remaining categories made up between 4.9% and 2% each. The number of candidates who engaged with their groupmates by challenging, supporting their opinions, modifying their own or persuading others numbered only 47 out of the 144 candidates. Since the purpose of the task was to elicit this kind of language, and candidates are apparently aware of the requirements, a mismatch was found between the design of the test and group discussion as a means of carrying it out. As He and Dai state, “conversational features do not appear in speaking tests just because we introduce speaking partners with equal social power” (2006, p. 393).

He and Dai (2006) put forward several possible reasons for the lack of use of the language functions that the task was supposed to elicit. Firstly, the results of their questionnaire suggest that the candidates were viewing the GOT as an assessment event rather than a “meaningful discussion with group members” (He & Dai, 2006, p. 388). The candidates seemed to be concentrating on producing their response to the question prompt rather than listening to another’s opinion and reacting to it, and this could be seen by the high proportion of responses like “I agree with what he/she says” (He & Dai, 2006, p. 389). The second reason they put forward is their students’ lack of confidence in their ability to react to another’s opinion instead of putting forward their own opinion, which may be less taxing. Thirdly, the topics were not interesting to the candidates. The questionnaire they conducted after the test showed that 60.2% of the candidates found the topics uninteresting or dull.

In contrast to He and Dai’s (2006) participants, Van Moere’s (2006) test-takers had a more positive impression of the multiple question prompts in their GOTs. In his survey, 80% of 1,088 students thought the multiple question prompts were “effective for making people talk” (Van Moere, 2006, p. 439). The different results from the two surveys can be partially ascribed to the different way the prompts were constructed in each study. The topics in CET-SET were a single question (“Is it desirable to live in a big city?”) whereas the topics used in Van Moere (2006) have the same format as the prompts for this dissertation. They have multiple questions and are presented in the form of a short paragraph, along with a translation in the students’ native language, and are related to topics that the students could be expected to discuss in their classes. Such a format may have more chance of being appealing as a test prompt.

The final reason He and Dai (2006) give for the tendency of CET-SET to produce long turns is that perhaps they interpreted the directions of the judge as requesting quantity rather than quality. This means fewer turns for each candidate to produce the interactions that were supposed to be elicited by the task. By contrast the students in the Van Moere (2006) study had up to twice as long in which to give their opinions (up to 10 minutes, as opposed to 4.5), and this might also help to explain the difference in opinion in the surveys of these papers. Qualitative research on GOTs subsequent to He and Dai’s study has found that the first turns of the GOT are often taken by longer turns arranged so that the participants get to hear every participants initial opinion (Luk, 2010; Nitta & Nakatsuhara, 2014; Van Moere, 2007), and following this a more natural conversation may develop. Since the entire length of the GOT in He and Dai (2006) was only four and half minutes, the possibility is that the test-takers were not given enough time to develop a more natural conversation. Nitta and Nakatsuhara (2014, p. 21) recommend a 5-minute span as suitable for eliciting a collaborative exchange from paired test-takers, and it seems evident that three or more participants would require more time. Another possibility is that the candidates were not at the developmental stage where they could have used some of the functions that the test was supposed to elicit (Galaczi, 2013).

The debate about the extent to which the GOT can elicit conversation-like interactions was answered to a certain extent by a CA analysis by Gan (2010) which contrasted a high with a low scoring group of four students. He found that where the higher level group engaged collaboratively with each other's ideas as they discussed the topic, the lower level group approached the task as one in which they should answer questions rather than as a conversation. Gan (2010) suggests that this could partly be ascribed to a difference in their prompts. Teachers have the freedom to adapt this Hong Kong school-based assessment according to the needs of the students, and so they gave the lower proficiency group additional support in the form of three supplementary questions to the usual single one. Gan speculates that the series of questions may "have contributed to some students' orienting to the task as predominantly one that calls for an 'answer' rather than actual discussion" (2010, p. 598). It should be pointed out that like the higher level group in Gan's (2010) paper, the prompts in He and Dai (2006) were single questions ("Is it desirable to live in a big city?"), and as pointed out above, these failed to elicit conversation-like behaviour. This weakens an inference that could be made from Gan (2010) that a single question prompt necessarily leads to a more collaborative discussion, rather than, for example, the attitude or confidence of the interlocutors, or, according to recent research, their development of the skill of interacting (Galaczi, 2013) and amount of planning time (Nitta & Nakatsuhara, 2014).

Also, according to Leaper and Riazi (2014), multiple question prompts need not preclude test-takers from having an authentic-like co-collaborated conversation, as an example in their study shows. This article examined the effect of prompt on 141 students in 41 group orals at a Japanese university. This study found that the four different prompts that were written to be parallel forms influenced significantly different patterns of interactions among the test-takers. Prompts that encouraged students to explain a back-story tended to have discussions that had fewer and longer turns, while those that were focused on more factual here-and-now subjects of the students life encouraged more and shorter turns. Although accuracy was not affected by the different prompts, fluency and complexity differed significantly, with prompts that required back-story having more complexity in their longer turns. Fluency was also negatively affected by a prompt, with those test-takers discussing a more personal subject on single vs. married life producing less fluent responses. The test-takers with this prompt had a significantly higher proportion of pauses in the longer turns of their speech as they attempted to answer this rather face-threatening topic in front of their peers. While the data of this study came from an actual test rather than more controlled conditions, it warns of the need to take care when creating prompts for speaking assessment.

In the qualitative section of her study, Nakatsuhara (2011) raises an issue that may well have had an impact on many of the earlier findings on GOTs. She contrasted the interactions of groups that had three students with those that had four, and found some interesting differences. Firstly, groups of three had a collaborative atmosphere that was lacking in groups of four. She noted a higher frequency of members collaboratively completing each other's sentences in groups of three. Also there were more successful attempts at supporting the quieter members of the group in threesomes than in foursomes. The second feature was avoidance in groups of four. In groups of four test-takers, a member could avoid contributing by simply agreeing with other members or by remaining quiet. It seems that members in groups of four felt less obligated to contribute than members of groups of three. Finally, in groups of four, turn-taking could be mechanical and unnatural. This occurred regardless of the extraversion level, and the turns were often managed by hand signals, or frequent, mechanical use of "How about you?" or "What do you think?" Nakatsuhara points out the similarity to Galaczi's (2008) parallel development of topic, when pairs talk about their own topic regardless of what the other person has said.

A reason Nakatsuhara put forward for this kind of unnatural discussion in a group of four is the lack of 'schisming', a feature of conversation in small groups in which the participants split into even smaller groups and carry on separate conversations simultaneously. It seems that students' awareness of the assessment situation and the need to present just one conversation to the raters may prevent this from happening. In a natural conversation this feature could be expected in groups of four, but it is impossible in a group of three. However, as pointed out in the section above on the quantitative findings of her study, one may speculate as to how much of the features found in her study were due to the low level of the students who took her tests. She has made some illuminating findings which affect all studies on GOTs in which groups of threes and fours were indiscriminately mixed. The predominantly low level of students in her study points to the need for confirmatory research to investigate if more advanced test-takers behave differently in groups of threes and fours.

The studies reviewed in this section thus far have focussed on the language produced by the test-takers, and while many researchers have acknowledged the importance of the role of gestural behaviour (Galaczi, 2013; Lazaraton & Davis, 2009), and raters seem to be deeply influenced by it (Ducasse & Brown, 2009; May, 2011; Orr, 2002), only one paper has focussed on its role in interaction in peer interaction tests. Gan and Davidson (2011) compared the gestures of the same higher and lower scoring groups as in Gan (2010). Although they only coded the first three minutes of each GOT, they found notable differences between the two groups. The higher scoring group's gestures were well synchronized to what they were saying, and this was particularly evident in the two participants who took the lead in this discussion, who used gestures not only to illustrate directly what they were saying, but also metaphorically and used beat gestures to introduce or emphasize their words. The lower scoring group lacked this synchronic association of gestures with meaning, with one member spending most of her time engaged in apparently reading it, and other participant's gestures seeming "irrelevant to his speech" (Gan & Davidson, 2011, p. 115). Even the most animated among the lower test-taker's gestures were more related to the struggle he had to produce language than as a way of enhancing his communication. One of the lower group used pointing gestures effectively to indicate others to take a turn or move on to another question, but few other effective uses of gesture could be found. Gan and Davidson (2011) sum up the gestural behaviour of the lower group as being "an outward sign of language difficulties, disfluency, tension and lack of confidence" (p. 116).

The research reviewed above can be seen to be contributing to our knowledge of the GOT in terms of the linguistic and interactional resources that Young (2009, 2011, 2013) identifies (see

Section 2.1). So far, only a single study on the GOT has focused on the remaining category by which interactional competence can be analysed, which are the identity resources. A study by Luk (2010) fills this void by investigating the 'impression management' of 43 female Hong Kong high school students. Luk shows how the test-takers consciously or unconsciously manage the interaction in order to position themselves as competent language users. In the eleven GOTs in this study, Luk used CA to identify three frames of talk among the data that were most common: task management, content delivery, and response. The task management frames were employed when the test-takers were starting, closing or engaged in routine matters of the talk, such as turn transitions. Following the start, comes the content delivery stage, in which the test-takers gave their answers to the questions in the prompt, usually with the aid of notes made previously. The most varied part of the group discussions were the response frames, in which the main functions of the discussion could be found.

Within this framework, Luk found evidence of stage managed interactions. Luk found the openings were 'ritualized' with one student announcing the start, the others agreeing and then the announcing student starting the content delivery stage. There followed an orderly co-constructed turn-taking session in which each student gave a prepared response, ending their speech with a handover question to someone who had not taken their first turn with ''What do you think Gloria?" (p. 37), for example. If the starting speaker had been the one sitting the left or right-most, then the turns proceeded either clockwise or anticlockwise around the table. The content delivery frame was when the students did most of their talking, usually around twice the amount that occurred in the response frames, and were nearly invariably at the beginning of test time rather than the middle or end, as is consistent with Nitta and Nakatsuhara (2014).

Luk suggests that this pattern was test-driven, a result of the need of all members to complete the pre-scripted component of the discussion which could perhaps be delivered with greater fluency and accuracy. In the response frames the most common were converging responses, such as "I agree with you," which Luk considers superficial since they were used regularly but not expanded on. The kind of responses that converged in the sense of test-takers' contributions building on what was previously said, were rare. Diverging responses that were disagreements were similarly rare, and Luk believes that this was due to the test-takers desire to maintain solidarity rather than risk a disagreement exposing another's lack of language skills. Luk also noticed that the students tended to avoid negotiation of meaning by being reluctant to use comprehension checks, clarification requests and so on, again perhaps because they did not want to expose their fellow classmates' incompetence in the language. Finally, Luk finds evidence of speakers using a disagreement phrase even though no disagreement is actually meant. The phrase seemed to be used more as a lead-in to a content frame in which the speaker could give their opinion, but apparently was used in order as a display of knowledge of the range of response acts the test-taker had.

Luk's (2010) study presents a picture of group discussion tests that should be familiar by now (see Gan, 2010; He & Dai, 2006; Van Moere, 2007). These lower level students have competing demands: on themselves to do their best in an assessment event, as well as having to work with fellow students whose similar desire may put them in competition with each other. It is no surprise that in this case, given that they are being tested with their fellow classmates, that they choose to collude with each other in the ways that Luk identifies.

2.2.5.5 Raters and the GOT

For the GOT there has yet to be an in-depth study which investigates the additional challenges faced by the rater of dealing with three or four test-takers (perhaps more depending on the format). Though the challenge presented by the rating context is indicated by Ockey (2009), whose findings suggest that raters find it difficult not to compare the performance of an individual independently of the assertiveness of his or her groupmates, as noted in Section 2.2.5.3 above. Ockey's findings of scorer variability are consistent with an earlier study at the same institution by Van Moere (2006). In a test-retest design that included 113 students, he found that although the inter-rater reliability coefficient of 0.74 was acceptable given the length of the test and the training of the raters, the G-study results revealed that most of the variation in scores was due to test-taker performance. Van Moere suggests that such factors may be related to interlocutors or group dynamics, one of which may well be the assertiveness of their groupmates that Ockey (2009) found to have a significant effect.

2.2.5.6 Development in the GOT

The only paper at the time of writing to investigate proficiency levels among students was a large scale study of 135 students from three different educational institutions by Negishi (2010). In this study she formed her groups by taking her participants from two junior high schools, two senior high schools and three universities in Japan, assuming that their ability would be in ascending order. For her task, students were put into groups of three and asked to plan what they would say on one of six familiar topics for five minutes. When the discussion started she gave them half a minute to introduce themselves and five minutes to discuss the topic. The transcripts were analysed according to a list of functions derived from He and Dai (2006) and Brooks (2009), adjusted for the low level of her students. She found that the junior high groups used a much greater proportion of 'asking for information or opinion' functions, tending to use more other-nominated turns instead of self-nominating, and in general could not develop a topic. The university students used questions to get further information about the topic and were capable of self-nominated turns, which were infrequent at the lower levels. When looking at supporting or agreeing functions, the middle and lower students had a limited range, overusing 'Me too', even in situations where it was not grammatical. The university students could use a much greater range, more appropriately, though 'Me too' was also noted as being frequent. The functions relating to disagreeing, challenging or persuading showed little difference between the groups, being equally rare in all of them, and Negishi provides cultural reasons for this lack, pointing out that Japanese prefer not to disagree directly. For Negishi, the category of modifying or developing shows "clear developmental characteristics" (p. 67). The lowest level students proving almost incapable of it, the high school students able to do it to a limited extent, while the university students could develop their talk over turns.

The biggest difference between the groups was found in the 'negotiation for meaning' category, which the higher level groups used with increasing frequency. Although the senior high school students asked for clarification in a simple way similar to the junior high schoolers, they persisted while the lowest level gave up, or the question caused a communication breakdown. The university students, on the other hand, while not much more sophisticated in form, could use them relatively smoothly in the conversation.

Although these are interesting findings, several limitations need to be pointed out. Firstly, this was not done in an actual assessment setting, but as a controlled experiment, and students may have conducted themselves less seriously in such a context. Secondly, while Negishi (2010) provides reasonable numbers to investigate a developmental sequence in interactional skills, it falls short in the analysis of the data by failing to provide measures of statistical significance, which would improve the power of the findings.

2.2.5.7 Summary of the literature on the GOT

This review of the literature on peer interaction in group oral assessment has brought to light various features of the GOT. It has found that the GOT:

- is generally appreciated by test-takers for the opportunity it provides to test takers to interact with peers rather than an interviewer by test-takers (Folland & Robertson 1976; Fulcher, 1996a; Hilsdon, 1991; Van Moere, 2006)
- measures different skills compared to other forms of speaking assessments (Gan, 2013; Pavlou, 1997; Shohamy, Reves & Bejarano, 1986)
- is capable of assigning grades to a large number of test-takers efficiently (Bonk & Ockey, 2003; Van Moere, 2006)
- may negatively affect shy test-takers (Bonk & Van Moere, 2002), and may result in unfair scores being awarded to test-takers, depending on the mix of assertive students in a group (Berry, 2004; Ockey, 2011)
- may have different patterns of discourse depending on the whether there are three in the group or four, and this may interact with the extraversion of the group members (Nakatsuhara, 2011)
- may be influenced by the prompt (Leaper & Riazi, 2014)

- may not elicit negotiating skills (He & Dai, 2006; Luk, 2010; Van Moere, 2007), and may elicit a relatively narrow range of interactional functions compared to other tasks (Van Moere, 2007)
- may elicit conversation-like interactions among proficient students (Gan, 2010; Leaper & Riazi, 2014)
- provides an interactional context in which test-takers can project an identity (Luk, 2010)

However, one thing all the above studies have in common is that they are based on single administrations (with the exception of Bonk and Ockeys' [2003] Rasch analysis). No study has yet examined how a test-taker's performance in a group oral test may change from administration to administration over a period of time, and given that one of the concerns is the variability of performance of the test-takers, this is notable omission. All the studies reviewed above have examined the test as a single instance in which variance in performance may be subject to a variety of factors. Many of the studies neglect to consider that the test-takers are also developing their interactive skills, and that until they are developed, may not be capable of producing the collaborative negotiating that administrators hope to encourage. This is perhaps a result of the notion of 'interactive skills' being such a recent concept that is has not been widely thought about or understood very well. An investigation of how an individual's skills develop over a number of tests taken over a number of years will go some way to providing guidance on this matter.

Having presented this review of literature on the assessment of peer interaction tests, namely the paired format and the GOT, the next section will provide a justification for their use according to research into the role of conversational interaction. If the literature shows conversation with peers to be an effective tool for learning a language, then an assessment in which students of a language interact with each other can be seen as an ideal instrument to motivate teachers and students to use and do peer interaction activities in class, as well as to evaluate their progress. That is, the next section will investigate the literature on conversation as a catalyst for language learning.

2.3 The role of conversation in language learning

The main objectives of this review of conversation in language learning are firstly to provide the theoretical basis for using interaction as part of a language program; secondly to investigate the particular aspects of conversation that benefit language learning; thirdly, to establish that non-native speakers (NNS) are capable of acting in a role that can support other NNS in learning the language; and finally to review the literature that investigates the nature of NNS-NNS interaction. The extent to which the literature can provide firm answers to these questions is the extent to which an administration can be justified in using a test such as the GOT for encouraging the positive backwash of teachers using peer interaction in classes and students talking to each other in the target language.

The first section starts with a brief review of the field of research that has developed around the role of conversation in language learning, and moves on to review criticisms of early developments of this theory, before explaining a framework that has emerged in the light of such criticisms and developments in the field of cognitive research.

2.3.1 Input, output and the interaction hypothesis

Up until the late 1970s much of the research into language acquisition had focussed on the acquisition of the form of syntactic structures (Larson-Freeman, 1991), and was mostly concerned with the language learner as an individual who generated data for analysis. In an influential article that pointed out the value of discourse analysis as a form of research, Hatch (1978) broadened the field by focusing on the crucial role of interaction in conversation, and by implication the assessment of it. She called attention to the underlying assumption in the literature of the time that the language learner builds up a store of linguistic features through 'input' which are then put into use as 'output'. Instead, she proclaimed the premise that "language learning evolves *out of* learning how to carry on conversations" (Hatch, p. 404, italics in the original). More tellingly, she asserted that "it is not enough to look at input and to look at frequency; the important thing is to… examine the interactions that take place within conversations to see how that interaction, itself, determines frequency of forms and how it shows language forms evolving" (p. 403), a statement that presages Van Lier's (1989) call to examine the discourse of peer interaction assessments.

These were criticisms that would be applied to the next major theory of language acquisition to be advanced, Krashen's (1981, 1985) 'input hypothesis', which claimed that language acquisition is an unconscious process and that if input is pitched just one step beyond the language learner's level, the target language could be acquired. While the value of 'comprehensible input' could hardly be denied, when the evidence was examined it became clear that input alone was not sufficient to explain acquisition. It was found that learners in the education system that had been exposed to comprehensible input in immersion programs could score as highly as similar aged native speakers in tests of the receptive skills of reading and listening, and yet still lag behind in the productive skills of writing and speaking (Swain, 1985). Indeed, even many years of living and functioning in the country of the L2, in a predominately L2 environment is not by itself enough, as a case study by Schmidt (1983) made clear. Despite years of exposure in the country of the target language, even advanced language learners may not incorporate vocabulary or grammar that NS children acquire from early stages (Pavesi, 1986). The empirical evidence from language learners has demonstrated that comprehensible input alone is insufficient for acquisition of the target language (Trahey & White, 1993).

One of the first to explicitly recognize the role of speaking was Swain (1995), who coined the term 'comprehensible output'. Production in the target language, she says, "may stimulate learners to move from the semantic open-ended, non-deterministic, strategic processing prevalent in comprehension to the complete grammatical processing needed for accurate production" (Swain 1995, p. 128). The production of the target language may require the learner to process the elements of language in a more thorough way to that required by merely comprehending the message. The comprehensible output hypothesis could be subsumed into the 'interaction hypothesis' (Long, 1996), which was stated as the "negotiation of meaning, and especially negotiation work that triggers interactional adjustments by the NS or more competent interlocutor, facilitates acquisition because it connects input, internal learner capacities, particularly selective attention, and output in productive ways" (p. 451–452). While the literature in the years since then has cast doubt on the primacy of interaction for language acquisition, there is little doubt over its facilitative role or value for priming acquisition (Gass & Selinker, 2008).

The ‘interaction hypothesis’ was arrived at independently from the ‘interactional competence’ described at the start of this chapter, though they are not incompatible. *Interactional competence* is the construct that explains how a participant’s performance incorporates both situational and generalizable properties through interaction of these various factors. For interactional competence, an individual’s experience of local contexts, other participants and the context interact in order to produce a response in a new context, like a peer-interaction assessment. If the ‘input’ and ‘internal learner capacities’ of Long’s *interaction hypothesis* are regarded as the ‘local context’ and ‘repertoire of resources’ of interactional competence, then the two constructs may be seen to be converging. The ‘selective attention’ of Long’s definition can be taken as the mechanism by which participants understand the new context appropriately and select from the repertoire of resources at their disposal. As long as the ‘interaction’ is seen as occurring between not only the input and output of speaker and listener, but also between the context, participants and the learner’s skills and abilities, the synthesis of the two constructs is a practical possibility.

Indeed, they can be seen to be converging in Swain (2008). Swain divides the functions of output in interaction into three: the noticing or triggering function, in which learners realize through interaction that their language ability is lacking in some way; the hypothesis testing function, in which output provides an opportunity for the speaker of the L2 to trial some language they are uncertain about in interaction with another speaker or reader; and finally a metalinguistic or reflective function, in which learners can consider the language used in the interaction to mediate their learning of the language. This final point was inspired by Vygotskian sociocultural theory of the mind in which speaking can be seen as a mediating tool that can develop and reshape experience. When Swain points out, “by studying the collaborative dialogue… according to a Vygotskian sociocultural theory of mind, we are observing learning in progress” (Swain, 2008, p. 70), the interaction hypothesis is moving firmly into the territory inhabited by interactional competence.

Other ways that conversation may benefit language learning come from the cognitive perspective. Output in a foreign language has also been theorized to aid language learners by accelerating 'automaticity' in the language (Schneider & Shiffrin, 1977; Swain, 2005). Second language acquisition is viewed as learning a skill in which learners are gradually asserting control over the various components that language consists of: The appropriate vocabulary, grammatical structures, pragmatic rules and pronunciation. As learners gain proficiency, these components are restructured into routines that become automatic, freeing the learner to concentrate on higher level processes (McLaughlin, 1987; Skehan, 2009). The implications for institutions seeking to build the communicative ability of their students in a second or foreign language are clear. By continued practice of speaking, language learners can assert more control over its various components, allowing them to move towards greater fluency.

According to this cognitive perspective, it might be thought that language learners could improve their automaticity simply by talking by themselves, without need for interlocutors of any kind. If so, improved automaticity might be considered a useful by-product of a curriculum with interaction at its core. Even so, interaction may indeed be a necessary or crucial ingredient to improving automaticity if 'asserting more control' means building up and expanding the boundaries of specific contexts of language practice so that they can be applied to other contexts, as asserted by IC. 'Automaticity' under IC may be considered as the developing ease with which learners can access their repertoire of resources. According to IC, the routines built up in the situated practice at an institute of language learning would be applied to the learners to contexts that they had not previously been exposed to.

The interaction hypothesis has naturally been subject to criticism on a number of grounds. Krashen (1998) observed that a problem with the output hypothesis is that output itself is rare, and the frequency with which negotiation over meaning incidents occur even more so. As evidence he quotes several studies and points out the scarcity of incidences of negotiated meaning in them, often as little as one incidence per hour.

In answer to these criticisms, it can be pointed out that although clear examples of negotiation of meaning may be rare in conversation it does not mean that they are unimportant to the learner, and nor does it entail that output does not play an important role in language learning. Interaction with other speakers may form some of the most memorable and powerful learning episodes that learners have, and there is anecdotal evidence (Piehl, 2011) and research (Sato, 2007) to support this. Additionally, it is through output that the learner is socialized with other participants of the speech community, and this in itself is seen in socio-cultural theories as an integral component of learning (Moll, 1990).

Moreover, the 'negotiation of meaning' that Krashen (1998) refers to may be only the most overt means by which output may facilitate language learning. Other research from a socio-cultural perspective concurs with Krashen that incidences of negotiation for meaning are rare, but language acquisition in conversational interaction can be found in the way learners collaboratively work together to construct meaning, and these have been found to be more common than incidences of negotiation of meaning (Foster & Ohta, 2005, reviewed in Section 2.3.4). Other examples of output for language acquisition may simply pass unnoticed by outside observers and thus is difficult to research: a learner's successful hypothesis about the use of vocabulary or grammar would be passed over in a transcript except by the learner retrospectively pointing it out (Schmidt & Frota, 1986). Finally, as mentioned above, automatization may be developed through output and this development may occur every time a learner engages in conversation, regardless of the kind of interaction, be it negotiation for meaning or otherwise. The very act of forming sentences in an L2 itself may well be beneficial, and thus, justifies its role as a core activity in the language education curriculum, and by extension, assessment in which interaction occurs.

For the role of conversation in language learning to be taken seriously it requires a theoretical foundation. Since the interaction hypothesis purports to provide this, criticisms of the cognitive grounding of it, will be reviewed in the next section. The criticisms reviewed here centred on the role of the Noticing Hypothesis, which is an essential part of it.

2.3.1.1 Criticisms of the Noticing Hypothesis

In the interactive hypothesis of Long (1996), the Noticing Hypothesis, which is incorporated in the definition as the 'selective attention' aspect became subject to scrutiny. The basis of the Noticing Hypothesis in the interaction hypothesis harks back to an influential article by Schmidt (1990) that reviewed the evidence for the role of consciousness in language learning, looking at its role as intention, attention and awareness. He concluded that while incidental learning is possible, paying attention is facilitative and is probably necessary in some instances. On subliminal learning though, he took a position directly opposed to Krashen (1985), by stating that it was impossible. For Schmidt, noticing is not only necessary for language learning but also sufficient for transferring input to intake: "if noticed, it becomes intake" (p. 139), a strong position that is open to counter-example and subject to debate.

In describing his claim, Schmidt defines 'noticing' as awareness, which he equates to consciousness. After surveying the literature Schmidt tells us that what the theories of consciousness have in common is that they are "compatible with the view consciousness separates mental life into two fairly distinct spheres" (Schmidt, 1990, p. 138): an unconscious and a conscious side, which he terms 'phenomenological awareness'.

Unfortunately for Schmidt, there is no general agreement in the literature on cognitive science that attention is the same as awareness (Truscott, 1998). Cognitive frameworks have been produced that completely omit 'awareness' as an operating function or as a process, and instead preserve it merely to describe the subjective experience of mental perception (Tomlin &Villa, 1994). Another problem is conceptually defining Schmidt's concept of 'noticing'. It must be a higher state than mere awareness of input on a global level and yet lower than awareness at the level of understanding. For if noticing is merely at a global level then the Noticing Hypothesis would state merely that 'awareness is necessary for learning', which few people would disagree with, and is not an interesting claim. More importantly, it barely distinguishes Schmidt's (1990) Noticing Hypothesis from Krashen's Input Hypothesis (1985). On the other hand, it cannot be fully conscious understanding, since it is unreasonable to suggest that complete knowledge of linguistic forms is necessary for learning. It must be some partial level of understanding, but there is no clear way to draw a boundary between what constitutes partial knowledge and full knowledge (Truscott & Smith, 2011).

The nature of attention itself has also been questioned. Under Schmidt's (1990) explanation, attention is a unitary concept. However, different aspects of attention can be distinguished; namely alertness, orientation and detection (Tomlin & Villa, 1994). Of these, the part in which noticing would logically reside would be as part of 'detection'. It is in the detection function of awareness that the memory registers exemplars and makes them available as a resource to other processes. Experimental work makes it clear that this can be done subliminally, for example semantic priming (Marcel, 1983) and sequences that can be learnt without subjects being aware of them (Carr & Curran, 1994). Research into relative clauses has shown that there is a hierarchy that diverse languages have in common, and once the lowest level relative clause is known generalizations can be made about higher level relative clauses without overt instruction or attention being drawn to non-target relative clauses (Eckman, Bell, & Nelson, 1988), which is contrary to Schmidt's (1990) version of the Noticing Hypothesis.

2.3.1.2 A cognitive framework for the Noticing Hypothesis

One possible way of avoiding these theoretical problems is to develop further the concept of 'awareness', by building a cognitive framework. Gass's (1997) model of second language acquisition created an additional level of 'apperceived input' which is described as "an internal cognitive act in which a linguistic form is related to some bit of existing knowledge (or gap in knowledge)" (Gass, 1997, p. 4). However, without an explanatory cognitive framework supporting it, it too falls prey to the same difficulty of distinguishing 'apperceived input' from understanding (Truscott & Smith, 2011).

Robinson (1995) proposed that awareness of noticing could be distinguished from the lower levels by it involving short term memory for rehearsal that could in turn lead to the formation of associations. For this he postulates the existence of an executive mechanism that allocates limited attentional resources. Robinson sides with Schmidt (1990) on the impossibility of noticing occurring without attention, claiming that the subjects of any research with contrary findings (in this case, Curran and Keele, 1993), could in fact be aware of their learning, but in a way the researchers did not assess. However, this counter argument can be made about any positive findings for unconscious learning, effectively making the Noticing Hypothesis unfalsifiable (Truscott, 1998).

If the Noticing Hypothesis is to be rescued, it needs to be built on the foundations of a framework that can account for consciousness in an integrated framework of how the mind works. A solution proposed by Truscott and Smith (2011) is to use the Modulated On-line Growth and Use of Language (MOGUL) framework. According to this processing oriented approach, the mind is composed of modules of expert systems that have evolved over time, and one of these systems is of language, which is itself composed of three major subsystems that handle the phonological, syntactic and conceptual processing required to understand and produce language. Each subsystem has a processor and a lexical store of primitives that pertain to that system, and work by forming representations of input. The subsections work independently, connecting with each other via an interface that processes the information. Thus, an item of vocabulary would have three representations, one in each of the subsections, that are connected by means of an index that the subsections have in common. Within this framework, the notion of activation plays a key role, and two states are distinguished. At *resting level*, there is no involvement in processing, whereas the *current level* is the activation in addition to the resting level. The level of activation is important since it is only the highest level that is selected for representation (Truscott, 2006).

Working from MOGUL, a hierarchy of processing-awareness was developed in which four levels could be distinguished (Truscott & Smith, 2011). The lowest level is subliminal perception, which is constituted of events that are not activated sufficiently to reach consciousness, and any use made of this kind of information falls into the category of 'subliminal perception'. The next stage is 'awareness of input', in which the activation level breaches consciousness. Although consciousness has been reached it is still not sufficient for learning and there is a lack of focus. Following this state is 'noticing-understanding' which occurs when a portion of the input has been processed by at least one section being represented as an example of a form. Finally, as more representations are formed from meanings or significance the last stage is reached, one of 'conscious understanding'. Under this framework, noticing occurs in an intermediate level between subliminal and full understanding, and is distinguished by the representations it forms in other parts above the level of perception.

For now, the MOGUL framework can provide an explanatory framework for the role of the Noticing Hypothesis, allowing it to be a meaningful sub-structure in the interaction hypothesis. It can also be adapted to interactional competence by providing a framework for the mental processes necessary for accessing the knowledge of how language works in specific domains. Thus it may be employed as the theoretical basis for incorporating interaction into the language learning curriculum, and justify the use of peer-assessment speaking tests such as the GOT.

Having established a theoretical background that will suffice for present purposes, attention can be turned to the empirical support for interaction in language learning. This is perhaps a more crucial question, since even if such a theoretical background as above was missing or found to be flawed, if it can be demonstrated through research that conversational interaction benefits language learning, this alone could be used to justify its value in a language learning program. The first question is about what particular aspects of conversational interaction benefit language learning. As the purpose of this dissertation is to examine the interaction that takes place between NNSs, this is naturally of more interest than the literature on NS-NNS interaction. However, it is important to define the terms and frameworks that have been used, which were mostly derived from the early research that investigated how NS adapted their language when talking to a NNS. As such, the next section will briefly review and summarize the relevant terms and features of interaction that have been identified through NS-NNS research, but could apply equally to NNS-NNS interactions.

2.3.2 Conversational interaction as language learning

The chapter by Hatch (1978) and research by Long (1981) triggered much interest in the role of interaction in language learning. Hatch’s declaration that conversation could be an arena within which language is learnt raised questions that Long (1981) was among the first to answer. Long and other researchers in the early period of interaction research were typically engaged in examining transcripts to show how NS could provide more comprehensible input to lower level NNS by adjusting their language to the level of the listener (Long, 1983a, 1983b). Later research demonstrated that interaction provides the participants a context within which they can form hypotheses, negotiate over meaning (Gass, 1997) and perceive gaps in their ability (Schmidt, 1990), thus playing an important role in the acquisition of the target language (see Mackey, Abbuhl and Gass, 2012 for an overview). With the consistency of the findings and advance of explanatory cognitive frameworks, researchers in the field have been confident enough to promote it from hypothesis to 'approach' (Gass & Mackey, 2006) and a framework that "contains not only elements of an empirically verified model… but also elements of a theory" (Mackey, 2012, p. 3).

The negotiation over meaning that occurs in interaction when a miscommunication occurs and is commented on by the participants has also become known as a 'language-related episode' or LRE (Swain & Lapkin, 1998). An LRE can be distinguished from the on-going discourse by its focus on how the message is being communicated rather than on the message itself, and includes any utterance in which participants in an interaction "talk about the language they are producing, question their language use, or correct themselves or others" (Swain & Lapkin, 1998, p, 326). LREs can be categorized as being either initiated by the speaker or by the listener in reaction to something that was said, although in practice it may be hard to separate or distinguish LREs on this basis (Foster & Ohta, 2005).

The interactional moves that take place in LREs were investigated by Varonis and Gass (1985) who investigated how miscommunications are signalled, uptake occurs and modified output result using their data from both NS-NNS and NNS-NNS interaction. They referred to the utterance that causes the misunderstanding as a 'trigger' after which follows the 'resolution', which may be composed of an 'indicator', 'response' and 'reaction to the response'. It should be noted that as with hypothesis forming (Shehadeh, 2003), triggers that are ignored will go unnoticed unless retrospectively pointed out by the participant. In this framework, the signal that draws attention the miscommunication is labelled as the 'indicator' and is recognized by its interruption to the flow of the conversation. Various devices can be used as indicators, as can be seen in Table 7.

Table 7: Categories of indicators and responses in LRE episodes

Categories of Indicators

1. Explicit indication of non-understanding: “pardon?” “what?” “I don’t understand”
2. Echo word or phrase from previous utterance
3. Non-verbal response: silence or mmmm
4. Summary: “Do you mean…”
5. Surprise reaction: “Really?” “Did she…?”
6. Inappropriate response
7. Overt correction

Categories of Responses

1. Repetition
2. Expansion
3. Rephrasing
4. Acknowledgement
5. Reduction

(Varonis & Gass, 1985)

It is possible for speakers as well as listeners to initiate a negotiation of meaning by overt use of an indicator, as in the example below taken from Varonis and Gass (1985, p. 78), where the speaker uses a comprehension check question:

A: I was born in Nagasaki. Do you know Nagasaki?

Following an indicator is the response, which may resolve the misunderstanding, or may add another layer of negotiation if it fails to be resolved. Possible responses are listed in the second section of Table 7. The final element is a reaction to the response, which is an optional part and typically closes the sequence before moving back into the flow of conversation, as can be seen in the final line of the conversation below, again taken from Varonis and Gass (1985, p. 78):

A: My father is now retire

B: Retire?

A: Yes

B: Oh yeah

When speaker B says ‘Oh yeah’, it represents a conclusion to this interruption and signals that the participants may now return to the topic of conversation.

The empirical findings into a range of interactional features in NS-NNS interactions are presented in Table 8. The research shows that overall such features are beneficial in various ways for language learners, though the disadvantage is that they would require a high ratio of NS to NNS to make such benefits pedagogically operational, and are therefore likely to be impractical in an English as a foreign language (EFL) context.

Table 8: Selected research into findings from NS-NNS interaction

Feature	Explanation	Context and Findings
Clarification requests	Listener asks a question about questionable element	NNS subject to clarification requests outperform control group with general request on past tense (Nobuyoshi & Ellis, 1993) Clarification requests positively correlated with development of question forms (McDonough, 2005)
Comp-rehension checks	Speaker questions interlocutors comprehension	Significantly more common in speech between NS -NNS dyads than in NS-NS (Long, 1981)
Hypothesis forming	Speakers trials language in interaction	On picture description task NNS, 1.8 hypothesis formed per min. 62% resulted in well-formed output (Shehadeh, 2003)
Negotiation of meaning	The above features collectively	Interactionally modified group performed better than modified input or control group, even students who do not take part can benefit (Ellis, Tanaka, & Yamazaki, 1994)
Recasts	Listener repeats misunderstanding, changing questionable elements	NNS with recasts produced more advanced question forms in information gap activity (Mackey & Philps, 1998) NNS with recasts outperform students exposed to input session (Long, Inagaki & Ortega, 1998) NNS with prompts with no corrected form perform significantly better than recasts (Lyster, 2004)

Subsequent research has investigated language acquisition from conversational interaction between NNSs. In a small scale study of 16 students in an EFL English class at a university in Thailand that investigated the link between target forms and small group and pairwork, McDonough (2004) found that students who were involved more in giving negative feedback and producing modified output demonstrated the most improvements in the target language, specifically the oral production of conditionals. However, she found that modified output was significantly more likely to be self-initiated, and that learners modified their output in response to negative feedback provided by their peers in only a minority of cases. The learners themselves, when surveyed, did not relate language acquisition to their experience in group interaction, preferring to ascribe it to explicit explanation of the teacher directed.

The ability of students to learn several general target structures from interaction with other NNS was researched by Adams (2007). Her study involved 25 participants who were each other's interlocutors on three tasks, each designed to elicit a different structure: locative prepositions, past tense and question forms. To investigate whether learning took place, the researcher made individualised tests that assessed points of language that came up in feedback episodes garnered from the recorded conversations, and found that on average the participants scored 60%. As Adams points out, this figure compares favourably with results from a study that used a similar methodology (Loewen & Philp, 2006) to investigate teacher-led learning and found a 50% learning rate. This suggests that learners can learn from each other to as great an extent as from teachers. It should be pointed out that the learning was not always positive: Adams also found evidence of the learning of non-target like forms from the interaction. Despite this cautionary example, Adam's study succeeds in providing empirical support for the ability of learners to learn a target language through interaction with each other.

The research reviewed in this section demonstrates conversational interaction can indeed benefit language learning, whether the interlocutor is a NS or NNS, and so justifies its place in a language program. Through clarification requests, comprehension checks, hypothesis forming, recasts and negotiation of meaning in conversational interaction, NSs have been shown to improve language learners' knowledge of the target language. The two studies reviewed above show that NNSs can learn from interacting with other NNSs, which demonstrates the potential for NNSs to play a similar role to sympathetic NSs. The next section will focus on studies that compare NS-NNS with NNS-NNS interactions. If it can be shown that NNSs can play a similar role as the NSs, then it is a major step to justifying curriculums that focus on conversational interaction and assessment between the NNS students enrolled in their programs, whether in an EFL context or not.

2.3.3 NNS as interlocutors for language learning in conversational interaction

With the bulk of this early research on interaction centring on NS interacting with NNS, it did not take long to be pointed out that in many language learning contexts, other language learners are, if not the most important sources of input, then certainly the most widely available (Porter, 1986). NNS interlocutors assume an even greater importance in foreign language contexts where access to NS interlocutors may be rare or non-existent. If it could be shown that NNS interlocutors provide useful input for other NNS learners, then it provides a strong justification for forms of assessment such as the GOT in which NNS interact with each other. For if the students know that their language will be assessed in a group discussion in which they interact with each other, it provides a powerful motivation to use the target language with each other to practice for the test, a vital consideration in foreign language contexts.

The first question concerned the extent to which negotiation of meaning could be found in NNS-NNS interaction in comparison to interactions between NNS and NS. From the beginning, there were grounds for optimism for NNS-NNS interactions as a context in which language learning could be enhanced. Varonis and Gass (1985) examined the transcripts of 14 NNS-NNS dyads, four NNS-NS dyads and four NS-NS dyads who were given the task to introduce themselves and find out about their interlocutor, and found that negotiation of meaning exchanges were significantly more common in NNS-NNS talk. Among the NNS pairs, there were more incidents of negotiation when they came from different L1 backgrounds or language abilities. The more different the background, the more incidents of negotiation for meaning took place. One reason they gave was that when two learners communicate they have less 'face' to lose when admitting their deficiencies (Varonis & Gass, 1985, p. 85). If there is more negotiation in NNS-NNS talk then there is more chance for the beneficial effects on acquisition to take place, thus in this regard NNS-NNS interaction may be considered beneficial in terms of learning potential than if the interlocutor is a NS.

As well as the question about the number of LREs that occur in NNS-NNS talk, there is the question as to the nature of the negotiation of meaning that takes place between NNS pairs, and how comparable it is to similar interactions taking place with a NS. An early study by Porter (1986) paid attention to this question by comparing interaction between NNSs at both advanced and intermediate levels, as well as with NSs. She found that even though NNSs would provide more examples of ungrammatical input to each other, the rate of the interactional features of repair and prompting were not significantly different, and that overall the input provided by NSs was no more comprehendible than that of other NNSs. There was a small difference in corrections whether paired with another NNS or a NS, with NSs making corrections to about 8% of the NNSs' grammatical and lexical errors, while learners corrected at a similarly low rate of 1.5% (Porter, 1986). Later research has confirmed that generally there is little difference in the opportunities to repair grammatical errors whether the interlocutor is another learner or a NS. For example, Sato (2007) found almost identical rates of repair according to the number of mistakes in his study.

One of the most important findings of Porter's study was that NNS made a mistake when correcting in only 0.3% of cases, lending to support to an earlier study that found that learners corrections of each other's language is usually accurate (Bruton & Samuda, 1980). In terms of output, Porter found that when paired with a NS, the NNS spoke less, showing that in this respect interaction with another NNS is more beneficial than when talking to NS.

Interestingly, although Porter (1986) did not include phonological mistakes as a category in her study, she noticed that when learners shared the same L1 background there were virtually no pronunciation miscommunications, and from this she concludes that it might be better to pair such learners together, since "their similar interlanguage phonologies will be comprehensible" (p. 209). Although reaching such a conclusion suits the purpose of this study (this dissertation's participants share the same L1) in terms of output 'pushing' L2 speakers to notice differences in their interlanguage (Swain, 2005), purposefully keeping those who share L1s in the same groups might well reduce the opportunities for negotiation since they may overlook their deficiencies and thus in the long run slow their overall development in the L2.

Subsequent research examining the quality of interaction that takes place between NNS compared to NS-NNS interaction confirms these early findings. A study that compared 10 NNS dyads with 10 NS-NNS dyads found that in terms of NNS producing modified output, there was little difference whether the interlocutor was a NS or NNS (Pica, Lincoln-Porter, Paninos, & Linnell, 1996). In terms of providing input, the results were more variable. Whether a NS or NNS provided more modified input as a result of negotiation depended on the task; on a jigsaw sequencing task NSs provided more, but there was no difference in a jigsaw story telling task. Overall, they concluded that while the input provided by NNSs was quantitatively less than that by NS, in terms of quality the feedback they provided would be beneficial to target language acquisition.

As opposed to the studies mentioned above that found more negotiation in NNS-NNS interaction, a study by Mackey, Oliver and Leeman, (2003) found that adult NSs provide significantly more feedback that indicated non-target-like forms were used than NNS. This study was on a larger scale, and included a mixture of NNS and NS among its 96 participants, half of whom were adults and the other half children aged between 8 and 12 (Mackey et al., 2003). These participants formed 48 dyads that included 12 NNS-NNS and 12 NS-NNS pairs per age group, with each dyad performing a one-way and a two-way information gap task. They found that adult NS interlocutors provided significantly more feedback than adult NNSs, but no significant difference among the dyads composed of children. When they examined the production of modified output in response to feedback, the child NNS dyads produced significantly more modified output, but there were no significant differences among the adults in this category. No other significant differences were found according to the age or interlocutor, but Mackey et al. (2003) noted that although adult NNS provided feedback on fewer errors, the feedback they did provide gave their interlocutor more opportunity to produce modified output. The authors single out the reason for the apparently contradictory finding that NSs provide more feedback as being due to their study investigating grammaticality rather than the overall quantity of negotiation, which in previous studies may have been over understanding as much as incorrectly formed grammar. Despite the greater quantity of feedback provided by NSs, when the researchers distinguished between occasions in which there was an opportunity for the interlocutor to modify their output on the basis of this negative feedback, or if the person who corrected continued speaking without such opportunity, they found that the feedback provided by NNSs provided significantly more opportunity for modified output than NSs. Although the quantity of feedback was less in NNS-NNS dyads, their research showed that the potential for learning through negotiation was potentially greater. In passing, it was noted that whether there was opportunity or not often depended on whether the feedback was given in the form of a recast or not, with recasts usually not providing an opportunity.

An earlier small scale study by Garcia Mayo and Pica (2000) focussed on comparing higher proficiency NNSs interacting with each other or with a NS. The researchers compared the interactions of seven advanced NNS-NNS and seven NNS-NS dyads on information gap and argumentative discussion tasks to investigate the modified output that occurred. Although they found that as much modified input, feedback and output took place in the interactions between NS and advanced NNS, the context in which they occurred differed. In NNS pairings the amount of negotiation was limited; instead the completion of each other's sentences and self-correction served to produce the input, output and feedback in their interactions. Advanced NNSs appear to be similar to NSs in regard to their preference for self-correction.

Unlike previous studies, Sato and Lyster (2007) use stimulated recall to gain a more detailed understanding of the interactions between learner interlocutors and with NS partners. They investigated the language elicited by the same eight NNS participants with either another of the NNSs or one of four NS participants on a couple of two-way information gap tasks. Feedback was categorized as either reformulation, which contained elements of the utterance called into question, or elicitation, which did not (for example clarification requests, or confirmation checks). The proportion of reformulated to elicited feedback differed, with the native speakers using a statistically significant greater amount of reformulation (59% to 41%), whereas NNS used a majority of elicitation (34% to 66%).

The other important finding was that with a NNS as an interlocutor there is a significantly greater chance of output modification, regardless of the type of feedback. The stimulated feedback investigation revealed that the high proportion of reformulations from NS was because they could guess what the NNS was trying to say. This was true on some of the occasions in the NNS-NNS dyads, mostly by virtue of their shared L1. Another effect of talking to "perfect speakers" (p. 139) felt by NNS, was that because the NNS knew that the NS understood, further modification of output was not felt to be necessary. The lack of modified output appeared to be partly due to the intimidation that NNS felt when talking to NS, relative to the easier atmosphere they felt when talking to a fellow NNS. Time was another factor that was reported, with learners having less when interacting with NS.

Using the same data as Sato and Lyster (2007), Sato (2007) found that the likelihood of an LRE being resolved in a grammatically accurate way did not differ whether a learner's interlocutor was a NS or a NNS (p.192). This is important since it might be assumed that talking to a NS would be preferential due to the grammatically correct model they provide, but Sato (2007) found that there was no significant difference in the amount of modification of erroneous utterances between dyads in which a participant was a NS or between two learners.

Sato's (2007) study also revealed a potentially important advantage of interacting with another learner verses a NS: it seems that the feedback received from another learner is more memorable than that received from a NS. When delayed modification was examined, errors that received feedback were much more likely to be modified at a later time if a NNS gave the feedback, whereas it happened much less frequently than when a NS gave the original feedback (Sato, 2007, p. 196). Although the small numbers of participants in this study mean that this finding should be treated with caution, it received support from the retrospective recall that the participants undertook. Here the participants reported being less careful about the grammar they used with their NS interlocutors because they felt NS would be able to understand what they felt to be their low level of English, and correspondingly more care with NNS interlocutors so that they would be understood. Overall, Sato concluded that negotiating with peers gave learners more opportunity to notice the gap between what they wanted to say, and their ability to retrieve and use it (2007, p. 199).

Another context in which interaction between NNSs has been compared with NS-NNS interaction is the classroom. This context is of particular concern for educators using a communicative approach who commonly use group work and need to know more about the value of the support that classmates can give each other. It is particularly relevant to this study since it uses groups of NNSs. Pica and Doughty (1985) compared a teacher fronted discussion with a single group of four students using a ranking activity, and found that there was very little conversational adjustment in the interaction of participants whether they took part in the class or the group discussion activity. Moreover, there was a wide variance of results in the three classes in the study, with one class reporting significantly more conversational adjustments in the teacher fronted discussion, one significantly more in the group discussion, and one showing no significant difference. Pica and Doughty ascribe this to the impact of the task, which did not require the meaning of the participants' utterances to be clarified or confirmed or mutually comprehensible or even participation. They also noted that the task did not require all participants to take part in the discussion, with a few dominant members in the group or the teacher to do the majority of the speaking. This tendency has been noted in studies that have examined the interaction of the GOT, which often uses groups of four (particularly note Nakatsuhara [2011] and Van Moere [2007] in this regard).

The researchers repeated the study with an information gap task that required all members to participate (Doughty & Pica, 1986) to compare teacher-fronted with dyads and small groups of four. The results showed, as expected, significantly more modified interaction in groups of two and four than in the teacher-fronted exercise, but no significant difference between dyads and groups of four.

The finding that the interaction of class discussions and groups of four can be dominated by a few students or individuals in these studies (Doughty & Pica, 1986; Pica & Doughty, 1985) needs to be seen in the light of the researchers' decision to only analyse the language of one group per class. In practice the entire class would be split into groups of four, meaning that the group interaction would be running concurrently with other students, while in a classroom only one discussion can occur at a single time, albeit witnessed by the entire class. To accurately reflect the quantity of interaction, the single class discussion should be compared with the on-going discussions that occurred in all groups, not just one. Quite different results would be expected had this approach been taken. Also, as research by Nakatsuhara (2011) and Ockey (2009) make clear, not all small groups include assertive individuals capable of dominating a group. To avoid using groups because a few of them may be dominated by assertive individuals would be unfair to such groups. Instead, particularly in classroom practice activities, measures can be taken by selecting appropriate tasks and giving guidelines that ensure that everybody in a group has a role to play. In this way, even shier students can build up an "architecture of the practice" (Young, 2000, p. 6). that they can transfer to an assessment situation such as the GOT.

To summarize, the literature shows that although interacting with a NS has advantages in terms of grammatical accuracy of input and a greater rate of correction or attention paid to non-target language in some contexts, there are also advantages of interacting with another NNS. In NNS-NNS interaction there is often more negotiation of meaning, there is little difference in the amount of modified output compared to interacting with a NS, corrections by other NNS are usually accurate, it may be easier to acknowledge mistakes, the quality of feedback may be easier to comprehend, it probably presents more opportunities for learning, it may be more memorable, and finally, interactions with other NNS may be less face threatening. The significance of these findings for this dissertation is that it shows there are just as many if not more advantages of having other language learners as interlocutors than NS. Administrators can feel justified that they do not have to import extra NSs to act as interlocutors. Similar benefits can be accrued from having language learners act as interlocutors for each other, and so setting up a system through the curriculum and assessment system to encourage this is justified as a valuable method for achieving the overall objective of producing graduates who are communicative in the target language.

2.3.4 Studies focusing on NNS-NNS interaction

Having established the value of NNS-NNS interaction in comparison to NS-NNS, the final part of this section will turn to research that illuminates the nature of interaction in NNS-NNS communication. Research in this area will allow a greater understanding for administrators who wish to place NNS-NNS interaction at the heart of their curriculum, and also provide insights into what can be expected of interactions that are elicited by the GOT.

Early research that examined language function in NNS-NNS interaction by Bruton and Samuda (1980) was inspired not by interaction studies, but by the error analysis that originated from the influential article by Corder (1967). Bruton and Samuda showed that NNSs do use a variety of techniques to provide corrective feedback to each other, whether requested to or not. In the ten hours of NNS-NNS interactions they recorded over a week, the corrections learners made to each other's utterances tended to centre on lexical more than syntactic, pronunciation or understanding errors, but the categories of errors checked were considered "very close to the types normally treated by teachers in the classroom" (Bruton & Samuda, 1980, p. 51). The students in the study treated errors by giving alternatives, offering choices, using repair questions, straight rejections, or by self-correction. In their corpus they noticed only a single occasion of a correct statement being mistakenly changed to an incorrect one. However, the data was not quantified, so results must be interpreted cautiously.

A common theme in the literature was to distinguish differences in negotiation for meaning that takes place between learners at different levels. A study conducted by Williams (1999) explored the relationship between level of proficiency and negotiation for meaning in eight language learners at four proficiency levels, with two students per level. The data was collected by a microphone as they took part in a variety of classroom activities, though only those that involved learner-learner interaction or requests to the teacher that were initiated from learner-learner interaction were included in the study. In the end about 25 minutes of interaction time per level was included in the study. The researchers aimed to identify LREs which were defined as interactions in which students talk about the language itself, or question their own language use either implicitly or explicitly. Among the data, the 255 LREs that were identified were not distributed equally among the difference levels: the bottom two levels had a similar number, and then there was a substantial increase in the number of LREs at the third level, and a further increase in quantity in the highest level of student. The reason they proffer for this was that lower level students are focussed more on the communicative act itself, and simply do not have the available mental resources to consider the form of what they are saying (Williams, 1999).

Among other trends noted was that as the proficiency level increased, the proportion of requests to the teacher dropped as the proportion of requests to other learners rose, as higher level students increasingly relied on each other for support. Also rising with proficiency level was the proportion of meta-talk and feedback that centred on form that occurred, while there was a decreasing proportion learner-leaner negotiation. Perhaps the clearest finding was that for all three proficiency levels, about 80% of the LREs were related to lexical rather than syntactical matters. An explanation for this can be found in Pica's (1994) assertion that negotiation for meaning is primarily about comprehensibility, and so tends to involve lexis rather than syntax, which will only be the focus if it makes a difference to the meaning.

A slightly different picture of NNS-NNS interaction at varying proficiency levels was provided by Iwashita (2001) in a Japanese L2 context. This study had 24 students in dyads that were composed of lower level learners, higher level learners, or a mix of higher and lower, with each grouping consisting of four pairs. Using one two-way and two one-way information gap tasks, Iwashita investigated the occurrence of confirmation checks and clarification requests. She found that the learners mostly relied on confirmation checks which did not allow as much opportunity for modified output. No significant differences were found between groups of different proficiencies, showing that for these learners the level of negotiation was the same no matter who they were paired with. In contrast to Williams (1999) she found that the majority of the modification was syntactic rather than lexical, which Iwashita ascribed to the low level of the learners not having a sufficiently large vocabulary to find a synonym. Comparisons are difficult to make with William's study, but it seems likely that even the higher proficiency Japanese as a foreign language students in this study had a low level of L2 compared to the participants in Williams (1999), who were in an English as a second language (ESL) context.

The importance of the role of the task when investigating interaction with different and same proficiency dyads was shown in a study by Yule and MacDonald (1990). In this study 40 learners undertook a single one way direction giving information gap task in which the maps given to the participants were mostly the same but had been altered slightly in order to promote incidents of negotiation for meaning. They arranged their dyads so that in 10 of them the sender of information was the higher level student, while in the other 10 the sender was at a lower proficiency level. They found substantial differences in the language elicited, with groups in which the lower level student sending the information taking on average twice as long, and including an "overwhelmingly" (p.454) larger amount of negotiation. In pairs with the high level participant sending the information very little negotiation was noted, as the higher level student frequently ignored indications from the lower level receiver that something had gone wrong, or the lower level student gave signs of understanding when in fact the directions were not working. They noted a tendency in such dyads when the higher level student simply abandoned part of the task, which was rare in the dyads in which the lower level dyad had the sending role. Dyads in which the lower level had the sending role, by contrast, were marked by a collaborative approach in which negotiating moves such as clarification requests and co-construction took place. Although this study presented no statistics in its support, it is a convincing portrayal of the need to consider the roles of proficiency and task in creating activities for collecting data.

The lack of negotiation in NNS-NNS talk that Yule and MacDonald (1990) noted in one of their tasks has been a matter of concern for some researchers. A study by Foster and Ohta (2005) concurs about the lack of actual negotiating that may occur in NNS-NNS talk. They point out that actual episodes of negotiation interrupt the flow of conversation, and so tend to be avoided. Using strict definitions for episodes of negotiation of meaning they found minimal examples of clarification requests, confirmation and comprehension checks amongst their English L2 and Japanese L2 subjects in a one-way information giving task. Even when they did occur, they found the actual function of them was not necessarily to negotiate meaning, but instead could function to signal interest in what their interlocutor was saying, for example. They found that 11 out of a total of 39 participants did not use a single one. This result corroborated a previous study that found little negotiation for meaning and students opting out of participation and that where it does take place, the overall statistics would have been obfuscated by a few individuals who dominated proceedings (Foster, 1998).

When Foster and Ohta (2005) examined the transcripts they found most of the modified output came from self-repair or in response to their interlocutor's expression of interest or encouragement. In their qualitative analysis they identified other means by which the interlocutors could collaborate to support each other that would be beneficial for language acquisition. Amongst them were co-construction, other correction, self-correction, and continuers, as well as linguistic aid given in response to a direct appeal or without, as summarized in Table 9. In the data, the researchers L2 Japanese were of a lower level than the L2 English students, and they found a correspondingly higher level of negotiation for meaning and collaboration than amongst the more advanced learners, who, they note, appear to have found the conversational task much easier.

Table 9: Interactional processes related to assistance

Feature	Definition	Example
Co-construction	Joint creation of an utterance	- completing another's sentence - adding phrases to create an utterance
Other-correction	Correcting another's utterance	- recast of an incorrect utterance
Self-correction	Speaker corrects own utterance	- repeating an element of own speech with correction
Continuers	Encouragement of speaker to keep talking	- listener repeats element of speakers utterance with rising intonation

- adapted from Foster and Ohta (2005)

Later research that supports Foster and Ohta's (2005) contention that actual negotiation for meaning is not that common is a study by Fujii and Mackey (2009). This classroom based study involved 18 Japanese students of English to investigate interaction in two decision making tasks; a survival task in which they should rank items, and a decision making one in which they were presented with a difficult situation in a homestay context. Given that these were not information gap activities in which the focus is on communication it is not surprising that examples of peer feedback were rare, with the frequency of feedback the learners receiving averaged just twice, ranging from none at all to nine times in about 25 minutes of total interaction time. The average per turn that contained a non-standard form was just 7% on the ranking task and 13% in the decision task. Although the amount was small, learners had an opportunity to modify their output in 50% of the ranking task and 81% of the decision task cases and of these occasions the modified output resulted in 46% and 62% respectively, both considerable proportions. When examining the data qualitatively they found that clarification requests and confirmation checks tended to result in complete changes to the original utterances, and that recasts tended to focus on mostly on lexis, but often contained non-target like language use. In the qualitative analysis, Fuji and Mackey also noted how the participants managed to avoid situations in which it might be necessary to give feedback altogether by just carrying on, or by helping their interlocutor before a communication breakdown is explicitly signalled. That the participants could do this relied to some extent on their shared cultural background (Fuji & Mackey, 2009, p. 288), the usual situation in EFL contexts.

Other research shows it may be that the frequency of negotiation for meaning is dependent to a large degree on the nature of the task, as a study by Gass, Mackey and Ross-Feldman (2005) shows. This research was carried in response to Foster's (1998) postulation that negotiation for meaning may be inhibited in classrooms compared to laboratories. To test this assertion, the researchers compared several interactive tasks in the laboratory with the same tasks in a classroom in a study that involved 74 L2 Spanish students. The researchers found that there was no significant difference in the quantity of negotiation for meaning (clarification requests, confirmation and comprehension checks), language-related-episodes, and recasts. However, significant differences were found in the quantity of interactional features according to the task the students took part in, with a consensus task producing significantly fewer examples of all three features of interaction.

Gender also appears to play a role in interaction, as a study that involved 64 L1 Spanish subjects, half female and half male showed (Ross-Feldman, 2007). Each participant performed three tasks that involved one- and two-way information gaps and open ended decision-making, in homogenous and heterogeneous pairs. Overall, there were significantly more LREs on the decision making task than the information gap activities, and although LRE episodes were typically initiated more often by females than males, the number did not reach significance in mixed gender dyads. However, when the resolution of LREs were examined, they were significantly more likely to be resolved if initiated by a male rather than a female. On the picture story task, it was found that male-male dyads produced less LREs than dyads in which a female was present. The significantly greater number of LREs on the picture story task is an aberration when it is considered that in the literature, this kind of task typically generates the fewest LREs. The authors say that this may have been because the activity necessitated not only speaking but writing, as the subjects were required to do, and this focussed them more on lexical matters. The greater number of LREs in the picture story task may also have meant that gender differences made more of an impact in this task than in others, where there were too few LREs to make the numbers significant. Task equality is another issue that the author believes may impact on gender balance, with some tasks having the potential for one person to take control. The conclusion the author reaches is that it is favourable for both males and females to be paired with a female.

To conclude this section, the research that has examined the nature of interaction of NNS has shown that learners do provide feedback to each other, whether asked to or not, that higher level learners are generally more actively engaged in negotiation for meaning, negotiation may be more commonly about lexis than syntax, the amount of negotiation may differ according to task and mix of proficiency levels, that gender may influence the frequency of negotiation, and that interaction may benefit the language learning of those who participate actively. For this dissertation, this section allows the interaction that is found to be elicited by the GOT to be put into the context of the studies surveyed here.

2.4 Summary of the literature review

In this section above, the literature on the role of conversational interaction in language learning was reviewed. It started with the genesis of the theory that output could play a crucial role in acquisition, and how it evolved into the interaction hypothesis. This was linked to the previously explained notion of interactional competence, and then some of the early literature on it was reviewed, particularly those who were critical of the noticing hypothesis. From this came the need of a cognitive framework which could encompass noticing in the interaction hypothesis, and the MOGUL framework was put forward as a solution at a cognitive level. With a theoretical framework established, the review proceeded to examine the empirical findings, which it divided into three sections. The first was to show what has been learnt about conversational interaction in the context of language learning, and this section covered some of the early literature which mostly investigated NS-NNS interactions, but also included some acquisition papers using NNS-NNS interactions. This section found that:

- LREs follow identifiable interactional moves (Varonis & Gass, 1985)
- Features which benefit language learners that NS use are clarification requests, comprehension checks, hypothesis forming, recasts and in general negotiation of meaning (see Table 8)

- In NNS-NNS interaction, those who most actively involve themselves in LREs are those who are more likely to benefit from them (McDonough, 2004)
- Negative feedback and modified output are those most likely to improve their language (McDonough, 2004)
- Even when students learn through interaction, they may not realize it (McDonough, 2004)
- NNS-NNS interaction is at least as efficient, and perhaps slightly better than teacher fronted classroom learning (Adams, 2007)

The second issue that this literature review addressed was whether NNSs interacting together could offer similar benefits that NS can do. This section found that:

- NNS-NNS interactions may be similar to NS-NNS interaction in the rates of prompting and repair (Porter, 1986; Sato, 2007); level of comprehensibility (Porter, 1986); chance of grammatically accurate outcome (Sato, 2007)
- NNS-NNS interactions may result in similar (Pica et al, 1996) or greater (Sato & Lyster, 2007) level of modified output than NS-NNS interaction
- There may be more negotiation for meaning in NNS-NNS than in NS-NNS (Varonis & Gass, 1985), or more opportunity for modified output (Mackey et al, 2003) though resulting quantity of input may depend on task (Pica, et al, 1996)
- NS may give as much feedback (Sato, 2007) or more feedback on non-target-like forms than NNS (Mackey et al, 2003)
- In classroom interaction, more modified interaction may occur in groups of students than in a teacher fronted exercise (Doughty & Pica, 1986)

It was concluded that the quality of feedback in interaction provided by NNS differs from that of NS, but that it may well offer more advantages to that provided by NS speakers.

The final section reviewed studies that have examined NNS-NNS interaction in its own right. The main findings here are that:

- NNSs rarely make mistakes when they do correct (Bruton & Samuda, 1980; Porter, 1986)
- LREs are more common among NNS at a higher level than a lower level (Williams, 1999)

- LREs among NNSs are more common for lexical issues than for syntax (Pica, 1994; Williams, 1999)
- Feedback by NNSs may provide interlocutor more opportunity for modified output (Mackey et al, 2003; Sato & Lyster, 2007); be more memorable (Sata, 2007)
- NNS-NNS interactions modified output may result from expression of interest, co-construction, self-correction, responses to requests for aid (Foster & Ohta, 2005)
- NNS may give feedback more by elicitation while NS do so more by reformulation (Sato & Lyster, 2007)
- Negotiation in NNS-NNS interaction may be limited; rather modified output may come in response to self-repair, interlocutor's expressions of interest, co-construction, other and self-correction, continuers and help on linguistic matters (Foster & Ohta, 2005)
- The quantity of negotiation for meaning is more dependent on task than whether it is classroom or laboratory based, with consensus tasks producing significantly less (Gass et al., 2005)
- In NNS-NNS interaction, the interlocutors may prefer to continue the interaction rather than give feedback (Fujii & Mackey, 2009)

When this evidence is taken collectively, the literature clearly demonstrates to administrators the need to develop communicative curriculums for their students that takes advantage of a resource that is inherent in nearly all language programs: other NNS students.

Once the value of interacting with other NNSs for learning the target language is accepted, it becomes necessary to find a form of assessment that encourages such behaviour. Indeed, the congruence of how language is learnt and the method of assessment is a cornerstone of testing textbooks (Bachman & Palmer, 1996; Brown, 2004; Hughes, 2003), with the objective being to test in a way that results in a positive effect on the teaching and learning of the course. Hughes points out that where the teaching and the testing do not agree with each other, the likely result will be felt negatively by teachers and learners (2003, p. 2), and this has been supported by research (for example, Choi, 2008).

This takes us back to the first part of the literature review, which examined the direct assessment of speaking from its mostly post-WW2 origins, when the role of interview-led tests was almost the unquestioned test of choice, to the current era, in which peer interaction tests have come to the fore. The first part reviewed the literature that critiqued the interview format for the language that it elicited, and found that it was far from what might be expected from conversation. Following this, the literature on peer interaction tests, first in pairs, and then in groups of typically three or four students was reviewed. This literature shows examples of these formats eliciting both conversation-like language and non-conversation-like language, at least showing the potential of having a test that encourages peers to interact with each other that would suit the purpose of the administrators.

The literature on the paired format has been encouraged by what is by some measures the world's largest testing organization (Cambridge Assessment) using it in its most important suite of exams, and has allowed increasingly sophisticated insights into interaction, many of which can be applied to the GOT. Particularly the recent research into the ability to interact at different proficiency levels (Galaczi, 2013), how raters rate (Ducasse & Brown, 2009; May, 2011, Orr, 2002) and the description of identity in paired interactions (Lazaraton & Davis, 2008) are applicable to the grouped interactions.

Research into the GOT itself has also developed considerably, with the most recent papers shining a light onto the language elicited to illuminate how various factors affect it. Particularly relevant to this dissertation are the findings about the impact of the assertiveness of group members on the raters (Ockey, 2009), the nature of the interaction and how various factors impact on it (Gan, 2010; He & Dai, 2006; Leaper & Riazi, 2014; Van Moere, 2007), and the possible identities that may be expressed within it (Luk, 2010).

If there is one thing that the literature on paired and group formats have in common is that there is an almost complete lack of research on how test-takers perform in the test over time. All research conducted so far has typically used data from single performances in the test, or within a short time, and thus leaves many questions about peer interaction tests begging. For every piece of research that has been done, questions may be raised about the consistency or stability of the phenomenon. It can usually be asked if the person would perform in the same way habitually, or if these aspects are traits or temporary issues of development. By exposing how students interact across a period of years, not only can it confirm much of the research that has already been conducted, but it can reveal aspects about the GOT that have been obscured thus far.

The next chapter will detail the quantitative section of the current study, firstly describing its methodology, and then explaining and discussing its results.

CHAPTER THREE

The Quantitative Phase

3.0 Introduction

This chapter presents the methods, results and discussion of the quantitative section of the study. In the methods part of this chapter I will first describe the design, then move on to the participants and procedure of the GOT at the institution where the data of the study was collected. Having persented this, the chapter proceeds by outlining the transcription process and describing how the temporal aspects of the group interactions were taken into account. The first half of this chapter ends with an explanation of the procedures for collecting the data and its statistical treatment. The second half of the chapter provides the results and discussion of the quantitative findings.

3.1 Design of the quantitative phase

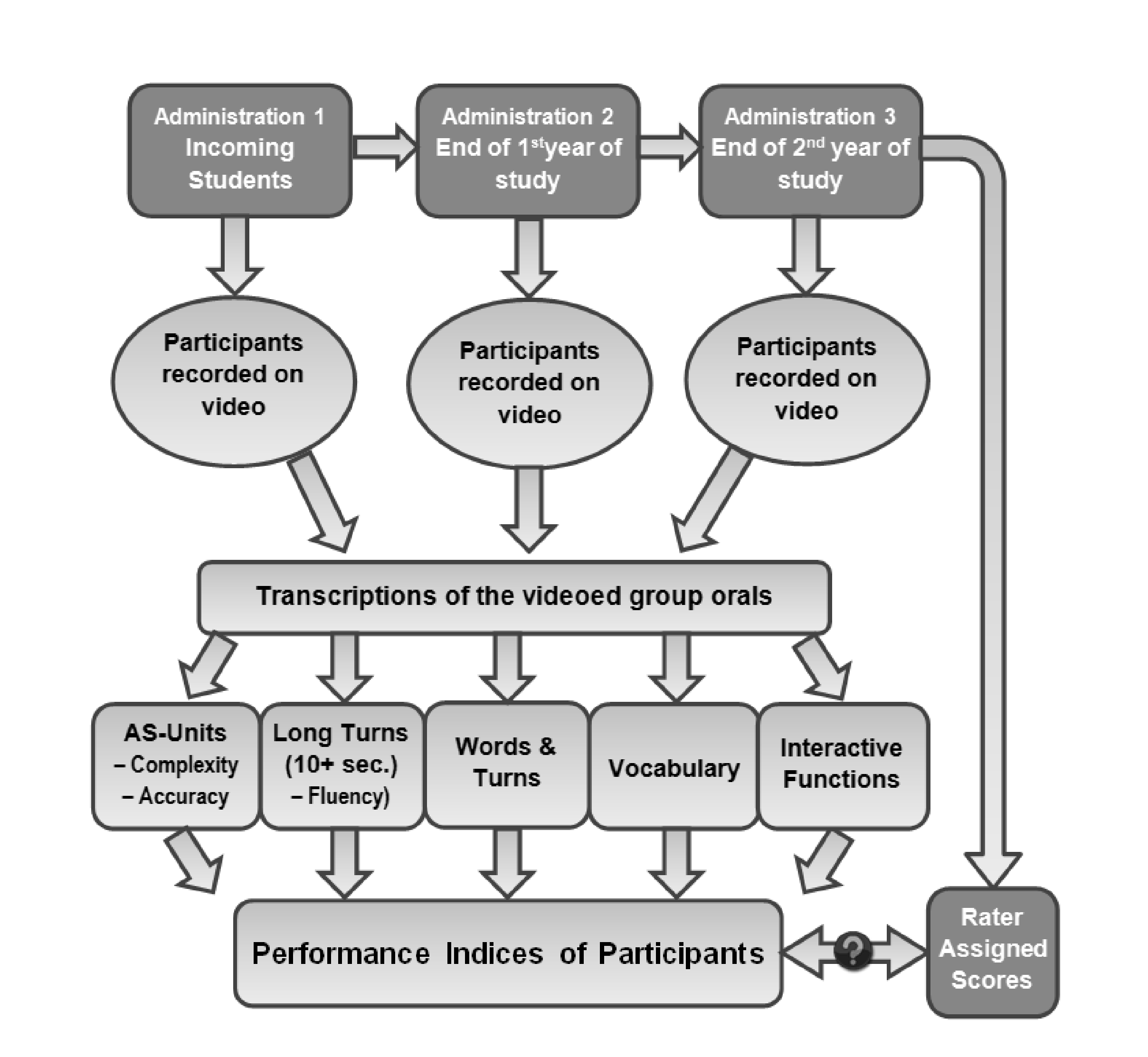
This chapter outlines the methods and procedures employed to chart the communicative and linguistic progress that 53 participants displayed in group oral tests over the three times they took the test in two years of study. Videos of the participants' GOTs were transcribed and timed, and it is from these that performance indices were created to measure the complexity, accuracy, fluency, vocabulary and interactive functions they produced in the test. The overall data collection and analysis is summarized in the flow chart presented in Figure 5.

3.1.1 Participants

The participants were drawn from a Japanese private university that specializes in communication and languages. This university has developed a four skills language test, with GOT as its speaking component, so that the entire student body is supposed to take it. Since 20% of the grade of the first and second year core subject scores come from their performance in this test, the freshmen and sophomores are the most motivated to take in the test. In the first two administrations it plays the additional role of placing students into streamed classes for the coming academic year, but this is not applicable for second year students taking the GOT the third time because the final two years of their degree they can choose elective courses. This removes one source of motivation for students to do well the third time they take the test, which may have an impact on the effort they exert to do well in the final administration of this study. Third and fourth year students have less incentive to take the test since it is not used for their placement, and most of their professors do not make it a requirement for their courses.

The number of students taking the test each year amounts to over 1300 who take it over two days in January, and an additional 700 or so incoming freshmen who take it in March immediately

Figure 5: Data sources for the quantitative part of the study



Administration 1 Incoming Students
Administration 2 End of 1st year of study
Administration 3 End of 2nd year of study
Participants recorded on video
Participants recorded on video
Participants recorded on video
Transcriptions of the videoed group orals
AS-Units – Complexity – Accuracy
Long Turns (10+ sec.) – Fluency)
Words & Turns
Vocabulary
Interactive Functions
Performance Indices of Participants
?
Rater Assigned Scores

before classes begin. The 53 participants of this study were English major students who were videoed as incoming freshmen in March 2004, before their classes start, and at the end of their first and second year in their program in January 2005 and 2006 respectively. The participants had a relatively uniform background that is typical for this university. They were all 18 years old when they first took the test, mostly female (79%), with a generally uniform educational background. Though some had started studying English earlier, they had all studied English for at least six years at Japanese middle and high schools. Before they were admitted to this university they had concentrated on entrance examinations that consist of multiple choice reading, grammar and listening items, so they had not needed to study or practice speaking English, and most of the students had not studied abroad before they started university, though some of them had done short home-stays of a week or two in English speaking countries during their middle or high schools. The university where this test takes place is a languages university, so students who choose to come here are usually motivated to improve their English in general, and especially their speaking skills. Certainly, in the first two years at this university their core classes expose them to more English only classes than most of them have ever experienced in their previous schooling: over the course of a 14-week university semester they have 15 hours of classes taught by native speaking staff per week, making for a total of 420 hours of classroom study before they take the speaking test a second time, and 840 hours in total before the third and final time.

3.1.2 Data collection instruments

3.1.2.1 The task of the group oral

The task for the three or four participants of the group oral test is to discuss a single multi-question prompt for up to 10 minutes or until the raters get a sample of their spoken ability sufficient for awarding a score. For each administration, the testing committee creates three or four prompts which are supposed to be of equal difficulty, and one of these is assigned at random immediately before each group is formed. The prompts were created according to the guidelines developed from previous administrations. That is they are created on subjects:

> (1) on which examinees were likely to have a ready opinion, (2) which are similar to the types of discussion questions students might have encountered on many occasions in classes, and (3) which do not require specialized vocabulary or knowledge. (Bonk, 2003, p.12)

The w ording of the prompts should follow a similar pattern, but they do not have to follow it exactly. The pattern is to start with an initial context-setting statement followed by a series of questions ranging from the concrete to more abstract, one of which requires test-takers to enumerate the advantages and disadvantages of the topic. A Japanese translation appears beneath each prompt to ensure students' complete understanding of the topic. The prompts used in the years of the test are included in Appendix A.

3.1.2.2 Rater training procedures

The raters of the GOTs are the approximately 40 native English speaking staff who teach at this private university. Rating this test is a contractual obligation for these teachers; they do not get paid extra for it, and they cannot opt out of it, which is not an ideal situation from a test administrator's point of view. The teachers could easily see this task as an imposition, and since the administrators cannot ask poorly performing raters to step down, it could lead to unacceptably low quality ratings in some cases. Yet the Rasch figures for the administrations have shown that for the most part the teachers carry out this duty diligently (Bonk & Ockey, 2003; Van Moere, 2006), and since there is no in-built motivation to do a high quality job, it is their natural diligence that the quality of rating in this administration depends on.

To train the teachers for rating, the testing committee (which is composed of teachers at the university who are interested in testing) chooses eight or so clips of GOTs from previous administrations that show a range of typical and interesting test behaviour, and which include test-takers at different levels. The test committee spends considerable time going over these videos individually and as a group to decide on the most suitable grades according to the scoring bands. The training takes place a few days before the test administration, with all teachers attending the session. As the videos are shown, the teachers assign grades which are then compared with the test committee's scores along with justifications and discussion of that particular grade. The teachers with diverging ratings gain an understanding why the testing committee's grades were chosen, and on what basis the scoring bands should be interpreted. Only in rare cases will the committee accept revisions to the grades they originally assigned. Every year a different selection of videos are chosen, leaving open the possibility that the standards of rating may be affected administration to administration, as the raters may be oriented to the particular features represented in the training session.

3.1.2.3 Procedure of the test administration

In the morning of the test day, the students take the reading, listening, grammar (tested using multiple choice questions) and essay writing section of the test, and in the afternoon they take the GOT. At the end of the morning's test, all students are given notice of the afternoon's oral test in a standardized announcement, and reminded that the purpose of the assessment is to sample their conversational ability. They do not find out the room, time or who they will be tested with until the afternoon of the test. After the test-takers are seated in the test room, the GOT begins when the raters inform them that they have one minute to read and think about how they will respond to the prompt. After this time elapses, a rater tells them that anybody can start talking, and they should continue until told to stop. Raters do not intervene until they inform the group that the test is over. The only exception is when a student does not speak enough to assign a score, in which case a rater has the authority to prompt the test-taker by asking a general question. This occurs more frequently the first time the participants take the test, only rarely is it necessary in subsequent administrations. The two raters independently give a grade using a five band rating scale which is comprised of fluency, pronunciation, grammar, vocabulary and communicative skills bands (Appendix I, p. 426). The raters are told not to compare their grades, and change partners regularly according to a schedule that was created in order to ensure the data is linked sufficiently for the Rasch analysis.

After the test, the mark sheets the raters used to record the scores are collected and passed to the members of the testing committee who process the sheets. The scores are scanned into an excel file and sent to the testing consultant who will perform a Rasch analysis in order to take into account the relative strictness or leniency of the raters and report quality measures. Finally, the Rasch adjusted scores are reported to the students and the administration.

3.1.2.4 Video recording the GOT

As a matter of routine, every administration of the GOT has some rooms equipped with video cameras to record a sample of the group discussions. The video recordings are used for rater training sessions as well as for research. For this study, rather than the usual random selection of students, it was planned to record the same students the three times they took the test in their first two years of study – as incoming freshmen in April 2004, then at the end of their first and second years, in January 2005 and January 2006. Each time the participants did the test, they had different interlocutors and different raters.

3.1.3 Data organization and coding

The first step in the analysis of the students' performances is to get an accurate transcription of the recording of their GOTs, and put them in a form that allows data analysis. The videos were initially transcribed using Eggins and Slades' (1997) coding system (see Appendix B), which given the quantity of data, was a suitable compromise between getting sufficient detail for the purposes of answering the research questions and being able to complete expeditiously. When transcribing the GOTs, the utmost care was taken to consistently record every utterance and ascribe it to the correct person, including all backchannels, hesitations, and other disfluencies.

Recording the time that each GOT took was important, because raters could call an end to it when they considered they had a rateable sample. Since longer tests potentially give participants more opportunity to speak, where appropriate the data was normalized by dividing by test time and the number of participants (since groups of either four or three were possible), allowing the relevant data to be calculated 'per opportunity to speak'. This was deemed the best method of normalizing, since it captured all of the language elicited without loss of data – an important consideration given that some test-takers talked more towards the end of the test than at the start, and vice versa.

The transcripts were copied into Microsoft Excel line by line to ensure that counts could be achieved with as little human error as possible and for efficient calculation of the indices. The turns in the transcribed data were checked to ensure they were consistent with the original oral data. Following this, they were analysed using Analysis of Speech Units (AS-units) (Foster, Tonkyn & Wigglesworth, 2000) by a research assistant before being checked by the researcher. AS-units were chosen as they were not only the most suitable unit for analysing conversational language, but also the most clearly described in the literature (Foster, Tonkyn & Wigglesworth, 2000). One complete test was selected to be audited independently for AS-Units and clauses by two qualified professional language education experts, each with over 20 years' experience of teaching at various levels. The agreement between the original and the first auditor's versions was calculated as 88.4%. Differences in interpretation were compared to the second auditor's version, which was used as a tie-breaker, and remaining inconsistencies were discussed. This resulted in several changes to the coding system to ensure consistency for the rest of the data analysis.

3.1.4 Measures

This study measures language improvement by means of indices for complexity, accuracy, fluency, vocabulary and interactive features. The first three of these are known collectively as CAF (Housen & Kuiken, 2009), and have been used in a large body of research investigating how varying a task's facets or parameters affects the language elicited (Ellis, 2009; Robinson, 2011; Skehan, 2009). Much of this research has focused on non-interactive monologue tasks such as narratives based on pictures (see Ellis, 2009 for a review). Only a few studies have used tasks in which the participants interacted with each other. Of these Taguchi (2007) investigated the impact on fluency and appropriateness in situations with a high or low degree of difference in power, social distance and imposition using role plays in pairs. Robinson used direction-giving (2001) and narrative tasks (2007) with paired subjects and comparison tasks were used in a study that used L2 Dutch dyads (Michel, Kuiken & Vedder, 2007). Although the tasks in the aforementioned papers were interactive, it is important to note differences between them and the task in this study. All of the tasks in the studies mentioned here were more constrained, like the comparing tasks of Michel, Kuiken and Vedder (2007) or were one-way information giving tasks conducted in dyads (Mackey, 2012). These tasks would more consistently generate longer turns in comparison to the free discussion task used in this dissertation in which the three or four participants are not assigned a role in the interaction. Despite these differences, there seems to be no reason why CAF indices should not be used to analyse a non-directional discussion task conducted in small groups of three or four. The difference is that there is no guarantee that the speakers in group orals will produce longer turns, since the task is not designed specifically to elicit them.

CAF measures have also been used to investigate language acquisition over time, though this use has not been as common as its predominant use discussed above. One example is a study of 14 students over a year on a study abroad program in a British university (Serrano, Tragant & Llanes, 2012). The researchers found that over the three times that the measures were taken the students increased significantly in the measures at various times. The largest increase in the spoken data came between the first and the second measures in fluency, and lexical richness of the participants also improved significantly over this period. From the second to the third measures accuracy and improvement in the accuracy figures was also measured. Over the entire period, accuracy, fluency and lexical richness improved significantly, and large effect sizes recorded for fluency (Cohen's d = 1.08) and accuracy (d = -1.33) with medium effect sizes for lexical richness (d = 0.65). It seems then, that using the CAF measures as an index of language gain is plausible and appropriate.

The measures that were counted or calculated to answer the research questions of this study are summarized in Table 10, and are explained in more detail in the sections that follow. Table 10 includes the core indices which are foundational in the sense that they are used to calculate many of the other indices. These core statistics are the number of words, clauses and AS-units, which are given as the overall raw count, the count in long turns, and per opportunity to speak given the number of participants in the group and total time of the GOT. These are all important figures: the number of spoken words indicates the total exposure of the test-taker to those who were awarding them scores;

the number in long turns is important for consistency since the fluency figures, as explained in Section 3.1.4.3, were only calculated for longer turns; and per opportunity to speech which allows these indices to be considered equally for their development over the three administrations of the GOT.

Table 10: Quantitative measures and related statistical analyses for research questions

Research Question 1: How does the language elicited by the GOT show the development of syntactic complexity, accuracy, fluency, and range of lexis in test takers' performance in the three GOTs taken over the first two years of study at university?	
Core indices:	Number of words, clauses, AS-units (raw counts, in long turns & per opportunity to speak)
Complexity:	Words per AS-unit, Clauses per AS-unit, Words per clause
Accuracy:	Ratio of error free clauses to total clauses, Error free clauses per opportunity to speak
Speed Fluency:	Speech rate, Articulation rate
Breakdown fluency:	Pause proportion
Repair fluency:	Maze ratio, Maze & sound ratio
Vocabulary:	Range, Vocab Profile
Statistical Procedures (except vocabulary): Freidman's ANOVA, Wilcoxon Signed Ranks	
Research Question 2: To what extent do the test-takers' number of words, frequency and length of turns show a consistent pattern of development in their performances in the three GOTs they take over two years?	
Patterns of discourse:	Turns per time, Turns per opportunity to speak, Words per turn, Words per opportunity to speak
Statistical Procedures: Freidman's ANOVA, Wilcoxon Signed Ranks	
Research Question 3: To what extent is the GOT able to elicit features of conversation that have been found to be beneficial to language learning?	
Interactive features:	Initiating, Responding, Developing, Collaborating
Statistical Procedures: Freidman's ANOVA, Wilcoxon Signed Ranks	
Research Question 4: How well do the speaking performance indices of test-takers who take the GOT over the three administrations in two years predict the scores awarded to them?	
Dependent variables GOT scores awarded in:	**Related Indices**
Fluency	Speech Rate, Articulation Rate, Pause Proportion, Maze ratio, Maze & Sound Ratio
Grammar	Clauses per AS-unit, Words per AS-unit, Words per Clause Error Free Clauses, Error Free Clauses*, Error Free Proportion
Vocabulary	Words*, Turns*, Words per turn
Communicative Skills	Initiating Features*, Responding features*, Developing Features*, Collaborating Features*
Statistical Procedures: RMANOVA, Multiple Regression	

*Normalized per number of participants and time of the test

3.1.4.1 Complexity measures

For this dissertation two measures of syntactic complexity were taken into consideration. The first one was the number of words per AS-unit, which were calculated to produce a mean length of utterance (MLU) index (Foster & Tavakoli, 2009). The second was to follow Skehan and Foster (1999) by quantifying complexity as the amount of subordination in speech by calculating the ratio of the number of clauses to AS-units, though unlike Skehan and Foster (1999) and Foster and Skehan (1996), AS-units are used instead of C-units. These studies found that this measure of complexity was sensitive to real time processing: The language produced by the participants was significantly less complex in a simultaneous viewing and describing of a video task (Skehan & Foster, 1999) and significantly more complex when more planning time was given (Foster & Skehan, 1996).

In addition to the measures that were calculated for every turn taken by the participants in the test, measures were also calculated for the ten plus second turns they took (*long turns*), so that they would be comparable to the fluency figures and to reduce the influence of minimal formulaic chunks of language that might influence the figures.

3.1.4.2 Accuracy measures

To measure accuracy, a ratio of error free clauses to total clauses in the speech of the participants was calculated in order to arrive at an index that represented global accuracy (Foster & Skehan, 1996). This statistic has been found to be a superior measure of general accuracy than the percentage of error-free AS-units, since it is less affected by the length of the utterance (Vercellotti, 2012). Errors were counted if mistakes were found in the syntax, morphology, word order or appropriacy of the words in the clause, and only counted if they were definitively wrong – if there was some doubt it was not counted as an error (Foster & Skehan, 1996). The analysis was performed on the pruned speech of the participants: after repetitions, false starts and self-corrections not necessary for communication, or ‘maze words’ (Loban, 1976) were removed. A similar overall measure of global accuracy was found by Iwashita, Brown, McNamara and O’Hagan (2008) to distinguish the higher levels from the two lower levels on a non-interactive speaking task which included 200 performances at five different levels. For the same reasons as for the complexity measures, the accuracy statistics for the long turns they took were calculated separately.

The second measure of accuracy was the number of error free (EF) clauses per opportunity to speak. This measure is sensitive to the total number of EF clauses that the participants used in the GOT, and it was thought that this might also be a relevant index since it is an indication of the rater's overall exposure to the accuracy of the test-taker.

3.1.4.3 Fluency measures

To research fluency it is necessary to sample the test-takers talking in more extended turns. Single turns that consist of such utterances as "how about you" or "I don't know" may be remembered as single fluently spoken chunks and thus may skew the data, especially if they make up a large proportion of speech. For example, it may well be that Nitta and Nakatsuharas' (2014) study was affected by not accounting for short turns being spoken more fluently than long turns (see Section 2.2.4.4). To eliminate such minimal responses, and along with considerations about collecting sufficient data, it was decided to include only turns that lasted ten seconds or more, (hereafter referred to as 'long turns'). This makes the data for long turns restricted in two ways: firstly the amount spoken in long turns is, with the odd exception, a subset of the amount spoken by each student; secondly, since some participants did not take a long turn in the first and third administrations (four and two participants respectively), the data set for the timed aspects of fluency also reduced in number. Despite this limitation, using only the long turns was felt to render a more realistic measure of speaking fluency.

Various aspects of fluency were investigated using the data collected from the long turns. For fluency as the speed of speech, measures of *speech rate* and *articulation rate* were calculated. For these measures, the procedure suggested by Towell, Hawkins, and Bazergui (1996) was followed. Within the turns each participant took, all understandable syllables in English, including repeated words and false starts, were counted, with non-lexical fillers like 'um' and 'er', and Japanese interjections being excluded. The total number of syllables is taken as a proportion of the total speaking time (including unfilled pauses) to calculate the speech rate (SR), which is then multiplied by 60 to get a measure of syllables spoken per minute. As has been pointed out by De Jong (2013) since the speech rate incorporates pausing it confounds the measure of speed fluency with breakdown fluency. Nonetheless, speech rate was included since it not only maintained consistency with previous studies, but also provided a more composite measure of fluency. The articulation rate (AR) excludes unfilled pause time to obtain a measure of average syllables per second (Towell et al., 1996) and is thus a pure measure of speed fluency.

To calculate an index of breakdown fluency, the amount of time spent in unfilled pauses within each turn was calculated. Unfilled pauses of one second or more were timed and taken as a proportion of speaking time, giving the measure ‘pause proportion’. Although Riggenbach (1991) considered unfilled pauses of 0.5 seconds or greater, out of practical considerations given the quantity of data, unfilled pauses of one second or more were chosen as a the minimum duration for measurement, a compromise that other studies have also made (Iwashita et al., 2008).

Measures of repair fluency (Tavakoli & Foster, 2008) were obtained via two indices: the *maze ratio* which was given by the proportion of repetitions, false starts and self-corrections not necessary for communication, or ‘maze words’ (Loban, 1976), to total words spoken, and the *maze and sound ratio*, which included not only maze words but filled pauses as a proportion of total words used. Filled pauses were counted individually if they were separated by an unfilled pause. To ensure that the data is consistent with the timed fluency data, these indices were also calculated for the long turns only, and are denoted by having long turns (LT) in brackets after them.

3.1.4.4 Vocabulary

Vocabulary gain has been found to play an important part in determining a test-taker’s score. In Iwashita et al. (2008), vocabulary was as important as fluency as the main determiner of language level in a semi-direct spoken test. Like other studies that have investigated vocabulary gain, this study will assume that as the test-takers develop in ability, as a group they will speak a greater range (types) of less frequent lexical items. Previous research into longitudinal data in vocabulary acquisition shows that caution is necessary when interpreting the results of vocabulary gain by word level. A study that examined vocabulary gains in written work amongst Francophone children learning English found that the number of words outside the 1000 most frequent did not increase after 300 hours of tuition (Horst & Collins, 2006). However upon closer examination, the researchers found that the learners at the first stage relied heavily on vocabulary borrowed from French, and that they improved over time by using a higher proportion of high frequency English words. While such a phenomenon is unlikely to be a factor among Japanese university students without a background in a language that English has borrowed as much from, this study shows the necessity of looking beyond frequency counts since language development can manifest itself in unexpected ways.

When researching lexical development in L2 acquisition, the objective is often to investigate the range of vocabulary that a learner can produce, or 'lexical richness'. One of the most widely sought after indices has been for a stable measure of lexical diversity (LD). Since the 1950s various methods of LD have been developed to measure language acquisition of first or other languages. The early standard was the type/token ratio (TTR), which is calculated by simply dividing the number of different words, or types, by the total number of words produced, or tokens. Although this ratio was widely used in earlier studies, it has been found that it is dependent on text length (see Richards [1987] for a summary of the problems associated with TTR). Since then there has been a search for an index of lexical diversity that is independent of sample size. The most recent measures that have claimed to be independent of text length are known as 'vocd' and the 'measure of textual lexical diversity' (MTLD).

These two measures are calculated in different ways. Vocd is derived from the TTR, but adjusted by selective sampling and curve fittings in order to minimize the effect of text length (McCarthy & Jarvis, 2007). The MTLD also uses the TTR, but in an entirely different way. In the calculation of MTLD, a cut-off value is used to compare with the TTR of each word in a string, and the final figure is produced by the average length of strings of words that maintain the cut off level. The default figure of 0.720 was arrived at empirically through testing on a range of narrative and expository texts, in which the TTR was found to have stabilized at this point (McCarthy & Jarvis, 2010).

A number of researchers (for example Harris Wright, Silverman, & Newhoff, 2003; Owen & Leonard, 2002) have used the vocd as a measurement of LD. In a study that compared 25 different measures of lexical density, sophistication or diversity in oral narratives by 408 Chinese English, Lu (2012) found that raters' scores were most related to lexical diversity, only to a small degree with lexical sophistication, and no relationship was found with lexical density. The lexical diversity index with the largest effect was the number of different words the participant used, in other words, the Type count; a count that is also dependent on sample size. The measurement of vocd had a small but significant effect, and was recommended as a different measure of diversity since it correlated the least with the other measures of diversity.

However, recent findings show that the vocd index is in fact affected significantly by text length (McCarthy & Jarvis, 2007). It was also found to correlate strongly (0.971) with the hypergeometric distribution function (HD-D), since it measures essentially the same trait (Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010). As such, recently research has come to focus on other measures of LD such as the MTLD.

Studies that have compared MTLD to other measures of diversity have been largely positive due to it being the least affected by text length. In an investigation that compared sophisticated measures of LD (like vocd or HD-D, Maas and MTLD) and flawed (like TTR), it was found that MTLD correlated strongly with the sophisticated measures and poorly with the flawed ones, and that the figures it produces are consistent for the various lengths of text that were included in this study (McCarthy & Jarvis, 2010). McCarthy and Jarvis also concluded that MTLD seemed to measure a different latent trait to vocd and Maas, and recommended that further research be conducted into these measures.

The text lengths included in the McCarthy and Jarvis (2010) research started from 100 words and ended at 2000 with the length going up in minimum steps of fifty, leaving an open question as to how LD indices performed with smaller text lengths. Studies by Koizumi and In'nami (2012) and Koizumi (2012) set out to investigate this by examining a 200 word spoken text by a Japanese learner of English that was broken into segments that increased in increments of 10 words from the fifty word level, and using six different LD measures, including MTLD, and the other two recommended by McCarthy and Jarvis (2010), vocd and Maas. They found that MTLD was the least affected by text size out of the measures included in the study. Also interesting was the finding the MTLD correlated the least to the other measures, suggesting that it measures different aspects to the other LD indices. The researchers conclude by recommending that MTLD be used in preference to the other measures for texts of between 100 and 200 words length. Below this length of text, MTLD was also affected deleteriously by variations in text length.

Given the above research findings, MTLD was chosen as the primary index of LD in this study. Since it captures different information to other measures, this study will include HD-D (vocd) and TTR for the sakes of comparison. The caveat is that the number of words spoken by the test-takers in this dissertation varies considerably, with many of them not producing 100 tokens in at least one of the administrations. As can be seen in Table 11, of the 53 participants, only 10 spoke over 100 words in all three administrations and a further 21 reached this mark in two administrations. Of the remainder, 14 breached the 100 word level in one administration, leaving eight who never reached this level in the three administrations. Since LD measures cannot give reliable measure to those who spoke such few words, the analysis will be restricted to those who reached 100 words in at least two administrations. These students will be divided into four groups: those who broke a hundred words in the first and second tests (N = 13 test-takers), in the second and third tests (N = 26), in the first and third tests (N = 12), and those who did it in all three administrations (N = 10), as detailed in Table 11.

The tool used to calculate figures for HD-D (vocd), MTLD and TTR was the software *gramulator 6.0* (McCarthy, 2011), since it could run the calculations expeditiously. The HD-D figures for *gramulator* have been scaled to a corpus such that a "0" score means that it has figures similar to

Table 11: Test-takers who spoke 100 words or over in more than one administration

Administrations - in which students who spoke 100+ words participated	**No. of students**
- first, second & third administrations	10
- first & second administrations	3
- second & third administrations	16
- first & third	2
Groups of students for LD analysis - those who spoke more than 100 words in more than one administration	
- all three administrations	10
- total first two administrations	13
- total last two administrations	26
- total first & third administrations	12

that expected in narrative and expository texts, though they should still correlate very highly with raw HD-D scores (McCarthy, personal communication), and indeed, vocd figures.

Viewing a test-taker's vocabulary improvement as a figure in an index may be a convenient way to visualize achievement, but a fuller picture of their achievement can be attained by investigating their vocabulary usage in terms of vocabulary frequency levels. Various methods have been developed to compare samples of language to lists of words that sample the frequency of the vocabulary in real world usage. Two tools that have been developed to do this are the software *Range* (Heatley & Nation, 1996) and the website based *Vocabprofile* (Cobb, 2002), both of which are based on the British National Corpus (BNC). *Vocabprofile* analyses the frequencies of the text into the two most frequent 1000 word families and the Academic Word List AWL (Coxhead, 2000), and has been used to analyse both the development of spoken and written vocabulary of learners. For example, Iwashita et. al. (2008) investigated the performance of 200 test takers at five different levels on the speaking section of the TOEFL iBT, a semi-direct speaking test, using *Vocabprofile*. They found that participants at the higher levels produced significantly more tokens and types, but they did not provide the results by frequency level, which would have been useful for comparing with this dissertation's findings.

For this investigation, *Range* will be used to calculate the coverage of frequencies to the first 14,000 most frequent bands. In addition to recording these frequencies, the *Vocabprofile* results will be calculated separately using the tool on *lextutor* (Cobb, 2002). As for the studies on the minimum number of words for consistent measurement of LD indices, deference must be paid to the minimum acceptable length of text for the stable measurement of frequency. In the case of *Vocabprofile*, for written texts the minimum length has been determined to be 200 words (Laufer & Nation, 1995), and there is certainly no reason to think that spoken texts would require less. If anything, due to a conversational text being made up of a sequence of turns, rather than one coherent text as a writing passage may be, it seems likely that if anything, more would be required. Rather than report figures that cannot be trusted, the *Vocabprofile* will be produced for each of the three administrations so that the development of the language they produce as a cohort may be tracked.

In order to produce frequency lists, txt files were created from the transcripts for analysis by *Range*. Extreme care was taken to remove all non-text artefacts to ensure the texts were compatible with the lists *Range* uses. The danger of failing to do so was made clear in an article by Neufeld, Hancioğlu and Eldridge (2011) which corrected a previous article by Li and Qian (2010), whose study miscalculated coverage by a factor of almost ten due to the use of lists that were not appropriately prepared. With this cautionary episode in mind all due diligence was maintained in order to ensure clean files were prepared for processing.

Since students should not be credited for using words from the prompt paper (as is made clear to the raters in their training), those used in the transcripts that also appeared in the prompt at the second two thousand most frequent band (k2) or lower frequency bands were separated and tallied independently, as were names of people, places, and food. Table 12 shows an analysis of the words from the prompt in each administration using *Range* and *Vocabprofile.*

The listed types on the AWL list or at k2 or less frequent were not included to the frequency counts of the corpus since the speakers should not receive credit for them. After this process, there were still some unclassifiable words, such as those formed by incorrect inflections (*unconvenient*), pronunciation mistakes (*clotheses*, *calshium*), and coinages (*younghood*), since they could not be added to the list, and yet were a valid part of the corpus, they were included in the *offlist* category of the count. That there were such a surprisingly little number of these probably reflects the test-takers' conservative use of language in an assessment situation where students may believe they would get penalized for incorrect usages, as well as run the risk of losing face by not being understood by their peers (Luk, 2010).

Table 12: Frequency analysis of words in the prompts

	Administration 1		**Administration 2**		**Administration 3**	
	Token	Type	Token	Type	Token	Type
k1	135	65	194	91	215	91
k2	10	6	12	7	12	9
k2 words	prefer, regular, tour, independent, tourist, tourists		Inside, phone, phones, prefer, earned, healthy, lots		alone, etc, factors, focus, Japanese, location, prefer, select, western	
k3	2	1	4	2	-	-
k3 words	abroad		traditional, computer			
k4	1	1	-	-	-	-
k4 words	overseas					
k5	-	-	1	1	-	-
k5 words			housework			
AWL	1	1	4	2	13	8
AWL words	overseas		traditional, computer		job, available, focus, select, affect, benefit, factor, locate	

3.1.4.5 Word and turn measures

The number of words spoken and turn length have been found to be valuable indices in peer interaction tests (Galaczi, 2010, 2013; Leaper & Riazi, 2014; O'Sullivan & Nakatsuhara, 2011; Van Moere, 2007). Tracking the number of words used is an indication of a participant's role in the GOT, and was used as an indicator of quantitative dominance by O'Sullivan and Nakatsuhara (2011). Words per turn is an important indication of the pattern of discourse. For example, Leaper and Riazi (2014) investigated the group oral tests in the second administration of the data in this dissertation (including data from some of the participants of this study), and found that significantly different words per turn between prompts was indicative of different patterns of discourse elicited by the prompts. Furthermore, an implication of Galaczi (2010, 2013) is that the length of turn may be related to development, as she found that more advanced learners tended to use more frequent shorter turns. As such, tracking the length of turns in words over administrations permits insights into various factors affecting their performance.

For these counts, non-words, such as ‘ums’, ‘ahs’, incomplete and immediately repeated words, were not included (Lennon, 1990, Taguchi, 2007). A turn was defined as an utterance that was responded to, or a coherent sequence of words that could have been responded to (Eggins & Slade, 1997). Care was particularly necessary when distinguishing between a turn and backchannel. Sometimes an utterance such as “Really?” which may have been intended to simply show support to the person taking the turn, was responded to, meaning that it should be counted as a turn. As for the collection of data for fluency (see Section 3.1.4.3), and to be consistent with Leaper and Riazi (2014), a long turn was defined as being over ten seconds in length. However, unlike the fluency figures, since this question is on the patterns of discourse, all test-takers were included in the data set, not just those who took turns over ten seconds in length.

The above collection procedures allowed the creation of the indices as shown in Table 13. While it does not necessarily reflect the quality of their performance, the raw number of words, turns and long turns are relevant given that they constitute the entire performance the raters were exposed to; the same figures when normalized provide an indication of their contribution to the discussion. The words per turn index allows an insight into the pattern of discourse the speaker typically uses, and the ratio of long turns to all turns and long turn words to all words give a measure of the proportional weight of the longer turns within a test-takers discourse.

Table 13: Indices used for analysing patterns of development in words and turns

Index	Method of calculation
Number of words	Word count
Number of turns	Turn count
Number of long turns	Count of turns over 10 seconds long
Words per Opportunity to speak	Number of words, turns and long turns normalized by time and number of participants in their GOT
Turns per Opportunity to speak	
Long turns per Opportunity to speak	
Words per turn	Number of words divided by number of turns
Long turns per total turns	The number of long turns divided by turns
Long turn words per total words	The number of words in long turns divided by the total number of words

3.1.4.6 Interactive features related to language learning

A key justification for the GOT is the claimed backwash effect on language learning. As seen in the literature review, research has found that conversation can play an important role in learning a language. If this is the case then having a form of assessment that encourages interaction among learners could provide a powerful incentive for students to practice these beneficial aspects of language learning, creating a virtuous circle. To justify the role of the GOT in this cycle, it is necessary to analyse the interactive features of the test to investigate what in fact it elicits.

The approach used in this dissertation was guided by Ellis and Barkhuizen (2005) who provide six principles to guide interactional analyses. Their first point is that since utilizing descriptive frameworks in the literature allows better comparisons with other studies to be made, adopting or adapting these should be the preferred approach. Studies that have examined communicative functions in speaking quantitatively have either done so by counting instances from a pre-existing list (He & Dai, 2006), using a pre-existing list but adapting it according to their data (Van Moere, 2007), by interpreting interactive function through a pre-existing system of grammatical analysis (Eggins & Slade, 1997), or by deriving a list of communicative functions from the transcripts of the data their study collected (Brooks, 2009). He and Dai's (2006) study worked from a list of interactive functions that the GOT was supposed to elicit according to the test-administrators, and as described in the literature review, found a disappointingly small proportion of them being used. This list had eight functions: agreeing and disagreeing, asking for opinions or information, challenging opinions, supporting opinions, modifying, persuading, developing and negotiating meaning. Van Moere (2007) found that considerable adjustments were necessary when he applied this list to the group orals in his study (p.294-5). Eggins and Slades' (1997) 'speech function' analysis was developed to complement a Hallidayan analysis of meaning in grammar, and attempts to classify every utterance in a conversation based on the speech function system outlined by Halliday (1994). However, the system outlined by Eggins and Slade (1997) was developed for describing NS speech, and is not entirely relevant for NNS contexts.

The alternative is to create a list of interactive functions by drawing them from an examination of the transcripts. This was the approach that Brooks (2009) used to investigate communicative function in paired interactions compared with interviews. While the list of functions in Eggins and Slade (1997), He and Dai (2006), Van Moere (2007), and Brooks (2009) overlap to some degree, there are considerable differences, which is not surprising given their different origins and purposes. This demonstrates the importance of investigating the language produced by the test takers, rather than having preconceived notions about what language those taking the GOT should produce. Thus, the approach I took was to examine the transcripts with an open mind, but refer back to terminology and definitions used in these previous studies in order to maintain at least a degree of comparability. While the codes are derived from the data, they are informed by previous coding schemes mentioned above (Brooks, 2009; Eggins & Slade, 1997; He & Dai, 2006).

Secondly, according to Ellis and Barkhuizen (2005), an interactional analysis should go beyond a mere 'list' of functions by developing a hierarchical framework. This will be done by deriving major categories from the functions that are found to exist in the data. It is likely that forming such a hierarchy would be a necessary preliminary step before statistical analysis in any case, since counting functions individually may result in finding few instances of many different types.

The final four principles relate more directly to examining the interactional functions: they should be operationalized fully; account for all the data; have no overlapping categories; and no unnecessary categories (Ellis & Barkhuizen, 2005). Accordingly, definitions were developed as the transcriptions were examined, and reviewed and adjusted as successive examples in the transcripts showed how the participants used them according to the changing context. This meant coding every utterance according to its function in the conversation, and when new functions are assigned, returning to previously coded transcripts to ensure consistency. The transcripts had already been divided into AS-units for the CAF analysis, so these were used as the basis for assigning interactive functions as far as possible, while realizing that the perfect alignment may not be possible. Thus, the list of functions was derived primarily from the transcripts, and the categories emerged using a ‘data-driven, iterative approach’ (Brooks, personal communication, August 2, 2013).

Following the above process, 30 functions emerged from the data, and these could be categorized as belonging to four different higher categories: initiating, responding, developing and collaborating, with an additional ‘Japanese’ category to account for the test-takers occasional use of their native language. In doing this, the following nomenclature was developed: the four higher categories have their first letter capitalized; the specific features full name is italicized, and are referred to as conversational ‘moves’; and the codes are derived from the moves, have the first letter capitalized, and are used in graphs and tables.

To verify and check the reliability of these functional categories a second rating was conducted. An experienced colleague with an MA in second language education was trained to use the system and code two complete tests chosen because all four participants happened to be included in the study; one test was from the first administration, the other from the last. After training, the auditor achieved exact matches in 79.86% of the AS-units. Many of the disagreeing codings were of functions that belonged to the same higher category, and the agreement of category matches was 89.93%, which was acceptably high. After discussing the conflicting codings, agreements were reached over changes to the wording of the definitions, and the changes applied to the data before the final statistical analysis.

One limitation of this approach stems from the nature of the data. Since the sole source is the transcripts of the tests themselves, it is not possible to access those features of interaction that are not publically viewable. For example, hypothesis forming has been found to be one method by which NNSs can learn from conversational interaction (Shehadeh, 2003), but there may be no evidence of it occurring in the transcripts. Regarding the possible results that could be achieved from interactive analysis, it will be interesting to observe how consistently the interactive functions are used across administrations. The literature on the development of interactive skills in any context is scanty and the few studies in an assessment context, namely Galaczi (2010, 2013) and Negishi (2010) were cross-sectional, so there is not much guidance as to what to expect.

3.1.5 Data analysis procedures

3.1.5.1 Statistical tests for performance indices

In order to test whether there were significant differences in the various indexes of language use, the non-parametric Friedman's ANOVA test was used because neither a normal distribution nor homogeneity of variance could be assumed from the collected data, and this was confirmed by significant figures being returned by the D'Agostino-Pearson K2 test of normality (D'Agostino, Belanger & D'Agostino, 1990) in most cases, as well as visually through the use of boxplots.

Friedman's ANOVA can only inform us of significant differences being found between the three administrations of data. Following the procedure recommended by Field (2005), Wilcoxon's Signed Rank tests were conducted on the significant Friedman's ANOVA results to compare the data of each administration to the other administrations in pairs. When making multiple comparisons there is a heightened risk of rejecting the null hypothesis when it happens to be true (or committing a Type I error), and this may be controlled by means of the Bonferroni correction, which sets a stricter level for p by dividing it by the number of comparisons (resulting in $p = 0.017$ for the three administrations in this study). However, the Bonferroni correction is known to be overly conservative and using it runs the risk of ignoring relationships that may be significant (Field, 2005), or Type II errors. Since this is exploratory research, the greater concern is to reveal the relationships in the data. As such it is accepted that there will be a greater risk of committing Type I errors, and this study will persist with the conventional level of significance of $p = 0.05$. Due to the possibility of incorrectly reporting significant statistics, caution should be taken when interpreting results that fall in the 0.017 – 0.05 range of p. In recognition of this, significant figures in this range will footnooted in the tables.

As pointed out by Field (2005, p. 565), effect sizes are not very useful for tests of general effect like Friedman's ANOVA, and so they were calculated from the results of the Wilcoxon signed-rank test, as being *z* divided by the square root of the sum of the number of observations from which the comparison was made.

3.1.5.2 Analysing the scores

Regardless of the actual performance of the students, it is the scores assigned to them that will colour the test-taker's opinion of it and through which their abilities will be evaluated by the administration. It is important then to analyse the scores over the three administrations so that trends in the data can be displayed. It is to these scores that their performance in the test over three administrations as represented by the indices of complexity, accuracy, fluency, vocabulary and communicative function will be compared.

The first question to answer is whether there are any statistically significant differences in the scores participants received as a group between the administrations, and this will be done by means of a within-groups repeated measures ANOVA (RM ANOVA) of their scores. To ensure the data met the conditions necessary for ANOVA, box plots were examined visually for normality and equality of variances, and it was discerned that the data varied somewhat from normality. The data was transformed using a log function, but did not result in improvement so the original figures were used. The ANOVA was run with the knowledge that statistical power may be lost by not exactly meeting the normality assumption (Larson-Hall, 2010, p. 340), though the general belief is that parametric tests such as ANOVA are robust enough not to be affected by data which may violate the normality assumption for parametric tests.

3.1.5.3 The relationship of the score to the performance indices

To answer the question as to how well the scores represent the students' actual progress as measured by the indices, multiple regressions (MR) was conducted, holding the score as the dependent variable and relevant indices as the predictors.

The MR can give a statistical answer to the question about the extent to which an index accounts for and can predict the score given. The standard MR shows how much each index uniquely contributes to the dependent variable, and a sequential MR shows how much each index contributes additionally to the previously entered independent variable. A test score is likely to be a result of the rater taking into account a number of different factors (indeed, some of which may not be on the scoring bands at all as Section 2.2.4.5 of the literature showed), and so it will be necessary to combine various indices to analyse the extent to which each accounts for the score. When deciding on the independent variables to include in the model, the maximum number of variables which can be included to ensure reliable statistics given the sample size available, must be taken into account. In MR the number of predictors depends on the desired size of effect and statistical power. According to Field (2005, p. 173), for a sample size of about 50 (which is close to the number of participants in this study) and using no more than six predictors, the MR should be capable of detecting large effects, but not medium or small ones.

The first stage of the statistical analysis was to determine which indices are relevant. This was done by running the regression analysis on the individual index relevant to each score. It should also be considered that some of the indices might generally relate to a rater's overall grading of the performance, specifically, the words spoken, the turns taken and the words per turn, and so these were included as well. The indices included in the analysis can be seen in Table 14. It can be seen that including the three common indices, the fluency, grammar and communicative skills scores had eight, nine and seven predictors respectively. Vocabulary only had the three common ones because, as discussed above, valid measures were not available due to many students not producing sufficient words for stable measurement. From these the six best performing indices were chosen for the standard MR analysis.

Table 14: Indices used as predictors for MR analysis

Skill	Fluency	Grammar	Communicative Skills	Vocabulary
Skill Specific Indices	Speech rate Articulation rate Pause proportion Maze ratio Maze & sound ratio	Clauses per AS-unit Words per AS-unit Words per clause Error free clauses* Error free proportion	Initiating features* Responding features* Developing features* Collaborating features	Words* Turns* Words per turn
Shared Indices	Words* Turns* Words per turn			

*Normalized per number of participants and time of the test

For the sequential MR, since each successive independent variable builds on the previous one, the order in which they are entered is crucial. For the scale specific indices, the order will be decided by the extent to which the individual regression shows they contribute to the score. Since there are three administrations in which this order may vary, their average contribution will be used to determine the order. Amongst the general indices, it is expected that the number of words spoken will be an important contributor in all indices, so it will be entered last to allow the contributions of other specific indices to be clearly seen. Similarly, the remaining two general indices will be entered second and/or third to last if they are found to be among the top six contributors to the score.

3.1.6 Summary of this section

The above sections described the methodology of the quantitative phase of this study. I led into this chapter by explaining the research design, and continued by describing the participants, context and procedure of the test from which the data were collected. Following this, the procedures for preparing the data for analysis were outlined, and then the specific measures used to create the performance indices were explained. The final section described the statistical tests that were used to analyse the data along with a rationale for their use. The second half of this chapter presents the results of the quantitative analysis of the data.

3.2 Results of the quantitative analysis

This section reports the results of the data analysis of the quantitative phase of this dissertation. These are presented under the respective headings for research questions 1-4.

3.2.1 Research question 1: Development of CAF and lexis

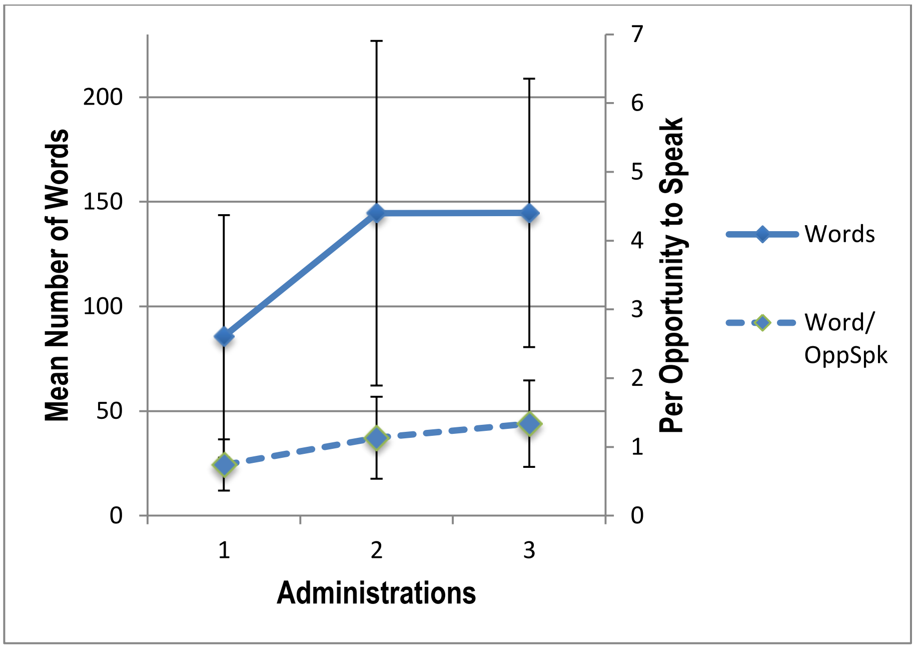
The first research question was about the development of the performance the test-takers display in the GOT in terms of CAF (complexity, accuracy, fluency) and vocabulary. The subsections follow the same organization: at the start of each subsection the main findings are highlighted in graphs, which are followed by tables of the descriptive statistics and the results of the tests of significance. The next subsection will explain the findings for the core statistics that were used to construct the indices before going on to the remaining subsections that directly detail the findings for complexity, accuracy, fluency, and lexis.

3.2.1.1 Development in the core measures

The foundational counts that are used to calculate many of the other indices are the number of words, clauses, and AS-units. Describing the results of these statistics will provide a fuller understanding of the findings for the other indices. Both the unadjusted counts of words, clauses and AS-units as well as the counts after being normalized for the time of the test and the number of participants will be described. Since the fluency figures are derived from the turns that are over ten seconds in length, for the cause of comparability and to describe the test-takers' performance thoroughly, the counts of words, clauses, and AS-units in longer turns will be analysed here as well.

The first graph, Figure 6, displays the average and standard deviation of the words spoken per administration. The top line represents the average number of words and uses the scale on the left hand side; the dotted line is the normalized figure, and uses the scale on the right hand side. The error bars show the standard deviation at each data point.

Figure 6: Words spoken: Unadjusted and per opportunity to speak



Immediately obvious in Figure 6 is the difference between the adjusted and unadjusted figures; the latter shows an impressive gain in the number of words made in the second administration followed by a slight decline in the final administration. However, the adjusted figures present a more realistic picture of the three administrations in terms of the number of words by showing that although the greatest gain is made in the second administration, the test-takers improved in the third administration as well. Also of note are the error bars for the unadjusted number of words spoken in the second administration that display the wide variability that occurred in this administration, as might not be surprising given the findings on the widely differing patterns of interaction that have been noted in that administration (Leaper & Riazi, 2014).

The next two graphs show a similar pattern for clauses in the raw (dark blue line), long turn (orange line), and normalized clauses (dashed line) in Figure 7 and AS-units in Figure 8.

Figure 7: Clauses: Unadjusted, in long turns, and per opportunity to speak

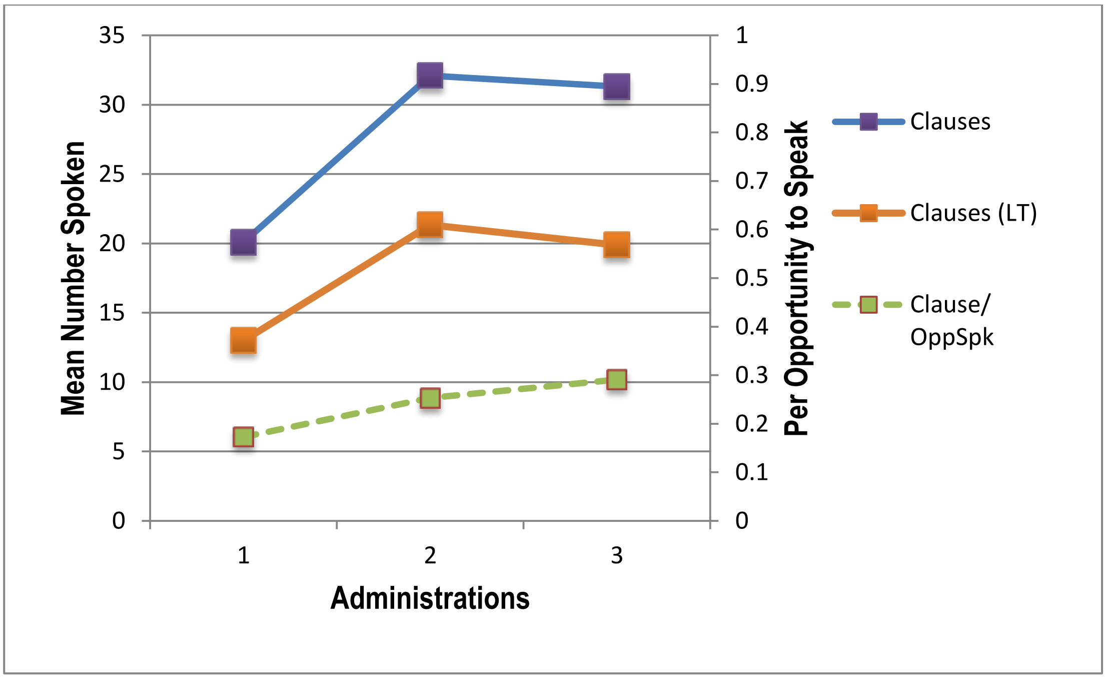
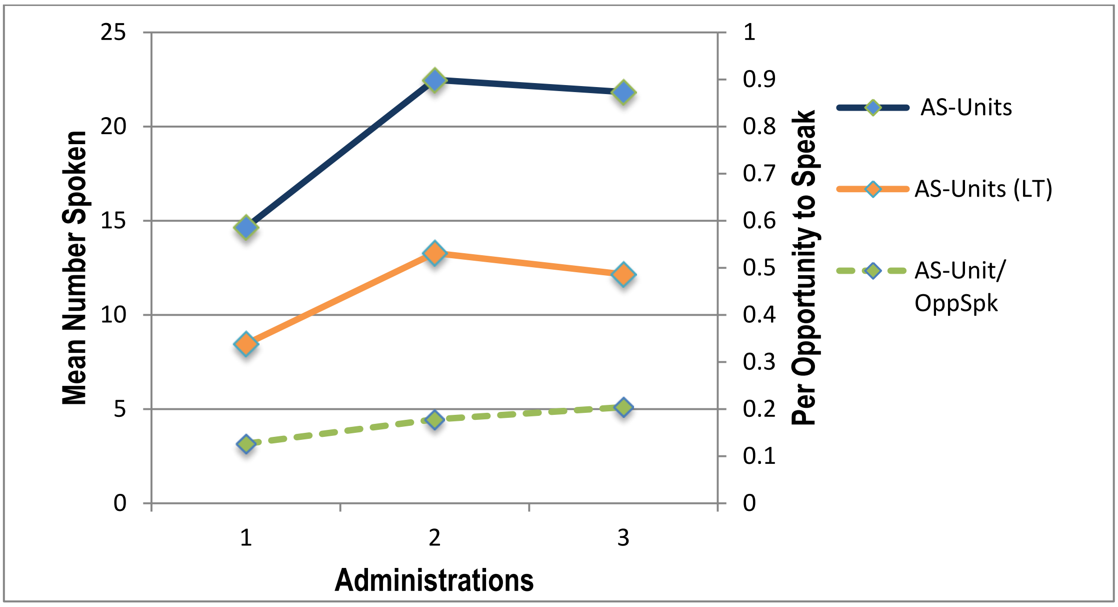


Figure 8: AS-units: unadjusted, in long turns, and per opportunity to speak



In both of these graphs, the unadjusted and long turn figures in the second administration show considerable gains followed by a slight drop in the unadjusted figures and a more pronounced decline in the long turn figures in the third administration, while the normalized figures show a gain in the second administration, followed by a smaller gain in the third.

Table 15 shows the descriptive statistics for these counts: columns 1-3 give the figures for words, clauses and AS-units, columns 4 to 6 the same features for long turns only, and columns 7-9 the features normalized by opportunity to speak (taking time and number of participants into account). It can be seen in this table that for the unadjusted and long turn figures, the standard deviation is greatest for the second administration, showing that it had the greatest variability associated with it. This may be due to the particular prompts that were used in that administration tending to produce GOTs with either many shorter turns or fewer longer turns, as described by Leaper and Riazi (2014). The median statistics in columns 1 to 3 and 7 to 9 of Table 15 show a consistent trend for AS-units and clauses: a substantial increase in the second administration, smaller increase in the third administration. The larger increase in the second administration from the first can be explained by the test-takers starting from such a low point when they first took the GOT before the academic year

Table 15: Descriptive statistics for core measures

Administration 1									
	1	2	3	4	5	6	7	8	9
	Words	Clauses	AS-units	Words (LT)	Clauses (LT)	AS-units (LT)	Word/ OppSpk	Clause/ OppSpk	AS-unit/ OppSpk
Test-takers	53	53	53	49	49	49	53	53	53
Min.	19	5	3	16	3	2	0.146	0.045	0.028
Median	71	19	13	56	12	7	0.784	0.170	0.122
Max.	337	75	57	180	37	27	1.880	0.418	0.318
Mean	85.642	20.075	14.642	62.306	13.000	8.449	0.739	0.173	0.126
Std Dev	57.958	12.806	9.566	40.013	8.101	5.485	0.373	0.081	0.060
Administration 2									
Test-takers	53	53	53	53	53	53	53	53	53
Min.	31	7	4	11	3	2	0.241	0.061	0.035
Median	126	27	18	95	19	11	0.922	0.215	0.146
Max.	395	90	76	282	51	41	3.223	0.734	0.620
Mean	144.566	32.094	22.472	107.774	21.340	13.283	1.133	0.253	0.178
Std Dev	82.333	17.128	13.387	61.045	11.363	7.373	0.598	0.126	0.102
Administration 3									
Test-takers	53	53	53	51	51	51	53	53	53
Min.	26	8	7	24	3	3	0.189	0.058	0.051
Median	130	29	20	92	16	11	1.230	0.275	0.181
Max.	299	65	44	254	42	27	3.371	0.767	0.553
Mean	144.623	31.302	21.830	104.490	19.882	12.157	1.338	0.291	0.204
Std Dev	64.155	13.282	9.553	57.177	10.447	5.917	0.628	0.138	0.102

commenced; they had more room for improvement in the second administration. After two years of study, the final administration in this study, they cannot improve in as many ways, and so the gains they make are more incremental. The figures for long turns show this pattern in the first two tests, but the decline is more emphatic in the third administration, showing that the participants relied on longer turns to a lesser extent in the third administration.

The initial test of significance, Friedman's ANOVA, showed significant differences in all of these indices (see Table 65 of Appendix C). The results of the Wilcoxon Signed Ranks tests that were used to specify which administrations were significantly different are displayed in Table 16. In this table, consistent patterns can be seen for the adjusted and unadjusted counts. For the unadjusted figures in columns 1 to 6, the second administration is significantly more than first and third administrations with moderate effect sizes ranging from 0.411 to 0.535 for the entire corpus, and from 0.326 to 0.493 for the data taken from long turns. The adjusted figures in columns 7 to 9 also show the same significant relationships as the raw figures. In addition they show that the remaining relationship between the second and third administrations is significant. The strongest effect sizes are for the second administration when compared to the first and third administrations, with effect sizes that vary between 0.404 and 0.558. The significant difference between the second and third administrations all had small effect sizes of between 0.209 and 0.270. Without adjusting for time and participants the test-takers' achievement when speaking more in every test would have been overlooked; this underlines how important it is to take into account these factors.

The development that these figures show is that in the second test the test-takers speak significantly more words which were organized into more clauses and AS-units each time they took the test. Between the second and third administrations, the finding that the normalized figures for words, clauses and AS-units showed significant growth, while the counts in the overall and long turn figures were similar or showed small declines might seem like a discrepancy. This can be explained by the fact that raters could call an end to the test when they got a rateable sample. It seems they reached this point earlier in the third administration than in the second, and this is most likely due to the overall

Table 16: Wilcoxon Signed Ranks tests for significant core feature results

		1	2	3	4	5	6	7	8	9
		Words	Clauses	AS-units	Words (LT)	Clauses (LT)	AS-units (LT)	Words/ OppSpk	Clause/ OppSpk	AS-unit/ OppSpk
Admin 1	Z	-5.511[b]	-5.493[b]	-5.067[b]	-4.979[b]	-4.767[b]	-4.243[b]	-5.281[b]	-5.033[b]	-4.161[b]
Vs	p	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
Admin 2	*r*	0.535	0.534	0.492	0.493	0.472	0.420	0.513	0.489	0.404
Admin 2	Z	-0.483[b]	-0.091 [b]	-0.264 [b]	-.309[a]	-.797[a]	-.807[a]	-2.780[b]	-2.386[b]	-2.151[b]
Vs	p	0.646	0.888	0.920	0.757	0.426	0.420	0.005**	0.017*	0.031*[c]
Admin 3	*r*	0.047	0.019	0.026	-0.030	-0.078	-0.079	0.270	0.232	0.209
Admin 1	Z	-4.993[b]	-4.496[b]	-4.233[b]	-4.072[b]	-3.557[b]	-3.260[b]	-5.750[b]	-5.136[b]	-4.710[b]
Vs	p	0.000***	0.000***	0.000***	0.000***	0.000***	0.001**	0.000***	0.000***	0.000***
Admin 3	*r*	0.485	0.437	0.411	0.407	0.356	0.326	0.558	0.499	0.457

a – based on positive ranks b – based on negative ranks c – falls between Boneffori correction and p = 0.05

faster rate at which the test-takers spoke. In the third administration the test-takers spoke a similar amount of words to the second administration but in fewer and therefore longer clauses and AS-units. Also, the length of long turns (turns that lasted 10 seconds or longer) was on average shorter than the long turns in the second administration. These figures point to the candidates speaking at a faster rate, in the third administration, and as will be seen in Section 3.2.1.1, this is borne out in the fluency figures.

The pattern of growth shown here, of considerable gain followed by little or no gain has been noted in the literature. For example, in a study on band score gain in the IELTS test, Elder and O'Loughlin (2003) found that the most important predictor was incoming proficiency: those who start from a low level could gain more, but those with a higher incoming level found it much more difficult to make further improvements. Similarly, after starting from a low level, the students could make impressive gains in the middle administration, but the same gains were not possible in the final administration.

Having described the overall trends of the words, clauses and AS-units that were the basis for many of the other indices, the next section will move on to describe the findings in the CAF indices, starting with complexity.

3.2.1.2 Development in complexity

The two measures for complexity are the ratio of words to AS-units and Clauses to AS-unit. As explained above, figures are given both for long turns and for the entire corpus. The main findings are displayed graphically in Figure 9, in which the figures for long turns are represented by the dashed lines and the straight lines show the entire corpus. The top two lines are the words per AS-unit measures and use the scale on the left hand axis, and the bottom two lines are the ratio of clauses per AS-unit which use the scale on the right hand axis. Both of these measures are consistent in showing that long turns are more complex in terms of having more words per AS-unit and a higher clause to AS-unit ratio. Long turns may be longer by virtue of the subordinating clauses as the test-takers express more difficult notions (Leaper & Riazi, 2014). It has also been found that the frequency of subordination may be boosted by the habitual tagging of "I think" to many of the test-takers' opinion statements in Hungarian (Bygate, 1999) and Hong Kong (Gan, 2013) contexts, and it is certainly a frequently occurring phrase amongst the transcripts of these Japanese students.

Figure 9: Complexity figures for entire corpus and in long turns

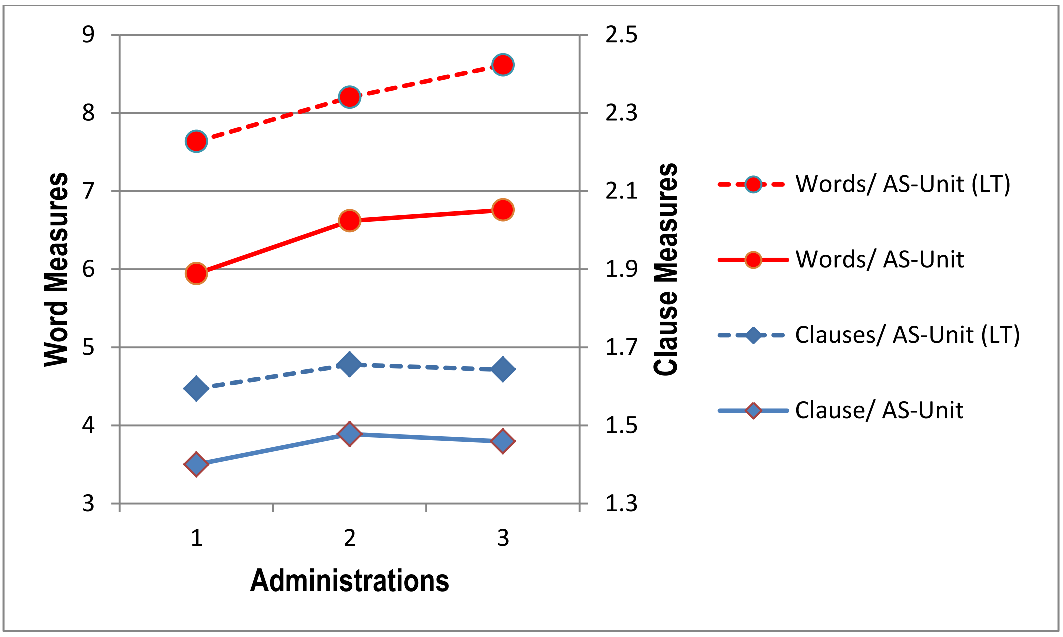


Figure 9 also shows that words per AS-unit improved in successive administrations for both long turns and the entire corpus, with the gain in the third administration for the entire corpus being not as great as that achieved in the second administration. The clauses per AS-unit figures for long turns and all turns are flatter, and show a small rise and decline over the course of the three administrations. From the figures in this graph it can be deduced that GOTs that feature a higher proportion of shorter turns are likely to bring the complexity measures down, and as such, prompts that tend to elicit long turns, as Leaper and Riazi (2014) found existing in the second administration of this data, may well have brought the overall averages up to some degree relative to the other administrations. It also helps to explain the finding of Nitta and Nakatsuhara (2014) that their raters assigned significantly higher scores to paired interactions which had a greater proportion of longer turns, as was discussed in the literature review (Section 2.2.4.4).

The descriptive statistics for the complexity indices in Table 17 reveal the numerical values for this graph, with columns 1-2 showing the figures for the overall counts, and columns 3-4 the long turn figures. The Friedman's ANOVA in (see Table 66 in Appendix C) found that the differences in the figures for clauses per AS-unit were not significant in neither the complete data set ($\chi^2 = 5.610$, $p = 0.061$) nor for long turns ($\chi^2 = 1.380$, $p = 0.502$). Whereas the words per AS-unit were statistically significant for the overall counts ($\chi^2 = 7.396$, $p < 0.05$) and in long turns ($\chi^2 = 8.576$, $p < 0.05$).

Table 17: Descriptive statistics for complexity in long turns

	Administration 1			
	1 Word/ AS-unit	2 Clause/ AS-unit	3 Words/ AS-unit (LT)	4 Clauses/ AS-unit (LT)
No.	53	53	49	49
Min	3.222	1.000	4.333	1.000
Median	5.912	1.357	7.304	1.500
Max	9.000	1.933	12.5	2.5
Mean	5.943	1.400	7.635	1.594
Std Dev	1.568	0.276	2.035	0.414
	Administration 2			
No.	53	53	53	53
Min	2.818	1.000	3.667	1.000
Median	6.364	1.500	8.000	1.667
Max	10.944	2.167	17	3.5
Mean	6.619	1.478	8.201	1.656
Std Dev	1.660	0.251	2.171	0.392
	Administration 3			
No.	53	53	51	51
Min	3.714	1.056	4.727	1.000
Median	6.385	1.400	8.667	1.600
Max	10.895	2.037	12.625	2.375
Mean	6.759	1.459	8.615	1.643
Std Dev	1.729	0.261	1.813	0.327

The Wilcoxon signed ranks figures (Table 18) showed that the significant relationships in the overall figures were due to higher complexity used by the test-takers in the second ($z = -2.439$, $p < 0.05$, $r = 0.237$), and third administrations ($z = -2.333$, $p < 0.05$, $r = 0.227$) compared to the first administration, with small effect sizes, but there was no significant difference between the second and

Table 18: Wilcoxon Signed Rank tests for the significant complexity results

		Words/AS-unit	Words/AS-unit (LT)
Admin 1	Z	-2.439[b]	-1.885[b]
Vs	p	0.015*	0.059
Admin 2	*r*	0.237	0.187
Admin 2	Z	-0.120[b]	-1.134[b]
Vs	p	0.905	0.257
Admin 3	*r*	0.012	0.111
Admin 1	Z	-2.333[b]	-2.656[b]
Vs	p	0.020*[c]	0.008**
Admin 3	*r*	0.227	0.266

a – based on positive ranks b – based on negative ranks
c – falls between Boneffori correction and p = 0.05

third administrations. The figures for long turns showed a similar trend, but the only significant difference was between the first and third administrations (z = -2.656, p < 0.01, r = 0.266). That the overall index of words per AS-unit for words are significantly higher in the second administration, but not in long turns is indicative of a general pattern of improvement. The lack of significance in the second administration's long turns suggests that they were constructed by stringing together relatively similar AS-units into longer turns: test-takers were improving by holding the floor for longer periods of time rather than by making their language more complex. Improvements in complexity in long turns appear to be incremental because it is only by the end of their second year that they show significant gains not over the second, but over their first test.

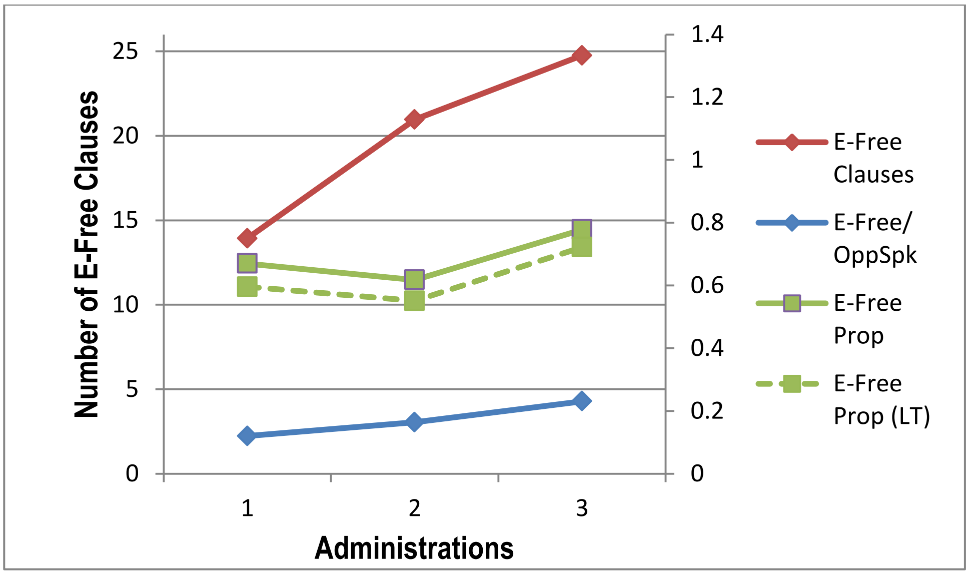
A factor that may offer at least a partial explanation is the influence of the prompt. Leaper and Riazi (2014) found that two of the four prompts (the 'Singles' or the 'Traditional' prompts, see Appendix A) in the second administration elicited significantly longer turns, that may in turn have influenced these figures to be higher. Consistent with this is the finding that of the 53 participants in this study, 31 of them had either of these two prompts, making the prompts that elicited longer turns, and hence more complexity, somewhat over-represented in the sample. The difference in the long turn statistics that the measure of the test-takers' complexity improved overall from the first to the third administration, yet did not improve in successive administrations perhaps suggests that after reaching a threshold level, students develop more in a qualitative way. The finding of no significance in the measure of clauses per AS-unit may indicate that this measure of complexity is more subject to change due to test specific factors such as the prompt, as was found by Leaper and Riazi (2014).

When compared to other longitudinal studies that have included subordination as a measure of complexity in oral language development, the results have been mixed. Other studies with the same finding of no significant gain in the complexity measure of subordination were on Catalan-Spanish (Mora & Valls-Ferrer, 2012) and Spanish university students (Serrano, Tragant & Llanes, 2012) on study abroad programs. These studies collected their data from peer interview and oral narrative tasks respectively. On the other hand, Ahmadian (2011) used subordination to measure the effectiveness of repetitive tasks, and found that it showed significant differences between the control group and the experimental group in their response to interview questions in a six month study. On a monologic task conducted over a year, Vercellotti (2012) found a slight but significant growth in the measure of clauses per AS-unit for participants. It might be thought the task would play its part, but all of the studies mentioned above would tend to elicit longer turns or monologues, which are quite different to GOTs in which there is no obligation to produce longer turns and there is likely to be a higher proportion of shorter turns with less subordination that are taken in order to maintain interaction. Another possible factor has been proffered by Norris and Ortega (2009) who contend that subordination might be "of greater value" (p. 564) for learners at intermediate and upper-intermediate levels. It is most likely that the explanation for the disparate results may be found in a combination of factors that would take further research to unravel.

3.2.1.3 Development in accuracy

The indices for measuring accuracy were the proportion of error free clauses to all clauses, and the normalized number of error free clauses, that is, error free clauses per opportunity to speak. These are displayed in Figure 10, where the top line (the red line, which relates to the scale on the left hand side) depicts the overall average of error free clauses and shows a steady increase from administration 1 to 3. The dashed green line is error free proportion in long turns, and the blue line at the bottom is the normalized number of error free clauses. The comparison between proportion of errors in long turns and all turns (the solid green line) shows that test-takers spoke with lower accuracy in longer turns, probably due to these being the turns in which they are attempting to construct meaning, as opposed to shorter turns that may include error free chunks such as "How about you?". The contrasting pattern for the proportion of error free clauses and the number of error free clauses is worthy of note: both the raw number (the red line) and per opportunity to speak (the blue line) increases in every administration, whereas the proportion of error free clauses declines in the second administration before finally improving. This suggests that students improve by speaking more before they improve the accuracy with which they speak.

Figure 10: Accuracy indices over the three administrations



The descriptive statistics for the accuracy indices that can be found in Table 19 give further information about the variability associated with these trends. It can be seen that the proportionate figures standard deviation declines in every administration, showing the narrowing of the range of the proportion of errors. The standard deviation of the error free clauses per opportunity to speak moves in the opposite direction in every administration, the greater variability perhaps a sign that some test-

Table 19: Accuracy indices over the three administrations

Administration 1				
	1	2	3	4
	E-Free Clauses	E-Free/ OppSpk	E-Free Prop	E-Free Prop (LT)
Test-takers	53	53	53	49
Min.	1	0.009	0.200	0.000
Median	11	0.097	0.667	0.619
Max.	62	0.346	1.000	1
Mean	13.925	0.120	0.670	0.596
Std Dev	10.878	0.075	0.204	0.236
Administration 2				
Test-takers	53	53	53	53
Min.	1	0.009	0.143	0.125
Median	16	0.131	0.629	0.571
Max.	88	0.718	0.978	1.000
Mean	20.962	0.164	0.618	0.551
Std Dev	15.282	0.114	0.153	0.194
Administration 3				
Test-takers	53	53	53	51
Min.	4	0.042	0.286	0.000
Median	24	0.216	0.800	0.762
Max.	54	0.591	0.946	1.000
Mean	24.755	0.231	0.778	0.722
Std Dev	11.713	0.124	0.136	0.187

takers are more confident about talking in English despite making mistakes, while others have improved their ability to speak more accurately.

The Friedman's ANOVA tests (Table 66 in Appendix C) showed that significant differences exist in all these indices. The Wilcoxon Signed Ranked tests in Table 20 show that the significant figures found for the proportion of error free clauses had the same pattern in the overall figures and for the long turns, with no significant difference found between the first and second administrations (z = -1.876, p = 0.062, r = 0.181 for overall; and z = -1.508, p = 0.132, r = 0.149 for long turns); and the third administration being significantly higher than the first (z = -2.922, p < 0.01, r = 0.284 for overall;

Table 20: Wilcoxon Signed Rank tests for significant accuracy results

		Error free clauses	Error-free clause per opportunity	Error-free Clause Proportion	E-Free Clause Proportion (LT)
Admin 1	Z	-4.180^{b}	-3.729^{b}	-1.867^{a}	-1.508^{a}
Vs	p	0.000***	0.000***	0.062	0.132
Admin 2	*r*	0.406	0.362	0.181	0.149
Admin 2	Z	-2.461^{b}	-4.165^{b}	-5.613^{b}	-4.752^{b}
Vs	p	0.014*	0.000***	0.000***	0.000***
Admin 3	*r*	0.239	0.405	0.545	0.466
Admin 1	Z	-4.897^{b}	-5.316^{b}	-2.922^{b}	-3.294^{b}
Vs	p	0.000***	0.000***	0.003**	0.001**
Admin 3	*r*	0.476	0.516	0.284	0.329
a – based on positive ranks b – based on negative ranks					

and z = -3.294, p < 0.001, r = 0.329 for long turns) and second administrations (z = -5.613, p < 0.001, r = 0.545 for overall; z = -4.752, p < 0.001, r = 0.466 for long turns). Meanwhile, the increase in the number of error-free clauses (both in the raw counts and when normalized for time and number of speakers per test) was significant between every administration with similar effect sizes, except for the relationship between the second and the third administrations, where there was a small effect size for the raw figures and a medium effect size for the normalized.

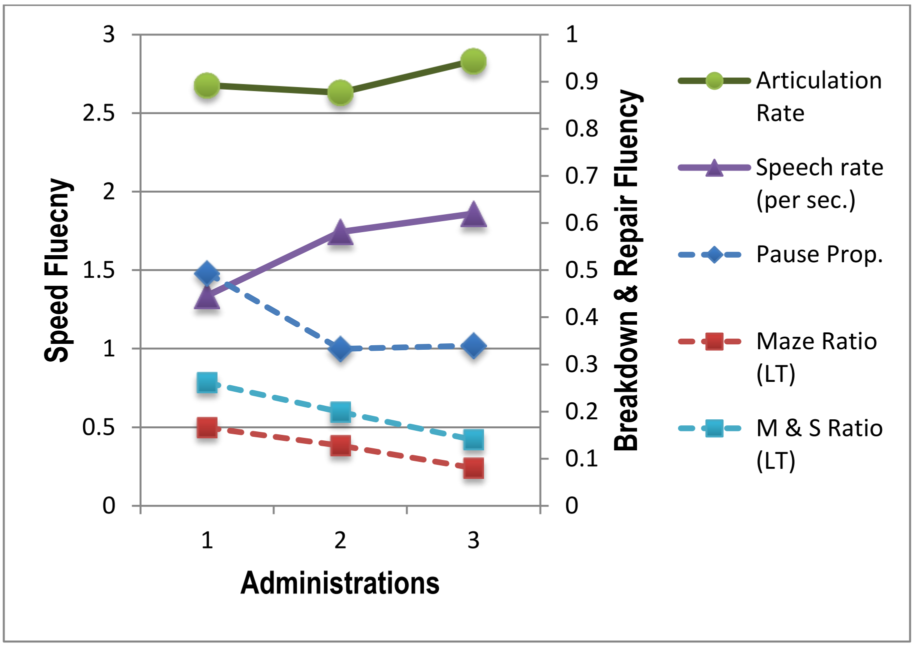
While it might seem paradoxical for there to be no difference found in the proportion of error free clauses between the first and the second administration when the number of error free clauses increased significantly, this is readily explained by the higher number of clauses spoken in the second administration (as was seen in Figure 7, which naturally included a higher number of error free clauses, though their proportion to the overall number of clauses was not significantly different. The finding that the test-takers do not show significant improvements in the proportion of error free clauses until the third administration seems to indicate that these students experienced a delayed onset in the improvement of global accuracy with which they speak. The continuing improvement in the number of error free clauses in the third administration from the second is impressive when it is considered that there was no significant difference in the number of clauses spoken between the second and third administrations. This finding illuminates the pattern of development in accuracy found in this data: that first the improvement in accuracy comes from the ability to use a greater number of error free clauses, and this is followed by a qualitative improvement in the proportion of clauses spoken without error.

This pattern of delayed development of accuracy in speaking is supported by the literature on language gain in longitudinal contexts. An investigation of 14 Spanish university students on a one year program at a UK university that also had beginning, middle and final data collection points, found the same delayed onset of improvement (Serrano et al., 2012). In the findings there was no significant improvement in the first semester, and it is not until the second semester that a significant gain is found in the proportion of errors. The difference in time span of one year for Serrano et al. and two years for this study might be explained by the study abroad context providing greater exposure over a shorter period of time. On the other hand, in a study conducted over six months in an EFL context, like this dissertation, Ahmedian (2011) found that although his participants improved significantly in complexity and fluency, accuracy was unaffected. If improvements in accuracy are delayed, the suggested reason for this is that Ahmedian's (2011) study was too short for the participants to show an improvement in the percentage of error-free clauses. It seems that accuracy may be a feature that takes more time for improvement.

3.2.1.4 Development in fluency

When describing the findings for fluency it is worth remembering that the fluency figures related to time, namely articulation rate, speech rate and pause proportion, are derived from the turns over ten seconds that the participants take, whereas repair fluency could be calculated for the entire corpus. As such, two measures, the maze ratio and maze and sound ratios, were calculated both for the entire corpus and for long turns only. Providing this data for long turns allows consistency with the time-based fluency figures, and providing it for the entire corpus allows a limited insight into one aspect of overall fluency. The main trends for the indices that were collected from long turns only can be seen in Figure 11.

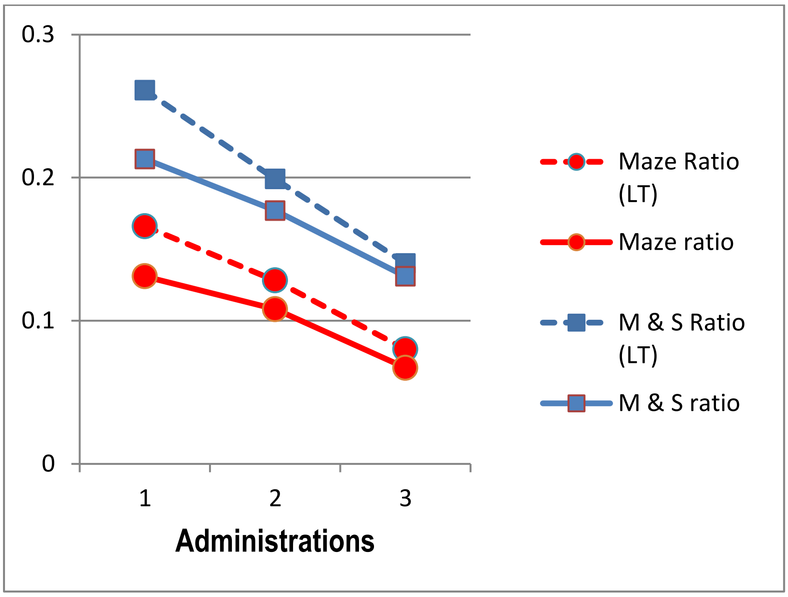
Figure 11: Fluency measures over the three administrations



The two top lines are the indices related to speed fluency, and use the scale on the primary axis on the left. The dashed lines all relate to the secondary axis on the right and represent breakdown and repair fluency, and for these indices, a declining slope represents improvement in fluency. Figure 11 shows that the various aspects of fluency do not develop evenly. In the measures for speed fluency, the articulation rate declines slightly in the second administration before improving in the third, while the speaking rate improves in each succeeding administration. The proportion of pauses, a measure of breakdown fluency, shows a considerable improvement in the second administration from the first, but a small decline in performance in the third administration. Finally, in the repair fluency figures, the maze ratio and sound and maze ratio in long turns show improvements in each succeeding administration (maze ratio: $\chi^2 = 15.906$, $p < 0.001$; maze and sound ratio $\chi^2 = 26.167$, $p < 0.001$).

Before continuing with an examination of descriptive statistics and tests of significance for fluency, the difference between the overall and long turn figures for repair fluency should be expanded on, and the graph in Figure 12 displays this. As for complexity and accuracy for the long turns, the figures for repair fluency in long turns show a slightly lower quality of performance than for the

Figure 12: Repair fluency in long turns compared to overall figures



overall figures, confirming that test-takers found speaking in longer stretches more challenging than in the shorter turns of conversational interaction.

Beyond the main figures that were explained above, in the descriptive statistics displayed in Table 21 the most consistent trend shown is the declining standard deviation for the repair fluency indices of the maze and maze and sound ratios in columns 4 to 7. This shows there was less variability in this kind of disfluency over the three administrations, and suggests a reduction in the difference between the most fluent and least fluent test-takers in this category. It is notable that contrary to what might be expected, the indices related to breakdown fluency, pause proportion and speech rate, the standard deviation spikes in the second administration, indicating increased fluctuation. One of the findings of Leaper and Riazi (2014) was that one of the prompts in the second administration in particular elicited a significantly higher proportion of pauses, and the most likely explanation was that this was due to test-takers reflecting on the potentially face threatening issues of getting married and having children. It could be conjectured that the impact of prompt is reflected in the elevated standard deviation figures here.

Table 21: Descriptive statistics for complexity, accuracy and fluency

	Administration 1						
	1	2	3	4	5	6	7
	Artic. Rate	Speech Rate	Pause Prop	Maze Ratio (LT)	M & S Ratio (LT)	Maze ratio	M & S ratio
Test-takers	49	49	49	49	49	53	53
Min	1.646	0.491	0.296	0.000	0.000	0.000	0.000
Median	2.699	1.360	0.461	0.120	0.224	0.093	0.190
Max	3.714	2.043	0.812	0.885	1.500	0.631	1.087
Mean	2.676	1.337	0.494	0.166	0.261	0.131	0.213
Std Dev	0.472	0.334	0.119	0.155	0.233	0.120	0.187
	Administration 2						
Test-takers	53	53	53	53	53	53	53
Min	1.640	0.864	0.077	0.000	0.030	0.000	0.023
Median	2.634	1.781	0.315	0.105	0.160	0.079	0.133
Max	3.354	2.370	0.672	0.545	0.636	0.390	0.613
Mean	2.630	1.744	0.333	0.128	0.199	0.108	0.177
Std Dev	0.396	0.371	0.125	0.105	0.137	0.091	0.129
	Administration 3						
Test-takers	51	51	51	51	51	53	53
Min	1.906	0.856	0.029	0.000	0.000	0.000	0.038
Median	2.839	1.860	0.339	0.079	0.122	0.067	0.121
Max	4.021	2.562	0.608	0.226	0.328	0.188	0.308
Mean	2.831	1.859	0.340	0.080	0.140	0.067	0.131
Std Dev	0.405	0.355	0.116	0.053	0.068	0.042	0.057

The initial tests of significance using Friedman's ANOVA showed that there are significant differences in all of these indices (Table 67 in Appendix C). The Wilcoxon signed ranks tests (Table 22) reveal that in the third administration the test-takers spoke with an articulation rate that was significantly higher than when they spoke in the second ($z = -2.489$, $p < 0.05$, $r = 0.244$) or first administrations ($z = -2.118$, $p < 0.05$, $r = 0.212$). In contrast the speech rate showed significant improvements in every administration: in the second administration from the first ($z = -4.974$, $p < 0.001$, $r = 0.492$), in the third from the second ($z =-2.301$, $p < 0.05$, $r = 0.226$), and in the third from the first ($z = -5.928$, $p < 0.001$, $r = 0.593$) and in doing so posted the strongest effect sizes in this analysis.

Table 22: Wilcoxon Signed Ranks tests for fluency

		1	2	3	4	5	6	7
		Artic.	Speech	Pause	Maze	M & S	Maze	M & S
		Rate	Rate	Prop.	Ratio (LT)	Ratio (LT)	Ratio	Ratio
Admin 1	Z	-0.602[b]	-4.974[b]	-5.397[b]	-1.149[a]	-2.049[a]	-0.943[b]	-1.598[c]
Vs	p	0.547	0.000***	0.000***	0.251	0.040*[c]	0.346	0.110
Admin 2	*r*	0.060	0.492	0.534	0.114	0.203	0.092	0.155
Admin 2	Z	-2.489[a]	-2.301[b]	-0.506[b]	-2.596[a]	-2.911[a]	-2.767[a]	-2.457[b]
Vs	p	0.013*	0.022*[c]	0.613	0.009**	0.004**	0.006**	0.014*
Admin 3	*r*	0.244	0.225	0.050	0.255	0.285	0.269	0.239
Admin 1	Z	-2.118[a]	-5.928[b]	-5.303[a]	-4.323[a]	-4.426[a]	-4.153[a]	-3.599[c]
Vs	p	0.034*[c]	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
Admin 3	*r*	0.212	0.593	0.530	0.432	0.443	0.403	0.350

a – based on positive ranks b – based on negative ranks c – falls between Boneffori correction and p = 0.05

The pause proportion, by contrast conformed to the familiar pattern seen in the core statistics: the second (z = -5.397, p < .001, r = 0.534) and third administrations (z = -5.303, p < .001, r = 0.530) being significantly lower than the first, but no significant difference between the second and third. The maze ratio had the same pattern in long turns and for the entire corpus by showing no significant difference between the first and the second administrations, but significant improvements in the third from the second administration (maze ratio [LT]: z = -2.596, p < 0.01, r = 0.255; maze ratio: z = -2. 767, p < 0.01, r < 0.01) as well as between the third and the first, and the maze and sound ratio for the corpus followed this pattern too. The maze and sound ratio for long turns was exceptional in that it improved significantly in successive administrations (second from first: z = -2.049, p < 0.05, r = 0.203; third from second: z = -4.426, p < 0.001, r = 0.443).

The results confirm that different aspects of fluency develop at different rates. The strongest trend, as shown by its large effect size, is in the speech rate, which improved significantly in the second and in the third administrations. This contrasts with the articulation rate which only had a significant improvement in the third administration, and the pause proportion, which only showed a significant improvement in the second administration with a strong effect size. The maze ratio improved significantly in the third administration, while the maze and sound ratio improved significantly in each administration. To explain this pattern of development it is important to point out that the articulation rate does not include pause time, while the speaking rate does, and thus it combines elements of breakdown and speed fluency (De Jong, 2013). Thus, amongst these learners, the pattern of development that is suggested by the data is that breakdown fluency improves first, as shown by the test-takers reducing their pausing in the second administration and the non-significance of the articulation rate. The next stage is that students improve their repair and speed fluency by increasing the rate at which they can produce the sounds of language whilst reducing the number of false starts, repeated and unnecessary words. Evidence can be seen in the improvements in the articulation rate and maze ratio and lack of improvement in the pause proportion in the third administration from the second. It seems that by their third administration test-takers are talking more quickly with fewer maze words and voiced pauses, but there is just as much pausing between chunks of language as the second time they took the test.

Other studies that have investigated fluency have typically done so in study abroad (SA) contexts in which there were only pre- and post-data collection points and use a multitude of fluency measures which limits their comparability to the current study. For example, Valls Ferrer (2011) found that all her measures of speed, breakdown and repair fluency increased significantly over a three month SA period, unlike the current study in which the first period saw improvements only in breakdown related fluencies (speech rate and pause proportion). An investigation that did include three data collection points over a longer period of time found that the significant gain in speed fluency (specifically, the articulation rate) took place at the half way point of their study between the first and second data collection points, and there was no significant difference between second and final data collection points (Serrano et al., 2012). Unfortunately this study did not include measures of breakdown or repair fluency, so the comparison is limited. Although this dissertation has found a clear pattern, further research in similar contexts are necessary to support its generalizability.

3.2.1.5 Development of Lexis

Before reporting and discussing the findings of the results by cohort, the first part of this subsection will describe the outcome of the attempt to measure the development of vocabulary displayed by those who spoke more than 100 words in two or more occasions in the test. The descriptive statistics for the groups that were formed can be found in tables 23 to 26. Table 23 gives the figures for the 10 students who produced 100 words in all three administrations.

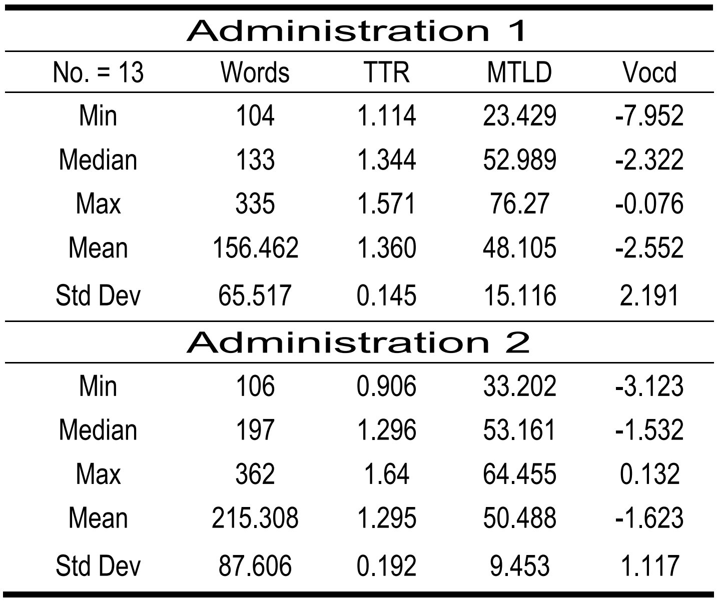
Table 23: Descriptive statistics of lexical indices for speakers of 100+ words in three administrations

Administration 1				
No. = 10	Words	TTR	MTLD	Vocd
Min.	104	1.114	23.429	-7.952
Median	133.500	1.411	51.812	-2.500
Max.	335.000	1.571	76.270	-0.076
Mean	156.462	1.360	48.105	-2.552
Std Dev	72.714	0.164	15.561	2.251
Administration 2				
Min.	106	0.906	33.202	-3.123
Median	225	1.294	48.869	-1.621
Max.	362	1.640	60.850	0.132
Mean	215.308	1.295	50.488	-1.623
Std Dev	86.178	0.212	9.248	1.239
Administration 3				
Min.	130	1.283	33.423	-4.004
Median	223	1.414	48.916	-2.282
Max.	290	1.607	64.784	0.423
Mean	185.538	1.420	53.241	-1.807
Std Dev	43.629	0.090	9.142	1.267

It can be seen that they spoke considerably more words in the second administration than the first, but in the third administration the absolute numbers declined (though if it followed the same trend as for the full sample of test-takers, it is likely that when normalized for time of test and number of participants they would show a slight improvement even in the third administration). The Type-Token Ratio (TTR) and vocd indices show the least consistent pattern of variation, and vary in different ways, as might be expected given that they are subject to variation in sample size. The MTLD, being the least subject to sample size (Koizumi, 2012; Koizumi & In'nami, 2012), showed the steadiest pattern as the test-takers improved in each successive administration. However, Friedman's ANOVA found no significant differences in any index.

The next table (Table 24) shows the descriptive statistics of those who used more than 100 words in their first two tests. Again it can be seen that in the second administration students speak considerably more in the second administration from the first, and according to Wilcoxon's Signed Ranks test, these figures are significant with a large effect size (z = -2.796, p < 0.01, r = 0.548). As before, no significant differences were found between the indices in the first two administrations.

Table 24: Descriptive statistics of lexical indices for speakers of 100+ word in 1st and 2nd administrations



Administration 1				
No. = 13	Words	TTR	MTLD	Vocd
Min	104	1.114	23.429	-7.952
Median	133	1.344	52.989	-2.322
Max	335	1.571	76.27	-0.076
Mean	156.462	1.360	48.105	-2.552
Std Dev	65.517	0.145	15.116	2.191
Administration 2				
Min	106	0.906	33.202	-3.123
Median	197	1.296	53.161	-1.532
Max	362	1.64	64.455	0.132
Mean	215.308	1.295	50.488	-1.623
Std Dev	87.606	0.192	9.453	1.117

Those who surpassed 100 words in the first and third administrations (Table 25) saw substantial increases in the median and mean scores in the third administration from the first. Among the indices, the MTLD shows slight declines and the TTR and Vocd figures record slight improvements, but once more none were enough to reach statistical significance.

Table 25: Descriptive statistics for students who spoke more than 100 words in 1st and 3rd administrations

Administration 1				
No. = 12	Words	TTR	MTLD	Vocd
Min.	102	1.114	23.429	-7.952
Median	133.5	1.366	51.812	-2.121
Max.	335	1.571	76.270	-0.076
Mean	161.083	1.343	49.352	-2.440
Std Dev	68.435	0.166	14.405	2.086
Administration 3				
Min.	130	1.230	32.202	-5.913
Median	214	1.393	46.067	-2.442
Max.	290	1.607	64.784	0.423
Mean	206.667	1.393	46.833	-2.464
Std Dev	47.262	0.103	10.023	1.621

Finally, Table 26 shows the descriptive statistics for the 26 students who spoke more than 100 words in the last two administrations. These test-takers achieved higher mean and median scores in the third administration, and the TTR reflects these improvements. No significant differences were recorded from the Wilcoxon Signed Ranks test performed on the data in this table.

There are likely to be several factors that contribute to the inability of these indices to show any significant differences. Most obviously, the number of test-takers is small and the number of words produced by an individual in the group oral varies considerably each time they take the test, as can be seen in Tables 23 to 26. Less obviously, it is likely that the nature of their response varied considerably depending on the relative proportion of speech in long and short turns, and this too could be contributing to the lack of consistency in these indices.

Table 26: Descriptive statistics of lexical indices for speakers of 100+ words in 2nd and 3rd administrations

Administration 2				
No. = 26	Words	TTR	MTLD	Vocd
Min.	101	0.906	28.572	-7.551
Median	142	1.214	42.108	-2.999
Max.	152	1.312	48.030	-2.006
Mean	267	1.333	53.519	-1.027
Std Dev	395	1.64	60.85	0.532
Administration 3				
Min.	116	1.06	33.423	-6.282
Median	186	1.386	44.427	-2.546
Max.	193.5	1.400	51.153	-1.937
Mean	229	1.464	52.703	-1.232
Std Dev	290	1.607	93.774	1.181

Finally, although it is indisputable that they spoke more in later tests, the possibility that their LD as measured by these indices did not significantly improve should also be entertained. Such a situation might stem from the nature of the GOT. As a communicative context in an assessment situation, it might encourage a conservative use of vocabulary that test-takers are sure their peers will understand, since to use more advanced language might lead to communication breakdown, loss of face (Luk, 2010), and, so they might believe, a negative impact on their test scores.

Rather than using LD indices, analysing the frequency counts may illustrate the development they show as a cohort. The results by administration are displayed in the frequency lists in Table 27. The most obvious trend that can be read from this table is a consistent pattern that had been noted above: while the second administration shows a large improvement over the first, the differences between the second and the third administrations appear to be minimal. In this case, a small decrease in the overall number of tokens and a slightly narrower range can be seen in the third administration compared to the second. However, qualitative improvements in the third administration over the second can be discerned. In particular, development can be seen in the percentage of tokens used in the k1 and k2 bands, which show a consistent trend over the three administrations. In the first band, the percentage of tokens decreases from 83.46% to 80.72% to 77.39%; at the same time, the percentage of tokens in the second band increases from 9.85% to 10.44% to 13.59%. This suggests that by the final administration there is a slight broadening of their vocabulary towards lower frequency words.

Looking further up the frequency bands, it might seem anomalous that in the third administration, the percentage of tokens in the k3 band declines compared to the second. One possible explanation for this aberration may be that the prompts in the second administration encouraged the use of certain words that happened to be in this band. For example, the prompt on mobile phones is very likely to have called for the use of the word *communication,* a k3 word; the prompt on entertainment may well have ushered forth the word *comedy,* and the prompt on getting married probably encouraged students to talk about *graduating*. Various inflections of these three k3 headwords alone account for 6 types and 14 tokens, and shows how small the difference is in word usage between the administrations. This could be another example of how the choice of prompt can influence the language elicited by GOT.

To get an alternative view of the test-takers' use of lexis, the results of the *vocabprofile* analysis (Cobb, 2002) for the 53 participants per administration the three times they took the test are shown in Table 28. The first four rows give the statistics for the number of 1k words, the number in the first 500 function words, and then the first and second 500 content words. Once more, it can be clearly seen that the absolute number of words used jumps from the first to the second administration, but there is little difference between the second and third administrations, and the overall proportion of function and content tokens in the first 1000 word band does not change much over the three times they took the test. The figures that show the development in their ability to use vocabulary is the number of types in the K1 band which can be seen to be declining as they use a greater range of less frequent vocabulary over the next two administrations.

Table 27: Frequency tables of the vocabulary in k1 – k14 frequency bands in three administrations

	Administration 1						Administration 2						Administration 3					
1k bands	Tokens			Types			Tokens			Types			Tokens			Types		
		%	Cum%		%	Cum%		%	Cum%		%	Cum%		%	Cum%		%	Cum%
k1	**4034**	95.71		**449**	83.46		**7192**	96.46		**603**	80.72		**7130**	95.85		**558**	77.39	
k2	**127**	3.01	98.72	**53**	9.85	93.31	**157**	2.11	98.56	**78**	10.44	91.16	**187**	2.51	98.36	**98**	13.59	90.98
k3	**30**	0.71	99.43	**16**	2.97	96.28	**50**	0.67	99.24	**32**	4.28	95.45	**34**	0.46	98.82	**27**	3.74	94.73
k4	**8**	0.19	99.62	**7**	1.30	97.58	**27**	0.36	99.60	**12**	1.61	97.05	**47**	0.63	99.45	**14**	1.94	96.67
k5	**3**	0.07	99.69	**2**	0.37	97.96	**5**	0.07	99.66	**7**	0.94	97.99	**7**	0.09	99.54	**7**	0.97	97.64
k6	**0**	0.00	99.69	**0**		97.96	**11**	0.15	99.81	**4**	0.54	98.53	**14**	0.19	99.73	**6**	0.83	98.47
k7	**1**	0.02	99.72	**1**	0.19	98.14	**1**	0.01	99.83	**1**	0.13	98.66	**2**	0.03	99.76	**2**	0.28	98.75
k8	**5**	0.12	99.83	**3**	0.56	98.70	**2**	0.03	99.85	**2**	0.27	98.93	**1**	0.01	99.77	**1**	0.14	98.89
k9	**1**	0.02	99.86	**1**	0.19	98.88	**4**	0.05	99.91	**3**	0.40	99.33	**0**	0.00	99.77	**0**	0.00	98.89
k10-14	**5**	0.14	100	**5**	1.12	100	**5**	0.07	99.98	**3**	0.40	99.73	**7**	0.09	99.87	**4**	0.55	99.45
Off list	**0**		100	**0**		100	**2**	0.03	100	**2**	0.27	100	**10**	0.13	100	**4**	0.55	100
Total	**4215**			**538**			**7456**			**747**			**7439**			**721**		

This finding is confirmed by looking at the increasing range of types in the k2 category, and to a lesser extent, the AWL list in the next four rows. Both of these bands have been broken into words that came from the prompt, and those that did not appear in the prompt, or ‘original’, as explained in Section 3.1.1.4 and shown in the frequency analysis of the prompt vocabulary (Table 12). As for the *Range* analysis, since raters were instructed not to credit the test-takers for their use of content words from the prompt, I shall ignore those used in the prompt and focus on analysing the ‘original’ words in these two bands. The students’ use of original words in the k2 band improved considerably in number in the second administration with a smaller increase in the third, but more importantly, the proportion of words that fit into this category increase steadily every administration.

Examining the AWL band, it can be seen that the token count increases greatly in the second administration but, surprisingly, in the final administration it decreases to less than the number used in the first administration. It is not until their use of lexis is examined more closely that a reason becomes apparent. The ballooning token count in the second administration is mostly due to a few AWL words being used many times, and this almost certainly resulted from the influence of the prompts used in the second administration. For example, even though the AWL word *job* does not appear in the prompts in this administration (see Table 12), this single word accounts for 22 of the 62 tokens, and every use of this lexical item was in response to the prompts on traditional families and singles.

Although these were two of four prompts, the impact was more than proportional since almost 60% of the test-takers in this sample responded to them. It seems that the token count of lower frequency and AWL words may be dependent on the prompt to a considerable extent. These findings indicate that frequency based analyses may be just as subject to prompt influences as index based accounts have been found to be (Vercellotti, 2012, p. 119). A better indication of development for these test-takers may be the Type count, which can be seen to be increasing every administration, albeit at a slower rate between the second and third times the students took the test.

Table 28: *Vocabprofile* of three administrations

	Administration 1				Administration 2				Administration 3			
	Tokens		Types		Tokens		Types		Tokens		Types	
k1 Words	3998	87.95%	421	67.90%	6767	88.28%	541	67.75%	6717	87.83%	509	60.96%
- 1st 500 Func.	2414	60.38%			4200	62.14%			4137	61.59%		
- 1st 500 Cont.	1296	32.42%			2097	31.03%			2168	32.28%		
2nd 500 Cont.	288	7.20%			462	6.84%			412	6.13%		
k2 Words - Original	107	2.35%	61	9.84%	241	3.14%	87	10.92%	300	3.92%	110	13.17%
k2 Words - Prompt	63	1.39%	6	0.97%	94	1.23%	5	0.63%	90	1.18%	5	0.60%
AWL - Original	30	0.66%	14	2.26%	62	0.81%	23	2.89%	29	0.38%	28	3.35%
AWL - Prompt	0	0	0	0.00%	28	0.37%	1	0.13%	85	1.11%	23	2.75%
Off list	348	7.66%	118	19.03%	473	6.17%	135	17.69%	427	5.58%	160	19.16%
Total	4546	100%	620	100%	7662	100%	797	100%	7648	100%	835	100%

NB: differences in the counts reflect differences in the Vocabprofile and corpuses used in *Range*, as well as the way they count words.

Overall, it can be seen that although the test-takers use slightly fewer tokens in the third administration, the type count has increased by proportionately more over the second administration, allowing the inference that they used a greater variety of words in the time they had available.

Such qualitative improvements are consistent with the literature. Among the few examples of relevant longitudinal vocabulary research, the trends noted above are congruent with what is known about the developmental path of vocabulary acquisition. Consistent with Schmitt (1998), Crossley, Salsbury and McNamara (2010) found that their participants who were studying in an ESL context over a year, improved by using more lower frequency words up to the fourth month of their study, but beyond this time did not improve in the same way. Following this period of growth, the participants of Crossley et al. (2010) developed their vocabulary by learning new senses of frequent words. In this dissertation, it appears that the test-takers' achievement in expanding the range of vocabulary is captured in their transcripts in the second administration, and they may be going through this second phase of qualitative improvement by the time of their third administration.

3.2.1.6 Summary and discussion of results for research question 1

The first research question was to be answered by analysing the language produced by GOT test-takers according to indices of complexity, accuracy, fluency (CAF) and lexical diversity (LD). The results of the CAF analysis are summarized in Table 29.

As Table 29 shows, in complexity, the test-takers could be seen to improve the length of their AS-units significantly the second time they take the test, but not the third time. Improvement in the complexity of turns longer than 10 seconds could only be seen in the difference between the first and third administration. For accuracy, the test-takers showed improvement in every administration by steadily increasing the number of number of error free clauses; but the proportion of error free clauses only improved in the third administration. Fluency also saw a delayed onset of improvements in some of its aspects. While their ability to speak with fewer pauses improved in the test-takers first year performance, it was not until their second year that articulation rate and ability to speak with fewer maze words improved significantly. Although their speaking rate improved significantly in each successive test, they showed their biggest improvement in their second test. Finally, for vocabulary it proved impossible to draw conclusions about their development by using lexical diversity indices due to the widely varying results of the indices of lexical diversity. However, in the analysis of frequency bands in addition to the impressive gains made the second time they took the test, more subtle qualitative improvements in the range of language in lower frequency bands were observable the third time they took the test. This concludes the findings from the CAF and lexis analysis; the next section looks at the overall patterns in turn-taking in the GOT.

Table 29: Summary of main findings from the CAF analysis

		Admin 1 vs 2	Admin 2 vs 3	Admin 1 vs 3
Core Features	Words	↑***	–	↑***
	Clauses	↑***	–	↑***
	AS-units	↑***	–	↑***
	Words (LT)	↑***	–	↑***
	Clauses (LT)	↑***	–	↑***
	AS-units (LT)	↑***	–	↑**
	Words/Opp Spk	↑***	↑**	↑***
	Clauses/ Opp Spk	↑***	↑*	↑***
	AS-units/Opp Spk	↑***	↑*c	↑***
Complexity	Words/ AS-unit	↑*	–	↑* c
	Words/ AS-unit (LT)	–	–	↑**
	Clauses/ AS-unit	–	–	–
	Clauses/ AS-unit (LT)	–	–	–
Accuracy	E-Free Prop	–	↑***	↑**
	E-Free Prop (LT)	–	↑***	↑**
	E-Free/ Opp Spk	↑***	↑***	↑***
Fluency	Articulation Rate	–	↑*	↑* c
	Speech Rate	↑***	↑* c	↑***
	Pause Prop	↓***	–	↓***
	Maze Ratio (LT)	–	↓**	↓***
	Maze & Sound Ratio (LT)	↓* c	↓**	↓***

Key

↑, ↓	- figures increase or decrease significantly
*, **, ***	- significant at the 0.05, 0.01, or 0.001 levels respectively
c	- falls between Boneffori correction and p = 0.05
–	- no significant difference

3.2.2 Research question 2: Development of turn taking in the GOT

This research question regards the extent to which words spoken and turns taken by the test-takers in the group oral are related to the development of language ability. The main findings are displayed graphically in Figure 13 (raw and normalized words spoken, and words spoken in long turns only), Figure 14 (raw and normalized turns taken) and Figure 15 (raw and normalized number of long turns). The descriptive statistics of these indices can be examined in detail in the following Table 30. These graphs are notable for showing the trend seen in the core statistics (AS-units and clauses) in Section 3.2.1.1; all three show an impressive increase in the unadjusted averages in the second administration, but small declines in the third administration. In Table 30, the statistics for the overall number of words spoken, words in long turns, overall number of turns and long turns can be found in the first four columns, followed by their normalized counterparts in columns 5 to 7.

An examination of Table 30 shows that the increase in median words spoken was greatest the second time they took the test, with only a small gain in the final administration from the second, but there was a considerable drop in the median figure for words in long turns despite the median number of long turns remaining constant in the second and third administrations, as is borne out in the declining length of long turn in average number of words in column 2. According to the statistical tests, Friedman's ANOVA (Table 70 in Appendix C) tells us that significant relationships exist in all these indices of words ($\chi^2 = 35.010$, $p < 0.001$), long turn words ($\chi^2 = 23.384$, $p < 0.001$), turns ($\chi^2 = 17.949$, $p < 0.001$), long turns ($\chi^2 = 12.409$, $p < 0.01$), words per opportunity to speak ($\chi^2 = 39.094$, $p < 0.001$), and turns per opportunity to speak ($\chi^2 = 6.491$, $p < 0.05$) except for the normalized number of long turns ($\chi^2 = 4.829$, $p = 0.089$).

Figure 13: Words spoken per administration: raw, normalized and in long turns

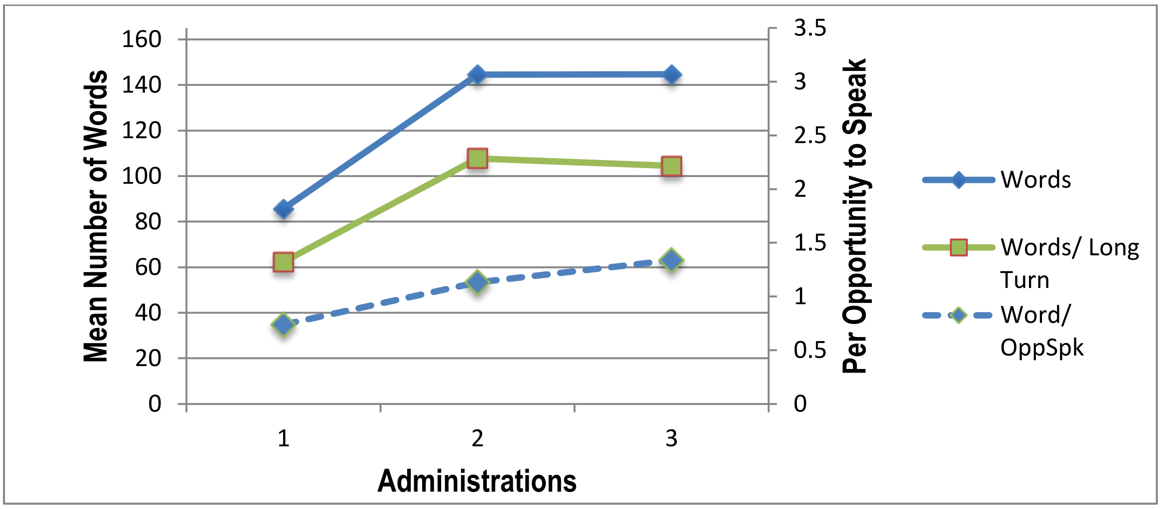


Figure 14: Mean and normalized turns per administration, with standard deviation

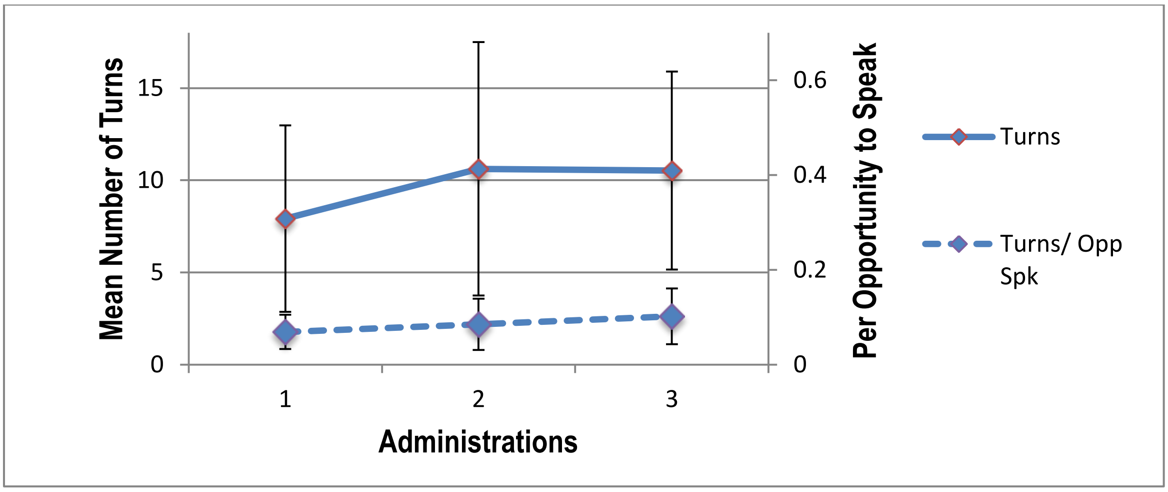


Figure 15: Mean and normalized long turns per administration with standard deviation

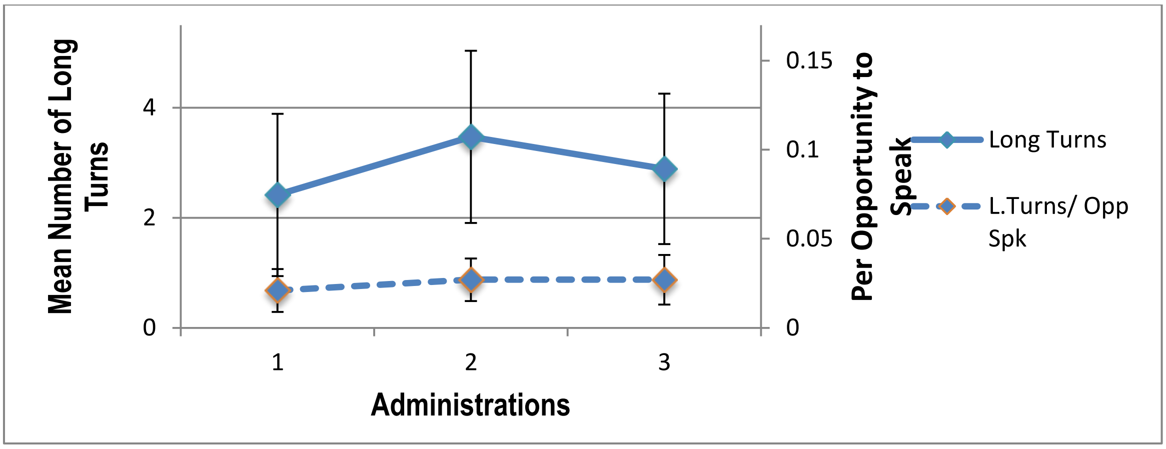


Table 30: Descriptive statistics for words and turns

Administration 1							
All administrations Test-takers = 53	1	2	3	4	5	6	7
	Words	Long Turn Words	Turns	Long Turns	Words/ OppSpk	Turns/ OppSpk	L.Turns/ OppSpk
Min.	19	0	2	0	0.146	0.019	0.000
Median	71	50	7	2	0.784	0.060	0.021
Max.	337	180	30	7	1.880	0.170	0.050
Mean	85.642	57.547	7.925	2.415	0.739	0.069	0.021
Std Dev	57.958	41.832	5.061	1.473	0.373	0.036	0.012
Administration 2							
Min.	31	11	1	1	0.241	0.009	0.009
Median	126	95	9	3	0.922	0.070	0.024
Max.	395	282	34	8	3.223	0.253	0.058
Mean	144.642	107.623	10.623	3.472	1.133	0.085	0.027
Std Dev	82.389	60.910	6.873	1.564	0.598	0.054	0.012
Administration 3							
Min.	26	0	3	0	0.189	0.021	0.000
Median	129	89	9	3	1.230	0.089	0.024
Max.	298	254	25	7	3.371	0.314	0.088
Mean	144.189	100.566	10.528	2.887	1.334	0.102	0.027
Std Dev	64.302	59.550	5.377	1.368	0.629	0.059	0.014

Following up on the significant results, the Wilcoxon Signed Ranks test, as displayed in Table 31, determined that the familiar pattern of the second and third administrations having a significantly greater number than the first administration held for the unadjusted indices of words, long turn words and turns. The number of long turns had a unique pattern: similar to other indices the figures saw a large increase in the second administration ($z = -3.910$, $p < 0.001$, $r = 0.329$), but unlike the others the third administration significantly declined from the second ($z = -2.094$, $p < 0.05$, $r = 0.203$), and was not significantly different from the first ($z = -1.707$, $p = 0.088$, $r = 0.166$). Also, as noted above, when normalized, the number of long turns turned out to be non-significant, unlike the normalized figures for words and turns.

Table 31: Results of the Wilcoxon Signed Ranks test for turn taking

		1	2	3	4	5	6
		Words	LT Words	Turns	Long Turns	Words/ OppSpk	Turns/ OppSpk
Admin 1	Z	-5.520[a]	-4.964[b]	-3.096[a]	-3.910[b]	-5.281[a]	-1.377[b]
Vs	p	0.000***	0.000***	0.002**	0.000***	0.000***	0.169
Admin 2	*r*	0.536	0.507	0.301	0.380	0.513	0.134
Admin 2	Z	-0.459[b]	-.281[c]	-0.488[b]	-2.094[a]	-2.727[a]	-1.890[a]
Vs	p	0.646	0.779	0.626	0.036*[c]	0.006**	0.059
Admin 3	*r*	0.045	0.029	0.047	0.203	0.265	0.184
Admin 1	Z	-4.958[a]	-4.103[c]	-3.084[a]	-1.707[b]	-5.741[a]	-2.873[b]
Vs	p	0.000***	0.000***	0.002**	0.088	0.000***	0.004**
Admin 3	*r*	0.482	0.419	0.300	0.166	0.558	0.279

a – based on positive ranks b – based on negative ranks c – falls between Boneffori correction and p = 0.05

As noted in the discussion of the core indices for the CAF analysis in Section 3.2.1.1, words per opportunity to speak are significantly more in each succeeding administration (column 5 of Table 31), with the effect size of the gain made in the second administration being medium and the gain made in their final year small. After normalization, turns per opportunity to speak are now only significantly greater between the first and third administrations (column 6 of Table 31).

Before discussing the meaning of these results, the final figures to examine are the ratios of words per turn, words per long turn, long turn words per words, and long turns per turn. The main findings for the ratios of words per turn and long turn are displayed in Figure 16. The contrast between the overall words per turn with the words per long turn clearly show that both of these statistics follow the already noted pattern of the second and third administrations being strikingly more than the first administration, but not much difference between the second and third administrations In particular, not only does this graph show the impressive gain in average length of the long turns in the second administration, but also the considerable variance associated with it, as indicated by the standard deviation shown in the error bars.

Figure 16: Words per turn and long turn over three administrations

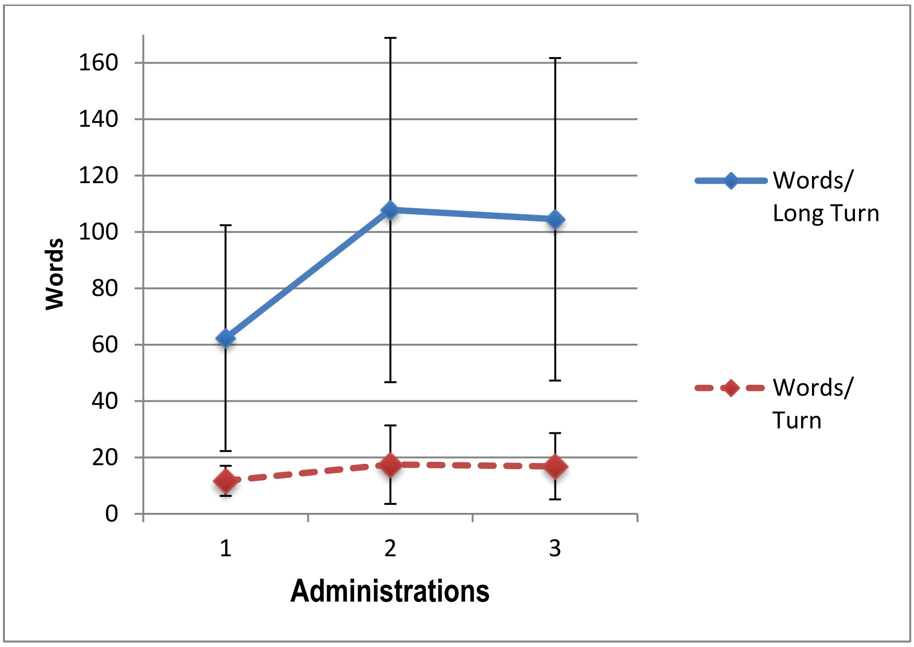


Table 32 shows the descriptive statistics for the final two ratios in columns 3 and 4: long turn words to all words spoken and long turns to all turns, and it can be seen that while they follow the same pattern of an increase in the second administration, the differences do not seem to be as pronounced (displayed in Figure 17). As can be seen in the descriptive statistics in columns 1 and 2 of Table 32, after the initial large gain, both these figures register small declines in average number of words in the final administration, and the median statistics follow this trend. Both of these ratios had significant relationships, according to Friedman's ANOVA and the Wilcoxon Signed Ranks test, for words per turn ($\chi^2 = 9.223$, $p < 0.05$), words per long turn ($\chi^2 = 23.384$, $p < 0.001$).

Table 32: Descriptive statistics of turn-taking ratios

Administration 1				
All administrations Test-takers = 53	1 Words/ Turn	2 Words/ Long Turn	3 LT Words/ Words	4 Long.Turns/ Turn
No.	53	49	53	53
Min.	3.867	16	0	0
Median	10.750	56	0.709	0.333
Max.	25.333	180	1	1
Mean	11.693	62.306	0.660	0.365
Std Dev	5.377	40.013	0.257	0.209
Administration 2				
No.	53	53	53	53
Min.	2.385	11	0.128	0.056
Median	13	95	0.783	0.375
Max.	94	282	1.000	1
Mean	17.444	107.774	0.753	0.422
Std Dev	13.914	61.045	0.181	0.239
Administration 3				
No.	53	51	53	53
Min.	4.333	24	0	0
Median	12.250	92	0.686	0.278
Max.	51.750	254	0.995	0.8
Mean	16.844	104.49	0.665	0.332
Std Dev	11.743	57.177	0.234	0.203

Table 33 shows that they both followed the familiar trend of the second and third administrations being significantly more than the first administration, though the effect sizes for words per turn were small compared to the long words per turn showing medium effect sizes. The index that measures the ratio of words in long turns to total turns is an indication of how important the words spoken in turns over ten seconds are compared to all words spoken. If students speak in fewer longer turns, or in a greater number of turns that are just over ten seconds in length, this proportion will rise, and conversely speaking in shorter turns will drop it. Similarly, speaking in more turns that are over ten seconds in length will increase the ratio of long turns to all turns, and with more turns that are shorter than 10 seconds in length, it will decline.

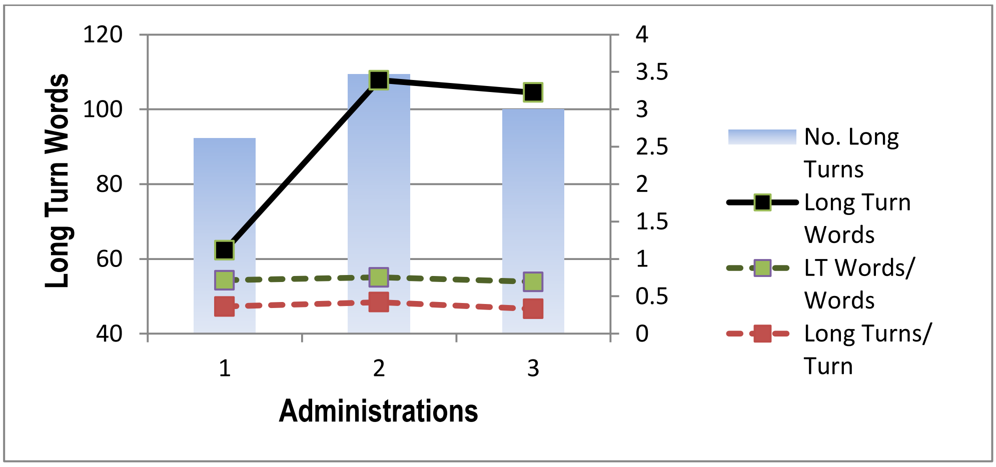
Table 33: Wilcoxon Sign Ranks tests of words per turn and long turn

		Words/ Turn	Words/ Long Turn
Admin 1	Z	-2.837[a]	-4.964
Vs	p	0.005**	0.000***
Admin 2	*r*	0.276	0.507
Admin 2	Z	-0.182[a]	-0.281a
Vs	p	0.855	0.779
Admin 3	*r*	0.018	0.029
Admin 1	Z	-2.536[a]	-4.103
Vs	p	0.011*	0.000***
Admin 3	*r*	0.246	-0.414

a – based on positive ranks
b – based on negative ranks

Yet, as shown in Figure 17, despite the increased number of words spoken in longer turns (the black line in Figure 17), and the greater overall number of longer turns (the bars in Figure 17) in the second administration, neither of these ratios is significantly different (long turn words to all words: χ^2 = 4.829, p = 0.089; long turn turns to all turns: χ^2 = 4.152, p = 0.125) according to Friedman's ANOVA (see Table 70 Appendix D). The reason for this is that in addition to the words spoken in longer turns that are represented in this statistic, test-takers also increased the words spoken in shorter turns. This trend continued in the third administration, which saw significantly fewer longer turns, and only an insignificant decline in both of the ratios.

Figure 17: Ratios of long turn words and turns compared to number of long turns and words

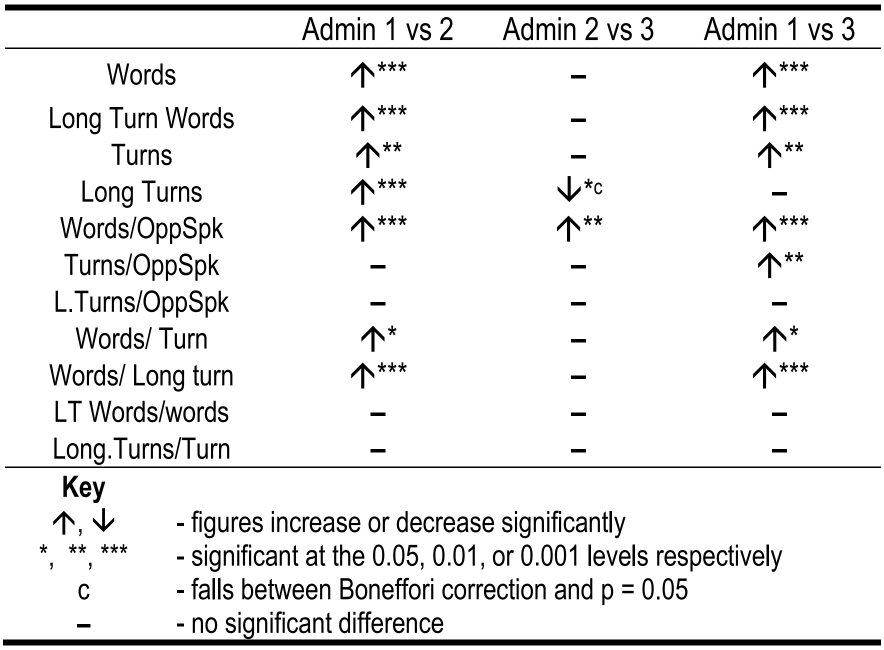


3.2.2.1 Summary and discussion of the results of research question 2

This question was concerned with the development test-takers show in the turn-taking system of the GOT. To do this the available figures for words and turns and long turns were scrutinized using a number of indices, as summarized in Table 34.

The most striking result from this set of data is the considerable improvement in the test takers' second performance in all facets over their first before normalization. After adjusting for the time of the test and number of participants, it becomes clear that the outstanding sign of improvement shown is in the continuous increase in the number of words spoken over the three administrations. The ability to take more turns seems to be more gradual, as the only significant development they show in the normalized figures is between the final and the first administrations, and there is no difference in the frequency of turns over 10 second turns at all. Their ability to take longer turns can be seen to improve along with everything else in the second administration as represented in the significantly more words per turn and words in longer turns, but development did not continue in a quantitative sense.

Table 34: Summary of findings in the development of turn taking in the GOT



	Admin 1 vs 2	Admin 2 vs 3	Admin 1 vs 3
Words	↑***	–	↑***
Long Turn Words	↑***	–	↑***
Turns	↑**	–	↑**
Long Turns	↑***	↓*c	–
Words/OppSpk	↑***	↑**	↑***
Turns/OppSpk	–	–	↑**
L.Turns/OppSpk	–	–	–
Words/ Turn	↑*	–	↑*
Words/ Long turn	↑***	–	↑***
LT Words/words	–	–	–
Long.Turns/Turn	–	–	–

Key

↑, ↓	- figures increase or decrease significantly
*, **, ***	- significant at the 0.05, 0.01, or 0.001 levels respectively
c	- falls between Boneffori correction and p = 0.05
–	- no significant difference

Drawing on other sources to explain this development, Galaczi (2013)'s cross-sectional study suggests that as students turn-taking skills develop they tend towards more frequent speaker changes, and although it is unlikely that the participants in this study approach the highest levels of speakers in Galaczi's (2013) paper, the significant finding in the turns per opportunity to speak are consistent. Secondly, from Leaper and Riazi (2014), whose participants overlap with those of this dissertation, it is apparent that two of the prompts in the second administration tended to encourage longer turns, and since 60% of the participants in this study had these two prompts, the figures for the long turns in this administration may have been inflated somewhat. At the same time, it must be accepted that it is in the second administration that students have the most to gain, since most of them had had little practice of speaking at the time of the first administration, so it would be premature to dismiss these results as due entirely to the prompt. Consistent with this are the lack of significance in the normalized number of long turns spoken and the ratios of long turn words per all words and number of long turns per all turns, that seem to be remarkably stable throughout the administrations, and the proportion of words in long turns, and number of long turns. Although the second administration saw a significantly greater average number of words per turn and length of long turn, these were amidst an increase in the overall number of words spoken. The fact that these did not continue to increase in the third administration is that the test-takers improve by speaking more in less time and using more turns to do so rather than longer ones, as noted by Galaczi.

Thus, the answer to this research question is that the length of the turn is more likely subject to the influence of factors other than the skill of the interlocutor, one of which, as contended by Leaper and Riazi (2014), is the nature of the prompt. Another, it seems is the development of the speakers ability, with the results presented here suggesting that the pattern of development may be initially towards longer turns before tending towards a greater number of shorter turns, which is consistent with Galaczi's (2013) findings.

3.2.3 Research question 3: Interactive functions in the GOT

As described in Section 3.2.3, the transcripts were analysed with the assumption that given the context, everything the candidates said has some interactive function. Using an iterative data driven approach (Brookes, personal communication, August 2, 2013) 30 features were identified, and were categorized into those related to initiating, responding, developing or Collaborative features. This section will go through each of the main categories, firstly describing the results of the coding (for a full account with illustrative examples taken from the corpus, see Appendix G), and then explaining and discussing the descriptive statistics of the count of the relevant feature, before going on to describe the results of the statistical analysis. The final section examines the patterns of development shown by the categories of interactive function to provide a broader view of the test-takers development of interactive skill and answer the question.

3.2.3.1 Identifying initiating features

Initiating features can be distinguished as those that are used to start or add to an interaction, and are produced on the test-taker's volition rather than in response to a question. Since the Initiating features are those put forward by the speaker rather than in response to another's question, test-takers who use these are actively engaging themselves in the conversation. As was pointed out in the literature review, those who do so are more likely to benefit in terms of language learning (McDonough, 2004). The most obvious way of doing this is to start a topic either by asking a question (*Opening Question*; Q-O) or making a statement (*Opening Opinion*; Opin-O), as well as questions that are asked to the previous speaker to get more information and expand on what they have said (*Follow-up Question*; Q-Fu) or to transfer a turn to another speaker (*Transfer Question*; Q-T). Although they are usually coded as the initial AS-unit of a turn, they may also appear as the final AS-unit of a turn, and the most common example were *Transfer Questions* in which 'How about you?' would often serve the dual purpose of marking a definitive end to a turn as well as passing the turn to the next speaker. Statements that were volunteered without being asked a question through speaker self-selection, are also classified as Initiating and it is through these statements that participants could inject their opinion. Included here are statements in which they added their opinion (*Opinion;* Opin), gave a suggestion (*Suggesting*; Opin-sug), appraise what has been said (*Appriasing*; Opin-Ap), summarize what has been said (*Summarizing*; Opin-sum) or challenge either the speaker or what has been said (*Controversial Opinion/Question*; Opin-con). These last three occurred when they were added as an opinion, and therefore meet the criteria for an Initiating feature. The final categories are those that are removed somewhat from the actual discussion: the *Function – Initiate* (Fun-I) category describe moves that are perform functions outside the conversation, for example by greeting their fellow test-takers at the beginning; *Manage Topic – Initiate* (Man-I) are moves that test-takers use to organize the task, for example by asking which prompt question to answer next; and *Japanese – Initiate* (Jpn-I) for those who started a move in their native language, regardless of what was said. The Initiating features are summarized in Table 35.

Table 35: Summary of Initiating features

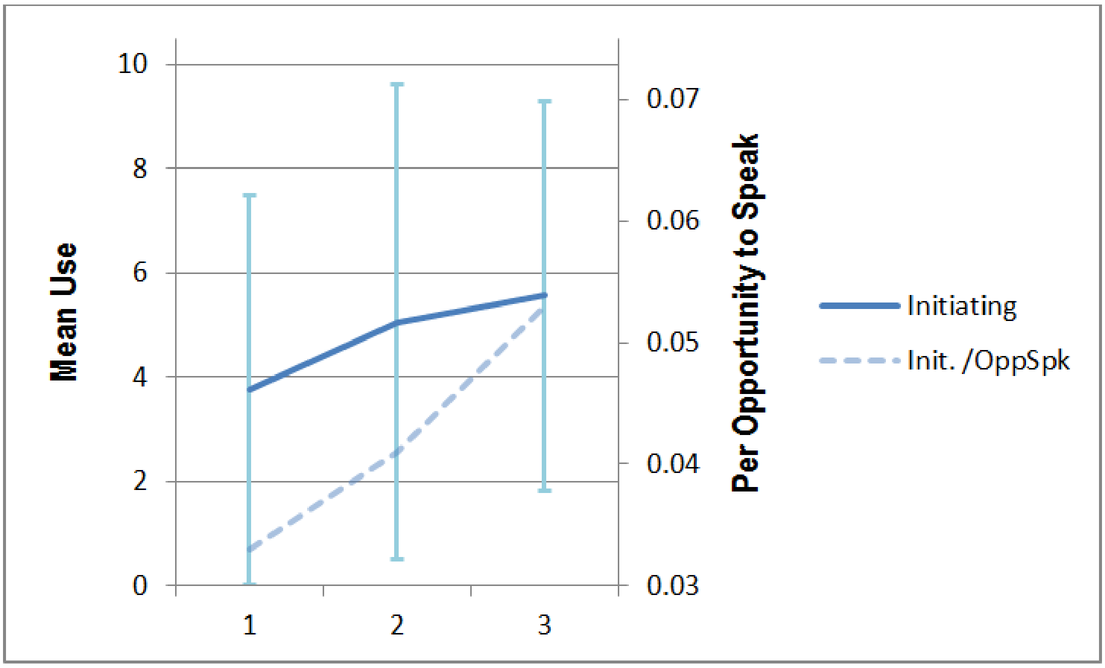
	Initiating feature	Code	Brief Explanation
1.	*Opening Question*	Q-O	Questions used to open conversation or restart a conversation.
2.	*Transfer Question*	Q-T	Questions used to develop the conversation by asking others for their opinion on the same topic.
3.	*Follow-up Question*	Q-Fu	Questions that seek further information on the topic of conversation.
4.	*Opening Opinion*	Opin-O	Statements that are used to begin a new topic or restart the conversation on a new topic.
5.	*Opinion*	Opin	Statements that are volunteered by the speaker not in response to a question to develop conversation on the current topic.
6.	*Controversial Opinion/Question*	Opin-con	Statements or questions that develop the conversation by challenging another speaker in a potentially face threatening way.
7.	*Suggesting*	Opin-sug	Statements that propose that the interlocutor take some action.
8.	*Appraising*	Opin-Ap	A statement in which a participant passes judgement on what an interlocutor has been talking about.
9.	*Summarizing*	Opin-Sum	A statement that synthesizes another speakers opinion or current state of the conversation
10.	*Function – Initiate*	Fun-I	Utterances like greetings that fulfilled a function rather than adding to the discussion
11.	*Manage Topic – Initiate*	Man-I	Utterances that had the purpose of organizing how the discussion should take place.
12.	*Japanese – Initiate*	Jpn-I	A candidate who starts talking in Japanese to somebody else had their utterance coded thus.

3.2.3.2 Initiating features in the GOT

The overall trend in Initiating features is displayed in Figure 18 in which the mean of all such features used in each administration are compared to the normalised figures. It can be seen that both figures grow considerably over the three administrations, though the unadjusted figures show a slightly pronounced gain from the first to the second administration, whereas for the adjusted figures it is from the second to the third that the slope is steeper. Thus when considering the graphs in Figures 19 and 20 it can be considered that the actual rate of usage would be slightly less for the second administration, and slightly more for the third.

Figure 18 displays the standard deviation in the error bars for the average use of Initiating features. An examination of this figure indicates that it is not until the third administration that the lower limit of the standard deviation is greater than one, showing that it is not until then that it can be expected that most test-takers are capable of initiating a topic in the GOT.

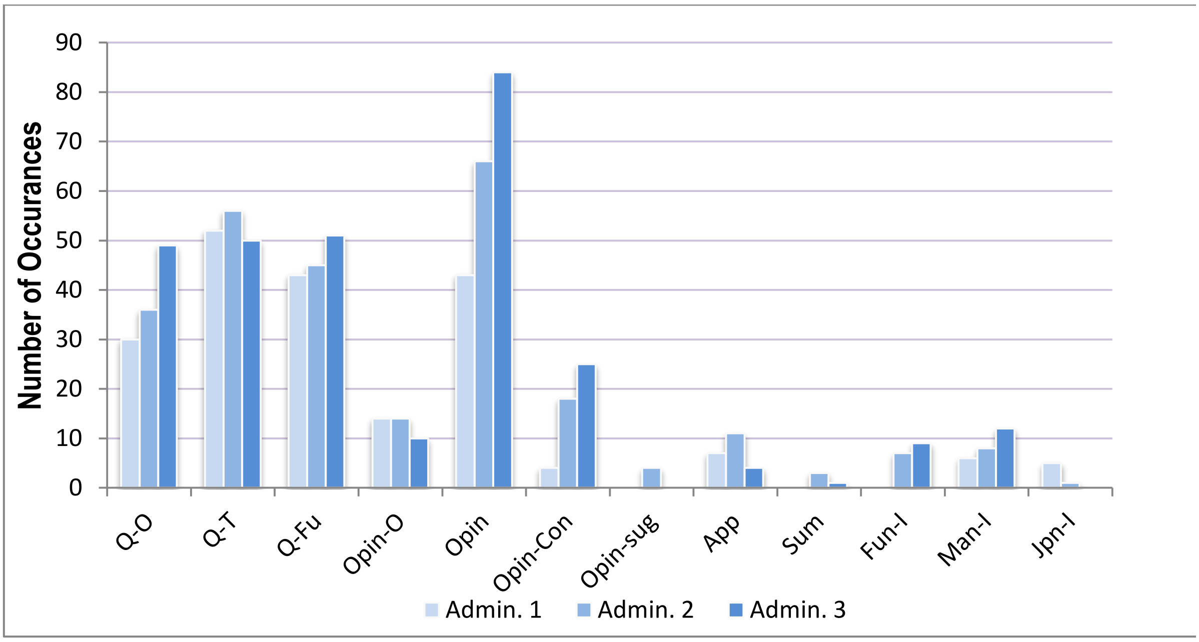
Figure 18: Initiating features: mean usage and std dev. compared with normalized



The counts of Initiating features are graphically displayed in Figure 19. In this graph it can be seen that the Initiating features that test-takers used the most were *Opening Questions, Transfer Questions, Follow-Up Questions* and *Opening Opinions*. Of these *Opening Questions, Follow-up*

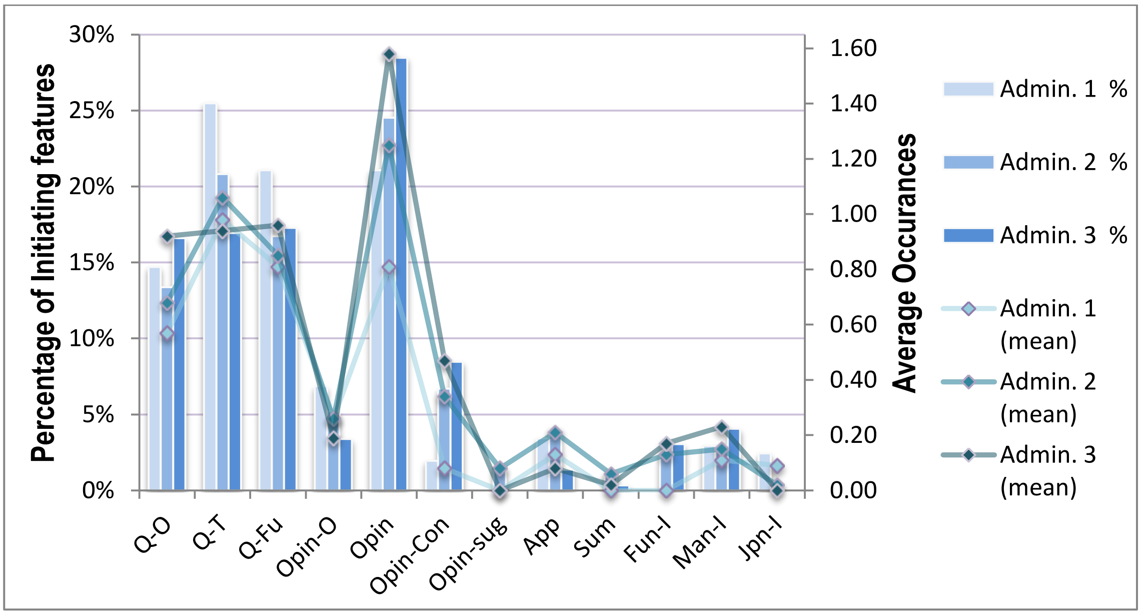
Questions and *Opening Opinions* clearly increase in successive administrations. Among the less frequently used functions, *Controversial Opinion or Questions, Manage Topic – Initiate* and *Function – Initiate* moves rise, perhaps indicating an increasing familiarity with the GOT as an assessment event. On the other hand *Japanese – Initiate* functions disappear by the third administration, possibly showing that their language skills have developed sufficiently for them to avoid the need for Japanese. Overall, the figures can be seen to indicate a broadening range of functions as they advance, though the numbers supporting this trend amongst the lessor used categories are thin.

Figure 19: Number of Initiating features in the three administrations



The next graph, Figure 20, shows the proportion of the categories to the total number of Initiating features in the bar graphs (using the scale on the left hand axis) while the line graph shows their average frequency (using the scale on the right hand axis). It can be seen that as a proportion, the students rely less on *Transfer Questions* in each successive administration (Q-T declines from 25.49% to 20.82% to 16.95%) to pass the move on to the next speaker, perhaps showing they have increasing confidence in the context of conversation that the next speaker will take the turn. A decreasing trend can also be seen in the *Opening Opinion* moves, as perhaps the test-takers are more able to continue

Figure 20: Proportion and average number of Initiating features in the three administrations



the same topic, or by asking *Opening Questions* – which increases by proportion in the final administration. The increasing proportion of *Opinion* in each administration (from 21.08% to 24.52% to 28.47%) is perhaps an indication of students taking it upon themselves to contribute to the current topic, rather than resort to a new one. It is also possible to discern an increase in the number of functional and topic management features (Fun-I and Man-I), which increase from very little in the first administration to collectively 5% and then 7% in the middle and last administrations, indicating a greater awareness of other speakers and the task as a vehicle of assessment. Finally, the proportion of *Controversial Opinion/Question* (Opin-Con) also rises with each administration, which perhaps indicates the test-takers' greater willingness to assert their own opinions, or the greater acceptance of dissenting voices. The other functions either do not show a clear trend (*Opening* or *Follow-up questions*), or do not have large enough numbers to be able to draw inferences from (*Suggesting, Appraising, Summarizing*).

The information about the average usage of these Initiating features in Figure 20 provides a valuable reminder of the number of times these functions were used that may be overlooked when discussing proportions. In the first administration, not one of these functions has a higher average than one, although transfer questions come close ($\bar{x}$= 0.98) showing how dependent most of the test-takers were on those few who did assert themselves by initiating topics. By the second administration, there are still only two functions that average over one: transfer questions and initiating opinions, suggesting that students still find it difficult to initiate. In the final administration, there is only a single function that is over one, *Initiating Opinions*, but it obtains the highest average ($\bar{x}$ = 1.58), and *Opening Questions* ($\bar{x}$ = 0.92), *Transfer Questions* ($\bar{x}$ = 0.94), and *Follow-Up Questions* ($\bar{x}$ = 0.96), are all well-used techniques for initiating topics, supporting the contention that as a whole these test-takers were broadening their repertoire of conversational skills. The descriptive statistics for Initiating features can be found in Table 71 of Appendix E.

3.2.3.3 Identifying Responding features

This category, as the name suggests, were those utterances that replied to another's conversational move. Usually they made up the second half of an adjacency pair in which the first part was an initiating move. Once a conversationalist has used an Initiating feature there is an obligation to respond, which was coded as a Responding feature. The answer to such a move was coded as *Answer* (Ans). Responses in which the speaker simply agrees to the current topic without adding anything new to it, even if they were not themselves asked directly (*Agreement;* Ag) were classified within this function. Also included in this category were responses to another's use of functional language (*Function – Response*; Fun-R) or management topic (*Manage topic – Response;* Man-R). Responses in Japanese could either be a response to another's initiating Japanese, or a response in Japanese to a question asked in English (*Japanese – Response*; Jpn-R). The Responding features are summarized in Table 36.

While the ability to use an Initiating feature is likely to be an indication of involvement in the process of learning a language, Responding features are their more passive counterpart. This is particularly so for the *answer* and *agreement* functions, which should be seen as the minimum participation necessary for maintaining the conversational interaction, and thus the engagement in language learning they show is limited. Amongst them, the feature that shows the most engagement is

Reacting, which also has the practical impact of giving the speaker more time to think of something else to say.

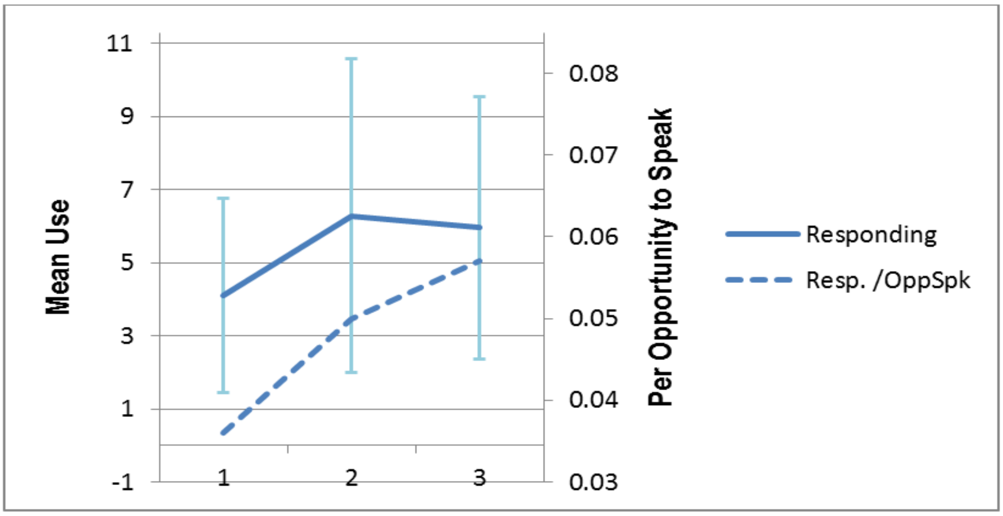
Table 36: Summary of Responding features

Initiating feature	Code	Brief Explanation
13. *Answer*	Ans	The initial response to any question was coded as *answer*.
14. *Agreement*	Ag	These are moves in which consists typically of a minimal token of agreement to the current topic.
15. *Reacting*	React	A short element at the beginning of a turn that often repeats elements of the previous utterance. These may express surprise or act as a device that buys the answerer thinking and floor time
16. *Function – Response*	Fun-R	These are functional responses to *function-initiate* moves. They do not add content to the discussion.
17. *Manage Topic – Response*	Man-R	These respond to the *manage topic – initiate* moves by expressing opinion about how to manage the topic of conversation, and so do not add content to the discussion itself.
18. *Japanese – Response*	Jpn-R	This is the Respond counterpart to the Japanese-Initiate function, and is coded thus if the response is in Japanese.

3.2.3.4 **Responding features in the GOT**

The overall use of Responding features can be seen in Figure 21. In this graph a contrast can be seen when the average use is compared to the normalized figures.

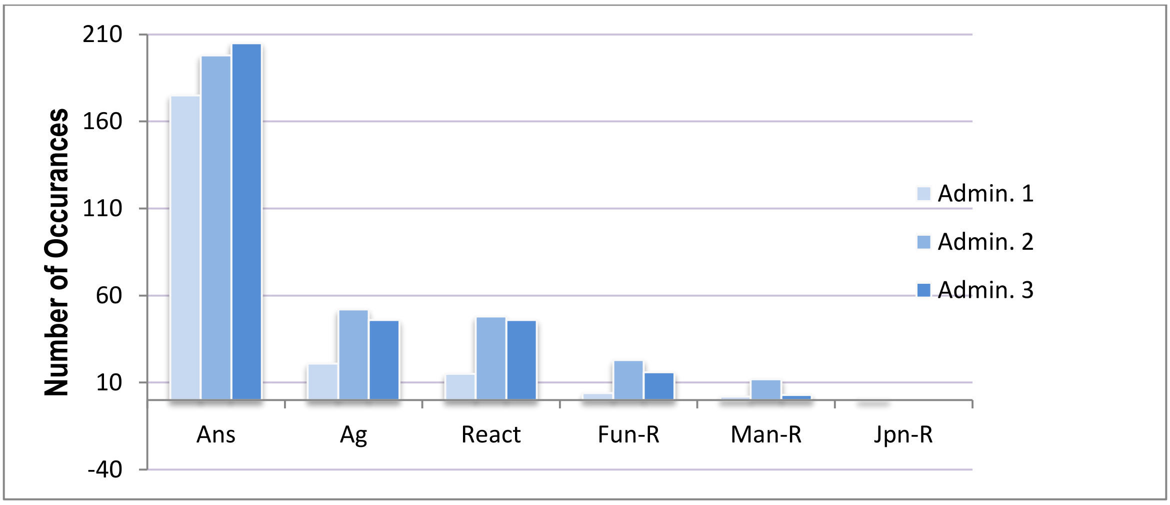
Figure 21: Responding features: mean usage and std dev. compared with normalized



Although the average count of these features declines abruptly, when adjusted for the time of test and participants, the figures continue to increase, albeit at a reduced rate compared to the growth seen in the second administration. Looking at the standard deviation, it can be seen that although there is considerable variability, Responding features are consistently used in all administrations.

The counts of individual responding features over the three administrations are displayed in Figure 22.

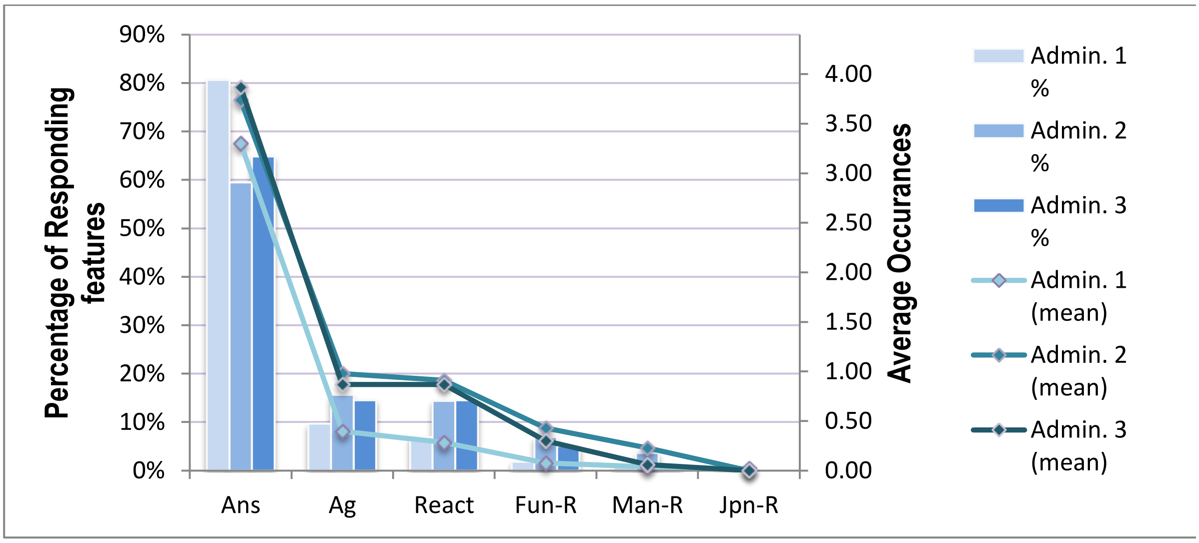
Figure 22: Number of Responding features in the three administrations



This graph shows that the main Responding features (*Answering, Agreeing* and *Reacting*), all increased in the second administration, but apart from a rise in the *Answering* move, their use decreased in the third. However, given the shorter average times and greater frequency of usage in the third administration, as pointed out in Figure 21, it is likely that the third administration figures would show growth when normalized.

The proportion of each feature to the total Responding features and their average usage are displayed in Figure 23. The first administration indeed stands out from the second and third administrations by the high number of *Answer* moves which reach almost 81% of the *Respond* functions in the test. This is indicative of a dominating interactive transaction of question and answer in the first administration, which is markedly reduced in the second and third administrations. Indeed, the figures for the second and third administrations almost parallel each other.

Figure 23: Proportion and average number of Responding features in the three administrations



The least passive of the Responding features, the *React* function was rarely used in the first administration, but gained an increase in proportion after the first administration, and held that gain in the third, and given the shorter average times in the final administration it was used at a higher rate. It seems that students learnt to use this technique in their first year of study, and continued to use it in the following administration. The descriptive statistics can be found in Table 72 of Appendix E.

3.2.3.5 Identifying Developing features

As explained above, the first AS-unit by a different speaker that addresses an Initiating feature was coded as a Responding feature. The speaker who responds has the option to either finish the turn, or continue talking. The AS-units that make up this continuance are classified as Developing. These features were occasionally also used for subsequent moves after other category functions so as to avoid counting them twice. For example, in an initiating turn classified as *Opening Opinion*, a student might use multiple AS-units to explain his or her opinion. To avoid over-counting the number of times a test-taker initiated, only the first AS-unit was classified as *Openning Opinion*, and all following AS-units counted as *Develop* moves.

Although it was not deemed necessary to code the type of expansion as some schemes do (Eggins & Slade, 1997), a few other relevant functions were identified, as seen in Table 37. Within

Develop features, students could make a *Reference* to another speaker in the conversation or outside, and this feature can be considered more advanced because it appeared more often in the second and third administrations.

Table 37: Summary of *Developing features*

Developing Function	Code	Brief Explanation
19. Develop	Dev	The standard coding for the succeeding AS-unit from the initial position *response* was coded as this.
20. Reference	Ref	Sometimes one of the AS-units within an extended turn was used to refer to something that another speaker said, either previously in the GOT or outside.
21. Finish – Summary	Fini-S	Test-takers could finish their extended turn by using an AS-unit to summarize what they said or felt and this acted as a signal to others that the turn was finished.
22. Finish – Yeah	Fini-Y	Another signal that the speaker had finished the turn was by ending on a short phrase, typically 'yeah' or 'yeah, I think so'.
23. Finish – Trail	Fini-T	Instead of finishing definitively, the test-takers sometimes just trailed off the sentence, leaving it unfinished. This is a particularly relevant coding for examining interactivity, since it allowed other speakers to participate by finishing off the sentence for them.

It was also interesting to note the various strategies that students had for signalling to other students when they wanted to complete their turn to speak. If they just finished with a final downward intonation then this was considered a standard, unmarked finish and was not coded. However, some chose to finish by rounding off what they said with a sentence that summarized their position (*Finish – Summary*), others used short words such as 'yeah' with a downward intonation as a marker (*Finish – Yeah*), and finally some speakers just let a sentence trail off, usually followed by a non-word verbal or nod of the head (*Finish – Trail*).

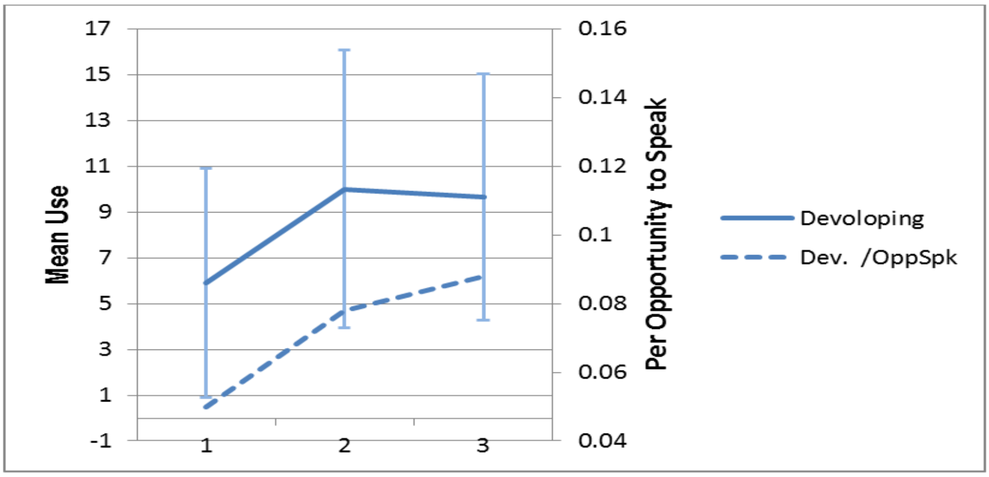
Among the functions included in this analysis, the Developing feature might seem at best indirectly related to interaction, and by extension language learning. Nonetheless, they may play an important role in several ways. Any utterance that is categorized as a Developing feature has the potential to carry information that may be questioned by the other participants, and these may include clarification or confirmation checking questions and other negotiation for meaning features that give rise to modified output (Mackey et al., 2003; Sato, 2007). Also, the ability to speak in extended turns may be beneficial by promoting automaticity (Swain, 2005). The practice of extending responses allows learners to assert control over the language producing processes and as they become routine, it frees up resources that allow them to focus attention on other aspects of the conversational interaction (McLaughlin, 1987; Skehan, 2009).

Finally, specifically related to this coding scheme, the method of ending the long turn by *Finish – Trail*, or trailing off without completing the sentence, is almost an invitation for other test-takers to step in and finish the sentence for the speaker, a co-constructed act that shows the test-takers' support for each other, sharing their meaning and in effect creating a zone of proximal development (ZPD) for language learning (Foster & Ohta, 2005, p. 425). It is through such means that the Developing feature can be related to language learning.

3.2.3.6 Developing features in the GOT

Figure 24 gives the figures for the average and normalized use of Developing features. The same pattern as that seen for Responding features in Figure 21 can be seen, with the average usage peaking in the second administration before a decline in the final one, and the adjusted figures showing successive gains, with the gain to the third administration not as great as the gain to the second

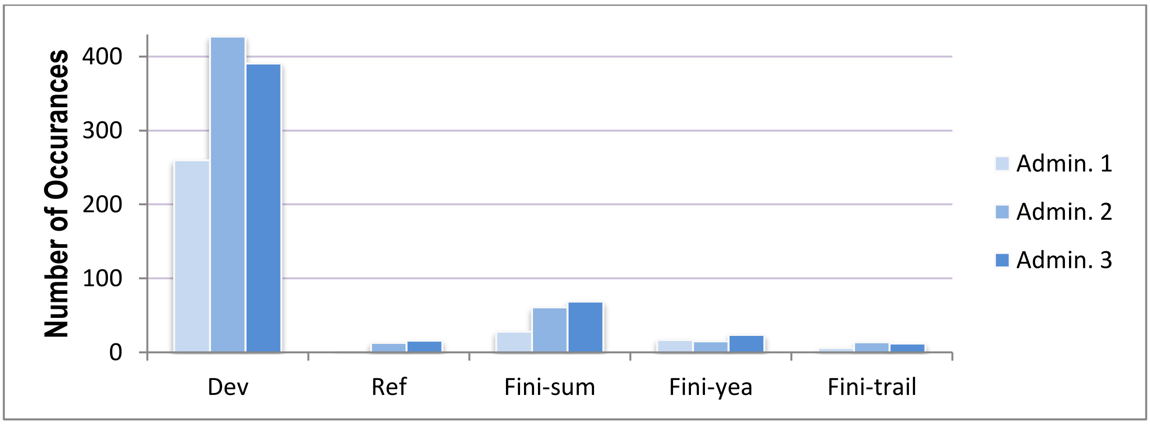
Figure 24: Developing features: mean usage and std dev. compared with normalized



administration. The figures for the standard deviation of the average usage in the graph show that most test-takers could do this to at least a limited extent in the first administration, but in the second and third administrations most test-takers were extending their responses multiple times.

The counts of the Developing features can be found in Figure 25. Since the *Develop* move also represents the quantity of speech, it is not surprising to see the *Develop* move dominate the figures in all three administrations. The investigation by Leaper and Riazi (2014) is relevant here, since one of the findings was that two of the prompts tended to elicit longer turns which would have more Developing features, and as has been pointed out, about 60% of the participants of this study had one of these two prompts, making them over-represented in the sample. Aside from the *Develop* move, the second most used category was for summarizing ends to their moves (*Fini-sum*), which saw increased use in each administration. This may be an indication that with increasing time spent practicing speaking, various elements of language are becoming routinized, enabling them to focus more on organizing their longer turns so that they finish with a summarizing statement. By contrast, Figure 25 shows only a small number of *Fini-Trail,* the function of ending the turn by trailing off. This function can be related to Collaborative features if other speakers collaboratively supply the end of the sentence, and is shown here to be rarely used, and although it rises slightly in the second, the final administration sees it decline again. A possible explanation may be that when it was used it really was because the speaker could not finish the turn, but by the final administration the test-takers had

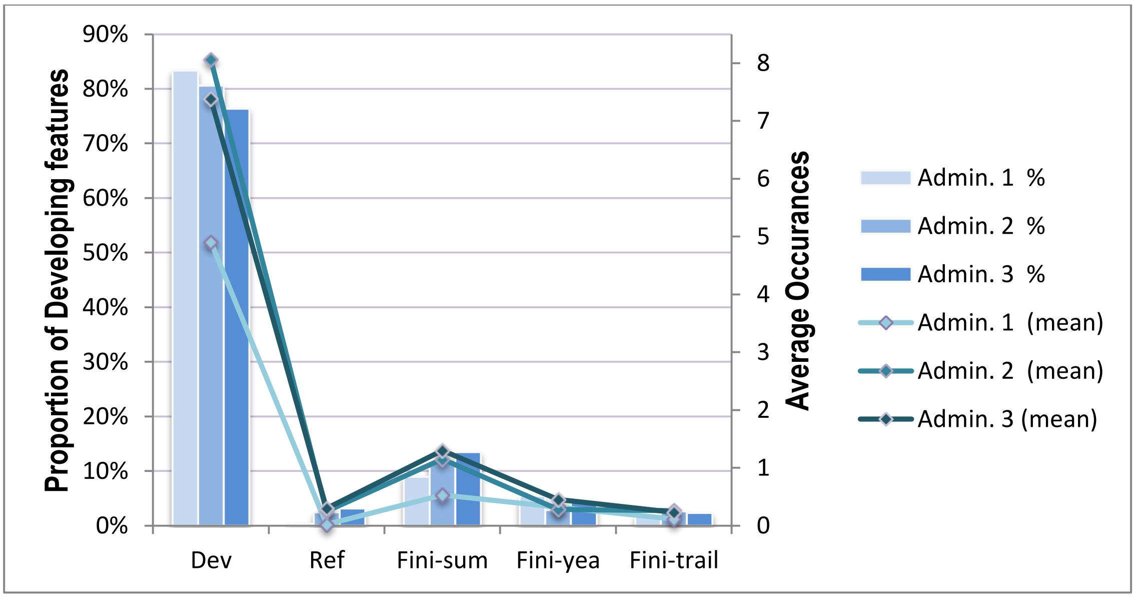
Figure 25: Number of Developing features in the three administrations



advanced to the stage where they could finish their turns in other ways than trailing off, by using a summarizing sentence for example.

The next graph in Figure 26 shows the proportion of all developing moves and average usage of this category of feature. This graph shows that over the course of the three administrations test-takers rely less on *Develop* moves. The increased use of *reference* within these turns can be seen here, as the participants more actively refer to what their discussion partners are saying, perhaps another sign of growing automaticity allowing them to focus more on what their peers have been talking about. Nonetheless, this graph also shows that the number of test-takers who used this technique was low. Amongst the methods of finishing long turns, Figure 26 shows that as well as an increase in the number of *Finish – Summary* moves, this category formed an increasing proportion of uses. For the descriptive statistics, see Table 73 in Appendix E.

Figure 26: Proportion and average number of Developing features in the three administrations



3.2.3.7 Identifying Collaborating features

This category is of particular interest since its functions are distinguished by the speaker paying enhanced attention to what the previous speaker said and interacting with them. As such it encompasses not only moves that question the meaning of what another said (Long, 1981;

McDonough, 2005; Nobuyoshi & Ellis, 1993), but also repair, negotiation of meaning (Ellis, et al., 1994), and co-constructing meaning in ways that have been highlighted in the literature as episodes of language use in conversation that may lead to language learning (Foster & Ohta, 2005) .

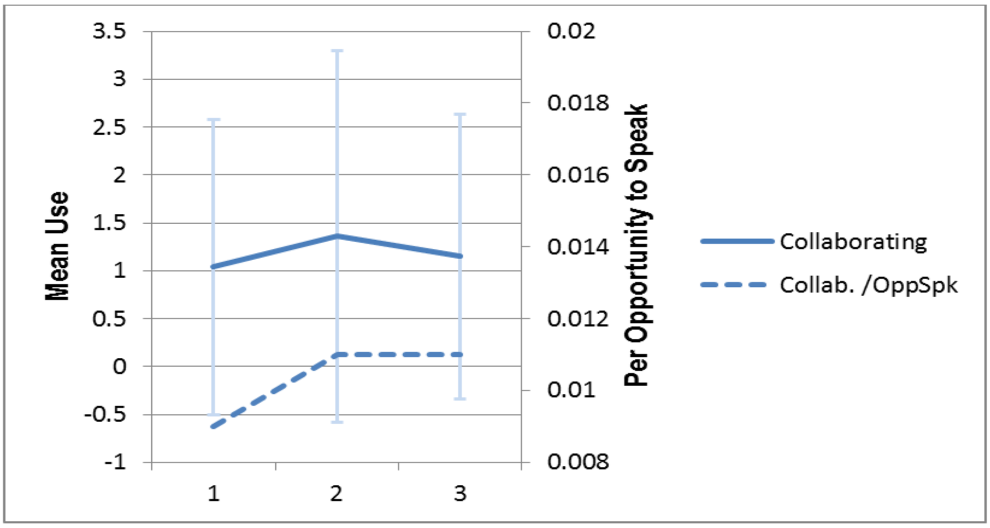
Table 38: Summary of Collaborating features

Collaborating Function	Code	Brief Explanation
24. *Question-clarify*	Q-Cl	A question that aims to clarify something about what the speaker just said.
25. *Question-confirmation*	Q-Con	These aim to confirm meaning and are distinguished by their preferred response being 'yes'.
26. *Correction*	Cor	If one test-taker suggests an alternative to what another said in the belief that what was spoken was a mistake in some way, then it was classified as a correction. This category includes recasts.
27. *Completing Sentences*	Compl	When another participant steps in and provides the next words that the speaker may have wanted to say, it is coded as completion.
28. *Suggest Words*	SW	Even when there is no obvious opportunity like a trailing sentence, sometimes speakers offered words that the speaker may have used.
29. *Incomprehension*	Incom	When a speaker admits to not knowing what was meant, or not knowing what to say, it invites others to collaboratively contribute by supplying an answer or clarifying meaning.
30. *Respond to help*	RespH	If a mistake was pointed out by another, then the recipient of this help could either acquiesce passively, or make the correction or repair, and if they did so it was coded as *Respond to help*.

3.2.3.8 Collaborating functions in the GOT

In Figure 27 a slightly different pattern of usage can be seen to the other functions. Although the actual incidences of usage similarly shows a rise followed by a decline, in this graph the adjusted figures are unlike the other features that showed increasing frequency of usage in the final administration.

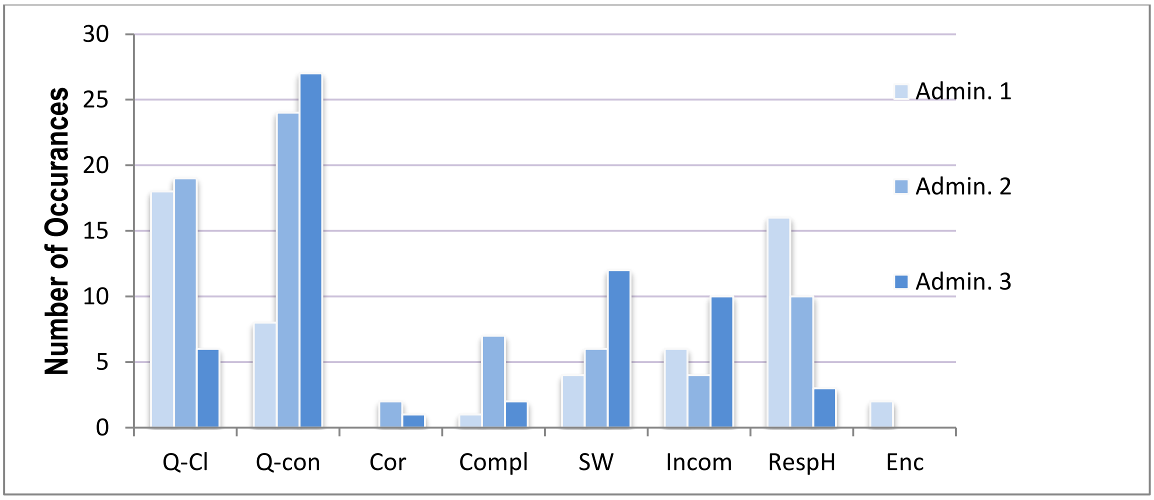
Figure 27: Collaborating features: mean usage and std dev. compared with normalized



For Collaborating features, the rate of change was flat between the second and third administrations, suggesting that the Collaborating feature might depend to a greater extent on the test-takers finding an opportunity to use them, rather than an individual's act of will to use such a function. The range of use confirms the rarity of these functions, as even by the third administration, many test-takers were not using any Collaborating features.

The graph in Figure 28 shows the counts of Collaborating features in each administration. It can be seen that a decline of *Question-clarify* (Q-Cl) corresponds with the rise of *Question-confirmation* (Q-Con).

Figure 28: Number of Collaborating features over the three administrations

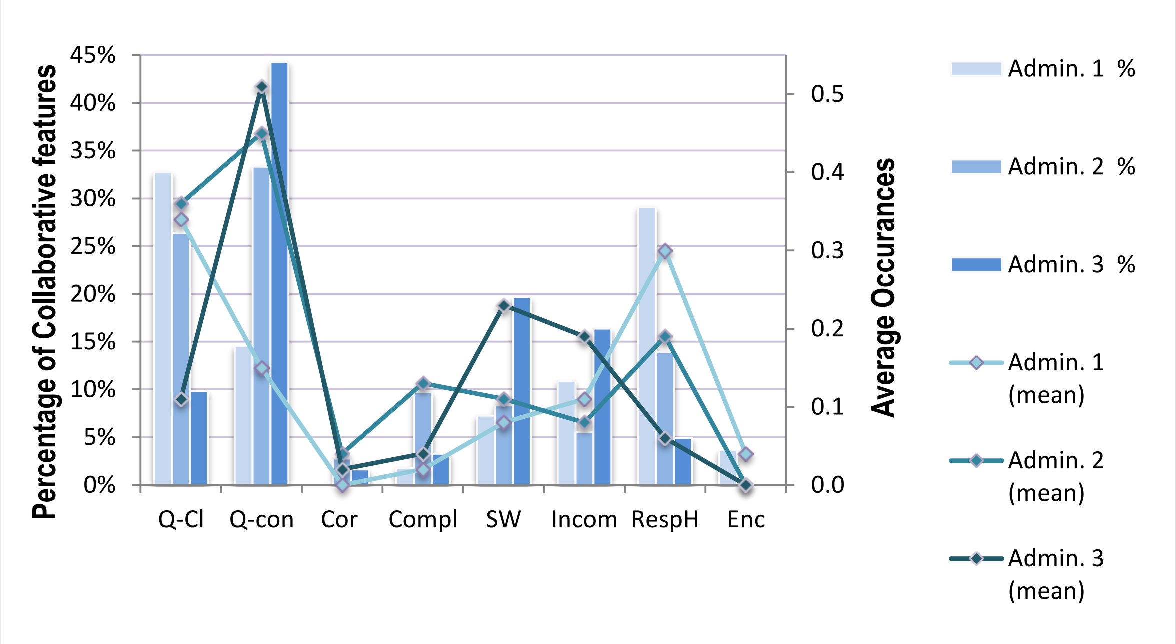


It seems that the collaborating features in the first administration relate to the difficulty of understanding as the test-takers are less likey to be accustomed to communicating in English. In this administration, clarification questions form the largest category, and are acknowledged in the *Respond to help* (RespH) category, which is the second largest. As the test-takers have become more accustomed to speaking in English after a year of instruction by the time of the second administration, it can be seen that confirmation questions overtake clarification questions as the most used feature in this category, though clarification questions are not far behind. By the time of the third administration, students are using clarification questions, with only six examples in the entire administration, and prefer to use confirmation questions instead.

Also rising to prominence in the third administration are *Suggest Words* and *Incomprehension.* The *Suggesting Words* features occurred when participants collaboratively built on what each other say by giving further examples, often in a playful manner. Admitting *incomprehension* might seem to contradict the image a student might want to build in an assessment situation, but it performs the useful communicative function of holding the floor while thinking of something else to say, showing the confidence a participant has to be able to say that he or she does not understand something, or finally, in terms of collaboration it may act as an invitation for other students to participate in the discussion.

Figure 29 confirms the above trends, but also points to the relative scarcity of these functions, since at no time do any of them approach an average of a single usage per test-taker. For the exact figures, the descriptive statistics may be consulted in Table 74 in AppendixE.

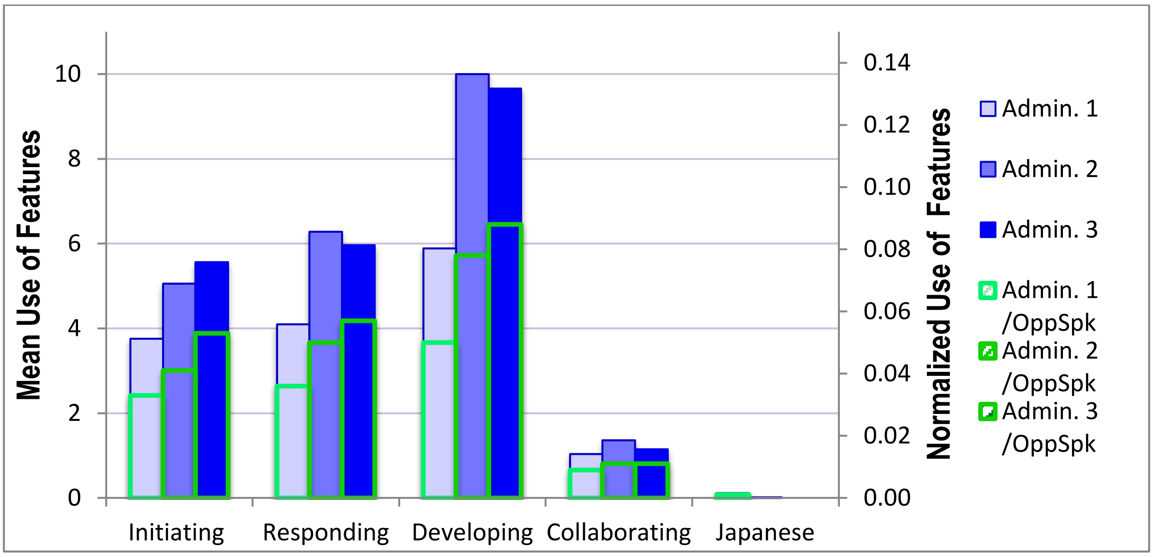
Figure 29: Proportion and average number of Collaborating features in the three administrations



3.2.3.9 Development in Interactive features

The bar graph in Figure 30 summarizes the overall findings of the interactive features by featuring the average and the normalized use of the main categories of functions as well as including figures for the use of Japanese. The blue bars represent the average figures and uses the scale on the left hand axis, and the green patterned bars show the normalized use using the scale on the right. This graph shows that for the categories of features, all except the Collaborating features and Japanese increased in successive administrations after time and number of test-takers per group were taken into account. The use of Japanese was minimal even in the first administration, but tailed off completely in the final administration, showing awareness by the test-takers that in this context English is the mandated medium of communication. As for the Collaborating features, it seems that test-takers could find no more opportunity to use them in the third administration than they could in the second, allowing the postulation of a threshold associated with their use that is related to the opportunity to use them.

Figure 30: Average overall and normalized use of features in the three administrations



The descriptive statistics are summarized in their major categories in columns 1 to 5 of Table 39; columns 6 to 11 show the figures normalized by the time of the test and number of participants. This table gives additional information about the median and range of their scores. It is worth pointing out that while the minimum count in nearly all cases is zero, this could be triggered by at least one of the subcategories not being used in an administration. The Friedman's ANOVA for the normalized features (Table 75 in Appendix E) found that Initiating (χ^2 = 10.106, p < 0.01), Responding (χ^2 = 6.577, p < 0.05), Developing (χ^2 = 21.981, p < 0.001), and the total features (χ^2 = 31.358, p < 0.001) contained significantly different relationships between the administrations, while Collaborating and Japanese features did not.

The results of the Wilcoxon Signed Ranks tests of significance that were conducted to identify the significant relationships can be found in Table 40. The gains the test-takers made in the second from the first administration were significant in the Responding (z = -2.532, p < 0.05, r = 0.246) and Developing categories (z = -3.801, p < 0.001, r = 0.369). However, in the third administration where there were no significant changes in the Responding and Developing categories, the Initiating features improved significantly (z = -2.313, p < 0.05, r = 0.225).

Table 39: Interactive features elicited in the tests by category

Administration 1											
	1	2	3	4	5	6	7	8	9	10	11
	Init.	Resp.	Dev.	Collab.	Jpn	Init./	Resp. ./	Dev. ./	Collab. ./	Jpn./	Total
No.	199	217	312	55	6	OppSpk	OppSpk	OppSpk	OppSpk	OppSpk	
Min	0	0	0	0	0	0.000	0.000	0.000	0.000	0.000	0.028
Median	4	3	5	0	0	0.025	0.033	0.044	0.000	0.000	0.122
Max	21	13	23	7	2	0.117	0.113	0.128	0.062	0.016	0.318
Mean	3.755	4.094	5.887	1.038	0.113	0.033	0.036	0.050	0.009	0.001	0.128
Std Dev	3.741	2.655	4.991	1.544	0.423	0.029	0.021	0.033	0.013	0.003	0.060
Administration 2											
No.	268	333	530	72	1						
Min	0	0	1	0	0	0.000	0.000	0.008	0.000	0.000	0.035
Median	4	5	8	1	0	0.033	0.039	0.066	0.007	0.000	0.146
Max	24	19	31	10	1	0.171	0.155	0.253	0.083	0.007	0.620
Mean	5.057	6.283	10.000	1.358	0.019	0.041	0.050	0.078	0.011	0.000	0.180
Std Dev	4.559	4.285	6.045	1.942	0.137	0.036	0.034	0.045	0.015	0.001	0.102
Administration 3											
No.	295	316	512	61	0						
Min	0	0	0	0	0	0.000	0.000	0.000	0.000	0.000	0.051
Median	4	5	9	1	0	0.037	0.048	0.084	0.009	0.000	0.185
Max	17	13	22	9	0	0.214	0.151	0.208	0.113	0.000	0.579
Mean	5.566	5.962	9.660	1.151	0.000	0.053	0.057	0.088	0.011	0.000	0.209
Std Dev	3.739	3.595	5.378	1.486	0.000	0.041	0.038	0.047	0.017	0.000	0.106

Table 40: Results of Wilcoxon signed ranks test on significant interactive features

		Init	Resp	Dev	Total
Admin 1	Z	-1.690	-2.532	-3.801	-4.046
Vs	p	0.091	0.011*	0.000***	0.000***
Admin 2	*r*	0.164	0.246	0.369	0.393
Admin 2	Z	-2.313	-1.142	-1.695	-2.289
Vs	p	0.021*^{c}	0.253	0.090	0.022*^{c}
Admin 3	*r*	0.225	0.111	0.165	0.222
Admin 1	Z	-2.807	-3.211	-4.449	-4.705
Vs	p	0.005**	0.001**	0.000***	0.000***
Admin 3	*r*	0.273	0.312	0.432	0.457

All figures based on negative ranks
c – falls between Boneffori correction and p = 0.05

3.2.3.10 Summary and discussion of the results for research question 3

In summary, the answer to the third research question is that the GOT can elicit a range of interactive functions that have been shown to be related to language learning, and that test takers develop unevenly in these features over the three times they take this test, as is summarized in Table 41 below. This table suggests that the extent to which they use the categories of interactive functions may be related to their developmental level and are indicative of a hierarchy in which students on the whole develop their ability to respond and develop before they improve their ability to carry out the kind of skills in the Initiating category that show more engagement in the conversation of the GOT. Within the category of Initiating functions, the test-takers gravitate from a reliance on transfer questions towards adding their opinions in the context of the conversation, showing their increasing awareness of the co-constructed nature of conversation as they build on each other's turns (Foster & Ohta, 2005). Along with this is an increasing confidence among some test-takers to raise contrary points of view.

Improvements in the Developing category may be indicative of their increasing automaticity, as their ability to expand quantitatively on their opinion improves in the second administration, before the organization of their turns can be seen to be improving in the third, as the increased number of summarizing features suggests.

Table 41: Summary of findings of Interactive features in the GOT

	Admin 1 vs 2	Admin 2 vs 3	Admin 1 vs 3
Initiating features per OppSpk	-	↑* c	↑**
Responding features per OppSpk	↑*	–	↑**
Developing features per OppSpk	↑***	–	↑***
Collaborating features per OppSpk	–	–	–
Total features per OppSpk	↑***	↑* c	↑***

Key

↑, ↓	- figures increase or decrease significantly
*, **, ***	- significant at the 0.05, 0.01, or 0.001 levels – respectively
c	- falls between Boneffori correction and p = 0.05
–	- no significant difference

Finally, the Collaborating features show some development from a reliance on comprehension checking questions to confirmation questions as they move from the basic problem of understanding what each other have said, towards confirming that they have understood them correctly. However, since the numbers are low it is difficult to draw firm conclusions from the features in this category.

The low level of Collaborating features elicited by the GOT might be seen as problematic since these are the features most directly related to language learning. However, whether this low rate is seen as sufficient for the role of the GOT in the administration needs to be considered in context. It should be pointed out that collaborating may be among the most advanced of the features of conversation (Galaczi, 2010, 2013), and when this is considered along with the low priority most of these students had placed on practicing conversation before entering university, the low rate of usage might not be considered surprising. At the earlier stages it should be recognized that test-takers may have more attentional resources devoted to to delivering and understanding the message than collaboratively interacting. Morevoer, using Collaborative features in an assessment situation may be thought of by test-takers as inherently risky (Luk, 2010), and so taking steps to encourage test-takers to use them more, such as by reforming the scoring system or curriculum, might see increased usage. Finally, the research conducted here shows that the test-takers are developing in language-learning relevant functions in other categories, such as Initiating and Developing features; and this may be thought sufficient from an administrative perspective. The GOTs additional ability to elicit Collaborative features could be considered a bonus. For these reasons, the expectation that Collaborative features be used simply because the format allows for it, as He and Dai (2006) pointed out, should be adjusted.

The results presented here are in broad agreement with other studies that conducted interactional analysis on the GOT, namely Van Moere (2007) and He and Dai (2006) by finding that the main functions in the GOT are asking for opinions and providing answers. However, the ability to examine the test-takers progress over time shows that there is a developmental aspect that their studies could not take into account: Over time the test-takers can be seen to be broadening their repertoire of functions and moving away from the staid question and extended answer pattern of interaction observed by Van Moere and He and Dai, and towards the more collaborative give and take of conversational exchange seen in the higher levels described by Galaczi (2010, 2013).

This section marks the last of the research questions related to the development that test-takers show in the GOT. The first research question used indices related to CAF and vocabulary to track the test-takers' progress. The second research question was about the development in the turn taking system, and the third section created indices for investigating the test-takers' development in interactive functions. The final research question in the quantitative section turns attention from development to assessment as it brings in the scores awarded by the raters and asks to what extent these represented the test takers' development.

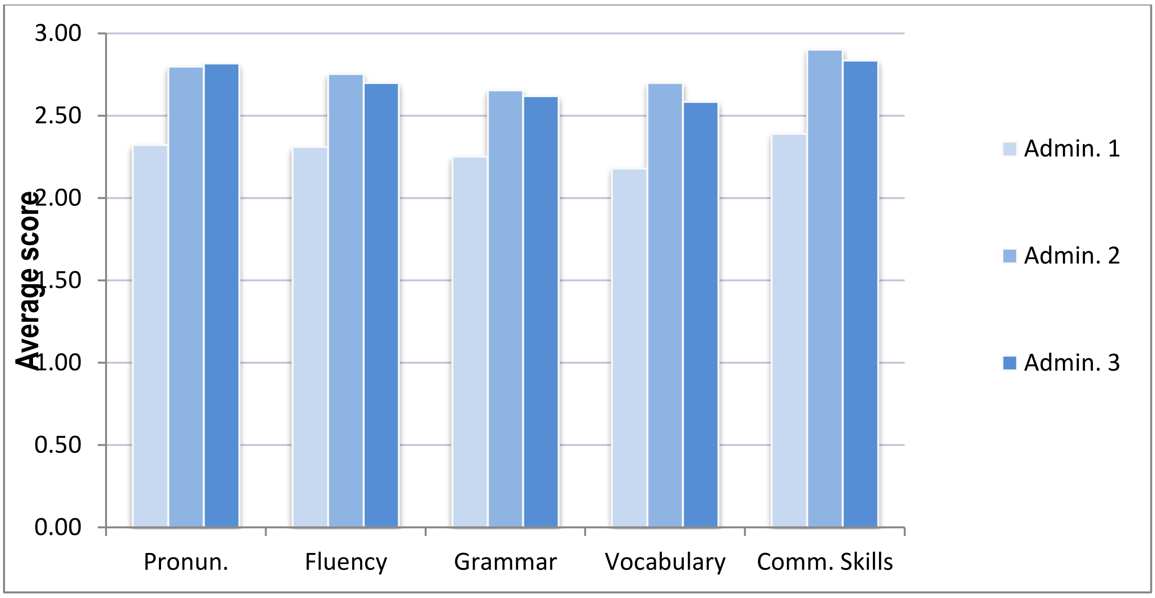
3.2.4 Research Question 4: The performance indices and the scores awarded

The fourth research question brings in the source of data that is provided by the other participants of the GOT, the raters, and ascertains the extent to which their assessment reflects the performance of the test-takers as represented by the indices. The first procedure for answering this question is to analyse the scoring of the administrations, which is covered in the next section. Following that section is a Multiple Regression (MR) analysis that relates the scores to the indices.

3.2.4.1 The scoring of the test

The average scores are displayed graphically in Figure 31, in which it can be seen that in all of the subscales except pronunciation, there is a large increase in the second administration, followed by a small decline in the third administration.

Figure 31: Average scores in the three administrations



The descriptive statistics can be found in Table 42. Here it can be seen that the standard deviation in all subscales except fluency declines in each succeeding administration, which is an indication of the gap between the best and worse students becoming narrower, even if the average and median for most of the scales follow the pattern of rise and decline. These figures were investigated using a one-way Repeated-Measures Analysis of Variance (RM ANOVA), the main results of which can be seen in Table 43 and the correlation matrix of test scores and performance indices can be seen in Tables 76 to 79 of Appendix F. The first row of the RM ANOVA table shows the figures for the test of the sphericity of data, which is one of the assumptions of this statistical test. As can be seen, this was violated in the data for pronunciation (χ^2 (2) = 15.011, p < 0.01), vocabulary (χ^2 (2) = 15.003 , p < 0.01) and the total (χ^2 (2) = 12.009, p < 0.001), and so the degrees of freedom were adjusted using the conservative Greenhouse-Geisser estimates of sphericity (pronunciation: $\varepsilon = 0.797$; vocabulary: $\varepsilon = 0.797$; total: $\varepsilon = 0.827$).

Table 42: Descriptive statistics for the scoring of the test

Administration 1						
	1	2	3	4	5	6
	Pronun.	Fluency	Grammar	Vocabulary	Comm. Skills	**Total**
Min.	0.6	1.2	1	0.9	0.8	**4.7**
Median	2.4	2.2	2.2	2.2	2.4	**11.6**
Max.	3.7	3.5	3.2	3.2	3.7	**17.0**
Mean	2.321	2.309	2.251	2.179	2.389	**11.449**
Std Dev	0.691	0.543	0.530	0.564	0.660	**2.843**
Administration 2						
Min.	1.3	1.3	1.2	1.3	1.1	**6.3**
Median	2.8	2.7	2.6	2.6	2.9	**13.8**
Max.	3.9	3.9	3.5	3.6	3.8	**17.9**
Mean	2.798	2.753	2.653	2.699	2.901	**13.806**
Std Dev	0.500	0.465	0.443	0.415	0.495	**2.159**
Administration 3						
Min.	2.1	1.9	1.8	2.0	1.8	**10.4**
Median	2.8	2.6	2.5	2.6	2.9	**13.3**
Max.	3.7	4.0	3.5	3.4	3.8	**17.7**
Mean	2.817	2.699	2.618	2.583	2.835	**13.551**
Std Dev	0.376	0.504	0.381	0.359	0.478	**1.834**
			N = 53			

Table 43: Figures returned by the RM ANOVA on GOT scores

		Fluency	Pro-nunciation	Grammar	Vocabulary	Comm. Skills	Total
Mauchly's Sphericity df = 2	χ^2	0.824	15.011	3.987	15.003	4.589	12.009
	p	0.662	0.001**	0.136	0.001**	0.101	0.002**
RM ANOVA	df	2	1.594	2	1.594	2	1.653
	F	29.827	26.457	23.753	37.923	23.995	37.195
	p	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
	H_p 2	0.365	0.337	0.314	0.422	0.316	0.417
Admin. 1 vs. Admin. 2	p	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
Admin 2 vs.Admin 3		1.000	1.000	0.483	0.045*	1.000	0.736
Admin 1 vs. Admin 3		0.000***	0.000***	0.000***	0.000***	0.000***	0.000***

Table 43 also indicates that all the scales had significant differences ($p < 0.001$), with medium effect sizes varying between 0.314 and 0.422. The pairwise comparisons shows that the increase in all scores noted in the descriptive statistics in the second over the first administration were significant ($p < 0.001$), and this was maintained between the first and third administrations. However, there was no significant difference between the second and third administrations for all figures except vocabulary, which registered a significant decline in performance ($p < 0.05$).

Although the scores can be seen to track the impressive improvements the test-takers made in the second test, they do not reflect the progress made in the third test, which although was not as rampant as the second time they took the test, resulted in some significant gains in the indices. As noted above, when adjusted for their opportunity to speak, the number of words spoken, AS-units and clauses, the test-takers continued to speak more in the third administration; although their complexity did not improve, their accuracy did both by proportion and in number of error-free clauses. For fluency, although the proportion of pauses was not significantly different, they improved in the rate at which they could talk and reduced the maze words and sounds in their speech. For vocabulary, they increased the percentage of lower frequency words somewhat. Finally the interactive analysis showed there were significant gains in Initiating features in the third administration. It seems then, that there may be a mismatch between the scoring in the third administration and the achievement of the test-takers as shown in the indices.

The next section looks further into this question statistically by using regression to quantify the extent to which the performance in the indices could predict the scores they did receive.

3.2.4.2 Regressions of single factors

The next step in answering this question regards the extent to which the indices represented their scores, and is answered by means of a series of regressions. When using regressions it is important to establish the relationship between the variables, and this can be done by means of viewing scatter plots. The scatter plots showed that the relationship between the variables was "reasonably linear" (Larson-Hall, 2010, p.187) albeit with some outliers. This was to perform regressions on all possible factors singly in order to ascertain their importance, and this would decide which to include in the standard multiple regressions (MR) and the order of their inclusion in the sequential multiple regressions. The results of these initial single regressions can be seen in Table 44 which shows the R^2 values of the indices and the scale in the scoring band they are related to.

Table 44: Results of single regressions

Fluency					
	Admin 1	Admin 2	Admin 3	Average	Order
SRpermin	0.320	0.212	0.320	**0.284**	1
ArtiRate	0.097	0.049	0.133	**0.093**	3
MazeRatio	0.043	0.111	0.001	**0.052**	-
PauseProp	0.019	0.130	0.049	**0.066**	4
MazeSndRatio	0.016	0.162	0.108	**0.095**	2
WordsTurn*	0.151	0.002	0.000	**0.051**	-
TurnsOpp*	0.223	0.083	0.174	**0.160**	5
WordsOpp*	0.504	0.385	0.358	**0.416**	6
Grammar					
EFOpp	0.529	0.362	0.424	**0.438**	1
EFProp	0.122	0.308	0.266	**0.232**	2
WordsClaus	0.068	0.120	0.003	**0.064**	3
WordsASUnit	0.047	0.082	0.004	**0.044**	4
ClauseASUnit	0.003	0.025	0.013	**0.013**	-
WordsTurn*	0.090	0.050	0.001	**0.047**	-
TurnsOpp*	0.294	0.022	0.078	**0.131**	5
WordsOpp*	0.502	0.339	0.330	**0.390**	6
Vocabulary					
WordsOpp	0.504	0.385	0.358	**0.416**	1
TurnsOpp	0.249	0.130	0.124	**0.167**	2
WordsTurn	0.141	0.000	0.001	**0.047**	3
Communicative skills					
InitOpp	0.426	0.250	0.325	**0.334**	1
DevOpp	0.313	0.341	0.177	**0.277**	2
RespOpp	0.031	0.108	0.249	**0.129**	3
CollOpp	0.030	0.090	0.168	**0.096**	4
WordsTurn*	0.131	0.000	0.034	**0.055**	-
TurnsOpp*	0.315	0.098	0.349	**0.254**	5
WordsOpp*	0.605	0.430	0.426	**0.487**	6

*Standard indices

As noted in the methodology section, three indices were common to all of them since they are foundational indices of their performance (normalized words spoken, turns taken and words per turn).

Among the fluency-related factors, the most stable predictor was the speech rate (SRpermin), a measure of the number of all understandable syllables spoken per minute. In the three administrations this accounted for more of the fluency score than the other predictors. The articulation rate (ArtiRate), which differs from the speech rate by excluding unfilled pauses, was also quite stable, but the others varied somewhat over the three administrations, with the ratio of maze words and other sounds such as audible hesitations and stutters, (MazeSndRatio) and the proportion of pauses in the speech (PauseProp), being more important in the second and third administrations, and the ratio of maze words alone (MazeRatio) being more important in the first two. Amongst the three standard indices, words per turn (WordsTurn) was only important in the first administration, but the number of turns taken and words spoken were important in all three. Thus, when their average rank over three administrations was taken into account, the maze ratio and words per turn were eliminated, leaving six predictors for the MR.

For the grammar related independent variables, the two top predictors were the number of error free clauses spoken when normalized (EFOpp), and the proportion of error free clauses (EFProp), which were consistently high over three administrations. Words per clause and words per AS-unit exerted some influence in the first two administrations but barely featured in the third; and the impact of number of clauses per AS-units was minimal across all three administrations. Amongst the standard indices, words per turn featured to a limited degree in the first two, but barely in the third administration, the turns taken featured strongly in the first, but not to the same extent in the last two, and words spoken was strong in all three. As such, clauses per AS-unit and words per turn were not included in the final MR.

For vocabulary, of the three measures, words per turn was important only in the first administration, while words spoken and turns taken were consistently strong performers, but all three were included since there were only three indices for this scoring band.

Finally, for communicative skills, Initiating features and developing features were both consistently strong performers, and responding features and collaborative features both showed the same trend of becoming increasingly relevant to the scores in each successive administration, and so were included in the order of mention. Of the three standard indices, words per turn only featured in the first administration, and so was omitted. The number of turns taken showed a surprising decrease in the middle administration, but was strong in the remaining two, while the number of words spoken was persistently strong across all three administrations.

The interesting point to notice is that with a couple of exceptions, the three standard indices (words per turn, words, and turns) show a general trend of declining in importance to the score of the band over succeeding administrations. This perhaps shows that as the candidates advance in ability these standards account for less of their performance. Presumably this is because their performance becomes more multi-faceted as their skills increase, and raters can take other factors into account in assigning the grade.

3.2.4.3 Multiple Regressions

Before discussing the results it is necessary to explain the extent to which the data met the assumptions required for MR. As noted above, the scatter plots seemed to show that a reasonable degree of linearity between the variables existed. Multicollinearity is the assumption that the variables are not overly inter-related to each other or the dependent variable, and is checked by examining the correlations. It would not be surprising if these variables breach this assumption since they are indeed quite closely related. In the literature, two thresholds are noted: Larson-Hall (2010) gives 0.7 and Field (2008) states that 0.9 is the point at which we should be concerned. While the 0.7 level is breached in most of the 12 MR analyses that were conducted, only once was a correlation higher than 0.9, which is Field (2008)'s recommended level. Collinearity is also assessed through the VIF (variance inflation figure) and tolerance columns in the SPSS readout. According to Field, if the largest value of VIF is over ten there is cause for concern, and if the average is well over one, the regression may be biased, and for Tolerance, if it is lower than 0.1 it can be considered a serious problem. While the MRs for vocabulary and communicative skills were always within the limits, those for fluency and grammar consistently breached them. For fluency, in all three administrations the VIF and Tolerance levels were exceeded for speech rate, articulation rate and pause proportion. This is not surprising given that the method of calculating these three figures involved some of the same core indices. For grammar, again in all three administrations, it was error free clauses and words spoken. Again, when it is considered that those who spoke more were probably more likely to be better at speaking and perhaps less likely to make grammatical mistakes, perhaps this need not be surprising.

There is not much more that can be done about these inter-related predictors than to acknowledge how multicollinearity may negatively impact on the findings of the MR. First, it can be expected that the size of R will be limited if they account for similar factors. Secondly, it makes it more difficult to estimate the value of the different predictors. Finally, it means that the estimate of *b*-value will lose precision, and will result in instability in the measurement of the regression coefficient (Field, 2008, p.175).

Another assumption is the independence of the errors of the MR, and this is indicated by the Durbin-Watson statistic that should be between 1 and 3. This assumption is met since this statistic varied between 1.464 and 2.134 for the data. The assumption of normality, which is indicated by the distribution of residuals in a P-P plot and by checking for outliers, was also checked. The distribution of the P-P plots showed some minor deviations that could not be considered severe enough to justify action being taken. Outliers were checked by examining the variance of the standard residuals, which in all cases was between -3 and +3, which is noted by Larson-Hall (2010) as being the thresholds. The final two diagnostics for influential outliers are Cook's distance and the Mahalanobis distance. Cook's distance should not be above one, and this was exceeded twice in the MRs conducted. For the Mahalanobis distance, the critical value was determined on a χ^2 table by reading off the number of independent variables, using an alpha level of 0.001 (Pallant, 2011). For the fluency, grammar and communicative skills MRs this was determined to be a value of 22.46, and for Vocabulary, 16.27. Nearly all MRs had at least one test-taker that could be considered an outlier, and one had four. The scores of those who were indicated to be outliers according to Cooks and Mahalanobis distances were eliminated and the MRs run again until these measures came within acceptable thresholds.

The subsections below are organized according to the scoring band that the indices are related to so that the results for fluency, grammar, vocabulary and communication skills will be explained in turn. The explanations will follow the same pattern: they will start by summarizing the trends noted in the correlation tables (placed in Appendix F, Tables 76 to 79 due to their size), followed by the results of the standard and sequential regressions.

3.2.4.4 Development of fluency and their scores

For a general overview of the results, the correlations among the fluency predictors in Table 76 of Appendix F will be described. In the table, both positive and negative relationships can be found. It seems the candidates who achieved higher scores in the first administration did so by having more frequent turns, with less hesitation, maze words and fewer non-word sounds. This is shown by the following relationships: the higher they scored in fluency, the more words they spoke ($r = 0.704$, $p < 0.001$) and the higher their speech ($r = 0.696$, $p < 0.001$) and articulation rates ($r = 0.306$, $p < 0.05$); and the lower their maze and sound ratio rates ($r = -0.387$, $p < 0.01$), proportion of pauses rates ($r = -0.455$, $p < 0.001$) and time taken between turns rates ($r = -0.472$, $p < 0.001$). The only exception in the pattern of negative and positive relationships comes in the correlation of words per opportunity and the maze and sound ratio. In the first administration it is a significant negative relationship ($r = -0.366$, $p < 0.01$), in the second it is non-significant ($r = -0.230$, $p = 0.054$), but still negative, but by the third it has become positive, though non-significant ($r = 0.037$, $p = 0.399$). This perhaps shows a general trend of students using fewer maze words and reducing the number of voiced fillers in their speech over the time of the test, as was noted in the section answering research question one.

The two most consistent figures over the three administrations are for speech rate and the score. In the first and the last administrations, the score of fluency is significantly related to all variables, but in the middle it did not correlate significantly with the maze and sound ratio or articulation rate, perhaps showing that something was different in the middle administration to break this trend, possibly the influence of the prompts in that administration (Leaper & Riazi, 2014). The rate of speaking showed a slightly more consistent trend over the three administrations, with it being significantly correlated with all other predictors in the first two administrations, and with all but one in the final administration, which was the maze and sound ratio. This may reflect the significant reduction in the number of maze words spoken between the second and the third administrations. The maze and sound index became less related to the other indices by the third administration, which is consistent with the finding that the Maze index was significantly reduced in the third from the second administration. Other indices do not show a clear enough trend in the correlations to comment on.

The standard MR (Table 45) showed that in all three administrations, when all six predictors were included in the model, the only significant one was the standard measure of number of words spoken in all three administrations. However, the sequential MR (Table 80 in Appendix F) showed that the speech rate could be a significant factor in models in which fewer predictors were used. In the first administration, the speech rate was significant in all models except the final one in which the number of words was included, and in the second and third administrations, it was significant in the first three models, but not in the last three.

The only other predictor to play a role was the number of turns, which was a significant factor only when the number of words was not included. This seems to suggest that when considering fluency, the rate of speech was more important to the scorers when judging fluency at lower levels, but perhaps other factors impress the raters as students improved their ability in the group oral.

The other noteworthy feature of Table 45 is that the R2 statistics show a substantially reduced in the middle administration. A possible explanation of this could be the role of the prompt, which was found by Leaper and Riazi (2014) to have a significant impact on the nature of the intereactions in this administration. One of the prompts resulted in significantly less fluent responses, most likely because it required the test-takers to reflect on personal issues related to their future.

3.2.4.5 Development of grammar and their scores

In the correlation tables for grammar scores, the statistics show a few consistencies. In the first administration, all the indices are significantly related to the score; in the second all but one is (frequency of turns taken), and in the third only the average length in words per clause and AS-unit is not significantly related. In the first test, higher scores were positively related to using more error free clauses ($r = 0.714$, $p < 0.001$) , a higher proportion of error free clauses ($r = 0.300$, $p = 0.05$), more words per clause ($r = 0.249$, $p = 0.05$) or AS-unit ($r = 0.\ 232$, $p = 0.05$) and by having less time per turn ($r = 0.513$, $p < 0.001$) and more words per minute ($r = 0.694$, $p < 0.001$). This was true also of the second administration, except that the frequency of turns taken was no longer significant. In the third administration, the frequency of turns became significant, but the length of the clause and AS-unit in terms of words lost significance.

Over the three administrations, the correlation figures for number of error free clauses is stable, always being significantly related to the error free proportion, a lesser amount of time between turns, and a greater number of words, and never being related to the length in words of the clause or AS-unit. It is interesting to note that these two complexity measures are unrelated: as their accuracy improves, their complexity of speech as measured by length in words may fluctuate between being negatively or positively related, but not enough to be significantly different.

The measures for length in words of AS-units and Clauses are consistent in administrations 2 and 3, showing that in these two administrations those who had more words per AS-unit and Clause did so by speaking more in fewer turns. The standard measure of words spoken is significantly related to all measures but the proportion of errors in the first two administrations, and by the third administration this is significantly related along with the rest of the indices. This suggests that as candidates spoke more they were awarded with a higher score, used more error free clauses, had a higher proportion of error free clauses (except for the first two administrations), had longer clauses and AS-units, and more frequent turns in the time of the test. The standard multiple regressions (Table 46) show the decreasing importance of the indices over the three administrations. It can be seen in

Table 45: Results of the standard MR for fluency

Admin	R2		Constant	SRpermin *B*	MazeSndRatio *B*	ArtiRate *B*	PauseProp *B*	TurnsOpp* *B*	WordsOpp* *B*
		sr^2		0.149	-0.106	-0.104	0.099	0.071	0.286
1	0.613	*B*	-1.372	0.042	-0.866	-0.928	4.486	0.008	0.802**
			(-5.58,2.83)	{-0.01,0.10)	(-2.50,0.76)	(-2.70,0.84)	(-4.53,13.50)	(-0.01,0.03)	(0.25,1.36)
		sr^2		-0.038	0.074	0.053	-0.061	-0.085	0.408
2	0.397	*B*	2.752*	-0.008	0.471	0.444	-2.004	-0.006	0.386**
			(0.02,5.48)	(-0.06,0.04)	(-1.04,1.98)	(-1.56,2.45)	(-9.85,5.84)	(-0.02,0.01)	(0.16.0.61)
		sr^2		0.029	-0.182	-0.009	0.005	-0.057	0.372
3	0.531	*B*	2.058	0.007	-1.401	-0.077	0.177	-0.003	0.294**
			(-1.00,5.12)	(-0.04,0.06)	(-3.02,0.22)	(-1.87,1.72)	(-7.90,8.26)	(-0.01,0.01)	(0.13,0.46)

Table 46: Results of the standard MR for grammar

Admin	R2		Constant	EFOpp	EFProp	WordsClaus	Words/ASUnit	TurnsOpp	WordsOpp
		sr^2		0.221	-0.038	0.235	-0.046	-0.014	-0.052
1	0.629	*B*	0.626	6.817*	-0.224	0.320*	-0.024	-0.002	-0.296
			(-0.30,1.56)	(1.19,12.45)	(-1.29,0.84)	(0.07,0.57)	(-0.12,0.07)	(-0.02,0.02)	(-1.33,0.74)
		sr^2		0.138	0.049	0.228	0.094	-0.116	-0.086
2	0.599	*B*	0.581	4.439	0.469	0.314*	0.048	-0.009	0.429
			(-0.30,1.46)	(-1.83,10.71)	(-1.39,2.32)	(0.05,0.58)	(-0.05,0.15)	(-0.02,0.01)	(-1.40,0.54)
		sr^2		0.080	0.128	0.089	-0.151	0.140	0.025
3	0.508	*B*	1.388**	1.572	0.673	0.121	-0.067	0.008	0.091
			(0.60,2.18)	(-2.50,5.64)	(-0.42,1.77)	(-0.16,0.40)	(-0.16,0.03)	(-0.00,0.02)	(-0.66,0.85)

Table 46 that in the first administration, both the number of error free clauses and number of words per clause are significant; in the second the latter is still significant, and in the third administration no single index is significant. The sequential multiple regression (Table 81 in Appendix F) shows clearly that the number of error free clauses is the most important index over the three administrations, followed by the number of words per clause, and in the second and third administrations, the proportion of error free clauses is also important. The number of error free clauses is significant in all administrations except for the final model in the second and third administrations, while the number of words per clause is significant in all models of the first and second administrations, but not at all in the third one.

3.2.4.6 Development of vocabulary and their scores

Due to the insignificant results from the indices of lexical diversity (see Section 3.2.1.5) the predictors for vocabulary are the least direct of the measures since they in no way take the quality of the words into account, which according to the scoring bands should be an important determiner of their vocabulary score (Appendix I, p. 426). Nonetheless, it can be seen that the number of words spoken correlates significantly to their vocabulary score in all three administrations. The negative correlation found in the turns taken per opportunity show that those who had more frequent turns also tended to score more highly in vocabulary in all three administrations. The length of the turn in words was positively correlated to all other indices in the first administration, but only to the number of turns in the remaining two administrations, which probably reflects that in the lower performances in the first administration, extended turns were a good deal less common, and when they students did they were rewarded for it. The frequency of turns taken was significantly related to the other indices in all three administrations, showing that those who had more frequent turns scored more highly and spoke more words, but also had turns with fewer words. The standard MR in

Table 47 showed that none of these measures contributed significantly in the first administration. In the second and third administrations only the number of words was significant.

Table 47: Results of the standard MR for vocabulary

			Vocabulary			
Ad-min	R2		Constant	WordsOpp	TurnsOpp	Words/turn
1	0.376	sr^2		0.164	-0.029	0.037
		B	1.578**	0.817	-0.005	0.009
			(0.63,2.53)	(-0.36,1.20)	(-0.05,0.04)	(-0.05,0.07)
2	0.401	sr^2		0.276	-0.100	0.011
		B	2.460**	0.343*	-0.011	0.001
			(2.07,2.85)	(0.06,0.63)	(-0.04,0.01)	(-0.02,0.03)
3	0.391	sr2		0.313	-0.013	-0.042
		B	2.127***	0.408**	-0.002	0.004
			(1.67,2.58)	(0.11,0.71)	(-0.03,0.03)	(-0.02,0.02)

The sequential MR (Table 82 in Appendix F) confirmed the importance of the number of words spoken by being the only measure that reaches significance in all models.

3.2.4.7 Development of communicative skills and their scores

In the table for correlations for communicative skills (Table 79 in Appendix F), it can be seen that in the first two administrations the score in this scale was significantly correlated to Initiating and Developing features, and the standard indices of number of turns and words, but in the final administration all the indices were related significantly to the score. The most consistent predictor was the index of Initiating features, which was significantly related to all other features in the administrations, except for Responding features in the first administration. Developing features were always related significantly with Initiating features, never with the other two communicative skills indices, and inconsistently related to the two standard indices. Responding and Collaborating features are consistently related to all other indices except developing features in the final two administrations, but not related significantly to the other communicative features at all in the first administration.

The standard MR in Table 48 shows that only the Initiating feature was significant in the first administration, apart from the standard words spoken, which was significant in the first and third tests. The sequential MR (Table 83 in Appendix F) affirms that Initiating features accounted significantly for the score in communicative skills all models for first administration, and shows that in models 2-5 developing features was also significant, though this was subsumed by the number of words spoken. In the second administration, the Initiating features are only significant in the first model, but once developing features are added they are significant until the final administration, where no index is significant. In the final administration, Initiating features are significant until the final model, in which the number of words spoken are introduced. Responding features are significant when they appear in the third model, and this significance continues in the fourth model, but disappears with the introduction of the first of the standard indices, the frequency of turns.

Table 48: Results of the MR for communicative skills

			Communication Skills						
Ad-min	R2		Constant	Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
1	0.644	sr^2		0.187	-0.035	0.126	-0.047	0.072	0.294
		B	0.906	7.657*	-1.204	6.342	-2.651	0.013	1.171**
			(-0.22,2.03)	(0.33,14.98)	(-7.35,4.94)	(-2.67,15.36)	(-12.71,7.40)	(-0.02,0.05)	(0.46,1.88)
2	0.443	sr^2		-0.019	0.017	-0.042	-0.084	-0.115	0.218
		B	2.467***	-0.480	0.522	-0.922	-4.746	-0.010	0.582
			(1.87,3.06)	(-6.51,5.55)	(-6.60,7.65)	(-6.05,4.21)	(-17.85,8.04)	(-0.03,0.01)	(-0.04,1.20)
3	0.576	sr^2		0.050	-0.175	0.051	0.049	-0.154	0.298
		B	2.330***	1.130	-4.087	1.007	2.752	-0.012	0.680**
			(1.84,2.83)	(-3.37,5.63)	(-8.71,0.54)	(-2.87,4.89)	(-8.36,13.86)	(-0.03,0.00)	(0.23,1.13)

Similar to the fluency figures in Table 45, Table 48 shows a substantially depressed R^2 statistic for the middle administration. A possible explanation may also be related to the particular prompts in this administration. Research by Leaper and Riazi (2014) showed that two of the prompts elicited significantly longer turns than the other prompts, which may have affected the ability of the test-takers to use techniques that counted in this category, and this may have been the obfusticating factor that reduced the R^2 statistic in this administration.

3.2.4.8 Summary and discussion of findings from the regression analysis

The first aspect that was investigated in this section were the scores, in order to understand the pattern of development that the raters saw by the test-takers. The results of the RM ANOVA, as summarized in Table 49 below, showed a familiar pattern of significant gains being made in the second administration but not different between the second and third administrations, with the significant decline in the vocabulary score being the only exception. This pattern may seem familiar because it follows the same pattern seen in the unadjusted indices for such measures as number of words, AS-units, turns and words per turn, and suggests that for some reason the raters are missing the more subtle gains that the test-takers showed in the third administration.

When examining the results of the regression analysis, it can be said that overall, there are a lot of inconsistencies from administration to administration. This may be due to different features of a learner's language being more salient at different stages of development. For example, in the first administration, the test-takers tended to speak much less than in later administrations, giving themselves very little opportunity to score in the communicative skills band, and thus, initiating a conversation might have appeared more prominent than in later administrations.

Table 49: Summary of findings in the scores awarded

	Admin 1 vs 2	Admin 2 vs 3	Admin 1 vs 3
Fluency	↑***	-	↑***
Pronunciation	↑***	-	↑***
Vocabulary	↑***	↓*	↑***
Grammar	↑***	-	↑***
Communicative skills	↑***	-	↑***
Total	↑***	-	↑***

Key

↑, ↓	- figures increase or decrease significantly
*, **, ***	- significant at the 0.05, 0.01, or 0.001 levels respectively
–	- no significant difference

Another feature to note is the role of the number of words spoken, which is a significant factor in most administrations in the fluency, vocabulary and communicative skills bands, but not at all in the scoring of grammar. It seems accuracy in terms of the number of error free clauses was more important in this regard, though in the first two administrations the complexity measure of length of clauses played a significant role in some of the models. The other standard index was the number of turns, but it seems that this had a very limited role to play, with it featuring only in one model of the fluency sequential MR in the second administration.

The R^2 figures also show some interesting trends. Firstly, for all indices for which there were specific indices, for fluency, grammar and communicative skills, the R^2 is highest in the first administration, and lower in succeeding administrations. This indicates that the indices account for more of their scores when the speakers are usually the least proficient: an inference that may be drawn from this is that raters may have found the lower level learners easier to score according to the core values of the scales. It may well be the case that as the test-takers' skills improve the raters find themselves taking a greater variety of other aspects of their performance into account when awarding the scores. Or perhaps, with the test-takers' increasing ability raters have to process an increasing amount of information in order to give a grade, and it results in the decreasing ability of the indices to predict the scores.

The second point was noted above for the fluency and communicative skills R^2 figures (Tables 45 and 48 respectively): the lowest figure by a considerable margin was for the middle administration. A possible explanation for this was the particular set of prompts in use for that year, which was investigated by Leaper and Riazi (2014). Their findings revealed that the prompts elicited significantly different patterns of interaction: one of the prompts elicited significantly less fluent responses, and two of the prompts elicited significantly longer turns, and this may well account for the depressed ability of the indices to account for the scores for fluency and communicative skills. Grammar was not affected in the same way, with the R2 showing declines in each succeeding administration; the largest drop being between the second and third adminsitrations. This may have been because there was no significant difference in accuracy between the prompts (Leaper & Riazi, 2014). Although complexity was found to be significantly affected by the prompt, it may well be the case that accuracy is more pertinent to raters than complexity, as was suggested by Leaper and Riazi.

The R^2 figure for vocabulary is the most consistent over the three administrations, but since the indices for these relate to the quantity of words, turns and words per turn, none of which is related to the quality of the words spoken, it is not surprising that this figures is so consistent over the three administraitons.

Overall though, the answer to research question four is that the indices are related to the scores to a limited extent. The limited extent being that the test takers' scores follow the pattern of unadjusted indices and limited ability of the indices to predict the scores assigned to them. One question that may be raised is, if raters are not paying much attention to the progress students make as represented in the indices, what are they attending to when they assign grades? Given the literature on this subject, the limited relationship may not be surprising. Raters of peer interaction tests have been found to incorporate aspects not on the scales (May, 2011; Orr, 2002) and to incorporate such aspects as physical appearance and comparisons to other members within a group (Orr, 2002).

This last point may be relevant to the foundational theory that this dissertation subscribes to. Interactional Competence declares that the performance is locally contingent and co-constructed (Young, 2009, 2011) and the suggestion from Orr (2001) that raters make comparisons to other test-takers in peer assessment seems entirely consistent with this. Despite being trained to only compare the performance they see in the GOT to the scoring bands, they may inevitably be affected by comparisons to their peers, and this would weaken the link between their actual performance as measured by the indices, and the scores that are awarded by humans, as may well be the case here. Whether this is in fact the case, will become more evident when the scores and indices are compared to a qualitative analysis of the test takers' performance, as will be seen in Chapter 4.

3.3 Summary of the quantitative results

The quantitative phase of the study and the results of the data analysis provided answers to the first four research questions. Significant differences have been found in the indices that represent improvements in the complexity, accuracy, and fluency recorded by the 53 test-takers in the GOT over the three times they took the test. The figures reveal that different aspects improved at different rates, showing the value of having multiple indices for each component of complexity, accuracy, and fluency (CAF) in this kind of study. The development of vocabulary in terms of lexical diversity, however, was hampered by inconsistencies in the sample, particularly in terms of the number of test-takers who reached the minimum number of words necessary to produce stable measures. Although the test-takers showed their improvement by speaking more words in each succeeding administration of the test, the results of the indices of lexical diversity failed to give any indication of development, and this was probably due to the nature of the context as much as the kind of data that was collected. When vocabulary was investigated at the level of the cohort, development could be seen in the second administration, but in the third was more subtle. The results here raise questions about the extent to which the scoring of vocabulary should be incorporated into the GOT, given the communicative context.

The second research question investigated the patterns of turns in the GOT and found that the development shown by the length of the turn is quite complex. In the second administration the test-takers did significantly increase the length of their turns, and this may have been influenced by the prompt (Leaper & Riazi, 2014), but further development does not necessarily lead to greater turn lengths, but rather a greater number of shorter turns. The result of an increased number of final summary sentences in the final administration (see Section 3.2.3.6) suggests that test-takers organized their turns better rather than made them ever longer.

The interactive features of conversation found in the GOT were the subject of the third research question; and to answer it a coding system was devised. As for the CAF analysis, this found that different interactive features develop at different rates, with the ability to respond and develop turns increasing significantly before the ability to initiate interactions, though the number of Collaborating functions did not increase significantly.

The final question regarded the extent the indices could predict the scores given by the raters to the performances. Although some of the key indices play a role in explaining the scores given, clearly the raters are affected by some other factors when awarding scores. The next chapter will examine the extent to which the scores are represented by the performance in more depth by means of a qualitative investigation.

CHAPTER FOUR

The Qualitative Phase

4.0 Introduction

Thus far in conducting this mixed methods research project, the quantitative data analysis in the preceding chapter has provided a statistical account of the test-takers' performance in the GOT. The quantitative findings illuminated some developmental patterns among the students, and in this section the qualitative phase of the study will be presented to complement the quantitative one. The qualitative analysis will focus on a sample of representative test-takers from the first part of the analysis in order to shed further light on the construct of the GOT. Before continuing it needs to be made clear that this section itself does not answer a research question; rather, it provides the complementary qualitative analysis that will allow the final research questions to be answered in Chapter 5, in which the qualitative findings in this chapter will be synthesized with the findings from the quantitative analysis presented in the previous chapter.

This chapter commences with a description of the methodology and continues by describing how the eight participants in this section were selected, how they were divided into three groups, and finally it presents the results of the qualitative analysis.

4.1 Methodology

4.1.1 Analytical framework

This section starts with a description of the analytical framework used in the qualitative analysis. This is a description of how the framework outlined by Young (2009, 2011) in his explanation of interactional competence (IC) will be used. This framework was intended by Young as an outline of the various elements that can be used "to construct a practice" of IC (Young, 2009, p. 217). For the convenience of readers, Table 50, which summarizes the resources of Young's framework, is reproduced from Section 2.1. These resources can be considered as a general list, and as such need to be tailored to the specific context in which they can be applied. Indeed, the examples that Young cites as examples of IC analysis compatible with this list of resources do not cover every aspect of them.

Table 50: Resources through which participants' interaction may be analysed

1. Identity resources	Participation framework	Identities of all participants, whether officially there or not, their role in the interaction, the identity they construct as participants
2. Linguistic resources	Register	Use of pronunciation, vocabulary, grammar that characterizes the practice
	Modes of meaning	Interpersonal, experiential and textual meanings participants construct for the practice
3. Interactional resources	Speech acts	The sequential organization of acts selected for the practice
	Turn-taking	The allocation of turns in an interaction, including who has a right to claim them, how they know a turn has ended
	Repair	How interactional difficulties are dealt with once they arise
	Boundaries	The means by which the various parts of the discourse are marked by the participants.

Young (2011)

Similarly, for this dissertation it is also necessary to explicate the resources before assessing which elements are relevant to the specific circumstances of this study. To this end, the sub-sections that follow correspond to the three resources on Young's (2009, 2011) list: identity, linguistic, and interactional. In these sections, I will review the explanation as provided by Young and the available methodologies to determine which elements are useful and how they will be approached in this phase of the dissertation. This will create a framework for analysing the participant's performance in the three administrations of GOTs they took part in.

4.1.1.1 Identity resources

These are described in Young (2011) as "the identities of all participants in an interaction, present or not, official or unofficial, ratified or unratified, and their footing or identities in the interaction" (p. 429). Identity can be seen on two levels. The first level is that of the overt roles of the individuals and agencies who are involved in an administration of the GOT, starting with the immediate and moving to the more distant. Foremost amongst these actors, and the focus of this dissertation, is the identity of the participants as students who are taking this test. Their role as test-taker is reinforced by the context of the situation. For the test the students received instructions to come to a particular room at a time and place, they were asked to write their identity numbers and names on the board and sit around some tables, were given a prompt and told that they have one minute to read and think about the prompt before the test begins. After the beginning of the test they collaboratively construct the discussion which became available for analysis in this dissertation.

The other participants in the immediate context are the raters. Their presence is only noted on the few occasions where they need to prompt a test-taker to say more because there was not enough language elicited for them to give a rating. Nonetheless, by the time the participants are ready to start talking for the test, the raters have asserted themselves by giving the directions, explaining what the students must do, and although they are mostly silent for the duration of the discussion, they must interrupt to end the test. Their very presence indicates a formality that is likely to be one of the major causes of 'exam nerves' that test-takers may be prone to in greater or lesser respects.

Another unseen audience in the immediate surroundings but would be noted by the test-takers is inferred by the physical presence of a microphone and video camera. There is nothing to suggest that the presence of these items makes any difference to the interactions – there were neither complaints from the test-takers nor reports from the raters about noticing anything different between the GOTs that took place in the video room and the vast majority that did not. Although peer interaction oral tests have been found to be less intimidating than other formats (Fulcher, 1996a), 'test day nerves' are a human condition. Since the additional presence of video cameras and microphones has passed uncommented on, it can be speculated that the nerve-inducing impact of taking part in a speaking test by itself overwhelms any additional impact of being videoed. Regardless, the presence of the camera and microphone must be acknowledged.

Further removed from the immediate environment where the GOT takes place, other participants exert some kind of influence over the interaction that was elicited. One of these is the test committee, whose members wrote the prompts. This aspect was investigated by Leaper and Riazi (2014), whose participants overlap to a substantial degree with the test-takers of this dissertation. The need to include such participants was pointed out by Norton (2013) who maintained that "the voices of test writers, whose personal interests, professional experience, and sociocultural background influence and inform the… tasks included in speaking tests, are embedded in the 'text' of the tests" (p. 313). However, it is possible go beyond Norton (2013) to those who decided on the type of test and parameters such as how many should take part in a GOT, the amount of time the test-takers have to consider the prompt and so on. This class of 'non-appearing but influential participants' can be extended further back to those who created the assessment system at the institute, and society itself from which comes the demand for graduates who are communicatively competent in English. These have been referred to in the introduction of this dissertation and it is not necessary to dwell further on them (and the other identities mentioned above) since it is the previously mentioned second level of identity that is more relevant to this dissertation.

The second level of identity is the more abstract sense that Young refers to as "their footing or identities in the interaction" (2011, p. 429). This is the identity that the participants create by their participation in the discussion. It is constructed through their interaction with the other participants, and may be evident from their particular lexical choices and actions such as turn-taking, topic initiation, body language and so on. This notion of identity has been implemented in several studies. For Yagi (2007), identity is encompassed within the participant's ability to "understand and express the roles and participation structures/frameworks" of "a given situated practice" (p. 21). The participation frameworks that are referred to here are from Goffman who describes it as the relationship of the hearers to the utterance:

> If one starts with a particular individual in the act of speaking – a cross-sectional instantaneous view – one can describe the role or function of all the several members of the encompassing social gathering from this point of reference (whether they are ratified participants of the talk or not)… The relation of any one member to this utterance can be called his 'participation status' relative to it, and that of all the persons in the gathering the 'participation framework' for that moment of speech. (Goffman, 1981, p. 137)

The participation status and identities of those taking part are revealed by their talk and this constitutes the participation framework.

Goffman's work (1959, 1981) on the social psychological notion of 'impression management' was drawn upon by Luk (2010), who as noted in the literature review (Section 2.2.5.4) authored one of the two papers that have thus far examined identity in peer interaction speaking tests. In this conception, any utterance is a representation of the speaker's self who "not only animates the words but is active in a *particular* social capacity" (Goffman, 1981, p. 147, italics in the original). A conversational interaction is a stage in which the participants present these performances of their self to each other. The actors' performances are informed by the environment and goal directed as they strive to manage the others' impression of themselves to be in accordance with their perceived expectations. In doing this, the participant of a discussion will marshal his or her linguistic resources to present a 'front' to the other interactants. The 'front' is the "part of the individual's performance which regularly functions in a general and fixed fashion to define the situation for those who observe the performance" (Goffman, 1959, p. 22). Luk gives examples of the "equipment" used to present their front as the "manner of speaking or the employment of linguistic resources such as conventional expressions or the choice of personal pronouns" (Luk, 2010, p. 27). The circumstances of the GOT make the desire to present the appropriate 'front' a useful analytic tool for conceptualizing the participants behaviour, particularly because they are not only performing in a way to project their own abilities for the benefit of the scorer, but also may "feel the pressure to display collaboratively idealized conduct, or to make prominent those characteristics that are socially sanctioned" (Luk, 2010, p. 27). The GOT participants may find themselves in conflict between acting in a way that makes them stand out to the raters, and breaking group solidarity by, for example, failing to understand or asking for clarification of something that was said.

The other paper that used identity to analyse peer interaction speaking tests was by Lazaraton and Davis (2008), though they do not specify a framework of interactional competence. In their literature review they refer to the concept of the 'positioning' by a test-taker in a paired assessment as 'competent' or 'proficient' by producing extended turns, nominating topics, self-correcting and demonstrating involvement with their partner. The less proficient partner might be positioned as being weaker by being provided words or responses, or ignoring his or her contributions. The less proficient partner might show resistance to this positioning by contradicting or insisting on his or her opinion. Lazaraton and Davis (2008) make the interesting point that identity is not necessarily consistently displayed in interaction, that it "may or may not be salient in all (or any) segments of that talk; it is up to the analyst to show that such identities are brought to bear in constructing and conducting interaction" (p. 317). This appears to be a convenient perspective for analysts, since it allows them flexibility in drawing conclusions from a transcript, however it makes it all the more incumbent on the researcher to find evidence to support any supposition about the identity being employed by the test-taker.

Such studies as these (Lazaraton & Davis, 2008; Luk, 2010) are directly relevant since they show how identity may be realized in the interactional competence displayed in peer interaction tests. It is also instructive to consider how identity has been explored by other IC studies. In Yagi's (2007) paper, the three participants were asked to telephone ten bookstores to request three books: one that would be commonly available, one that would be difficult to find and finally one that had yet to be published. The first participation structure that Yagi identifies is that of the telephone conversation, from which the participants could draw on experience from performing the same interaction in their native language. Yagi explains that the 'hello' or 'excuse me' uttered by his subjects when there was a silence by their interlocutor indicated their familiarity with their identity as a telephone caller. On the other hand, he also points out that since silences might be expected as the store clerk is busy looking up information, it also signals their lack of familiarity with their respective roles in this situated practice. As Yagi's subjects made more phone calls, they "came to a new understanding about the expected roles for the caller and receiver in this situation" (2007, p. 26) with respect to the possibility of extended silences from the clerk of the bookstore. Another identity that one of the study's participants invoked was that of a member of the non-English speaking community, and this was done by explicitly stating his nationality and stating that he could not speak English very well. This was not done in all telephone calls, but seemed to be reserved as a strategy to overcome or bypass linguistic difficulties.

Identity may also be signalled by a participant's choice of specific linguistic devices as Young and Miller (2004) described in their study. In this analysis of the discursive practice of 'revision talk' in which an ESL writing instructor interacts with a NNS student, several elements that signal their participatory roles are noted. For Young and Miller (2004), the instructor's use of 'you' is a signal not only brought about by the whole point of the interaction of revising the student's writing, but it:

> helps to construct her institutional role as instructor…. She has license to identify the student's problems and successes in his writing. The absence of *you* in the student's utterances positions

> him, not as an equal, but as the recipient of the problem identification and the suggestions for revision. (Young & Miller, 2004, p. 526)

Young and Miller (2004) identify a sequence of eight acts (*sequencing* being another interactional resource explained in Section 4.1.1.3 below) that remained relatively constant in these interactions over the time of their study, and describe the development of the NNS participant in terms of how the student gradually initiated more of them. That their identities as 'instructor' and 'student' were never in doubt is shown by the two acts that are at the core of the instructor's role, the issuance of directives and evaluations, never being encroached upon by the student.

In a similar way to these studies, this dissertation will analyse the choice of language of the participants for evidence of the identity or impression that the participant may be attempting to project. This being an assessment event, the participants will presumably be attempting to project an identity that gives the impression of themselves as competent language users while at the same time maintaining solidarity with their group members (Luk, 2010). The particular identity they manifest might not be continuously maintained, but may well be applied on a turn-by-turn basis, if at all (Lazaraton & Davis, 2008), and is thus "complex, contradictory and multifaceted" (Norton, 2006, p. 26). Above all, since identity is presented through participation in the interaction, and this is achieved through their choice of what, when and how to express something, it is impossible to discuss 'identity' without also discussing the language produced. As such, the 'identity' the test-takers display will be bound to the discussion of the linguistic elements of the test-taker's contributions to the GOTs in the data. It is these that will be addressed next.

4.1.1.2 Linguistic resources

The *linguistic resources* that participants bring to bear are the primary means by which the message is communicated, and can be further categorized into two. Firstly, there is *register*, or the knowledge and ability to use pronunciation, vocabulary and syntax that are typical of the practice in which the interaction takes place. To describe this, the analysis conducted in the quantitative section of this dissertation can conveniently be drawn upon. The second part is the "interpersonal, experiential and textual meanings in a practice" (Young, 2011, p. 429), which are envisioned as being analysed using systemic functional linguistics (SFL) (Young & Miller. 2004). This element has been featured in Young's account of linguistic resources for some time now (Young, 2009, 2011; Young & Miller, 2004), but to the author's knowledge it has yet to be fully incorporated in an account of interactional competence in a testing context, and is rarely included elsewhere. For example, while Young and Miller (2004) explicitly point out the metafunctions above, they chose to focus on other elements in the framework (p. 521). Yagi (2007) does not call upon Young and Millers' (2004) framework, and also has a relatively restricted analysis of the grammatical aspects of his participants' language. Finally, Dings (2007) uses the framework of He and Young (1998) which does not mention these metafunctions. Given that the objectives of this chapter are to investigate the basis for awarding scores to the participants and to investigate the interactions of the GOT qualitatively, grammar will be examined from the comparatively limited perspective of the extent to which it can be used to justify the raters scoring decision. As such, this dissertation will leave a full SFL investigation of grammar in interactional competence for other researchers to pursue.

4.1.1.3 Interactional resources

In this category, Young (2011) proposes four resources. The first of these is the sequential organization of *speech acts* used in the practice. The example given in Section 4.1.1.1 above by Young and Miller (2004) showed how identity was displayed through the sequence of events that were found to comprise the *revision talk* in the ESL writing conferences their participant took part in. In Young and Miller (2004), 'acts' seem to be analysed by characterising a series of turns by a more general functional label, and appears to be based on the notion of 'script', a collection of typical actions that together make up an everyday event (Bower, Black & Turner, 1979; Schank & Abelson, 1977). Although such scripts were first conceptualized to describe cognitive processes, they have also been applied as a framework to explicate language events in which NNSs take part, for example, where there is a conflict over expectations due to different scripts being prevalent in different cultures in medical consultations (Ranney, 1992), school routines (Saville-Troike & Kleifgen, 1986) and in recalling narratives (Harris, Lee, Hensley & Schoen, 1988). The approach of Luk (2010) to categorizing sequences of turns is more overtly embedded in the literature, drawing on the notion of ‘frames’ (Goffman, 1974) and other aspects of Goffman’s work (1959, 1967). As elucidated by Tannen (1993), a frame is a “culturally determined familiar activity” necessary for communication: “In order to interpret utterances in accordance with the way in which they were intended, a hearer must know what ‘frame’ s/he is operating in, that is, whether the activity being engaged in is joking, imitating, chatting, lecturing, or performing a play, to name just a few possibilities” (p. 18). With the notion of frames forming an organizing principle, Luk (2010) identified the recurring functional speech acts (Austin, 1962) that characterized them. In this way, the sequences that Luk’s described served the purpose of distinguishing the ‘architecture’ of a context of practice (Young, 2000). Since the purpose of this dissertation is to track the progress of individuals within the practice, the interactive analysis can provide a parallel service.

Another of the interactional resources is *boundaries*, which are described as “the way the participants open and close a practice and differentiate a given practice from adjacent talk” (Young, 2009, p. 216). The specific nature of GOT discussions creates some limits to the applicability of this resource. While it is certainly the case that there is an interesting variety in the patterns of the openings exchanges of GOTs (Leaper & Riazi, 2014) that may serve to inform the development of IC in this context, the way that GOTs are summarily halted by the examiners removes the possibility of analysing closing sequences. The focus of the interactional resource of *boundaries* then, falls entirely on the opening sequence of the GOTs. For the purposes of this dissertation, this period of time is a useful marker of the participants familiarity with the expectations of the GOT, and so particular attention will be paid to the individual’s role in the opening sequences. Moreover, the participants’ willingness to engage in the GOT is often signalled within the initial turns, and so the analysis of the opening period is often bound with their identity as a test-taker in the GOT.

When the question arises of how to analyse conversational interaction, it seems axiomatic that CA be the methodology of choice. Indeed, for the analysis of *boundaries*, and the other three interactional resources that Young (2009, 2011) identified, *turn-taking* and *repair*, IC studies have invariably relied on CA (for example, Dings, 2007; Ishida, 2009; Nguyen, 2008; Young & Miller, 2004). How this will be applied in this dissertation will be discussed in the final part of this section.

The CA most relevant to this study is that which has been applied to conversation in institutional settings, or ‘applied CA’ (ten Have, 1999, p. 8). The institutional talk that occurs in such applied settings may differ from naturally occurring conversation by being oriented to a set of institutional goals, having some constraints imposed on the participants and by having normally occurring interactional features modified or suspended to some degree (Drew & Heritage, 1992); the more general term of ‘talk-in-interaction’ was coined to cover both contexts (Schegloff, 1987). As an ‘applied’ methodology it has been used to analyse the language elicited by speaking tests, such as interviews (Brown, 2003; Lazaraton, 1992, 2002), paired assessments (Galaczi, 2008; Sandlund & Sundquist, 2011) and the GOT (Gan, 2010; Luk, 2010; Van Moere, 2007), as discussed in the literature review.

At the core of CA is the assumption that the orderly sequence of talk produced in a conversation is a result of the participants’ co-ordinating their contributions, and this is exactly what Young (2009, p. 216; 2011, p. 428) is referring to in his explanation of *turn-taking* as an interactional resource. The order that is achieved can be seen across speakers, not just in the individual, and so consists of a collaborative shared understanding (Liddicoat, 2011). This aligns precisely with the notion of IC as “distributed across participants” (Young, 2011, p. 430).

The collaboration achieved by the participants is displayed through their conversational interaction, and it is this which is available to be analysed (Ellis & Barkhuizen, 2005). A useful definition that encompasses this, was given by Markee (2000) who described CA as a form of analysis that:

> accounts for the sequential structure of talk-in-interaction in terms of interlocutors’ real-time orientations to the preferential practices that underlie, for participants and consequentially also for analysts, the conversational behaviours of turn-taking and repair in different speech exchange systems. (Markee, 2000, p. 23)

This definition is particularly relevant because it incorporates another of Young’s interactional resources, *repair*, or “the ways in which participants respond to interactional trouble in a given practice” Young (2009, p. 216; 2011, p. 430), and repair, turn-taking and sequence organization have been identified as the three types of organization in the talk-in-interaction at the core of CA (Ellis & Barkhuizen, 2005).

Within the turn-taking system, turns consist of *turn constructional units* (TCUs), which may consist of single words, phrases or sentences of a participant. The point where a listening participant may take a turn in the conversation is known as a *transition relevant place* (TRP). In everyday conversation the system whereby a transition takes place was described by Sacks, Schegloff and Jefferson (1974) as occurring by either the speaker selecting the next participant or self-selection if nobody was appointed by the speaker. If nobody speaks, then the current speaker may continue, and if so the same routine may occur at the next TRP. In the context of the GOT, as conducted by NNSs of varying ability in the target language, it will be interesting to investigate the extent this mechanism is maintained, and how it develops over the successive administrations of the test that form the corpus of this dissertation.

The sequence of turns is a main focus of CA. A conversation consists of a coherently ordered sequence of talk, and at the heart of this is the *adjacency pair* which is formed when one turn typically follows another, as an answer typically follows a question, with the question being the first part of the adjacency pair (FPP) and the answer being the second part (SPP). Where this occurs normally, the expected turn is known as the preferred response, and how this operates can be referred to as *preference organization* (Schegloff, 2007). In addition to such sequences, in NS conversation complex systems of pre-sequences, sequence expansions and post-expansions have been described (Schegloff, 2007). The extent that such complexities can be found in the NNS talk of GOT discourse is questionable: studies such as Van Moere (2007) have found some sequences in the GOT to be considerably more straightforward, while others seem more comparable (Gan, 2010).

Finally, 'repair' is the term in CA for what happens when there is a misunderstanding of some sort, perhaps a mishearing or, as may happen among NNS, inappropriate word choice, misuse or malformed grammar. In CA studies the effort to correct the trouble source can be *self-initiated*, by the speaker, or *other-initiated*, by another participant. Likewise, the actual repair can be carried out by the original speaker, *self-repair*, or the person who was listening, *other-repair*. Whoever carries out these actions, the act of repair necessitates the interruption of the previous direction of conversation until the repair has been resolved. In this study, identifying and describing episodes of repair is a matter of interest, since it has been contended that it is through such episodes that language can be learned, and thus justifies making conversation an important part of language teaching and assessment (see Section 2.3). Finding a format of test which elicits such examples would allow the students' abilities to be assessed. Against this, it must be pointed out that since this is a speaking test, test-takers will be presumably trying to show how well they can speak, and this could result in them attempting to avoid language breakdown, which they might believe would result in them getting a lower grade.

In addition to the four resources Young (2009, 2011) points out, the study by Gan and Davidson (2011) shows how body language and gesture plays an important role in the communication of a group oral. It is an element that most studies ignore, and since the primary resource of this study consists of videos it is one that this study can incorporate, depending on the camera angle. As such, careful attention will be paid to these visual aspects of the test-takers' behaviour.

Before finishing this section, it is important to point out some key assumptions of CA that may be transgressed to some degree in this dissertation, and how such violations may be at least acknowledged. Firstly, it is often recommended in CA textbooks that the transcript be produced by the analyst. "There is no benefit to be gained… by recruiting the assistance of another transcriber to do the job" since it is through the process of transcribing that the analyst can notice the fine details of the talk (Ellis & Barkhuizen, 2005, p. 209). The CA conducted in this dissertation has already departed from this. As described in the methodology of the quantitative section (Chapter 3), the transcripts were originally produced by assistants, and I have since re-transcribed them all in greater detail and improved them still further with repeated listenings for other parts of this dissertation. As such, the transcripts were analysed once more in the spirit of CA by maintaining a wide view of the interactions while focussing on the part played by the individual whose progress was being tracked.

A second departure from more orthodox CA studies is that since the original participants had to co-create the discussion on a turn by turn basis, and CA seeks to analyse how it unfolded, the analyst should avoid preconceived notions about the participants or the discussion. Here, my pre-existing familiarity with the video and transcripts and modes of analysis that they have already been subject to is obvious. As I have already analysed the participants' other performances both pre and post, it may mean that unconsciously I am predisposed to interpret the performance in a particular way. For example, in this section alone I have already pointed out some intersections between CA and the group oral, and because the CA is being done after the CAF and interactive analysis, I already know details about the performances of the participants in terms of the indices and their scores. Additionally, I am familiar with other CA studies on group orals in the literature, especially those that were conducted at the same institution (for example, Van Moere [2007] employed CA under similar conditions), and this knowledge too may affect my interpretation even unconsciously. All that can be done with regard to this objection is to admit the possibility of this occurring, while attempting to maintain a CA perspective by allowing understanding to emerge from the data.

Finally, CA generally avoids the practice of coding and quantification, since to do so runs the risk of being "led seriously astray" by the danger of it "free(ing) us from the need to demonstrate the operation of what we take to be going on in singular fragments of talk" (Schegloff, 1993, p. 102). For CA, a single instance of language use is a significant number since it "is built on routines of various sorts, but it is, at the same time a unique achievement here and now" (ten Have, 2007, p. 38). Although I will not quantify the CA findings themselves, especially in Chapter 5 I will overtly draw attention to how the participant's CAF statistics, interactive analysis and scoring are realized in the participants' performances. A major part of this dissertation is to compare the performance of participants taking part in the same test in different administrations, and making comparisons across their performances will lead to useful insights that cast light not only on the interpretations made by the raters who scored them, but allow insights into the nature of the GOT as a tool of assessment and learning. Indeed, the research question that CA is being partially used to answer is to illustrate how the performances as described by the scores and quantitative statistics of the CAF and interactive analysis are actually borne out in the GOT.

In making such compromises, this dissertation is not alone in using CA to the extent that it suits the researcher. In her analysis of oral proficiency interviews, Johnson (2001) draws on features of CA such as those outlined above for her coding system and subjects counts of features to statistical analysis, although she does refer to her study as a 'discourse analysis' rather than CA. Galaczi (2008) also admitted that she did not adopt a "purist CA methodology" (2008, p. 94) by using terms from other perspectives, coding and quantification. As this section makes clear, the CA employed in this dissertation is part of an adaptation of an IC framework and is used to provide depth to the quantitative analysis. As such it does not claim to be a full CA analysis, and will be referred to as a 'qualitative analysis'.

In summary, in the qualitative section of this study, the objective is to examine the transcripts in sufficient depth to assess the extent to which their qualitative performance matches their performance indices, justifies their scores, and informs us of the construct of the GOT. Since the performance in question is a discussion test in which three or four participants take part, and discussions are collaboratively co-constructed, it is in fact the interactional competence of these participants that is being investigated. The first part of the current methodology section has examined features of Young's (2009, 2011) framework to show how it can be investigated to answer the question. Firstly, the participation framework will be examined to provide an insight into their footing and identity as test-takers, and how they position themselves in the discussion. Although 'identity' is not directly assessed in the rating scales, their entire performance is bound directly to their ability to assert their identity (Lazaraton & Davis, 2008; Luk, 2010). Thus, their identity is displayed through their linguistic and interactional resources. In terms of the resources in Young's framework, this is a holistic combination of the register with which they speak, the interactive functions that they perform, the turn-taking skills they display and the repair they take part in. Within this performance, their engagement in the boundaries of the starting sequence and then continuance will be investigated.

Having explained how the framework of IC is being operationalised for the analysis of the qualitative data, the next section will describe how the participants of this part of the study were selected.

4.1.2 Participants

To identify a range of test-takers to examine qualitatively, it is important to consider not only those with extreme scores but also those who represent more commonplace experiences. The multiple regression (MR) analysis from the quantitative section generated statistics that enabled outliers to be identified, specifically through the Mahanalobis figures. To integrate these figures with a selection of more ‘usual’ experiences, the following steps were taken. Firstly, the gain scores of the bands that the indices were related to (Fluency, Grammar, Vocabulary and Communicative Skills in the scoring bands) were calculated as the difference between these scores in the third and the first administrations and turned into z-scores using SPSS. The difference between the third and first administrations was chosen because it was felt that this best represented the difference in their skills at the start of the program and the end of it. These standardized gain scores were ranked and were divided into four groups: those whose gains were below -1 (8 test-takers), those between -1 and 0 (21 test-takers), those between 0 and 1 (17 test-takers) and those above 1 (7 test-takers). From each of these four groups, two test-takers were chosen: the first was the one with the highest number of influential outlying scores according to the MR analysis, and the second was chosen using a random number generator from the test-takers with no outliers. The choice of test-taker also depended on having a good view of the participant, since factors like gaze and body language are an important aspect of communicative behaviour (Gan & Davidson, 2011) even if not specifically part of Young’s framework. If a clear view of the participant was not possible in the video footage, the next closest participant to meeting the conditions outlined above was chosen. Of the eight original candidates, three had to be chosen again due to less than adequate camera angles in at least one of the videos.

The total scores in the three administrations for the eight students who were finally chosen are given in Table 51 below (their scores in each scale can be found in their respective sections).

Table 51: Qualitative sample participant’s scores over the three administrations

Group	Participant	Admin 1	Admin 2	Admin 3
Low Initial Group	Kabuhito (M)	5.4	6.3	10.5
	Saaya	6.4	10.9	10.8
	Machiko	6.8	12.7	14.1
Medium Initial Group	Taeko	10.7	14.0	14.5
	Aemi	10.7	15.3	12.1
High Initial Group	Naeko	16.6	15.7	13.5
	Hamako	16.7	16.5	15.7
	Yamahiko (M)	17.0	16.4	16.9
Highest possible score = 20				

Here it can be seen that according to their scores in the first administration they fall neatly into three groups: three whose initial totals were between 5.4 and 6.8; two with scores of 10.7; and a high scoring group with scores between 16.6 and 16.7. Since a purpose of this dissertation is to track student progress, it was deemed appropriate to analyse the students according to these groups, which will henceforth be referred to as the low, medium and high initial groups.

4.2 Qualitative Analysis of the Group Orals

In this section the chosen eight participants' performances in the three successive administrations of the group oral will be qualitatively analysed. The results will be organized into three subsections according to the low, medium and high initial groups, as explained above, with each subsection consisting of a qualitative analysis of the group members' performances. The findings will be synthesized in the summary, allowing a perspective to emerge of how these participants progressed over the three administrations.

4.2.1 Qualitative analysis of the initial low group

4.2.1.1 Initial low group in the first administration

The participants in the initial low group display similar negative interactive behaviour in the first administration by showing reluctance to communicate or finding ways to minimize their contributions in one way or another. The identity they display in this test seems most aptly described as a 'reluctant embarrassed participant'. This reluctance is clearly shown in the interactional resource of their body language. When they are not talking, they typically make minimal eye contact with the other speakers, usually by keeping their head down, with their eyes darting rather furtively toward the speaker. They particularly tend to avoid eye contact at Transition Relevance Places (TRP), usually by keeping their head down, and their eyes on the prompt. For example, for almost the entire first test, Kabuhito sits holding the prompt paper in front of him in both hands, with head bowed as if he is reading it. The only sign he gives of following the discussion is that his head is slightly angled towards the current speaker, eyes rather furtively darting from the prompt to the speaker. Saaya has a very similar posture, spending almost the entire test with head bowed, eyes on her paper, one hand grasping the prompt paper, the other she rests her head on, fingers covering her mouth. Likewise, Machiko keeps her eyes on the paper, with occasional glances at another speaker while they are in the middle of their turn, and a TRP does not seem imminent. This is very similar to the behaviour noted by Gan and Davidson (2011) of a participant in their lower scoring group, who appeared to be engrossed in her notes for long periods of time.

These participants' embarrassment at having to talk is displayed when they are addressed directly and they cannot avoid remaining silent. For example, when Kabuhito is addressed by another test-taker in the opening moments of the first test, he giggles and uses Japanese. Later on in the test, perhaps due to feelings of group solidarity after witnessing the other group members struggle in English, Kabuhito does attempt to use the target language (Excerpt 1), but here too he uses embarrassed smiling or giggling to cover his lack of ability.

As noted above, Kabuhito responds to the first question he is asked (at 1:22.7) in Japanese in a whispered conversation to another group member, thus avoiding the use of English linguistic resources altogether. He only responds when he is addressed directly, and from this evidence it seems likely that he would be silent for the entire GOT if he were not asked a question. The second time he is asked a question, he responds in English, and the exchange that takes place is in Excerpt 1 below.

Excerpt 1: Kabuhito in the first administration (Person C)

Turn	Time	Person	
17	6:40.0	B	(4.3) How about you, have you ever been (.2) to another country?
18	6:44.8	C	(2.4) Pardon me (1.1) pardon me
19	6:48.0	B	(.9) Have you ever been to another countries?
20	6:53.7	C	(2.0) No (2.1) (cough) (1.8) I (2.7) have-(.8)not-(.9)to (1.7) I have not been (1.7) other country but (.2) I want (.3) to (1.5) want to (.8) Australia (4.3) because (8.0) my (3.6) my father's …
-	7:45.7		(13.1) <B Japanese whispered to C – C Japanese whispered to B>
21	7:54.7	B	(5.2) People who he knows.(1.8) People who he knows
22	8:08.3	C	(8.2) my (.7) father (5.1) pe-ople (3.3) who (1.1) my father (.2) know (5.5) is (1.9) there

In this excerpt, Kabuhito shows that he has the ability to use a communicative strategy (turn 18, using 'pardon' to elicit a repetition), and that he can self-repair his heavily accented pronunciation of "havunotu" to a more target-like "have not" (turn 20). That his linguistic resources are being stretched is shown by the disfluencies in his speech. Kabuhito manages some basic verbal constructions, but even when there are no pauses of more than .2 of a second in between, he annunciates each word separately. He soon finds that his available resources are not sufficient to express himself, and turn 20 of Excerpt 1 trails off into a long 13.1 second silence. It is notable that it is not Kabuhito who resorts to Japanese on this occasion; it is B who initiates an exchange in Japanese to diagnose the problem so that she can offer a repair in line 21. Kabuhito manages more complex grammar in the form of a relative clause in turn 22, however, this is mostly repeated from B's repair in the immediately preceding turn, though Kabuhito can be given credit for sufficient linguistic awareness to substitute the pronoun.

On the video, the struggle that Kabuhito has with expressing himself in English is evident in his body language. In turn 20 when he starts talking his eyes are looking straight ahead, then at his paper, and when he reaches "my father's" his face contorts into a smile, his hand moves in a circular motion and he puts his hand on his brow in a display of embarrassment at his inability to find the word, before B rescues him by whispering to him in Japanese and interpreting for the group. The hand movements here are an outward sign of his struggle to use the limited linguistic resources at his disposal, as Gan and Davidson also found for their lower level participants (2011, p. 115).

When Saaya is directly asked a question, she uses minimal agreement to avoid giving a full answer, as seen in Excerpt 2.

Excerpt 2: Saaya in first administration (Person C)

Turn	Time	Person	
5	1:18.8	A	(2.5) Ah I I think so too (.8) because (.5) unusual places (.3) can be experienced (1.0) mm eto (1.2) usual (.2) huh? ah unusual things so (.4) I want to go to unusual places. (.8) How about you?
6	1:49.2	C	(13.0) I think so.
7	1:51.8	D	(1.8) I think so too. [ABCD: huh huh huh]
8	2:03.1	A	(6.7) For example where do you want to go?
9	2:12.6	B	(7.2) I want to go (.5) ah (1.5) I want to go (.5) this Africa. (.4) Um I (.3) I (.3) I don't know er (.3) I don't know about Africa many things so (.2) ah I (.5) uh (.2) I (.3) I go ah I visit (1.6) visit (1.6) I visit there (.3) ass- and I learnt (.5) many things about Africa. [A: mm]

The 13 second pause before she answers the question in turn 6 could almost be described as 'theatrical': Saaya's hand drops from where it had been propping up her chin to the prompt paper, which she stares at intently for a while, then she raises a hand to her mouth and lifts her eyes to her interlocutor before giving her minimal answer. The length of time that passes and the signs of struggle in her body language seem to be signs of inner turmoil as she is attempting to construct an answer. This forms a striking contrast to the simple agreement that she eventually produces. This indicates that her response is an example of the "surface converging responses" noted by Luk (2010), or agreements given without elaboration (p.38). In this case at least, the purpose seems to be to avoid talking further and pass on her turn to someone else. The next speaker in Excerpt 2, Participant D (turn 7), is also reluctant to participate, and is apparently a quick study: her response in line 7 is so overtly an act of avoidance that all participants (including D herself) laugh at its obviousness.

Saaya's verbal consent would seem to indicate that she, like A, prefers "unusual places"; but this is shown to be insincere at the end of the test when the rater asks her a question in order to get a rateable sample, as shown in Excerpt 3 (the rater's intervention here can be considered justified since before turn 22 in Excerpt 3, Saaya's only other utterances were the three words of turn 6 in Excerpt 2).

Excerpt 3: Saaya in first administration (person C)

Turn	Time	Person	
21	5:55.8	T	(11.9) Erm (.5) where do you want to take a vacation? (.3) Where do you want to travel?
22	6:00.3	C	(1.0) ah (.9) I want to (3.0) I want to go to (3.3) New York (3.3) Because (1.4) I (.7) I want to see (.6) the statue of liberty (2.4) huh huhhuh

Since New York could hardly be described as an "unusual place" that A was talking about in Excerpt 2, Saaya's words in Excerpt 3 contradicts what she had previously assented to. In Excerpt 3 Saaya uses a single main verb twice ("want") with frequent long pausing, indicating a limited range of linguistic resources.

The final person in this category, Machiko, also shows limited linguistic resources in her long and frequent pauses, as can be seen in Excerpt 4. As is common among many of the groups in the first administration, how to start the test and who to initiate a turn is a confusing period for them. After the rater has told them to start it takes 16.9 seconds before the first test-taker speaks. In this period, there is some non-verbal communication between the four test-takers. Initially there is a shared nervous laughter and eye contact between the three females (throughout this period the other participant, a male, has his head down and does not interact). Machiko, in seat A makes eye contact with B, and C moves her head to take in A and Bs' shared eye contact while D (the male) is steadfastly staring at his prompt, his head buried in his arm. Following this brief spurt of activity, Machiko and B bow their heads to the prompt and are still. Test-taker C does so for a second and then looks up to see all the others with their heads bowed, looking at their prompt. Finally C removes the uncertainty by taking action, and begins talking, lifting her head to talk to the two bowed heads of A and B. After she begins talking, Machiko lifts her head to make eye contact with her, and responds. The opening verbal exchange is shown in Excerpt 4.

Excerpt 4: Machiko in first administration (person A)

Turn	Time	Person	
1	12:22.5	C	(16.8) Do you want to live (.8) another country?
2	12:27.4	A	(1.2) Yes
3	12:28.7	C	(.7) Why?
4	12:29.8	A	(.9) I (.7) I would like to (.8) get (.7) the (6.8) other country business
5	12.47.8	B	(5.1) What about you?
6	12.49.9	C	(1.0) Me too (.7) so …(1.4) uh I went to Australia (.8) last year (.8) to (1.3) study English (.6) and …(1.6) I spent (1.3) there (1.8) with my host family (1.1) so …(2.6) un I had …(4.2) really (.2) good time (.4)there (.4) and (.9) so (.8) if I (.6) can do that (1.1) I (1.6) want to (.8) take my (.4) parents (2.1) with (.3) me

Minimum answers to polar questions, invariably “yes”, have been noted in GOTs in other contexts. Gan (2010) points out that they do not encourage mutual development and are more common among lower ability participants. Test-takers soon realise that if they do not voluntarily provide an explanation their peers will ask them for it, as C does in turn 3. The minimal response succeeded by a follow-up “why” question is a sequence that most frequently occurs in the first administration, and thus shows the test-taker’s lack of experience of what constitutes a sufficient answer in this “interactional architecture” (Young, 2000, p. 5).

The remainder of this discussion is dominated by B and C who turn out to be the most capable speakers of the group. However, later on, when the conversation falls silent, Machiko takes it upon herself to initiate a new topic by reading out a question directly from the prompt. This is a key difference between her performance and the other two participants in the initial low scoring group: Machiko shows a willingness to engage in the interactional resource at a higher level, unlike Kabuhito and Saaya, who seem reluctant to participate in this GOT.

In terms of Young’s (2009, 2011) framework, these students are yet to build up an interactional resources that they could transfer to the context of the GOT that would enable them to take part in it more fully. Of the three test-takers in this group, only Machiko shows any inclination to be anything other than a passive member who takes part only when asked. Of course, part of their difficulties may well be their lack of ability to adequately comprehend the flow of conversation. Kabuhito’s response in Excerpt 1 suggests that he has problems following the conversation. However, from the evidence in the video, from the visible struggles that Machiko and particularly Saaya go through to respond suggests they understand it but have problems forming a response.

4.2.1.2 Initial low group in the second administration

In this administration, the three participants in this group experience very different patterns in their discussions. In Kabuhito’s GOT, another participant assumes an identity of ‘conversation leader’ and dominates the discussion with her questions. Of the 16 questions that are asked in this GOT, 14 are asked by participant C, and the remaining two questions are asked to C, one each by Kabuhito and participant A. There is no verbal interaction between Kabuhito and participant A, making this group oral more closely resemble an interview than a discussion.

The insight into Kabuhito’s interactional resources suggested by his body posture is similar to that in the preceding administration: in both tests he is hunched slightly forward, his head down, though angled towards the speaker, eyes down on the prompt paper, though darting towards the speaker. The only difference here is that in the first administration he held the prompt paper in both hands throughout most of the test, but this time he keeps his hands off the desk by putting them by his sides. In the second test his contributions are considerably more than in the first, mostly because of his response to what almost amounts to an interrogation by participant C, as seen in Excerpt 5.

Excerpt 5: Kabuhito in 2nd administration (person B)

Turn	Time	Person	
7	50.3	A	(1.4) Because ah… …(10.6) um (1.1) men (1.0) men is working everyday(.9), and so (1.5) women (1.5) almost I think almost almost almost (.7) women (.9) be (.9) doing (.3) housework (.9) so (1.2) women (1.1) don’t touch money (B: ohmmm)(1.1). Just do the houseworking (1.3) so…
8	1:31.5	C	(2.7) How about you?
	1:37.9	B	(5.6) uh err
9	1:38.9	C	(2.0) Do- (.3) ahem (.3) - does your father work everyday for (.8) for you?
10	1:48.4	B	(2.5) Yo- (.2) ah (.3) My father work (2.6) outside (2.9) my mother (1.2) care (1.9) me
11	2:02.9	C	(1.4) If, (.8) if you want uh if you got (.2) married (.9) uh do you want to (.7) work (.5) for (.5) her? (.7) Or do you want to do housework (1.0) in the future? [B: I…] (1.4) Which do you prefer?
12	2:20.1	B	(1.1) I (1.4) I I want to (1.3) work [C:Work] (2.7) for (.9) my family.
13	2.33.6	C	(3.3) Don’t you like (.6) housework?
		B	< slight shake of head>
14	2.40.0	C	(4.4) Do you help out (.2) your mother?
15	2.45.0	B	(1.6) Uh,(1.0) no [AC: huuh]

16	2:47.2	C	(.7) Maybe you should do. (1.7) Maybe you should (.2) help your mother. [B: Ah] (2.2) So she, (.2) she may be glad…(1.5) and maybe (.5) she will surprise

In this sequence, the dominating test-taker asks five questions to which Kabuhito can only respond in minimal comments, if at all. He is even subject to some chiding, as can be seen in turn 16, which puts C in a morally superior position and against which Kabuhito lacks the linguistic resources to defend himself – on the video at this point he smiles and nods his head in agreement. However, later in the script at potentially the most face threatening point, he does show the ability to ask a clarification question, as shown in Excerpt 6, in turn 33.

Excerpt 6: Kabuhito in 2nd administration (person B)

Turn	Time	Person	
32	5:36.4	C	(4.5) Uh (.8) which do you think (.8) um (.4) which is (1.0) more happy (1.5), to (.5) work (.6) or to have a baby? (1.4) Which is happy (.7) what do you think?
33	5:58.2	B	(5.5) About woman?
34	5:59.5	C	(.6) Uh… (.8) yes for women. (1.5) So (.3) please think if you are woman
35	6:07.3	B	(.8) oh (.8) If, (.3) if I I (.6) I were (.6) if I were ah I I woman (2.0) I (1.6) I want (.3) to (1.0) I want baby
36	6:26.3	C	(2.6) So you choose (1.0) um baby (.3) than (.4) work?
37	6:32.0	B	(.4) Yes

Participant C's position here has been that she would like to work even after marriage and having a baby, whereas Kabuhito and A are in favour of the woman giving up work to look after the child. Here, she appears to be trying not only to get Kabuhito to understand her point of view, but seems to want to persuade him that he would do the same if he were in her position. Rather than simply agree and allow the conversation to continue around him, and despite his limited linguistic resources, Kabuhito resists in turn 35, by stating the opposite point of view. Still, as for the first administration, Kabuhito's participant in the second is reactive, and he seems to continue his identity as 'reluctant participant': the single question he asks may well have been borne out of the need to maintain his face.

The limited linguistic resources he has are shown by in turn 35 which is Kabuhito's longest in this test. He manages a conditional in this turn, but the difficulty he has in assembling and producing it is shown by the high proportion of pauses and maze words in it. His other long turns are in Excerpt 5. In turn 10 he has two short independent clauses with no connective, though in turn 12 he manages a more complex verbal construction “want to work for my family”. It is noticeable that grammatical structure is an echo of that used in C’s questions in line 9 where she uses “work for” and line 11 where she uses it as an infinitive “to work for her”. It could be conjectured that C has ‘primed’ Kabuhito to use this grammatical form (McDonough & Mackey, 2008). Nonetheless, despite the many hesitations and maze words, Kabuhito should be given credit for producing a grammatically accurate sentence here.

Saaya’s discussion in the second administration took a very different shape to the interview-like GOT of Kabuhito’s. In the eight turns in the entire GOT, there were only two questions. The interaction consisted of participants exchanging short speeches in which they gave their opinion of the topic, with little reaction or development of what the others had said. Throughout, Saaya’s body language is similar to what she showed in the first administration, though sometimes she takes on a more open stance in which both hands are palms down on the desk, parallel with the sides, perhaps indicating when she has something she could say. Her greater willingness to participate in the second administration, is shown by her occasional use of the verbal backchannel, though it is limited to a voiced “mm” and a single quiet “yeah”. However, at TRPs she still tends to avoid eye contact by keeping her eyes on the prompt. In this test, all Saaya’s words come in one long turn that lasts almost one minute and can be found in Excerpt 7.

Excerpt 7: Saaya in the second administration (person C)

Turn	Time	Person	
3	2:45.8	A	(1.3) ur (.9) My family is not traditional. (.7) Uh…(1.5) umm (1.7) both of them (.2) work (.2) and (1.0) my father (.3) cook…[BC yeah] (1.9) so (.3) uh (.5) not traditional. (.8) Uhm (2.8) I think (1.1) uh…(2.4) uh same (.8) opinion… [B: ah mm mm]..(7.9) mm (.5) Japan traditional style (1.4) mm should (.3) be change.
4	3:33.2	C	(7.2) Um my family and my family is (.7) um (1.5) traditional (.2) family (.7) um because (.4) my mother (.3) work hou- (.2) ah housework (.6) and (.2) my (.2) father (.7) ur (.7) worked (.7) ur (.7) for company (.6) [A: nn] and (2.1) earn the money (.5) so (1.3) um (.6) I think (.5) Japanese (.3) family (.9) um ah Japan (1.9) have (1.3) many ah Japan traditional (2.6) um (.5) Japan have traditional family formerly (.5) but now (.6) nn (.3) Japan is (.9) changed (.6) the (.5) not traditional (1.3) mm(.6) ja- family. [B: nn yea]
	4:36.5	B	(5.7) And I have more one reason (.9) why I think (1.1) ah it (1.2) ah why I think (.4) the Japanese style (1.5) the Japanese traditional style is going to change (.6) because (.2) uh (1.9) so for example the same (.4) company (3.0) like canon (.5) and (1.0) man and (.9) women (2.3) is employer(.3) [D: nn] so (.3) work hard and hard (.5) but (1.5) many of company (.3) is now ah (.9) same (1.0) money of salary now (.5) not…

Excerpt 7 is from the opening turns of the group oral where the turn-taking is often such that each person gets a turn to express his or her opinion which often results in an “orderly turn-taking mechanism” being co-constructed (Luk, 2010, p. 37). This conversation seems to be an exemplar of the kind of GOT that Van Moere (2007) noted as being particularly prevalent in the opening moments of GOTs. According to Leaper and Riazi (2014) this may be partly due to this particular prompt giving them sufficient material to compose long sequences of talk in their minute of preparation time, and in this case, the participants were able to extend them over the duration of the test. Before Saaya speaks, the other three participants have had one turn each, perhaps building expectation that her turn must be next. Since the preceding speaker, A in turn 3, does not overtly pass the turn on, and both B and D have their heads down and are focussing on their own prompts, it seems that peer pressure is building for Saaya to speak next. In the absence of a clear signal from the previous speaker that this is a TRP, the long pause of 7.2 seconds means that Saaya can be certain that A has finished speaking. This long pause and the reluctance of anybody else to fill it provides evidence of “dramaturgical loyalty” (Goffman, 1959, p. 212) of the other participants to give Saaya the space to have her turn. The fact that she is the last of the group to speak shows that her identity as ‘reluctant participant’ may again be applied. Though, as noted above, her body language on occasion seems to indicate that she wishes to communicate, but cannot find an opening, means that she could perhaps be taking on the identity as a ‘frustrated participant’.

The linguistic resources displayed in this turn still show limited ability. There are some multi-clause AS-units that show evidence of complexity as “and earn money”, “I think”, but the remaining clauses are strung together using connectives “and” “because” “so” “but”. The accuracy of the grammar used here is low, with missing prepositions, uneven use of tense, and so on. The lack of fluency is shown by the many maze words, hesitations, and voiced pauses. However, she still manages to put together a long turn, which she could not do the first time she took this test. For her vocabulary, she uses mostly high frequency words, with the exception of “formerly”.

Unlike Saaya and Kabuhito, who remain mostly passive in their second test, Machiko is willing to take on the identity of ‘leader’ this time through her interactional resources. In the opening sequence she is the first to speak. Machiko starts by giving her opinion and ends the turn with a question to transfer unambiguously to the next person. Later on she uses a follow-up question and she refers to what other speakers said before in the discussion. When she is not the speaker she uses eye contact and frequent and varied verbal backchannels to show she is engaged in the discussion, and only occasionally glances down to the prompt. Excerpt 8 below shows an example of her speaking and listening skills.

Excerpt 8: Machiko in the second administration (person C)

Turn	Time	Person	
4	8:43.9	D	/ / (2.0) Traditional, or (1.0) man and women (.6) both both man and (.2) women (1.2) get the job (.7)…is better? C:[ahm]
5	8:54.1	C	(.9) I think (1.9) both (.6) both person (.4) get the job is good (1.0) so (1.1) um (.7) recently (.7) there are ah (2.6) mm a lot of (.8) ta- (.5) ah situation (1.1) mm (1.5) I think (.5) it is good.
6	9:20.0	A	(1.5) ah ahem I think (.3) mm recently (.7) mm house (.3) housewife (.3) wife (1.0) is bored (.2) of (.2) her life (.4) and get the (.4) job outside (.5) So I think (.4) un (2.2) traditional (.4) Japanese traditional (.2) family is (.4) little old style. [C:mmm] (.5) So [C: I think so] (.7) we have to change (.2) the style. [C:mmm]
7	9:49.1	D	(1.6) but I think it's difficult
8	9:52.2	C	(1.2) why?
Z	9:53.1	D	(.7) because (1.1) I said I said before but (.4) I I think (1.9) ah women can't (.2) get high position (.2) in (.3) their work
10	10:04.0	A	(1.0) umm (.2) for example (.3) what?
11	10:06.9	D	(.9) for (.9) ah office lady, [C: OL] (1.6) they can't (.6) maybe (.2) I think (1.1) they (.2) they do (.4) ah (1.6) ah (.6) they (1.7) they can't (.5) do (1.3) high ah high responsibilities work. [C: mm-mm-mm] (2.9) So (.8) I think (1.3) women can't get (.4) earn (.3) much money [AC: mmm]

Turn five is the second time she speaks in this group oral, and here she shows improved linguistic resources. This turn is in answer to the question posed by D at the end of his first turn. As is often the case, the last person to give a verbal backchannel in the previous person's speech is the next speaker, and in this case it is Machiko. In her turn there are still frequent unfilled pauses that would try the patience of a listener, and it seems like she avoids expressing herself more fully by being vague "recently there are ah a lot of ta- (.5) ah situation", but she finishes emphatically with a summary statement, and her gesture of dropping her arms to her sides synchronizes with her words to make it clear that she has finished speaking (Gan & Davidson, 2011). She does not signal overtly who should be the next speaker, but after a brief pause, A fills the gap. When Machiko is not speaking, she maintains eye contact with the speakers and supports them by constant backchanneling, not only her standard “mm” sound, but more sophisticated agreements (“I think so” in turn 6) and clarification (“OL”, which is Japanese for ‘office lady’, in turn 11).

The final example, Excerpt 9 below, shows both her improved ability as well as some further limitations in her linguistic resources.

Excerpt 9: Machiko in the 2nd administration (person C)

Turn	Time	Person	
16	10:43.4	A	(.2) in fact (.5) ah (.7) after getting married (.3) so (.2) women have their (.8) baby (.4) or have to (.9) take care of baby (.4) but (.8) um…(.5) it’s (1.0)mm (.3) fair [C: ahm] (1.8) if woman work (1.4) work so it’s fair (.2) should (.2) fair (.5) to income [AC umm]
17	11:08.8	C	(.9) I think so. Ah (.5) soci- (.2) Japanese social (.5) [A; ahm] should change change mind. (2.8) [AD: mm] For example (.5) your idea, (1.4) I think. [D: ahum] (1.3) How about you?
18	11:23.5	B	(1.9) I think it’s (.4) un Japanese traditional (.4) situation (.4) [C:mm] is (.4) changing (.4) [C:mm] for (.8) in our future. (1.5) umm (.6) Recently (.7) uh many (1.2) uh eto (.7) increase (.2) the women (.3) who (.3) not (.3) marriage (.4) not marry (.7) [ABD: mm ahum] (1.6) to keep (.5) to (.3) to (.3) keep working. (.2) [ABD ahmm]. Not (.5) un (.3) so this situation (.8) become (.2) change (.3) and (.5) should (.4) change this (.5) type.

Here she is expressing agreement with the sentiment that everybody in the group appears to agree with. In turn 17 she seems to reach a TRP after “mind”, but self-selects to continue by adding the referral to participant D (she clearly gestures to him), an interactive function that is rare and so could be considered a mark of a more advanced speaker. However, it is not entirely clear which of D’s words she is referring to. The first time D spoke he stated that the traditional Japanese family was better because men can earn more money, and following that, in turn 11 (Excerpt 8 above), he pointed out that women have difficulty earning more money because they cannot get high responsibility jobs, neither of which seems relevant to what Machiko is agreeing with. That she cannot explain the relevance, or that she misunderstands what D was saying, may be an indication that she is at the limits of her ability here. Ultimately, by espousing her agreement with previous statements of group members, she is reinforcing group solidarity in an act of “dramaturgical discipline” (Goffman, 1959, p. 216) to maintain the fiction that the group members align with each other’s opinion, even though it seems unlikely that they would agree according to their actual words.

The first time they took this test as incoming freshmen, a dimension of novelty was evident in the general atmosphere of levity; this has largely dissipated the second time they take the test. After one year of study there is a greater awareness of this as an assessment context. In particular the students know their performance in this test determines the class they will be placed in next academic year, thus a more serious atmosphere prevails in the second administrations. By the time of their second administration they have had a year of study in which they have been encouraged to participate in group discussions in their various classes, and have various opportunities at the institution to help them to accumulate their interactional competence and improve their speaking performance. The resources they should have been able to build up throughout the academic year should have been available to be transferred to greater participation in this test. On the evidence presented in the second administration, although improvements have been noted in all their performances, Kabuhito and Saaya are still passive participants, and only Machiko can be seen to be taking on a fuller performance in the GOT.

4.2.1.3 Initial low group in the third administration

In the third administration, Kabuhito and Saaya show a limited ability to take the initiative. As a listener, Kabuhito shows improved interactional resources by occasionally joining in the laughter that takes place in the vivacious group oral that he finds himself in, and occasionally even backchannels. His identity seems to be one of 'fringe participant'; as somebody who can be brought on to give an opinion, but will otherwise merely observe and follow the discussion. There are two periods in the group oral where he is called on to take turns to speak. His participation in the initial opinion exchanging sequence of the opening is shown in Excerpt 10.

Excerpt 10: Kabuhito in the 3rd administration (person D)

Turn	Time	Person	
1	46.1	C	(5.2) Would you prefer to live in the city or (.4) in the country? [A: ah]
2	51.3	B	(1.3) I prefer (.5) living in the country [A: mm] (.9) because (.9) haa I'm from (.5) very very countryside [ABC: hahaha yeah] (.3) when I (1.1) when I am walking in the street [C: mm] (.6) uhm (.8) if (.2) there are many (.4) crowds [C: mm] (.5) I feel stressed [A: ahh]
3	1:15.0	C	(4.4) I (.3) I'm from Nagano (.3) I think (.6) mm (.2) Nagano is very (.4) country countryside (.2) and (.5) when I (.6) came (.2) to (.3) Chiba (.4) first (1.1) my (1.0) mm (.7) my heart beatedhaha [A: hmm] (.3) because (.5) Chiba is (.7) very (.7) umm (.8) Chiba is a big city and there (.3) were many people [ABD: mm]
4	1:46.3	D	(5.0) I (.7) I (.8) live in (.2) Chiba [C: mm] (1.4) Do you know Yachibatta?
5	1:51.3	B	(.5) no I don't know (A:C: no ABC huh [huh]
6	1:53.4	D	(1.0) Yachibata is famous for (.7) peanuts [ABC: ah] (5.6) I (.6) I (.8) like (.6) my (.2) hometown [B: mm] (1.5) If (.2) we (.5) we live (.2) in the country (1.2) we can see (1.8) many stars in the sky [B: ahh C: nn] (.9) it's beautiful (.6) and I (1.1) I (.4) was (.4) I I am so(.5) relaxed
7	2.25.5	A	(2.6) I have lived in Chiba [C: nn] (.4) for (.7) about (.8) for mm 20 years [C: nn] (1.0) but I (.4) don't like (2.0) to live (.5) here because (1.0) I (.4) hate the place where people (.2)

> are crowded [B: ahh] (.5) so (.5) I also feel (.2) stress [C: nn] (.9) and (.5) so I want to- (.3) if I can do – (.2) I want to live (1.0) in the countryside [C: nn] (5) because (.8) so (1.2) because there are a lot of nature [BC: nn] (.7) and beautiful views [C: nn] (.6) so (.5) that's – (.5) there are few (.4) views (.4) in (.2) city (.3) now [B: yes C: nn] (1.3) I envy [BC: huhuh] (.7) I envy [B: yes] people who live in (.4) the countryside. [C: nn]

In the initial rounds, Kabuhito sits with the same basic posture as the previous administration: hands by side, tucked under his thighs or on his lap, slightly hunched over, head bowed, eyes glancing towards the current speaker. The turn he takes in Excerpt 10 is clearly transferred to him by C, turning her head and making eye contact, so he has no need to create space for his participation. The question he asks in turn 4 is a pre-sequence that serves "to check whether a certain condition for a possible next action exists" (ten Have, 2007; p.131), namely whether the other participants have heard of his home town, rather than a question that shows he can initiate a topic in conversation. It is also an example of listener initiated negotiation of meaning, as noted by Varonis and Gass (1985), and is a feature of conversation that is rarely found in this corpus. This move successfully reserves the floor for him to continue his turn. Since it is his first turn this well-organized turn is likely a result of the planning he could do in the preparation time, as well as while the other speakers were talking (He & Dai, 2006). However, as before, his limited linguistic resources are shown by the number of unfilled pauses in his halting speech, even on the subject of his hometown, a topic that might be expected to be familiar to him. His struggle with it can be seen in the long gap of 5.6 seconds in turn 6, in which it appears as if he is subvocalizing to rehearse what he will say next. Still, developing a response is something that he has not previously done, and so does show development in his linguistic resources.

For most of the remainder of the test, as the conversation flows around him he joins in with laughter but makes minimal eye contact, and does not change his pose until almost the fifth minute when he folds his left arm behind his right arm, and he steeples his fingers of his right hand, perhaps a sign that he is ready to talk again. However, he lacks the ability to make space for his turn in the conversation, and it is not until the last few turns of the test that he is brought into the conversation, again by a direct question, again by C who asks for more information about his home town, as shown in Excerpt 11.

Excerpt 11: Kabuhito in the 3rd administration (person D)

Turn	Time	Person	
24	6:37.8	C	(5.3) The most problem th-(.3)thing (.9) which I experienced after I moved (.3) Chiba (.9) that (.6) mm is ah was (.7) I have to (1.3) buy wa-water [A: nn](.9) I have to buy water [B: ahhh]=
25	6:54.7	B	=Ahhh water is stinky=
26	6:56.3	C	=Yes stinky very stinky and in Nagano we can (.5) drink water (.4) from the (.9) tap [AB: nn] (1.5) but (.9) I don't want to drink (.9) wa-water (1.0) in Chiba [A: nn B:yea]. (1.9) I-(.3)is (.3) does your family (.5) buy (.6) water?
27	7:18.1	D	(.6) Yes my (2.1) my (.9) water is (1.0) come from (.8) ur [C: well?] [B: underground?] (2.8) underground [B: ah] [C:reallyhuhuh?](1.3) the (1.5) the(.2) the water (1.0) includes (3.0) groundo (C: ah) (4.3) sto-so-[AB: stones?] (.5) stones yes [ABCD: uhuhuh] small stones. [C: ahhh] (1.3) I-we couldn't (.4) drink (.4) the water [A: ehh]
28	7:53.8	C	(1.9) Were insects (.4) included?
29	7:56.5	D	(.3) No nono [ABCD: hahaha A:insects!]
30	8:00.3	B	(.2) So what (.2) do you use the water for?
31	8:07.4	D	(1.2) Uhh (2.2) baths [ABC ah] or (.8) um (.8) bathwater or (.8) wash [B: wash ABC haha]
32	8:16.9	C	(.7) Do you use (.3) that (.2) water (.4) for cooking?
33	8:21.2	D	(1.2) No [C: no ABC huhuhuhuh]
34	8:31.9	A	(10.3)I (.7) I've lived (.4) here for a long time [BC: nn] so (1.1) I (.3) didn't know the difference (.4) the (.8) water [B: ahhh] (1.2) for (.3) c- in the city and the (.2) countryside (.7) but after I (.8) grew up [C: nn] (.6) I [C: How did you eat?] (1.5) I (1.0) can know (.8) the difference [C: mm] (.3) so (.5) such as sweet (.2) or (.6) soft, (1.0) the hard. [D: nn]

As in the earlier administrations, he requires the support of the other participants in order to be understood. He makes heavy use of gesture in turns 27 and 30 by miming so the others can guess what he is trying to say. His laughter seems genuine, and immediately after this interaction even gives one of his very rare audible backchannels in turn 39, perhaps showing that he feels more like he has taken part in this conversation.

Saaya also shows some development in her interactional resources in her third group oral, but also some similar tendencies. She is still slow to engage, as can be seen by her being the last in the group to contribute her first turn, as she was in her previous administration. Her third GOT is dominated by two females who set all the topics, and of the five turns that Saaya takes in this test, three of them come after being asked "How about you?" as in Excerpt 12.

Excerpt 12: Saaya in the third administration (Person A)

Turn	Time	Person	
11	1:07.8	C	(1.9) I think the important thing is the contents of the (.4) work (2.) job (B: aha D: mm) (.5) so (.8) hmm (.3) I heard that Japanese student ah focus on ur (1.4) earning money (.2) too much [D: mm], (.3) so its not okay (.8) how urr (.2) how do you think so – (.9) about it, (1.8) – ah (1.5) important thing is (.4) – what is [important thing…
12	1:32.8	B	Important thing] (.7) I think important thing is (.2) of course money (.3) but um (.6) nn surrounding people,[D: mm] (.4) you know, (.6) work together, (.5) so it's important to (.3)

			cooperate with other people (.5) (C: um) yeah (.5) so surrounding people is, (.4) yeah important (.9) I think. [looks over at A]
13	1:52.8	C	(1.8) How about you?
14	1:52.5	A	(.3) Um (.8) I think (.6) um it is important that (.7) um (.4) money (.7) um and (.4) if (.7) um (1.0) one person don't have part time job (.6) no working part time job (.5) um (1.0) people (.6) um (.4) get (.7) money (.3) by their (.3) parents (.4) so it is not (.4) good things (.4) so (1.0) yeah, [D: mm] (1.8) how about you?
15	2:22.1	D	(.7) Um (.4) The important point for me (.5) is (.3) whether I can enjoy (AC: mm) (.3) um (.6) ah my (.3) my former (.3) part time job it was so boring (.3) and I didn't like it (.3) but (.4) now my job is so fun (.5) and I can (.5) I could (.2) I could meet (.5) many people in (1.0) different ages or (.6) different (.2) type of students (.3) so its very (.6) um (.4) good for me [B: mm]

After B takes her turn, the expectation is that Saaya will have her turn, but it is not until C asks the transfer question in turn 13 very quietly that Saaya starts talking. As in her previous GOT, she shows a low level of accuracy and many hesitations and false starts. She shows development in her interactional resources by improving her ability to engage in the conversation as a listener as she maintains eye contact with the other speakers more often, nods her head at what they say, and gives more frequent verbal backchannels. The evolution in her identity as a more active participant can be seen in her attempt to introduce a topic into the conversation, as shown in Excerpt 13.

Excerpt 13: Saaya in the third administration (person A)

Turn	Time	Person	
24	4:28.7	B	(7.042) How do you use your money?
25	4:36.2	C	(3.590) Mm Shopping [ABCD: huh huh] or for food [B: aha] I'm living with I live with my family [B: yeah] so [B: hmm] I don't need too much money [B: ahum] for myself [B: yea] so the problem, point is er if you living with your family or yourself is er point I think so huhuh [B:huhuh]. What is your purpose?
26	5:19.7	B	(1.474) Shopping, maybe same things. But er I have a car (C: oh
27		D	your own car?
28		B	yeah so I have I have to go you know gasoline (A:nn D: un)
29		C	so you have to pay=
30		B	=yeah money for car
31	5:43.8	A	(1.7) Maybe I want to buy car [B: ahuh] (.6) yeah
32	5:48.2	B	Yeah I want new one but it's so [expensive huhuh
		C	[so exp—huhuh
		A:	[huhuh]
33	5:51.1	C	you have to earn money more
34	5:52.9	B	yeah

In the exchange in Excerpt 13, Saaya initiates a turn for the first time in three administrations. After C tells the group in turn 26 that she has a car, at the next available TRP Saaya takes the opportunity to state that "maybe" she wants to buy a car, which she does while turning to talk to B.

There is the potential for this to become the next topic of conversation; group members could ask what sort of car she would like to buy and so on. However, the very tentativeness of her statement is characteristic of the ‘powerless’ style of speech (Bradac & Street, 1989, p. 203), and perhaps because it is not original enough (since ‘desiring a car’ is probably not unusual among university students) it fails to pique her groupmates’ interest sufficiently for them to ask a follow-up question. Rather, it is overshadowed by B whose preparatory ‘yeah’ is not an acknowledgement of the establishment of Saaya’s topic (Geluykens, 1993, p. 192) but a prelude to her own new topic, in which she raises the stakes by stating her desire for a *new* car, and this is ratified by C’s acknowledgment in turn 33. B’s turn 32 is said after turning to make eye contact with C, effectively cutting off Saaya, and C is the first to respond by chiming in with a helping move (Gan, 2008). Although Saaya’s initiating of a topic was not sanctioned by her GOT groupmates, this greater willingness to engage in the conversation allows her the identity of ‘potential active participant’ who can be called on to contribute by the more active group members, and can put forward topics herself that with more sympathetic group members could sustain discussion.

It is interesting to note that she synchronizes some of her gestures with participant B’s, following a second or less after her. This occurs while another participant is talking: at 0:48 B uses her left hand to brush her hair and Saaya’s left hand goes to her hair in a synchronized movement. Later at 2:40, again while listening to another participant, she mimics B’s pose of having both arms on the desk in front of her. Since person B is one of the two dominant participants, it could be speculated that this mimicry is an attempt to collude or indicate affinity with B, perhaps in order to gain floor space from her. As further evidence of this interpretation, Saaya’s initiating move, as described above, came on the tails of B’s car topic, and she finishes her turn with her gaze on B. However, B effectively ignores these approaches, and as B occupies the topic with her over-riding turn 32, her gaze turns to C, who gives an acknowledgment token and the conversation develops from that point.

The final person who was among the low scoring students in the first administration is Machiko. As for her GOT in the second administration, she starts the discussion by setting the first topic. It appears that she is attempting to continue her identity as ‘leader’ from the second administration. However, it seems she does not quite have the linguistic resources to be able to dominate in this GOT, as can be seen in Excerpt 14.

Excerpt 14: Machiko in the third administration (person B)

Turn	Time	Person	
7	1:12.3	B	(.5) Where do you buy?
8	1:13.8	D	(.8) Mmmm (.4) I often go to Chiba. [B: ah]
9	1:16.9	C	(.7) Oh yeah me too! =
10	1:17.8	D	= Oh really? Huh [huh]=
11	1:18.3	C	=[Yeah] because there are so many shops, [D:oh yah-yah-yah] and department stores [D: ahum] yeah [B: mmmm]
12	1:23.9	B	(.3) Yes yes yes…(.8) mmm so…
13	1:27.8	C	(.5) How often (.2) do you go (.3) shopping?
14	1:29.9	B	(.2) I (.7) I go to shopping per month, (.6) once per month [CD: mmm C:me too]
15	1:37.5	A	(2.7) Per..(1.0) In my case (.7) I (.6) go shopping (.3) per week [D: mm] (.6) but (.2) mm I (.4) don't (.3) buy (.4) many things (.6) because I don't (.3) have (.2) eno- (.2) enough money [BCD: mm] (.5) so (.2) I often buy (.6) very cheap CDs (.2) second hand [BCD: mm]
16	1:58.1	B	(2.8) How about you?
17	1:59.2	C	(.6)Um I go shopping maybe (.4) twice a month [B: oh really] or some-(.2)thing like that (.5) but (.9) ss um (.4) I don't buy sometimes I don't buy anything (.2) just go and look at the shops (B: yeah D:mm ha ha) (.6) and (.2) yeah (.2) and if I find something really good or something (.2) I (.4) like (.6) I buy yeah
18	2:18.2	B	(.4) Yes me too, I like just looking (.5) clothes [C:yes D: ha ha)

In turn 7 Machiko sets the topic of location for shopping. In the exchange between turns 8 to 10 it becomes apparent that C and D are proficient speakers who quickly develop a rapport with each other. Their proficiency is displayed by their speed of speaking, the latching of turns 9 and 10, the overlapped words between turns 10 and 11, and the exuberant backchannels they supply each other – a commendable performance of "doing interactive" (Lazaraton & Davis, 2008, p. 321). In turn 12, Machiko creates space to take her turn again but lacks the linguistic resources to generate a new topic quickly enough to sustain her leadership, and C steps in to set the next topic by asking Machiko a question. However, when A does not clearly indicate the next speaker at the end of her turn, Machiko is the one who returns the favour by asking C a transfer question to maintain the topic. Throughout this exchange, eye contact is maintained with the other speakers and in the entire test recourse to the prompt paper is far less often than in previous administrations.

Another element that seems to have improved markedly in the third group oral is the elapsed time between her response and the previous speakers. In the second administration, in conversation within a topic, her reaction times most often ranged from 0.7 seconds to 1.2 seconds, but here the gap is usually between 0.2 and 1.3 seconds, showing her improving ability to access and use appropriate linguistic resources in conversation. However, in this test she does not show in this test the ability to predict what the other speaker says and overlap turns, as C and D clearly show they can in lines 9 to 11 of Excerpt 14.

4.2.1.4 Summary of the qualitative analysis of the initial low group

In summary, amongst the three who scored low in the first administration, there is a marked difference between Machiko and the other two members, as can be seen in the summary in Table 52. In the first administration, the most obvious sign of this difference is that Machiko shows her willingness to participate by initiating a topic. In the second and third tests she shows that she is inclined to take on a leadership role in the discussion, firstly by starting it, and then throughout the discussion by initiating topics, asking questions and transferring turns. That she can sustain it in the second administration but not the third is due to the higher proficiency of her groupmates in the final administration. The other two participants, Saaya and Kabuhito, show improvements in each administration in the way they participate. The first time they take the test they are reluctant to speak, but they improve their ability to respond to their groupmates. However, Kabuhito shows that even by the third administration he is still heavily dependent on his groupmates to make meaning.

Table 52: Summary of IC analysis of low initial scorers

	Admin.	Identity	Linguistic	Interactional
Kabuhito	1st	"Reluctant embarrassed participant"	Limited fluency: long pauses & hesitations, can self-correct pronunciation limited grammar & vocabulary, body language shows struggle	Avoids eye contact, responds, shows limited communication strategy
	2nd	"Reluctant participant"	Limited: long pauses, many hesitations, limited grammar & vocabulary but sufficient for short responses	Mostly avoids eye contact, responds, asks clarification question
	3rd	"Fringe participant"	Limited: Many pauses, uses mime to elicit support from groupmates, with support can produce longer turns	Some eye contact, limited backchannel, responds, asks confirmation question as part of pre-sequence
Saaya	1st	"Reluctant embarrassed participant"	Limited: only short responses, long pauses before speaking	Avoids eye contact, limited to minimal responses to avoid interacting

	2nd	"Reluctant participant"	Many pauses, shows ability to construct a long turn, many grammatical errors	Mostly avoids eye contact, limited verbal backchannel, single response
	3rd	"Potential participant"	Frequent pauses & hesitations, can construct long turns, many grammatical errors	Maintains eye contact, some verbal backchannel, responds, initiates a topic.
Machiko	1st	"Inexperienced, but willing participant"	Many pauses, can develop response in longer turns	Some eye contact, limited verbal back-channel, initiates topic
	2nd	"Leader"	Some pausing, gives extended opinions, asks questions, does not always understand what others say	Maintains eye contact, uses verbal backchannel, transfers turns, refers to others, asks follow up question, initiates topic
	3rd	"Leader"	Some pausing, gives extended opinions, asks questions, takes short turns	Maintains eye contact, uses verbal backchannel, transfers turns, asks follow up question, Initiates topics

Saaya does not require support to take a turn as Kabuhito does, showing that she can speak in an extended turn in the second administration, and continues this in the third administration by taking two long turns. In terms of her ability to manage conversation, in the third administration she shows that she is capable of initiating a topic, albeit unsuccessfully on this occasion. The next section investigates how the quantitative statistics gathered in Chapter 3 played out in the individuals in the medium initial group.

4.2.2 Qualitative analysis of initial medium group

4.2.2.1 Initial medium group in the first administration

The two females in the medium group show a greater willingness to engage in their first discussion than the low initial group members, though in the first test they too spend a large proportion of their time with their heads down staring at the prompt. They both show their ability to speak in longer turns and engage with the other test-takers in the group by asking questions. The first participant, Aemi, asks four questions in the discussion in order to transfer the turn (twice) and asks follow up questions (twice).

Excerpt 15: Aemi in the first administration (person C)

Turn Time Person
8 2:16.5 A (11.1) Have you (.8) have you even been to abroad?
9 2:20.4 C (.2) no, no I haven't. (2.0) But I'd like to (.8) go overseas (.7) some day. [A: huhuh]

(1.0) How about you?
10 2:31.6 A ah I went to (1.2) I went to (1.2) America -New York (1.4) um… (1.9) last (.6) last month (3.0) I… (2.1) went there with my boyfriend um (1.3) umm… (2.4) travel's (.4) purpose (.5) is (.3) too meet (.5) his (1.4) his parents (1.8) um I (.6) we stayed (.6) at (.8) his (.6) parentses (.6) house. (2.7) Ah seven days.
11 3:16.2 C (2.6) how long (.6) did you stay in there?
12 3:20.6 A (1.8) ah (.2) 10 days.(1.9) (huhuhuh)
13 3:31.4 C (7.4) Yuko, how about you? (breathy laughing)
14 3:33.9 B (1.0) ah (1.0) I never been overseas.

The topic starts with A's question, which Aemi shows she has sufficient linguistic resources to answer with accurate grammar, even if it has considerable pauses in it. She then displays her interactional resources by transferring the turn back to the questioner, asking a follow-up question to the answer and then transferring the question to the other participant. At .2 seconds, the elapsed time between being asked the question and the start of her answer is relatively quick, especially when compared to other low initial scoring students in the first administration. By contrast, the longish times between the preceding speaker and her questions in turns 11 and 13 show the lack of competition for the speaking floor that is typical throughout the first administration. In any event, Aemi asserts an identity as a leader over this section of the conversation, and is particularly shown by her authoritative use of B's name in line 13. This section shows her potential, though since she does not maintain it, nor initiate any new topics in this test, her overall identity can be described as 'active participant.'

Of particular interest with Aemi, is very clear body language that is consistent over the three times that she takes the test. When she is not speaking, her hands are by her sides, under the desk, and when she is speaking, or has something to say, her hands are on the table; as soon as she has finished speaking, they drop to the sides again. This consistent pattern is replicated in all her tests, and is useful for gauging when she has something to say but does not get an opportunity to say it. It probably acts as an unconscious signal to the other participants that helps them realize when she has something to say, or when she has finished speaking.

The other participant, Taeko, does not display her identity consistently in the first test, but does contribute two long turns and ask one transfer question, and in the limited time of this test, it is sufficient to sustain the identify of 'active participant'. Her first longer turn is in shown in turn 2 of Excerpt 16 below.

Excerpt 16: Taeko in the first administration (person B)

Turn	Time	Person	
1	34.6	C	(20.5) if you are (3.0) if you are going to a new place (.8) would you prefer to go on group (.6) tour (.7) or (.2) as an independent tourist?.. Eh
2	58.4	B	(9.9) I prefer to go on a group tour [C: mm] (1.5) because (.8) mm (2.0') if I (1.2) go to (1.7) ah (.3) if I travel (.5) alone (.8) I (2.7) mmm (4.9) I…
	1:26.2	C	(2.4) feel lonely [C:huh huh huh]
	1:26.9	B	(1.3) I feel lonely so (.3) I will (.5) go (.6) somewhere with my friends.
3	1:34.8	C	(1.1) me too. Me too
4	1:36.5	B	(1.4) how about you?
5	1:38.3	A	(.7) me too. (huh ahhh)
6	1:40.6	C	(1.3) how about you?
7	1:40.9	D	(4.0) me too.
8	1:42.8	A	(1.9) ah (3.5) hum? (1.0) Me too. (.3) When I went to Germany (.5) I wento with my (.8) cousin and (1.2) I didn't feel (.4) homesick (.9) or lonely (.7) so (.9) I prefer (.4) to go on [B: ahah] (1.2) with (1.3) a… group tour.

The start of this test was particularly awkward, with over 20 seconds of silence between the rater telling them to start and C's first question. In this silence, Taeko and C have exchanged glances and giggled, and since C is brave enough to start with a question read directly from the prompt, that eye contact established Taeko as the person to answer. As C started talking, Taeko is grasping the prompt paper with both hands, and does not look up at C until she says "tourist", then, her head goes down again and she speaks while looking at the prompt paper, with occasional glances up at her peers. She does make eye contact when she asks her question to A in turn 4, perhaps for the first time since the test began. As is common in the GOTs in the first administration, Taeko makes little eye contact with the other test-takers in her group, spending most of the test with her head down, grasping the prompt paper in both hands, showing her lack of interactional resources to know appropriate communicative behaviour. In terms of linguistic resources, the language she uses is reasonably accurate, though she struggles for the appropriate words and incorporates C's collaborative help into her turn.

4.2.2.2 Initial medium group in the second administration

In the second test, Aemi shows impressive improvements as she is involved from start to finish. She takes the first turn in the test, asks follow up questions, initiates a new topic, takes a couple of long turns, and even corrects one of her peers, as can be seen below in Excerpt 17.

Excerpt 17: Aemi in the second administration (person B)

Turn	Time	Person	
7	21:55.2	B	(1.7) I want to (.3) get married (.2) in the future (.7) [A; hm] because (.2) I have a boyfriend [AC: hm] (.4) wh- (.7) mm (.3) his family and my family is (.7) have a good (1.0) relationship [AC haha] (.8) mm (.4) I often go to his house [A: hm] and he often come to (.5) comes to my house (.7) so (1.2) I want to live with (.5) him (.4) and I want to (1.0) have a baby [A: nn] so (.5) I think its (2.2) ah ah (.5) it's my dream (.5) so (.9) I want to get (.4) married [C: ahum]
8	22:40.8	C	(3.8) Sooo …
9	22:43.5	A	(2.6) I want to baby too
10	22:46.4	B	(1.8) Have a baby
11	22:47.2	A	(.2) Have a baby yes

The long turn that she takes here shows improvements in her linguistic resources over the long turn in her first test in terms of the length of the pauses, which with one exception are a second or under. Also, this excerpt shows her correcting A's ungrammatical use of 'baby' as a verb. This is a positive sign since research has found that language learners who give negative feedback are more likely to make improvements in the target language (McDonough, 2004). Corrections of any kind in the administrations of this test are extremely rare, as might be expected given that it is a potentially face threatening thing to do in an assessment situation with one's peers (Luk, 2010). To do a correction in this assessment situation is a display of "doing assertive" (Lazaraton & Davis, 2008, p. 326) for the raters to take note of, and hopefully reward with a higher grade.

Throughout her second GOT Aemi's interactive resources have developed substantially from that witnessed in the first test. She is engaged throughout the entirety of this test, her eye contact is nearly always on the other speakers, she supports the speaker by verbal backchannels and nodding, and her body language is open and positive. The trait noticed in her first administration of having her arms on the desk in front of her when she speaks or wishes to speak is noticeable here: her arms are mostly on the desk, only falling to her side after she finishes a turn. With this performance, Aemi positions herself as 'leader' of the conversation, with an authority over her linguistic resources that allows her to correct her peers.

The other member of medium initial scoring group, Taeko, also participates more actively in the second administration than her first by contributing three long turns and asking two follow-up questions, as well as engaging more with her fellow participants by non-verbal and verbal backchanneling, with much improved eye contact and more open body language. Excerpt 18 shows her follow up questions and her longest turn.

Excerpt 18: Taeko in the second administration (person C)

Turn	Time	Person	
4	1:49.9	A:	(.7) Uh me too I want (1.1) to (.2) I want marry (.6) but ... (2.2) if (.6) if I marry someone (.9) um (.4) I want to marry (1.3) when I ... ...(4.6) twenty (.2) eight (C: twenty-eight) (1.3) more than twenty years
5	2:15.7	C:	(.6) why? (B:mm)
6	2:17.0	A:	(1.1) Because I (.5) if I (.4) marry (.9) marry (1.3) I (1.2) always (1.4) home (.5) home homeworking ...(1.4) homeworking? [C: mm]
7	2:33.1	B:	(2.9) House - housework
8	2:34.9	A:	(.9) Ah housework (.9) for example (.6)[washing
	2:37.7		C: so [you
9	2:39.8	C:	(1.4)You don't want to do (.8) such a thing?
10	2:43.0	A:	(.3) Un [BC hhh hh]... ... (8.2) How about you (.7) when (.6) do you want (.5) want to marry?
11	2:56.5	B:	(1.1)Me?
12	2:57.7	A:	(1.0)Yes
13	2:58.0	B:	=(1.8) Uhh ... Uhh.. (.5) I think (4.4) ... ... uh (1.6) about thirty (1.5) maybe (.6) so ...(1.9) uh ... (2.7) if if I get married (1.2) so so early ... (3.9) that marriage (1.9) may (1.6) may become ...(3.2) get worse (3.0) uh uh (.4) if the marriage (.7) was mistake (.9) uh (2.4) its fucked [AC: hahah]... [C: mm] ... (8.0) how do you think?
14	3:54.2	C:	(.2) Ah I think I wanna (.9) marry (.6) about thirty (.8) because (.4) er (.9) I'm (.2) in here to study (.4) and (.6) I wanna (.7) get a job (.6) and (.5) if I marry (1.0) nn (.5) after eh? (.9) if I marry (.4) bout (.2) twen-twenty or twenty (.6) one or twenty two (.5) um (1.7) ah (.5) eh-h-h-toh (1.1) it (.4) it has no meaning (.7) to come(.6) to university (.4) I think (1.0) un (2.1) I wanna (1.1) continue (.7) the job (1.0) after (.2) ah (1.6) ehh (.3) nothing [AB hahah] sorry

In this excerpt, Taeko's interactive and linguistic resources are deployed as she effectively uses both simple and more complex forms of follow-up questions in turns 5 and 9 respectively. Perhaps because she is under some competitive pressure to get her turn in, she simply asks "why" in turn 5. Although being the last to backchannel in the previous speakers turn, as she does when she echoes "twenty-eight", is often a signal that the speaker will take the next turn, it is not definitive. Thus, she may well have felt it necessary to speak quickly lest she lose her turn to another. By contrast the floor for her turn 9 was guaranteed by her false start which overlapped with the end of A's turn 8. The false start is a clear indication that she would take a turn so she did not need to come out so quickly with turn 9.

Later on in turn 14 her long response gives an indication of the difficulty that the test-takers had marshalling their linguistic resources to deal with the 'singles' topic (Leaper & Riazi, 2014). As can be seen, this turn is marked by many pauses as she struggles to state reasons for her choice of appropriate marriage age, and it ends in an unconvincing way as it seems she abandons something that she was going to say. This excerpt provides a good sample of her interaction during this GOT, as she capably manages to state her opinion without attempting to lead or control the discussion, as such she continues her identity as an 'active participant' in the GOT.

4.2.2.3 Initial medium group in the third administration

The third time the two participants in the medium group take this test they seem to take different paths. While Aemi seemed to be enthusiastically engaged in the second administration, her body language in the third test is much more relaxed, with her arms dropping to her sides rather than being mostly on the desk. Her participation is somewhat sporadic: she takes the first turn in the test, but then for almost four minutes in the mid to late periods of the test does not take a turn, although she does use verbal backchannels to show support for the speaker. Excerpt 19 below shows her in one of her active periods.

Excerpt 19: Aemi in the third administration (person B)

Turn	Time	Person	
14	2:17.4	C	(2.3) So..(.9) what (.2) so what do you focus on, (.3) to (.4) find your (.2) part time job?
15	2:23.3	A	(2.5) Salary, (C: salary - yeah of course ABD: hahaha) (2.2) and my location (1.1) location (C:aha D: nn) (1.3) I if (1.7) I would do (C:uniqlo) (1.1) by bicycle [C: aha D: nn] (.9) but it takes only 3 or 4 minutes (C:eh that's quite good) (1.6) so it's good. (1.7) How about you (to B:)
16	2:43.8	B	(.1) Ohh (.5) I start (.2) work (1.5) er (.2) start work (.5) yesterday (A: oh C:okay) (1.2) I (.2) work at (.2) clothes shop (C:ahem)(1.0) so it's so interesting (.4) and (.2) until then I don't have (.3) I didn't have any jobs [C: ohh](.4) because I don't have (1.3) free time (.4) to work (A: oh)
17	3:06.3	C	(1.1) so you're a hard(.6) hard worker
18	3:08.3	B	(.1) I want to focus on (.2)studying C: [studying]
19	3:12.2	D	(1.7) So (.2) is this (.3) the first time to (.2) work?
20	3:15.8	B	(.8) Er no, (.7) so, (.2) when I was a (.9) urrr freshman [D: nn] (.3) I work (.6) at (.5) shoe shop (1.3) [D: nn] so (.2) but (.4) I can't (.5) I can't make (1.2) time (.2) to work [C: I see] (1.7) so I retired
21	3:38.7	A	(3.1) Ehh so…
22	3:39.4	C	(.4) So (1.1) what do you think about your owner
23	3:43.4	B	(.8) owner?
24		C	=This time (AD: hahah])
25	3:46.0	B	(1.4) owner?

26 3:46.5 C (.2) yeah=
27 3:46.5 B = ahh (1.0) I think I thought (.7) she is good

Unlike the second administration, her participation here is mostly reactionary, though she does find it necessary to ask a checking question to C to query what he means by "owner". Although she clearly understands that he means "boss", she does not make a correction, as she did in the previous administration. Instead she avoids giving feedback, something that has been noted as common in NNS-NNS interaction in the literature (Fuji & Mackey, 2009). Also noteworthy in turn 16 is her strategy of making an initial verbal reaction in her answer before starting to talk. Not only does this give her more time to formulate her reply, but her initial "ohh" acts as a placeholder that signals that she is taking the turn. This is a common technique that occurs more frequently in the later administrations than in the earlier ones, and is likely an interactive resource that the test-takers have transferred from discussions that took place in various classes at university. For example, in Aemi's GOT in the first administration, this feature can be found in turn 10 of Excerpt 15 by participant A, who is the highest scoring student in that group.

Although Aemi's participation in this test is somewhat less than in the second test, she shows improvement in her linguistic resources by her faster reaction times between her turn and the preceding turn. Though even in her first administration she was capable of reacting within a short time of the previous speaker's response, in the third administration she more consistently responds within a second of the previous speaker's question. However, she still shows no evidence of being able to predict the current speaker's words and complete the sentence for them, as C almost does for her in turn 18 above. In her linguistic resources, she shows again that she can use complex sentences to create meaning, though she shows some inconsistency in tense use in turn 16 of Excerpt 15, though she manages to self-correct herself once.

In terms of identity resources, in the final test, Aemi retreats from her 'leader' role, back into the 'active participant' role that she played in her first administration; Taeko, on the other hand, engages far more actively in her third administration. She dominates the conversation in this GOT, being constantly involved throughout in numerous ways. A good example of this is in Excerpt 20 below.

Excerpt 20: Taeko in the 3rd administration (person B)

Turn	Time	Person	
9	1:55.4	B	(.8) But why do you work?
10	1:58.0	D	(1.3) What?
11	1:58.4	B	(.2) Why (.2) do you work?
12	2:00.2	D	Cook
13	2:01.7	B	(1.0) Eh (.3) Why(.5) Why (.6) do you have a part time job? Sorry.
			D: [Ah part time..]
14	2:05.7	D	(.3) Oh (.6) because (1.7) it's (.6) very hard to (1.9) to (1.4) buy or er (.8)
15	2:13.6	B	something to eat huhuhh
16	2:17.4	B	(.2) so you need money
17	2:18.5	D	Yeah (1.1) or I have so (.7) as I said (.6) I joined to soccer club [B: ahum] so (1.2) after the club we (.7) go restaurant or (.5) bar (.7) so
18	2:31.1	B	(.6) very busy=
	2:36.4	D	=[yeah] ah, very (.3) yeah. (2.3) Why (.6) do you work?=
19	2:38.3	C	=umm (2.0) you know, as you know the school fee is (.2) very expensive (.4) and (.7) um (.6) my parents (1.3) pay (.4) all of the school fee [B: mm] (.4) so (.7) I think (.5) its (.4) I'm (.6) really sorry for (.4) if I (1.0) sorry for my parents if I (1.2) ask them (1.6) some (.2) money (.2) to (.2) for me. (2.5) So (.5) to (1.2) to earn money then I (.2) can use (.4) for myself (1.0) I (9.4) have a part time job [B: mm]
20	3:21.5	B	(4.2) How about you!
21	3:22.2	A	(1.8)Umm (1.1) I'm also (.2) to (.4) I'm also in club (.2) and (1.1) yeah we go (.2) to after we finish we go to restaurant and (.7) and I have to (.2) earn some (1.0) money for playing
22	3:39.1	B	(.6) playing?=
23	3:39.5	A	= Mmm (.7) yeah so (1.7) I'm doing (.2) part time . (3.3) How about you?=
24	3:47.8	B	= I (.7) I also work for money. (.5) I live in Yokohama. (.4) So (.9) er (2.2) I need (1.4) a lot of money (1.7) um (1.0) for (1.0) for (.2) nanyakore for train, (.3) train fee (1.0) and (1.0) ah I have a sister (.3) so (1.0) ah my (1.1) I (.3) can't ask (.5) my parents (.7) to give me money so I have to (1.4) earn (.4) money (1.0) by myself mm.

This excerpt shows an uncommon occurrence in these tests: the verbal expression of a misunderstanding. Turn 9 is the first topic after the initial opinion giving stage, and it is initiated by Taeko. When participant D fails to understand it, Taeko successfully deploys her linguistic resources to rephrase the mishearing in turn 13. Turn 16 also performs the function of summarizing D's situation, another feature that appears more often among advanced learners and later administrations of this test. Indeed the "summary recap and reformulation" has been found to be a feature of "educated discourse", which teachers often use when leading class discussions (Mercer, 1995; p. 95). Taeko also shows she has developed an ability to predict what her interlocutor is saying in turns 15 and 18, though not always successfully as turn 15 shows. Her response times are often less than one second and she demonstrates she is capable of latching on to the previous speaker's utterance in turn 24. This is a clear example of a test-taker "doing interactive" (Lazaraton & Davis, 2008).

In her long turn 24, it is notable that when her linguistic resources momentarily fail her when she cannot access the word 'ticket', she uses a Japanese filler ("nan ya kore") rather than an English one. Taeko shows she has the linguistic resources to engineer meaning by finding a substitute that successfully conveys the meaning, "train fee", even if it is not exactly correct. Throughout the test she shows well developed interactional resources by maintaining eye contact with the speakers and giving verbal backchannels, as evidenced in turns 17 19 and 21 to her peers D, C and A respectively. In the final administration, Taeko successfully displays her competence as a language learner, taking on an identity of 'leader' in the process.

4.2.2.4 Summary of the qualitative analysis of the initial medium group

The performance of the initial medium group tests is summarized in Table 53. Both members of this group show strong improvements in the second administration, but Aemi does to a greater degree than Taeko by taking on the identity of leader. In the third administration Taeko improves markedly, showing the ability to use some advanced turn-taking features, while Aemi does not seem to put the same effort into her performance as she did in her second test.

Table 53: Summary of IC analysis of medium initial scorers

	Admin.	Identity	Linguistic	Interactional
Aemi	1st	"Active Participant, potential leader"	Some pausing in speech, quick reaction times. Sufficient grammar to construct longer turns with good accuracy & question others opinions	Some eye contact, responds, transfers turn, asks follow-up questions, uses other's name
	2nd	"Leader"	Little pausing, can develop opinion using some complex grammar, can form variety of questions and lead discussion	Maintains eye contact, uses verbal backchannel, transfers turns, follow-up questions. initiates topics.
	3rd	"Active Participant"	Little pausing & quick reaction times, uses complex time clauses, but not always with accuracy, asks checking question.	Maintains eye contact, uses verbal backchannel, asks clarification question

Taeko	1st	“Active Participant”	Considerable pausing, Can speak in longer turns with some accuracy, with support forms a conditional, forms a simple question	Little eye contact, limited use of backchannel, transfers a turn
	2nd	“Active Participant”	Considerable pausing in speech, can speak in long turns, uses variety of question forms for a different purposes,	Maintains eye contact, uses verbal backchannel, can transfer turns, ask follow-up questions
	3rd	“Leader”	Some pausing in speech, can speak in long turns, can ask questions, can rephrase questions	Maintains eye contact, uses verbal backchannel, summarizes, completes others sentences, initiates topics

4.2.3 Qualitative Analysis of the initial high group

4.2.3.1 Initial high group in the first administration

Of those in the high group in the first administration, the most immediately noticeable feature of their participation is early involvement in the discussion. Of the three participants in the initial high scoring group, two are the first to speak, and the other is the second to speak in their respective GOTs. Compared to first administrations of the other groups, they are more likely to initiate new topics and their turns have fewer long pauses and longer runs of words between pauses. When they listen they have more positive ‘ready’ body language, better eye contact, and provide some verbal feedback.

The first of this group, Naeko, shows limited interactional resources by having just four long turns throughout the test, of which one is used to start the test, and another to restart the conversation after a silence of 21.4 seconds, and the other two are answers to questions. However, she produces no follow up questions, and has no short turns at all. Excerpt 20 shows Naeko restarting the test after a long silence. In the silence there was an exchange of eye contact with C, and then Naeko looks down to her prompt with her hands under the desk. Without raising her head, her left arm comes out to be on the desk in front of her, palm down, and a few seconds later she starts talking in turn 12. This body language here clearly signals that she is about to start talking. When she says “culture” she finally looks up and makes eye contact with C who gives a small kind of laugh. Apparently C requires a question to engage, and after a silence of 6.6 seconds, Naeko obliges. Just before D starts speaking in turn 17, Naeko’s hand once more goes to the surface of the desk, perhaps indicating that she had something to say, but after D says “difficult” it disappears under the table again, perhaps showing that she considered her speaking opportunity was lost. Naeko shows initiative by starting topics, but she hardly seems to enthusiastically embrace the situation. Her role seems to be more one of ‘leader by default’ since nobody else in her group is willing to do it.

Excerpt 21: Naeko in administration 1 (person A)

Turn	Time	Person	
12	9:37.0	A	(21.4) Working (.2) abroad is (.6) good but (.6) ahm (1.1) little (.3) bad (.2) point (.6) mm (.2) is (.6) ah (.8) we we may forget our (.3) Japanese (.2) own culture. [C: hahem] (6.6) Do you think (1.0) do you think (1.5) mm studying (.4) studying? (.2) working abroad is (2.1) …a- abroad has bad points? [C:hm]
13	10:15.3	C	(5.5) I think no. [A: aha] (2.0) Through working (.2) in overseas (.6) we can (.6) tell the Japanese culture (.6) to other countries’ people.
14	10:27.8	B	(.7) but (.7) there are many crime [C: mmm]
15	10:32.5	D	(2.6) and some (.4) dangerous.=
16	10:34.7	B	= yes, (.7) terrible.. (4.1) There… (2.2) many (.2) terrorism (.7) in (.2) all over the world. (2.9) What do you think?
17	10:54.3	D	(4.1) um (1.2) difficult question. (1.3) Actually in these days terrorists (.3) are very dangerous (1.4) and (.4) hmm… (6.1) so we must protect (.7) myself [B: hhmm] (1.9) but (.3) it’s very difficult… (13.5) But ah (.8) working abroad (.8) have many (.3) good (.3) experience.

Some of Naeko’s linguistic resources can be seen in this extract. The number and length of the pausing, and the frequent maze words show the challenge she had composing this turn. Also of note in turn 12 is her ability to self-correct both lexical and grammatical items as she changes “studying abroad” to “working abroad”, and then “is” to “has”. In both cases her correction results in the correct form.

Yamahiko, by contrast, shows he has well-developed interactive resources by being active throughout the discussion by maintaining positive body language and eye contact with the other participants, and only occasionally referring to the prompt paper. This group oral is characterized by having many short turns, which is unusual in the first administration. As can be seen in Excerpt 22, Yamahiko seems to have a relaxed attitude to this assessment situation.

Excerpt 22: Yamahiko in the first administration (person C)

Turn	Time	Person	
1	10:47.1	B	(1.1) so (.7) do you want (.3) want to live (1.3) and wowo-work abroad?

2 10:55.1 C (1.4) uh (1.2) I actually want (.2) to (.3) work (.5) somewhere (.8) not in Japan. (1.4) But urr (.7) depend like family. (1.1) I don't (.4) want (1.2) for my (.4) wife. (B:ahh how about.) (.8) huh (1.0) I like Japanese girls. (B: indistinct)

3 11:14.8 A (1.5) I (1.1) I want to live and work in Japan. (3.7) I find (1.2) very lonely (.2) for me.

4 11:27.9 C (1.7) Us two can make (.3) good friend. (5.6) ohm (1.9) heh (1.5) so what (.5) how about you?

5 11:35.9 B (.4) ah (.6) ah actually (1.0) I don't want to (1.5) work abroad (.8) because (4.3) Ja- ah(.3) I I like Japan =

6 11:50.7 C = Yeah..(B: hahah ha) (3.4) hmm (4.3) So (.7) if (1.3) you go to (.9) right (.3) let's say (.8) you get (.3) work holiday, (B: mm) (1.5) mm (1.0) which country you wanna go?

7 12:13.3 B (1.8) Mmm (1.7) ah (.2) last year I've been (1.3) in Australia [C: yeah] (.8) uh for (.2) ten months. [C: yeah] (.7) For studying abroad. So (.6) I also want to go (1.5) to (1.0) Australia again (1.0) for (.2) study abroad or (1.0) something.

8 12:35.4 C (1.8) do you wanna go somewhere?

9 12:37.0 A (1.0) I want to go Australia. [C: okay] (.5) My friend is (.3) home staying (.2) there [C: oh yeh] (.6) so I want to see her. [sniff]

10 12:43.9 B (.7) how about you?

11 12:45.2 C (.5) I (1.0) actually like to go (1.6) Ireland or somewhere [b: oh] (1.9) I've been to (.5) America (1.9) so [sniff] (.5) I've been to Australia. (2.3) I don't like (.2) I don't like (1.4) the (.6) United Kingdom that much (B: ahh)(.4) so huh (.5) I wanna go to Ireland.

12 13:09.7 B (.7) Why why don't (.5) don't you like (.8) umm United Kingdom?=

13 13:13.8 C = ahah no (1.1) they just seems (1.1) little (2.4) ah Huhuh [B: ah haha] (1.3) I think (.4) it's (.3) hard to get along with them. [B: ahh] (.5) I haven't even met them so (.7) Just my guess…huhuhuh [B: ahemm]

Yamahiko's overall ease of using his linguistic resources is shown by his casual manner. In Yamahiko's first turn (turn 2) he reveals personal information of his preference of Japanese nationality for a partner, and when the next speaker, a female, admits that she prefers to stay in Japan, he flirts with her "us two can make good friends" (turn 4), to which A does not seem to react, nor does it do much to lift the overall serious atmosphere. His advanced linguistic resources are also displayed by his use of conversational contractions such as "wanna" (turn 8) and fillers such as "right, let's say" (turn 6) and "like" (turn 35 below), as well as correct use of present perfect in line 11. It is also noticeable that rather than use the typical and generic "How about you?" to exchange turns in turn 8, he makes the transfer question relevant to the topic to perform the same function. In his longer turns he shows the ability to put together runs of 5-6 words between pauses, though he still has some long pauses of over two seconds when he talks. What he says is also perhaps deliberately atypical to what most other students say in these tests. For example, in turn 11 he gives Ireland as the place he would rather go, and takes the virtually heretical position that England is not a country he favours. When B asks him about it, he is forced to admit that this is based on conjecture.

As noted above, this conversation has an unusually large number of turns, and as the test continues the pattern of Yamahiko asking a question, receiving a short reply, then Yamahiko transferring the question to another speaker asserts itself as is shown in Excerpt 23 below.

Excerpt 23: Yamahiko in the first administration (person C)

Turn	Time	Person	
33	16:09.8	C	(4.0) ah huh huh (3.4) so (2.7) what country have you ever been?
34	16:13.8	B	(1.2) Ah just Australia.
35	16:15.9	C	(.8) I went to Australia like (.8) two weeks ago. (B: oh) (1.4) yeah it's cool.
36	16:21.3	B	(.6) oh (.8) so (.6) where in Australia?
37	16:25.1	C	(.3) I went to Sydney and (.5) Wollongong (.8) Wollongong?
38	16:30.2	B	(.5) Ah I don't know.
39	16:31.7	C	(.2) ah (.3) it near Sydney. [B:oh] (1.2) How about you?
40	16:34.7	A	(.3) I have never been abroad.
41	16:36.9	C	(1.2) never?
42	16:37.5	A	(.2) never.
43	16:38.7	C	(.9) English pretty good though. [AB ahaha] (4.9) So you went to study abroad?
44	16:46.3	B	(.3) Yeah
45	16:47.1	C	(.4) Australia
46	16:47.9	B	(.2) yeah (.2) ah do you know (.3) Adelaide city? (1.2) Adelaide?

Yamahiko dominates this section by asking questions or by making statements that his groupmates ask questions about. He takes every second turn here which is the prevailing pattern in this discussion. The only exceptions in the entire test were when B transferred the topic to A twice. This is a very clear example of "sequential dominance" whereby the discussion is controlled by initiating moves by the speaker and the responses comply by providing the information requested (Itakura, 2001, p. 1864). Yamahiko further shows self-confidence in this role of 'leader' in turn 43 by assessing A's English as being "pretty good" despite not having the opportunity to study abroad. Especially by his use of this qualifier, Yamahiko suggests by implication that his linguistic resources are superior, and thus has the right to pass judgment on other's language ability. As such he asserts his identity as a leader.

The final person in this group, Hamako, also dominates her first GOT. She shows she has well-developed interactive resources by being active throughout her discussion. Not only does she take two long turns, but she also has shorter turns which she uses to transfer the floor, agree with other speakers and ask questions that are related to the topic but not merely read from the prompt, as can be seen in Excerpt 24.

Excerpt 24: Hamako in the 1st administration (Person C)

Turn	Time	Person	
7	2:32.3	C	(1.8) how about you?
8	2:34.3	A	(1.3) I prefer to go on trips (1.0) group tour (.9) because (.5) I don't know (.2) where (.2) to go, (1.7) the famous place. (2.9) So (.3) the guide… ..(11.6) when you (.2) new place is very dangerous (2.3) and… (4.0) if you go ah if we go (.2) new place… (6.6) language is different (.2) and…(4.4) cost much money. (mm)
9	3:32.5	C	(2.5) but (.2) I also (.2) I also think (2.0) traveling al- alone (.2) by myself is (1.2) sometimes good (.4) because (.2) if we (.4) you know (.6) that (.2) place or that country (.3) you can you can go (.9) wherever you want (.2) to (1.4) sometimes. (.9) [B:mm] If you if you really know (.7) where it is (1.5) and (.9) I (.4) sometimes feel (3.1) shopping by myself is comfortable. (.6) So maybe (1.3) it will nice to go (1.3) by myself (2.1) So I (1.4) I can't say which is (.8) better (1.7) it depends on the case (.4) I think B: [mm].
10	4:27.5	D	(1.3) I want to go on a on a group tour but (.6) I think (.7) if (.4) I think (1.3) ah independent tour is (.2) good (.5) to learn (.8) new language.[C: mm] (.8) I think.
11	4:45.8	C	(1.5) yeah I think so (.2) too.
12	4:49.1	A	(2.0) I think so too.
			…
13	5:20.6	C	(31.3) have you ever (.2) been to another country (.2) with (2.0) ah (.7) as a group (.5) tour? (mm) (1.9) Do you have any experience?
14	5:33.6	D	(1.5) I have been to (.2) Australia (.4) ah (1.1) studying language (.7) and home stay (1.2) un (.6) it was good experience to (1.0) to-o (1.0) to meet (.9) another new people (.3) and to (.4) study (.2) language and (1.1) Australian culture. (.8) It is (.4) it was good experience.
15	6:05.7	B	(1.3) I also have been to Australia (.8) with ah (.9) my family.
16	6:22.5	C	(10.7) Can you go (1.3) to Australia next time (.2) by yourself? (.6) Or do you want to go there with (.2) your friends?
17	6:31.9	B	(.6) humm (.7) with your friends hurh.
18	6:38.9	A	(3.6) I haven't been abroad yet. (.5) But (.3) I want (.4) go (.5) abroad (.2) and study.

This excerpt covers the middle portion of the test, and demonstrates Hamako's interactive resources. The topic is whether it is better to travel overseas in a group tour, and so far the consensus has been that group touring is preferable to traveling alone (including Hamako). The excerpt starts with Hamako using a transfer question to move the turn to A. The previous speaker finished her turn making eye contact with Hamako, who uses this speaking opportunity (turn 7) to transfer the question to A. When A speaks she makes eye contact with Hamako to transfer the next turn back to her, and

Hamako uses turn 9 to put forward her contrary opinion. The linguistic resources demonstrated in this turn are considerable. It is clearly signalled as a contrasting statement by “but”, and her available linguistic resources are shown by accurate use of the conditional and subordination. Although there some disfluencies (pausing and maze words), she shows she can consistently include five word runs to maintain the flow of her opinion.

Her opinion forces D to make an exception to her previous opinion that a group tour is preferable in turn 10. Having exhausted their opinions on this topic, there is a long silence of over 30 seconds before Hamako restarts on a related topic in turn 13 and develops it by asking a follow-up question related to the theme in turn 16, which also comes after a long break of over 10 seconds. Not only does she direct who gets to speak (turn 7), but she shifts the topic in turns 9, 13 and 16. The ability to shift the topic has been found to be related to the leadership role (Palmer, 1989). The other quality that marks her leadership role is her long utterances (turn 9 above was one of two in this GOT), which her groupmates do not come close to matching. The number of words spoken is linked to the quality of leadership (Stang, 1973), and in the literature has been used as a measure of ‘quantitative dominance’ (Itakura, 2001; O’Sullivan & Nakatsuhara, 2011). Such behaviour is a clear indication of Hamako’s identity as a ‘leader’.

4.2.3.2 Initial high group in the second administration

Naeko shows improved interactive resources in the second administration by contributing not only long turns but also short turns during the test. As for her first GOT she takes the first turn and continues to show initiative by setting most of the new topics in this group oral. In Excerpt 25, Naeko restarts the topic after a long silence of over 10 seconds by asking a question which is read out from the prompt. When she asks the question she spreads her eye contact to the three other participants equally. She agrees (turn 18) and then adds some factual information that supports D’s answer. These shorter turns function to maintain the conversation and were not present in her previous GOT. Before her long turn (23) she gives a gesture that signals her readiness to contribute which is to put the fingertips of her right hand on the desk and steeple her right hand.

Excerpt 25: Naeko in the second administration (person B)

Turn	Time	Person	
16	12:18.3	B	(12.4) Do you think (.2) the situation in Japan is changing?
17	12:22.2	D	(.6) yeah I think so.[B: nn] (1.3) Many woman (.4) woman (.5) get job .[B: nn] and (.6) try to (.3) enjoy their life (.3) as you said.
18	12:33.1	B	(1.3) I think so too
19	12:34.1	C	(.2) I think so too (.2) nn
20	12:37.8	B	(2.2) age of (.5) ma-marriage [A: nn] (1.5) uh (2.9) more (.2) older [A: yeah] become becoming more older.
21	12:49.3	A	(2.4) because many women (.6) um has (.9) has a pride of her (.4) job (.5) so (1.7) that situation (1.1) is happen.
22	13:05.8	D	(3.4) Um I think more and more (.6) women get (.4) get a job and (.7) after university your (.8) school (.6) and (.8) they will (.7) do (.6) um (.9) really great job (.7) in the office or (1.0) just (.4) their (.3) position will change. [A: ohh] (.5) I think so.[C: mnn]
23	13:33.4	B	(5.2) but I think Japanese system is not so good because (1.0) working mother (.9) um working mother (.4) have to (.4) ah always have to work (1.0) worry about (.6) their (.3) children (3.0) um (.5) If (.7) if they (.4) if they (.9) enter this (.2) their children to the (.4) kindergarten or (.7) something like that (.6) it (.2) tak- it cost (.6) much money.[A:nn] (1.3) So (.5) so (1.1) uh many (.2) wo- (.2) many many women (.5) don't want to have a (.2) child.[ACD nn]

Although Naeko is the leader of this discussion, there are long pauses before many of her turns of between 2 and 5 seconds. This is indicative of a group with a low level of competition (Galaczi, 2008), suggesting that Naeko's leading role may be due as much to the lack of assertiveness of her groupmates. She can lead here because no other group member is willing to step into the role.

Naeko's improved linguistic resources can be seen in the better developed long turns, though like her first administration she is constantly self-correcting herself, which produces a high level of maze words and reduces her fluency. Still, as before, her self-corrections result in more accurate formulations. For her vocabulary, although she uses natural sounding phrases like "something like that", and comes up with the lower frequency "kindergarten", she also mistakenly collocates "enter" with "kindergarten".

The next person in this group, Yamahiko, takes part in a group oral that has a similar pattern to the discussion he had in the first administration. He is the first to speak, and starts with enthusiasm, as can be seen in Excerpt 26.

Excerpt 26: Yamahiko in the second administration (Person A)

Turn	Time	Person	
1	9:47.1	A	(2.0) Let's get started (1.0) Ahm (.6) I'm sure you do but do you have mobile phone?
2	9:51.8	B	(.2) yeah,of [course.] (1.7) You too? C: [yes B:mobile] phone
3	9:55.0	A	(.4) I do too yeah (1.2) Ahum (1.2) Do you often use it?
4	10:00.0	B	(.5) yeah everyday.(.3) I love it. A: [you do] <?>
5	10:02.7	C	(.4) um me too.
6	10:04.5	A	(.9) so how much cost does it? (.8) How much cost <muttering under breath>
7	10:07.0	C	(.2) Cost three [thousand yen. A: [per per month]<check this sequenct
8	10:10.6	C	(.8) per [month] A: [three thou]sand yen=
9	10:11.0	C	= yes.
10	10:12.0	A	(.6) that's pretty cheap.
11	10:13.3	C	(.9) yes I don't often use it.
13	10:16.4	A	(.7) You don't. (1.7) Do you?
13	10:20.2	B	(1.4) un (.4) fif(.2)teen tho-(.3) thousand (.4) yen [A: eh?!] (.7) a month. (1.5) Because I love (.4) to [A: fifteen thousand per month] email (.2) to my friends.[A: ah] (.3) I love to contact with my friends (.2) always. (.3) All every (.3) every time (.5) yeah.
14	10:37.5	A	(1.2) un (1.2) My phone costs like (.9) seven thousand yen.
15	10:44.9	B	(1.9) Do you pay your…(1.0) bill?
16	10:48.0	A	(.6) yeah, (.5) I do (.3) pay mine. (1.5) Do you?
17	10:52.0	B	(.5) no
18	10:52.8	A	(.4) or your parents do=
19	10:53.7	B	= yeah hahaha [A: aha] [C: okay]

Yamahiko's initial turn is worth commenting on, since it shows the well-developed linguistic and interactive resources at his disposal. Particularly in the first administration, speakers often start by simply reading a question from the prompt aloud, or by stating their answer to it. His preface of "I'm sure you do but" is unusual in that it considers the state of knowledge of his peers, and is evidence of more advanced speakers being able to use available linguistic resources to consider the status of other actors beyond the immediacy of the 'here and now' of the "I think" statements that can be commonly found in these excerpts, and in the literature (Gan, 2012). Overlapping speech can also be found in Excerpt 26 in turns 7 to 9. It appears that this is caused by A answering Yamahiko's question before he specified the time limit, and then Yamahiko repeats the sum of money while A is confirming that it is per month. As such it seems more like a misunderstanding than an example of collaborative co-construction.

The opening turns of Excerpt 27 show that this conversation is falling into a similar pattern to his GOT in the first administration of short turns in which Yamahiko is the centre of the conversation, and he is talking mostly with B. However, a curious sequence takes place later on in turns 26 to 31, as Excerpt 27 shows.

Excerpt 27: Yamahiko in the second administration (person A)

Turn	Time	Person	
23	11:46.3	A	(1.1) I started to use it (.3) since (1.2) high school [C: high school?] (.6) high school yeah (.4) When I got in high school (.9) I bought mine (.9) Since then I've (.8) I have (.2) been paying mine.
24	12:00.4	B	(1.2) really?=
25	12:00.9	A	= yeah. [B: ah] (4.0) So wh-what do you think those guys (.2) like (.3) like Naomi (.8) uh like the guys whose addicted to (.6) use (.3) mobile phone [B: ahaha] (1.1) watching display all the time
26	12:18.3	C	(.8) nnn (1.2) are you living alone now?
27	12:21.0	B	(.3) no (.2) living my family.
28	12:22.9	C	(.2) Oh really? (.9) And (.8) are you (.2) ah do you have part-time job?=
29	12:28.0	B	=yeah I have.
30	12:30.1	C	(1.4) um… (1.4) nn…(2.4) how do you use your money.
31	12:37.7	B	(.9) I'll save (.2) I save my money [C: oh really?] (.7) in my bank. (.9) Hun. [C: that's good]
32	12:45.1	A	(2.7) yeah (1.5) yeah so why don't you pay yours?
33	12:50.4	B	(1.5) I don't know [A: ha] my parents pay mine. [while laughing]

Until this exchange the conversation has been mostly between Yamahiko and B, with C occasionally being brought in. By the end of turn 23 Yamahiko and the female sitting in place B are mirroring each other's body language, both sitting leaning forward with their arm closest to each other along the front edge of the desk, and the arm furthest behind their ear, exposing their inner wrists to each other. However, the topic becomes exhausted by turn 25 and after the longest pause so far of four seconds, Yamahiko attempts to start a new topic on mobile phone addiction, and appears to make eye contact with B, who hesitates, looking down at her prompt. Into this TRP comes C in turn 26, who completely ignores Yamahiko's question and starts questioning B, who answers quite minimally, without giving extra details, or asking questions to C. It appears as though C is trying to take control of the conversation, however, she is not able to sustain it for long, and Yamahiko reasserts himself by referring back to the information B gave earlier about not paying for her mobile phone (turn 17), and challenging her on it to provide a reason. It is a provocative statement that focuses the interaction on B's answer.

This incident underlines the collaborative nature of dominance. As was pointed out by Itakura (2001), sequential dominance relies on the willingness of the respondent to comply with the speaker by providing information (p. 1866), and Excerpt 27 shows that in a GOT, unlike for example, an interview test of oral ability, there is no obligation for the participants to be compliant. However, C's exertion of control is temporary, and following this brief exchange, Yamahiko shows he has the interactional resources to be able to continue with his sequential domination to the end of the test.

The excerpt above provides a sample of Yamahiko's linguistic resources as he correctly uses the present perfect in turn 23 to talk about his personal history. He had used the present perfect in the first administration to talk about where he had been. Since there is no equivalent in their language, Japanese speakers may find the perfect aspect challenging. In general, in this test he appears to be able to speak with less lengthy unfilled pauses and longer runs between pauses compared to his first GOT.

The final participant of the high initial group, Hamako gives a commanding performance in the second administration. Excerpt 28 shows Hamako's first turn, in which she gives the opinion on the topic that she had the preparation time to develop. Student A had started the test by asking a question which she directed at both Hamako and C, between whom eye contact is made. Hamako smiles and this, apparently is permission for C to answer the question. Since A has not given her opinion on the topic yet, Hamako's "How about you?" at the end of turn 3 is directed at her.

Excerpt 28: Hamako in the second administration (person B)

Turn Time Person
3 18:48.6 B (2.5) I also want to married (1.0) someday [A; ahum] (.4) but now I want to work
(A:aha) (.8) yes (.2) and then my mother (.3) told al-always told me that (.5) sh-e
(1.1) got married very (.5) when she was early [A; ahum] (.2) ah-whe- when she
was young (1.6) not very young but (.4) twenty two or something (.3) so she
always told me that (.4) she I don't have to (.6) married [A; ahum] (.6) when I was
young because (.4) she said that (.9) if I married (.3) get married (.2) I (.4) can't
do (.4) what (.6) I want to do (.6) so (1.2) I (.4) agree with her [A; ah] (.3) so I want
to do (.2) I want to work (.3) or I want to do what I want to (.5) do (.3) now but (.4)
someday I want to get married. [A: ahhh] (1.1) How about you?
4 19:43.1 A (.6) Uh me too [B: haha] uh I want to get married and uh I want to have a children
[BC: ahh] (.7) because childrens are very (.3) cute=

5 19:49.7 B = Yes do you like children?=
6 19:50.8 A = Yes…huhhuh [BC huhuh] and uhah (.5) but I want to (.4) have a job [BC: mm]
(1.1) because ah housework is very boring=
7 20:00.9 B =yes yes (.6) that's what mum told me always [laughing]=
8 20:05.1 A =I want to have a good job [B: yeah] and (.8) I (.2) I want to spend [B: mm] (.7) my
ah so (.3) happy life=
9 20:13.4 B =Yes (.6) yes (.7) I- think (.8) only getting married [A: mm] is the (1.3) happy [AC
mm] (1.0) good things for women (.8) yes (.8) yeah[AC: mm]
10 20:28.2 (1.9) A:[Bu-] … sorry huhuhuh=
B: [Bu] ah
11 20:29.2 B =Sorry (.9) but (.4) do you think (.4) do you think of any advantage of marrying
(.2) marriage (.6) being married (.3) married
12 20.38.1 A (.7) Ahh (.8) mm
13 20.45.5 B (4.7) I think (.4) one of the advantage is (.4) have a baby [A: mm] (2.7) yes Huhuh

As in the first administration, Hamako displays her interactional resources by asking appropriate follow-up questions, as in turn 5, but something not seen in the previous administration is the interaction between turns 6 and 9 where turns are collaboratively built on top of each other with very little time between them by two participants who are in agreement with each other, a feature similar to that noted in L1 women's speech (Coates, 1994). Hamako's reference to what another person says, even if it is her mother, not only brings levity into the conversation, but is something that is rarely found in the first administration, where test-takers are typically concentrating on expressing what they personally think. In turn 10 they run into some interactional trouble when Hamako and A try to speak at the same time. However, Hamako wins right of way by taking advantage of A's embarrassed laughter: she times her restart at the point where A's laughter finishes by apologizing and then continuing. This gives some indication of the assertiveness with which Hamako speaks.

Turn 3 is the longest that Hamako takes in this discussion, and it shows the linguistic resources at her disposal. It is a well-organized turn: it starts by stating her preference, it provides support for her position by quoting an example, and finishes by reiterating her preference. In this excerpt, it is worth noting her use of the phrase 'get married'. She starts using it erroneously without the auxiliary, and about half way through she switches to the correct usage. This can perhaps be explained by prioritizing attention (Schneider & Shiffrin, 1977). Her attention at the beginning may be on the construction of her long turn, and it is only after she passes the halfway point of her turn that she has the available resources to pay attention to the form of vocabulary. Another improvement that is noticeable is the length of her runs of words between pausing. Runs of six words are frequent here, and there is one run of nine words in “or I want to do what I want to”, though she did arrive at it through rehearsal when correcting herself.

4.2.3.3 Initial high group in the third administration

The third time Naeko takes the test she finds herself in the same group oral as Saaya (person A in Excerpt 13 above). In this test, unlike in the previous two group orals Naeko is entirely passive, not initiating any topics nor asking any questions. For a stretch of time of two minutes in the middle of the test she does not contribute any turns. In the previous two tests she was the first speaker, but in this one she is the third to provide information about the topic in the beginning of the group oral, with Saaya being the last. This might be considered surprising since the topic of this test is part-time jobs, and she announces in her first turn that she works at Disneyland, as shown in below in Excerpt 29, which is considered a desirable job for university students.

Excerpt 29: Naeko in the third administration (person D)

Turn	Time	Person	
1	32.9	B	(2.3) Do you [have a part time job…= C: [I...oh]< check overlap onset>
2	34.0	B	=Sorry
3	34.9	C	(.2) It’s okay
4	36.1	B	(.8) Do you have a part time job? [C: Do you] < check overlap>
5	37.5	A	(.4) [Yes]
		C	[Yes]
		D	[Yes]
6	38.5	B	(.4) What kind of?
7	39.7	C	(.6) Ahh I am working at (.2) a clothes store, (aha) and selling clothes. (1.1) How about you?
8	47.0	D	(.6) I working at (.3) Disneyland (ABC: ah wow) (1. 5) and its (.4) its so fun. (.5) How about you?
9	53.6	A	(.4) Uh My tim -part time job is (.5) um (.4) drug store, (.4) to sell the drugs [B:yeah A:Ahem[?]
10	1:02.5	C	(3.6) How about you? Hahah [D: hm])
	1:03.8	B	(.8) I work at McDonalds

In this excerpt the competition to start the discussion can be seen in turns 1 to 4. In the end B manages to take the first move by virtue of taking C’s “It’s okay” in turn 3 as permission for her to go first, when it appears from C attempting to ask the question again in turn 4 that it was supposed to be an acknowledgement of C's apology. This competition between B and C sets the tone for this discussion as they are similarly assertive in conversation and most of the interaction ends up taking place between them. It could be conjectured that having a prestigious job like working for Disneyland would be considered as a possible topic for conversation, but this never happens in this conversation. By contrast, that B has the least interesting job is shown by the beat of silence and the bringing of this topic of the conversation to a pre-emptory close after she admits to it. Before the next turn there is a gap of almost two seconds before C starts a new topic which is about the important aspects of having a part time job. This allows Naeko to talk about the interesting aspects of her job, as shown in Excerpt 30 below.

Excerpt 30: Naeko in the third administration (person D)

Turn	Time	Person	
13	1:52.5	A	(.3) Um (.8) I think (.6) um it is important that (.7) um (.4) money (.7) um and (.4) if (.7) um (1.0) one person don't have part time job (.6) no working part time job (.4) um (1.0) people (.6) um (.4) get (.7) money (.3) by their (.3) parents (.4) so it is not (.4) good things (.4) so (1.0) yeah, (1.8)um [D: mm] how about you?
14	2:21.2	D	(.6) Um (.3) The important point for me (.5) is whether I can enjoy (AC: mm) (.6) um (.6) uh my (.3) my former (.3) part time job (.2) it was so boring (.4) and I didn't like it but (.5) now my job is so fun (.5) um (.2) I can (.5) I could (.3) I could meet (.5) many people in (.9) different ages or (.5) different (.4) type of students (.3) so it's very (.7) um good for (.4) me [B: mm]
15	2:49.6	B	(1.4) What did you have, (.3) part time job, before?
16	2:53.3	D	(1.0) (Huhh) I worked at (.6) Lawson
17	2:55.6	B	(.4) Ahh, convenience store [BC: huh huh] (C: convenience store)

However, as can be seen, the follow up question that she does get is about her previous job, which is at a mere convenience store, and which they can have a chuckle about. There are two quite plausible explanations for Naeko's 'passive participant' role in this test. The first explanation is that similar to Aemi in the medium initial scoring group, she simply seems to care less about her score in this administration, possibly for the same reasons given to explain Aemi's lack of engagement. Another explanation is one revealed by a close examination of her 'leader' role in her previous GOTs. In the previous two administrations the leadership role fell to her because no one else was more willing than her to take it on. The long gaps before she took the floor to initiate a topic in her previous two GOTs are evidence of lack of competition, and something that does not happen in this discussion since gaps in the conversation are almost invariably taken up by B or C (and it makes Saaya's ability to find a gap in this discussion to add her potential topic all the more noteworthy). It is only in this third administration in which she has had to compete with two other test-takers who are assertive and capable of dominating the discussion that her own lack of willingness to take on this role is revealed.

The second person in the high group, Yamahiko finds himself in a very interactive group. After the opening turns, a familiar pattern emerges, as can be seen in Excerpt 31.

Excerpt 31: Yamahiko in the third administration (person B)

Turn	Time	Person	
4	2:06.5	C	(.3) So you mean you like to live in uh (.6) ur-urban city (yeah) (1.6) compared to your (.2) uh home town?
5	2:14.8	D	(.9) Yeah
6	2:15.3	B	(.2) Is that because you are from (.2) countryside?
7	2:17.7	D	(.3) Yeah
8	2:18.2	B	(.3) Yeah (.6) So you don't wanna (.4) go back to your (.3) home town?=
9	2:21.6	D	=Ahhh but someday maybe (.7) if I get accustomed (.2) get accustomed to the way of living in the cities (.2) may-maybe I miss my hometown and (.3) sometime..[B:um]
10	2:31.5	B	(.6) So when you get old (.5) maybe you go back.=
11	2:33.6	D	=Yeah (Huhuh ahah (B:mm)
12	2:33.7	AB	(4.2) I..
13	2:39.7	A	(1.7) Go ahead=
14	2:40.0	B	= I think (.4) pretty much the same. (.6) I (.3) want to live in the city (.2) when (1.6) 'til I get (.8) like forty (.3) or fifty (.9) then I want to go back to my hometown (1.0) with my family (.2) huhuh (.5) How about you?
15	2:56.3	A	(.3) Um (.7) I like living in the (.5) countryside (.4) because I'm from Ibaraki so (.7) maybe originally I like countryside (.9) um (.5) take a walk in nature [B: huhhu]

In turns 6, 8 and 10 the pattern resembles Yamahiko's previous two GOTs which he dominated by asking short questions to other group members. It soon becomes apparent that the group members in this GOT are of a higher level than Yamahiko encountered in previous administrations. Participant D shows he is of high ability by latching his answer (in turn 9) to the end of Yamahiko's question in turn 8, that is, by leaving no silence between the two turns. The competitive nature of this GOT becomes evident in turn 12, where both A and Yamahiko vie to claim turn 12. Participant A gives permission for B to take it in turn 13, but Yamahiko's turn 14 follows so quickly it is latched to A's turn in which he gives permission. This may indicate that Yamahiko was going to go ahead anyway, regardless of A's granting of the turn to him, and is indicative of Yamahiko's assertiveness in discussion.

The next excerpt affirms Yamahiko's determination to have his say, as he finds himself having to compete for the floor. In Excerpt 32, D starts the topic by asking the question to C directly, and they carry on the conversation between the two of them for over a minute.

Excerpt 32: Yamahiko in the third administration (person B)

Turn	Time	Person	
26	4:26.9	D	(1.0) So you have been living in cities for (.3) your entire life?
27	4:31.4	C	(.7) No hh…actually er (.7) I went (.4) - when I was an elementary school student I went (.4) to (.2) for (.3) different (.6) ur elementary schools because of my father's job (.3) so I I.. (.7) I had a (.3)I I I.. I've had (.2) I've lived I've lived in uh (.3) Nagano Hyogo [B: mm] (.3) and many places around Japan, (.6) so..
28	4:54.8	D	(1.0) But [C:yeah] (.3) you are living in cities longer than us (.3) right
29		C	(.7) yeah that right=
30		D	=so how do you think like living in (.6) like (1.0) ruruko (.3) prefecture (2.0) like Nagano (.7) or yeah Akita prefecture=
31	5:10.0	C	=ahhh (1.4) the air is (.2) very clear, [B,D: huhahah] (1.3) the water is delicious,[D:haha] (.5) and I'll ever…(.8) I was surrounded by (1.8) nature (.6) eh so (1.8) um [B: mm] that was reallynice but (.7) it's not convenient (.5) to live (.8) close the[re=
32	5:31.7	B	=[Bu]t like for (.2) city (.6) guys, that I think (.6) like my friend (.4) who's (.7) living (.2) who's been living in Tokyo for (.5) maybe for (.8) his entire life (.4) he (.5) thinks (.8) um (.2) he is just (.2) like (.2) us (.2) that we have home town (1.0) in (.4) separate prefecture
33	5:57.5	A	(3.5) Ah yeah (.2) when I (.4) went back to my home town (.4) I feel more comfortable um (1.8) than (.7) stay in there =
34	6:05.9	B	=[Um]m (.5) but (.7) what if you (.4) have your family (.3) in (.3) here
35	6:11.1	A	(.2) Um (1.9) living (1.9) um I think (.7) maybe I would think (.2) I want to move to (.4) another place, (.4) maybe countryside.[B: mm]
36	6:24.2	D	(1.4) When I go (.3) back to my home town (.5) the first few days I feel really happy to be there because (.3) I really want I really want to my (.3) I really wanted to meet my family [B: umhm] (.5) but (.7) a few days later I feel<?> [B: you get bored] (.3) I want to go back (.4) to return to (.5) here to our prefecture

In Yamahiko's previous GOTs he has never been quiet for this amount of time. In turn 27 Yamahiko makes space for himself in the conversation by starting his turn before C has finished, and he repeats this technique in turn 29. This use of overlaps to restrict other speakers from taking part is a feature of "participatory dominance" (Itakura, 2001, p. 1867). Also, he shows the ability to predict his conversation partner's words accurately in D's turn 31. Although his role is undeniably less prominent in this test, the abilities he shows to claim his turn here arguably show a greater level of conversational ability than in his previous two tests.

As for the linguistic resources that Yamahiko demonstrates in this GOT, it is noticeable that the time it takes him to start his turn after the previous speaker is considerably quicker than in previous GOTs; this may be as much an indicator of the greater competitiveness of the participants as it is of development. In grammar he again demonstrates his ability by using such complex forms as the present perfect continuous ("who's been living" in turn 32), which he arrives at through a self-correction, accurate use of time clauses ("when" in turn 10 and "till" in turn 14) and conditionals ("what if" in turn 34). In fluency, he impresses with his use of contracted forms ("wanna" in turn 8), fillers ("like" in turn 14) and runs of up to nine words between pauses ("then I want to go back to my hometown" in turn 14). When these considerable resources are combined with his interactive resources, it tends towards a leadership role. Given the quality of his fellow participants, Yamahiko can claim a co-leadership identity in this GOT.

Finally, Hamako finds herself in a very interactive group with much laughter and overlapping speech. Still, Hamako continues her identity in the 'leader' role by setting the topics and playing a central role throughout the discussion, as in Excerpt 33 below.

Excerpt 33: Hamako in the third administration (person D)

Turn	Time	Person	
32	3:52.7	D	(4.0) Do you have any ideas (.4) for (.4) last (.4) question [A:Hmmm] (1.5) Or do you think fast food (.3) has affect-affected Japanese (.2) society?
33	4:04.6	B	(1.3) Ah (.2) Japanese people (.7) get (.3) got to be more (.4) busy, (.5) too busy I think. (.6)It (.5) It doesn't take (.5) much time to eat fast food (.4) so (.5) we can (.4) we can work more more (.3) and more lot of times (D: yeah that's right) (1.0) we are very bu-(.2) (.4) very busy now I think (D: yes) (mm)
34	4:28.2	D	(3.0) It's convenient (.2) to for us to (.8) get food (.2) in the morning (.3) especially it's really busy in the morning. (mm) (1.2) They provide us (.4) really fast (yeah right yeah)
35	4:46.0	C	(6.4) Ahem (.5) but recently (.3) Children (.8) mm (1.4) is (.5) very (.3) easy to (1.5) get angry. [B: ah~] (1.0) It's (.3) because (.5) of (.4) fast food (.5) really (.3) so (1.1) at home (2.9) mother or (.8) um (.3) mother (1.4) have to (.2) um? (.3) Mother should cook (D: yes) (1.3) slow food eat slow food is (.2) very important (.4) for us
36	5:19.9	D	(.2) Yes with having a conversation with your parents= [Parents]?
37		D	=Yes
38	5:28.7	B	(5.3) And (.9) In Japanese (.3) -many Japanese people get (.8) fat (yeah) (3.1) its problem for us (.4) I think

39 5:42.9 D (3.7) I think we have (1.4) fast food doesn't have (.5) a lot of (1.4) I don't know how it calls but (2.4) white things or calshium

40 5:56.2 A (.5) Calcium?

41 5:57.0 D (.2) yes or (.3) I don't think it has a lot of variety of (C:value?) (B: Vitamins) (3.1) Vitamins (huhuhuhu) (C: ahhh! Huhuhu)

In this part of the conversation Hamako's interactive resources are displayed by the way she sets the topic in turn 32, and then takes every second turn, so that by the end of this period of just over two minutes she has interacted with every other group member, however briefly. In this segment of conversation, rather than being built on question and answer sequences, most of these turns are built on the top of the previous turn by being related by topic and connective ("but" in turn 35, "And" in turn 38). This is similar to the expanding sequence found by Gan (2010) in which the participants build on each other's ideas collaboratively. This can be found to a limited extent in her second group oral, but it is much more prominent in the final GOT. That she is among other advanced speakers is shown by the overlapping turns and quick response times between turns by all participants. Again, Hamako shows her assertiveness when student C attempts to collaboratively complete Hamako's sentence turn 41. It would have been easy for her to submit to this ending, but she insists on her original ending, much to the groups' amusement.

Although Hamako does not have an opportunity to use such a variety of functions as summarizing and reference to other speakers that she did in the second test, it can be argued that leading this discussion demonstrates a higher degree of interactive and linguistic resources than what she showed in the second administration, since the overall level of the participants is higher.

4.2.3.4 Summary of the qualitative analysis of the initial high group

It is clear from this analysis that the students who comprise the high group had a double advantage of not only being able to deploy their more advanced English skills in the first administration, but also not having to compete for floor space with groupmates who lacked an "architecture of the practice" (Young, 2000, p. 6) that they could apply to the GOT. This is particularly obvious in the case of Yamahiko and Hamako, whose skills allow them to dominate in the first two administrations, as can be seen in the summary in Table 54 .

It is not until the third administration that they find themselves grouped with peers who approach their level of ability, and yet they still are able to control the conversation for periods of time, which is revealing of their ability. The question that cannot be answered is if they would have been capable of a similar performance in an earlier administration, had their peers been of higher ability, and this is surely a disadvantage for this format of test. The final member of this group, Naeko, presents an illustrative contrast. She plays a large role in the first two administrations, but it seems more likely that this is due to the reluctance of her particular group members to take the lead. In the final administration she finds herself in a group in which there are two assertive members, and despite having relevant experience to contribute to the topic, plays a passive role.

Table 54: Summary of IC analysis of high initial scorers

	Admin.	Identity	Linguistic	Interactional
Naeko	1st	"Leader by default"	Develops opinion in longer turns, self-corrects, uses some simplified vocabulary, Can form questions	Some eye contact, initiates topics, can transfer turns
	2nd	"Leader by default"	Develops opinion in longer turns, uses some simplified vocabulary & grammar, uses short turns to maintain conversation	Mostly maintains eye contact, uses backchannel, initiates topics, asks follow-up questions, transfers turns,
	3rd	"Active participant"	Develops opinion in longer turns, responds appropriately in shorter turns.	Mostly maintains eye contact, uses backchannel, transfers turns
Yamahiko	1st	"Leader"	Develops opinion in longer turns & contribute shorter turns, can ask questions, uses conversational fillers & contractions	Maintains eye contact, initiates turns, transfers turns, gives contrary opinions, transfers turns, assesses others ability,
	2nd	"Leader"	Few pauses, uses fillers, takes long turns using complex grammar, short turns to maintain conversation, asks questions	Maintains eye contact, initiates topics, transfers turns, refers to previous information, ask provocative question
	3rd	"Co-Leader"	Few pauses, uses fillers, takes long turns using complex grammar, short turns to maintain conversation, asks questions	Maintains eye contact, initiates topics, transfers turns, asks follow-up questions assertively takes a turn, can overlap turns

Hamako	1st	"Leader"	Takes long turns that develop topic with complex grammar, takes short turns, can form questions, little pausing or maze words	Maintains eye contact, transfers turns, gives contrary opinions, asks follow-up questions initiates to restart conversation
	2nd	"Leader"	Well organized long turns, complex grammar, takes short turns, can form questions, little pausing or maze words	Maintains eye contact, transfers turns, assertively takes a turn, collaboratively build turns, ask follow-up questions
	3rd	"Leader"	Well organized long turns, complex grammar, takes short turns, can form questions, little pausing or maze words	Maintains eye contact, transfers turns, assertively takes a turn, collaboratively build turns, ask follow-up questions

4.3 Summary of the qualitative chapter

This section set out to provide a qualitative insight into how individuals in the GOT perform over three administrations that they take the test in two years. The participants were divided into three groups depending on their initial score in the GOT. It could be seen that among the three who started the test with an initial low score it was possible to improve in certain respects in each successive administration, and as they improved it was possible to define their identity in a different way. Of the three, only Machiko managed to play more than a participatory role in the group discussion, and she showed signs of willingness to do this in the first GOT she took, despite lacking knowledge of the practice of group discussion, as was evident in the GOT. It was impressive that she felt confident enough play leading roles and display a range of interactive functions in her second and third GOTs. The other two members, Kabuhito and Saaya were overall passive in their GOTs, with only Saaya showing she could initiate a topic in the final administration, despite it not being taken up by her fellow group members. This perhaps shows the difficulty for those who start with low ability to catch up relative to their group members. Since the level of their cohort is improving at the same time as them, they remain the least able to play a more active role in the GOT.

The two members of the middle initial scoring group, Aemi and Taeko showed contrasting developmental paths. In the first administration, they show enough ability to express an opinion and ask questions, but to a limited extent. Aemi shows that she can lead for a part of her discussion, but does not sustain it, while Taeko does little more than answer questions and transfer the move to another. In the second test, Taeko improves by showing a greater ability to interact with her fellow groupmates, while Aemi does enough to be identifiable as the leader of her group. Their roles are reversed in the final GOT, as Aemi falls back from a leadership role while Taeko leads her group using a variety of interactive functions. The example of these two participants emphasized the point that the ability to lead a group is not just a matter of having the skills of leadership, but also of having the will to do so. This point is reinforced by the example of Naeko in the high initial scoring group. In her first two GOTs she displayed leadership skills, but it seems like she did so because nobody in her group was willing to put themselves forward. There was a distinct lack of competition for the conversational floor amongst her group members in the first two GOTs. In her final GOT she was faced with two co-test-takers who were quick to take on the leadership role, and as a result, Naeko slips into a participant role. The other two members of the high initial scoring group, Yamahiko and Hamako, are leaders in all three of their GOTs. Yamahiko reinforces the point about the will necessary to be a leader. In the first two administrations he dominates through his ability to ask questions, but in the last one he finds himself in a group where he has to compete for the floor by overlapping his turns and speaking more quickly. If he had not done this, he too would have been identified as a 'participant' rather than a 'leader'. Similarly, Hamako had to display her assertiveness in the final GOTs in order to maintain her identity as leader.

The purpose of this chapter was to complement the quantitative data and prepare the way for the research questions related to how individuals progress in the GOT relative to their scores. It has done this by analysing their performance in the GOTs within an IC framework. The next chapter will continue by comparing what has been learnt through this qualitative examination by synthesizing it with their progress as seen through the indices from the quantitative side and the actual scores that they were awarded by the raters.

CHAPTER FIVE

Quantitative – Qualitative Synthesis

5.0 Introduction

The previous two chapters have covered the quantitative and the qualitative phases of the study, and since this is an mixed methods research (MMR) study, this chapter will bring these two phases together to answer the final research questions, which are repeated below:

RQ5. How well do the test-takers' speaking performance indices and scores awarded to test-takers of the GOT over the three administrations of the test taken in two years represent their performance compared to a qualitative analysis of their performance?

This question relates to the objective of this dissertation to investigate stable measures of the GOT so what it in fact measures can be clarified. Answers to this question will contribute to the literature on the construct that this test of oral proficiency measures. The incorporation of the qualitative aspect answers the questions regarding the objective of the GOT as a means of measuring the student's speaking ability, and this will have implications for the rating of peer-interaction tests.

Before presenting the findings, it is necessary to clarify the position taken on the contradiction between the distributed nature of IC and this dissertation's focus on tracking individual progress. The conflict between IC and the need in testing to award scores to individuals is not a new conflict in the field of language testing, as Section 2.1 in the literature review has made clear. To explain the stance taken here, it needs to be pointed out that the raters of this test are instructed to give their scores according to the criteria on the rating scale, which are intended to be assigned on an individual basis. Thus, it is assumed that the test-takers' scores were assigned by the raters on the performance the test-takers displayed within the context of their performance in the GOT despite this performance being contingent on their group members, according to the theory of IC. As such, the question of whether the score itself is a fair reflection of the candidates' actual proficiency level will not be answered here.

While it is certainly possible that a performance by any individual could be better or worse in different conditions, it is beyond the scope of this dissertation to investigate representativeness of their score. The question that this dissertation will answer is the degree to which the score that they did receive was justified, given the evidence provided by their performance in the GOT. The results will contribute to and provide implications for the GOT rating scales and rater training, as well as the construct of the GOT as a whole.

This chapter will, like the previous chapter, be organized by the grouping of the students that were chosen for the qualitative analysis: the low, medium and high initial scoring students. These three sections will each be divided into two: a subsection that discusses the qualitative findings in relation to their performance as depicted by the indices from the quantitative chapter; and another which compares the qualitative and quantitative findings to the scores they received.

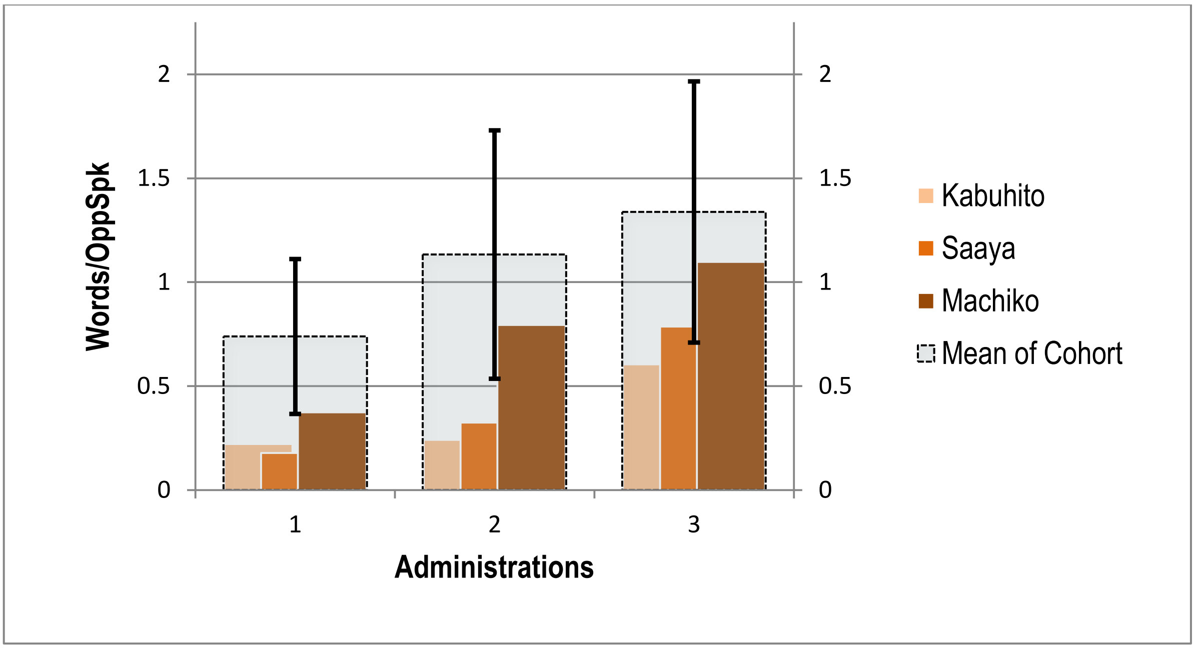
The discussion of their performance as indicated by the indices will start by describing their performance in words spoken per opportunity to speak and words per turn. The former was chosen because it is foundational in the sense that many of the other indices are based on it, and the number of words spoken gives an indication of their overall performance. The latter was selected since it is a composite index that combines information about the turns the participants take with the overall words spoken, and thus gives an indication of the interaction that the individual was involved in that particular GOT. Following this, the test-taker's progress in the complexity statistics of words per AS-unit and accuracy indices of error-free clause proportion and number of error-free clauses per opportunity to speak will be analysed. These indices should be related to grammar scale of the scoring bands. Next will be the fluency indices of articulation rate, speech rate, pause proportion, and the maze and maze and sound ratio, which should be of relevance to their fluency score. Following fluency will be a description of their performance in vocabulary. Although the indices of lexical diversity were not found to be able to distinguish between their performances, a breakdown in terms of the less frequent words used and the academic word list (AWL) may provide an insight into their progress on the vocabulary scale of the scoring bands. Finally, the interactive function analysis of their discussion will be explained, which will provide a basis for evaluating the raters' scores on the communicative skills (CS) scale. At all stages, the individuals performances are compared with the average for the cohort to provide a standard that they might be compared to. In addition, this section will focus on the language development the GOT shows they make over the three administrations.

5.1 Comparison of low initial group with quantitative analysis

5.1.1 Comparison of low initial group with indices

This section, like the following parallel sections for the medium and high initial scoring groups, will start by comparing the qualitative findings with selected core indices. Graphs will be used to display the relevant statistics in this chapter; the descriptive statistics can be found in Appendix H, Tables 84 to 87. The first graph, Figure 32, shows the words per opportunity to speak by the three participants over the three administrations compared with the average of the group of 53, for which the standard deviation is shown by error bars.

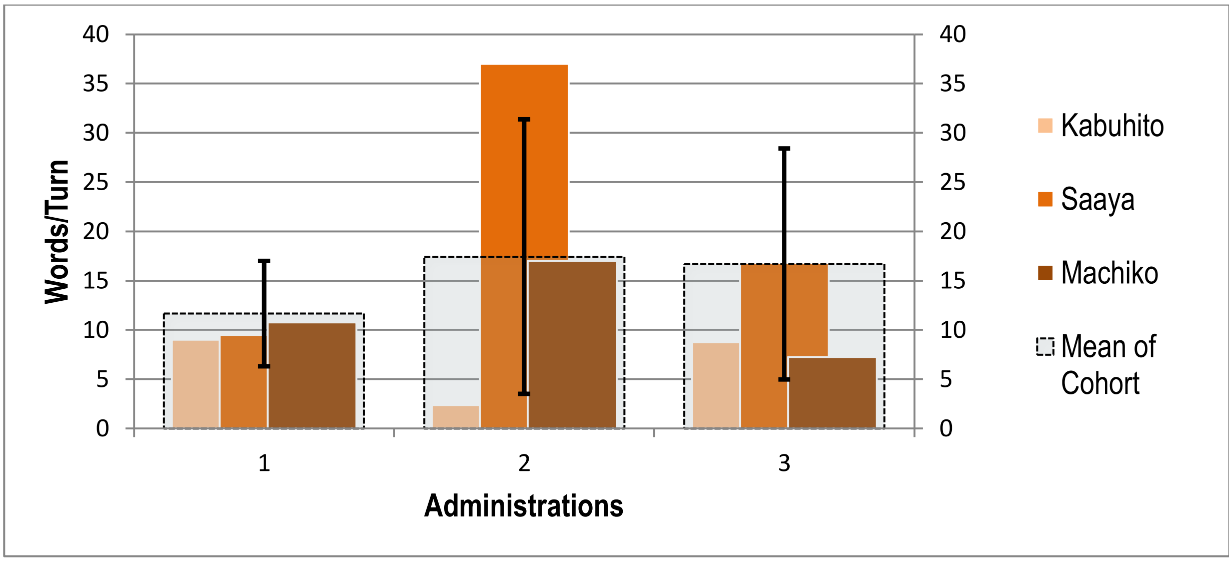
Figure 32: Words per opportunity to speak of low initial group over three administrations



This graph makes it clear how little these students were engaged in the first administration, with all three of them being under the lowest range of the standard deviation from the average of this cohort. Indeed, Kabuhito was so in all three administrations, and it is not until the third administration that the number of words Saaya speaks falls within the range of the standard deviation. Figure 32 also shows that Machiko's contribution in the GOTs she took part was always more than the other two in the low initial scoring group, and she comes closest to achieving the average of the cohort. Despite falling short, through the qualitative analysis we know that the second and third administrations her contributions were sufficient for her identity to be described as that of 'leader'.

Figure 33 displays the words per turn over the three administrations by the three participants, compared with the average and standard deviation for the cohort. With the number of words spoken being low, especially for Kabuhito and Saaya in the first two administrations, inferences based on their statistics must be treated with caution. Perhaps this accounts for the fluctuating figures for these participants, particularly in the second administration. In the qualitative analysis, we saw that Saaya took just one long turn in the entire test, and this accounts for the spike her figures shown in the middle administration of Figure 33. In the final administration this declines to the average of the cohort.

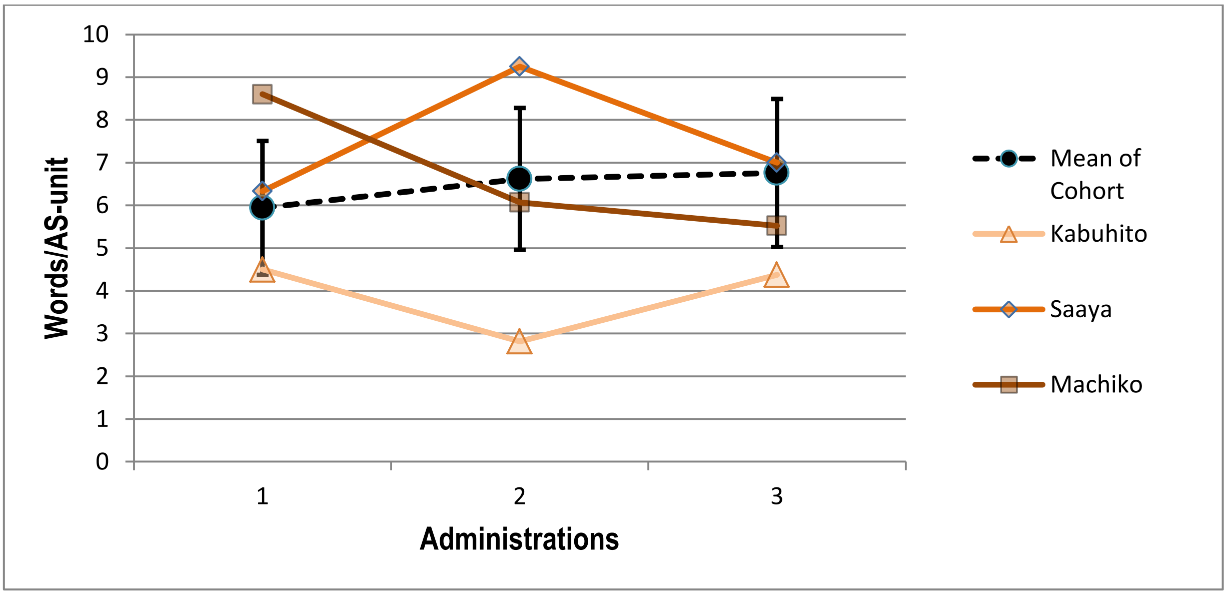
Figure 33: Words per turn of low initial group over three administrations



As was seen in the qualitative analysis, Kabuhito was subject to what could be described as a cross-examination in the second administration, and being heavily involved in answering questions to which he gave minimal answers had the effect of spreading the words he used over a greater number of turns. In the third administration, as he both took more turns and used more words when answering questions that his group mates were kind enough to direct at him, and his figures for words per turn return to a similar level to the first administration.

The complexity index of words per AS-unit for the low initial scoring group can be seen in the graph in Figure 34, which also displays a variety of patterns. For Kabuhito and Saaya, this index follows the same pattern as words per turn, though it does not vary nearly as widely, showing how this measure is influenced by turn length. The figures from the first administration are based on very few words and turns, though they are comparable in the sense that they were based on three similar performances in terms of their low participation.

Figure 34: Words per AS-unit of low initial group over three administrations



Nonetheless, Machiko spoke considerably more than the other two in this group, and the longer turn she took in the first administration is reflected in her higher complexity score. As was commented on in the qualitative section, Kabuhito's complexity figures were boosted by his incorporation of a repair from a groupmate which included a relative clause (see turn 22 in Excerpt 1. In the middle administration, Saaya's single turn included full sentences, boosting this measure of complexity. This is in stark contrast to Kabuhito, who spoke a similar number of words spread over several short turns in which he responded to questions, and this lowered the number of words per AS-unit. In the third administration, these extreme patterns are moderated by speaking more words and taking more turns, bringing Kabuhito's complexity up and Saaya's down. In the middle and final administrations, Machiko's performance in terms of complexity was lowered by the multiple turns she took as she led the discussion. It is the qualitative analysis that tells us that the shorter turns Kabuhito takes are a result of answering questions, whereas Machiko's figures come from leading the discussion.

The next two graphs show the proportion of error free clauses and the error free clauses per opportunity (Figures 35 and 36). Again these figures need to be discussed in conjunction with the low

Figure 35: Proportion of error free clauses by low initial group over three administrations

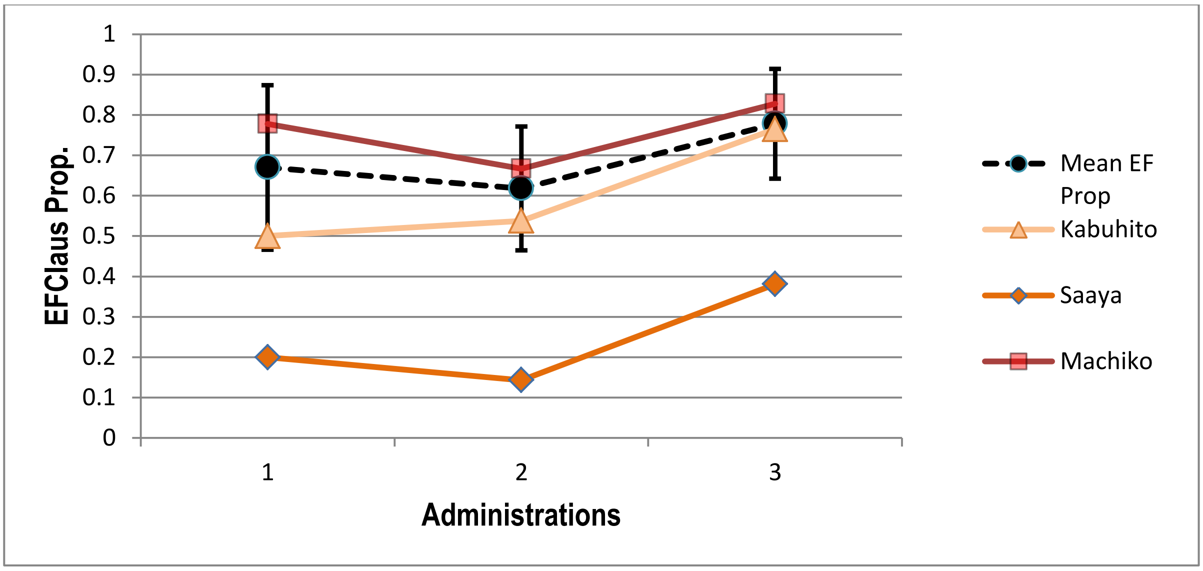
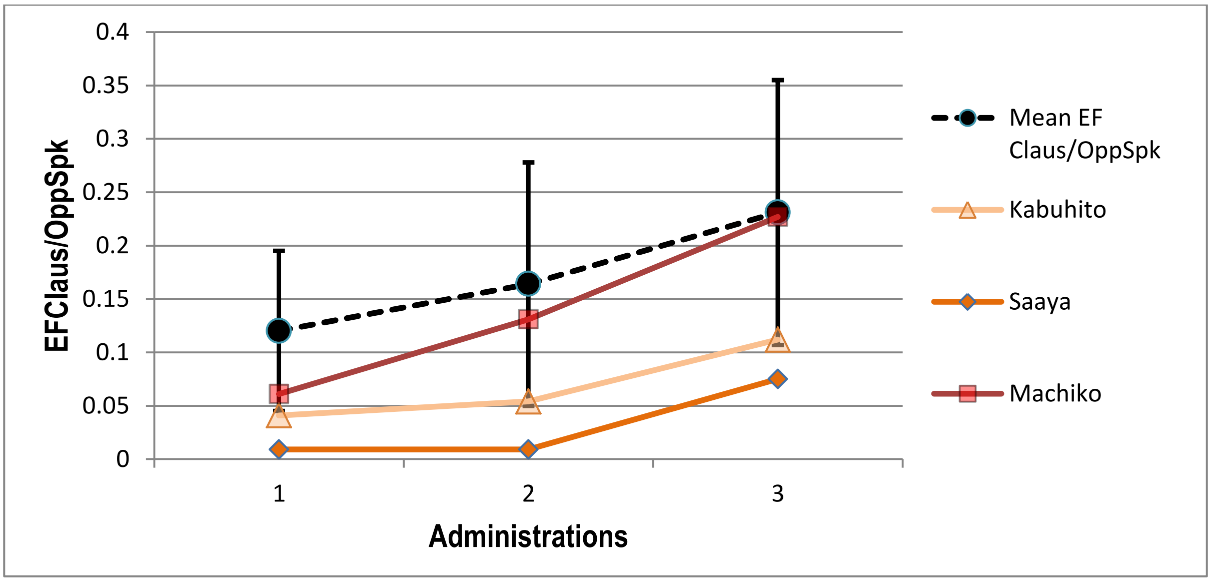


Figure 36: Error free clauses per opportunity to speak by low initial group over three administrations



number of words spoken in the first two administrations. Machiko's high proportion of error free clauses in the first administration might be considered as partly due to being skewed by reading aloud a question from the prompt; whereas her performances in the second and third administrations, in which she speaks far more frequently, are probably more representative of her ability, and in both of these Machiko's proportion of errors was consistently slightly higher than average. For Saaya, in all three administrations her accuracy is well below the average of her cohort in both of these measures, and though she shows considerable improvement the final time she takes the test, she is still well below the range of most of her cohort. The steadiest index is for error-free clauses per opportunity to speak, which rises mostly according to the increasing number of words they spoke in each administration, with the exception of Kabuhito, whose index was static in the first two administrations. It can be inferred from this statistic that the test-takers exposed the raters to an increased number of error-free clauses in each successive administration.

The fluency indices in Figures 37 to 39 vary widely for this group. For the first two indices for speed fluency, the articulation rate[5] and speech rate[6], a higher figure means more fluency, whereas for the proportion of pauses[7] (breakdown fluency), maze[8] and maze and sound ratios[9] (repair fluency), a lower figure indicates there are less disfluencies. For Kabuhito the most consistent figures are for speech rate and proportion of pauses, which show improvements in every administration. It should be pointed out that since the speech rate includes pause time, it is influenced by the proportion of pauses statistics. Kabuhito's articulation rate, maze and maze and sound ratios all vary from administration to administration, most likely influenced by the small samples in the first two administrations, and in particular the number of short responses in the second. Kabuhito's third administration figures are the soundest since he spoke the most in this administration, and they show improvement in speed fluency, but it is difficult to compare his progress with the previous two administrations due to the lack of

[5] Syllables spoken per second exclusive of pause time
[6] Syllables spoken per second inclusive of pause time
[7] Proportion of time spent in unfilled pauses to time of turn
[8] Ratio of maze words (false starts, repetitions and so on) to all words
[9] Ratio of maze words and voiced fillers to all words

Figure 37: Fluency indices of low initial group in the first administration

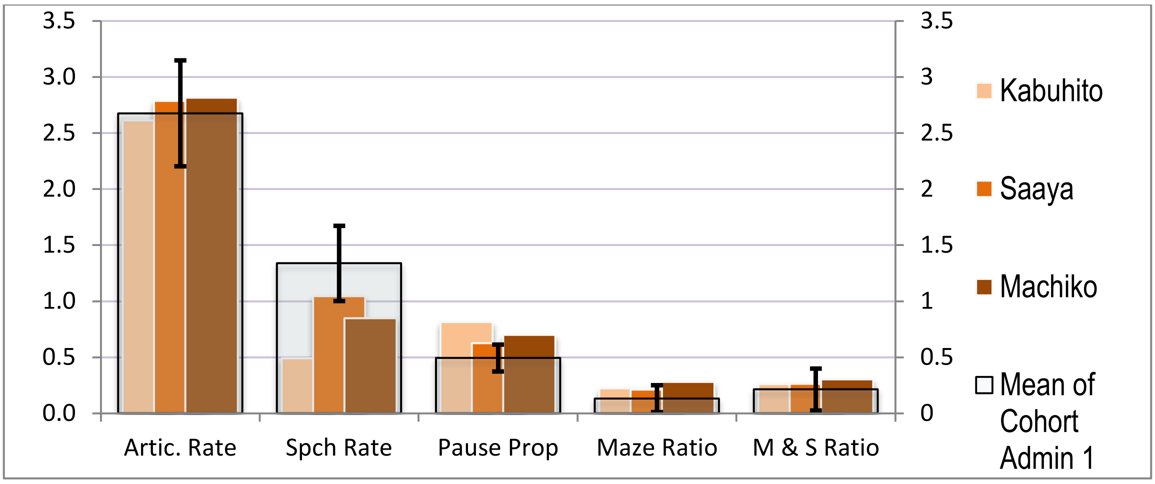


Figure 38: Fluency indices of low initial group in the second administration

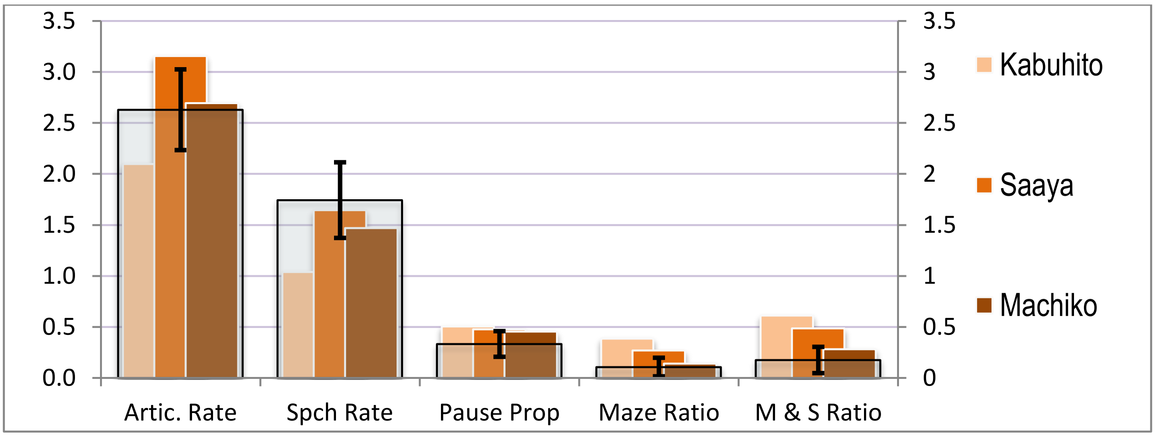
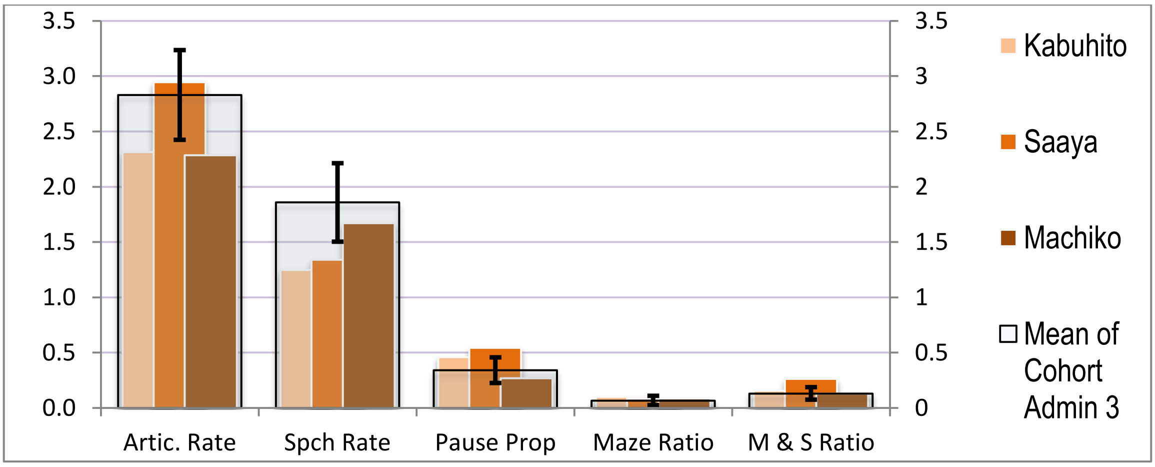


Figure 39: Fluency indices of low initial group in the third administration



words he spoke. The most aberrant figures are Saaya's, for whom all indices vary from administration to administration. For her, the articulation rate, speech rate and pause proportion all improve in the second, but all decline in the third, while the maze and maze and sound ratio show a decline in the second from the first, before improving in the third. Similar to Kabuhito, the most reliable figures are likely to be those in her third administration since the number of words spoken is greater in this test.

In Machiko's statistics, the indices of articulation rate and speech rate show a consistent trend of the articulation rate declining while the speech rate increases. This is a result of the particular mix in her utterances of pause proportion declining as the number of syllables spoken increases as the time in long turns declined in the final administration. Her figures for the maze and maze and sound ratio are more consistent in that they show continuously improving performance in each administration.

Although the vocabulary indices could not distinguish between the levels, an indication of the level of the lexis they used in their tests can be shown by analysing their pruned utterances using the general service and academic word list and the software *Range* (Heatley & Nation, 1996). The results are summarized in Table 55 below, and show some interesting features of their performance. Kabuhito's performance in terms of variety of words shows little difference between the first two administrations, but in his final test he improves considerably, with an impressive number of lower frequency types. For the most part these lower frequency words are concrete nouns, and the qualitative analysis showed that they were elicited from his group members' sympathetic questioning (see Excerpt 11). Saaya's token count more than doubles in each succeeding administration; though the number of lower frequency words does not improve from the second to the third administration. Machiko also makes improvements in each succeeding administration, and in the final administration produces a few lower frequency words.

Table 55: Development in lexis of low initial group over three administrations

Administration 1						
	Kabuhito		Saaya		Machiko	
	Type	Token	Type	Token	Type	Token
k1	22	27	11	18	33	44
k2 ≤	1	1	0	0	2	2
k2 ≤ words	Australia		-		communicate, abroad*	
AWL	0	0	0	0	1	1
AWL words	-		-		communicate	
Administration 2						
k1	21	32	22	31	48	86
k2 ≤	0	0	5	8	2	2
k2 ≤ words	-		formerly, Japan, earn*, housework*, traditional*,		Japanese, traditional*	
AWL	0	0	1	3	2	4
AWL words	-		traditional*		job*, traditional*	
Administration 3						
k1	38	58	50	82	60	106
k2 ≤	10	11	4	4	4	6
k2 ≤ words	bath, baths, chemical, famous, peanuts, relaxed, sky, stars, stones, underground,		drug, drugs, opinion, store		credit, exchange, goods, opinion	
AWL	1	2	1	4	1	2
AWL words	chemical, relaxed		job*		credit	
*word from prompt						

The final set of graphs, Figures 40 to 42 show the interactive statistics for this group. The first graph, in Figure 40, clearly shows the lack of participation of this group of students as they all fall well below the average in the three major categories. In the other category, Collaborating features, the exception is the single instance of Kabuhito incorporating a repair initiated by a group mate coming to his aid. Kabuhito is exceptional in being the only one in this group to take part in Collaborative episodes, as in the third administration he also incorporates a suggested word into his turn, and in the second administration he asks a clarification question, as was described in the qualitative analysis (Excerpt 6). Apart from this category, the only other test where Kabuhito outperforms the average is for Responses in the second administration, where he answered a higher than average number of questions. As pointed out in Section 3.2.3.3, Responding features are indicators of passive involvement, and should be considered the minimum participation participation necessary for maintaining interaction.

Figure 40: Interactive indices of low initial group in the first administration

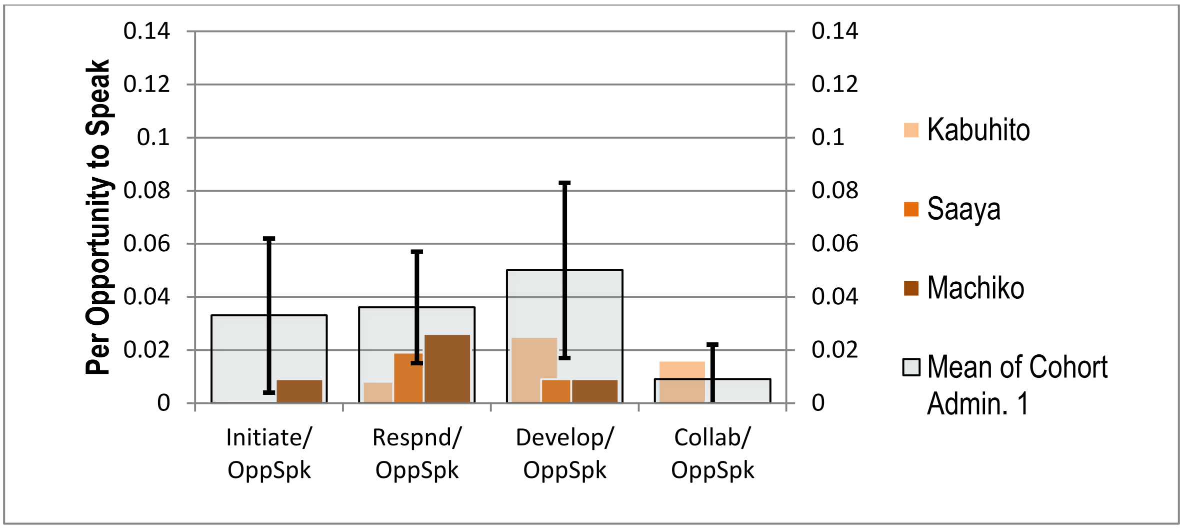


Figure 41: Interactive indices of low initial group in the second administration

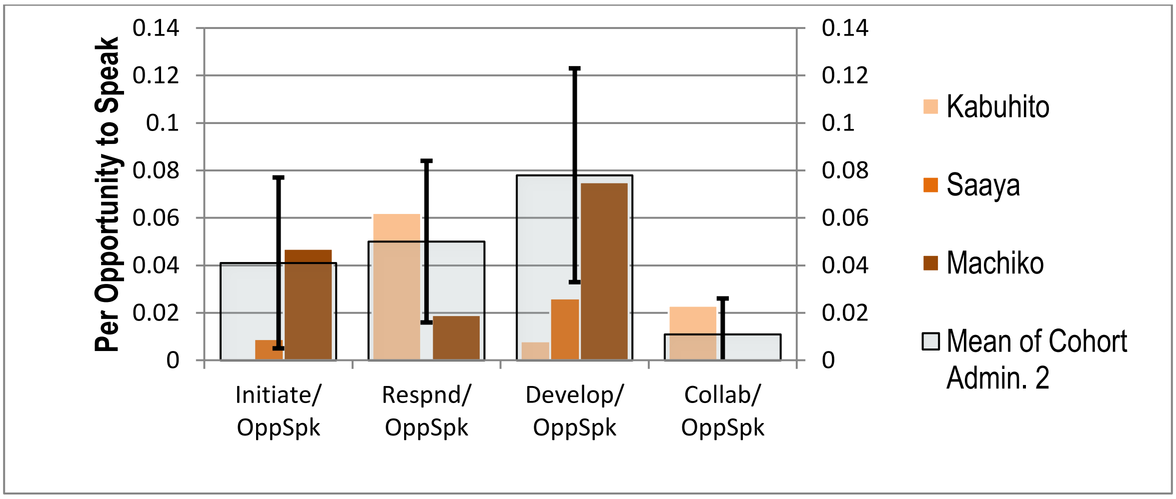
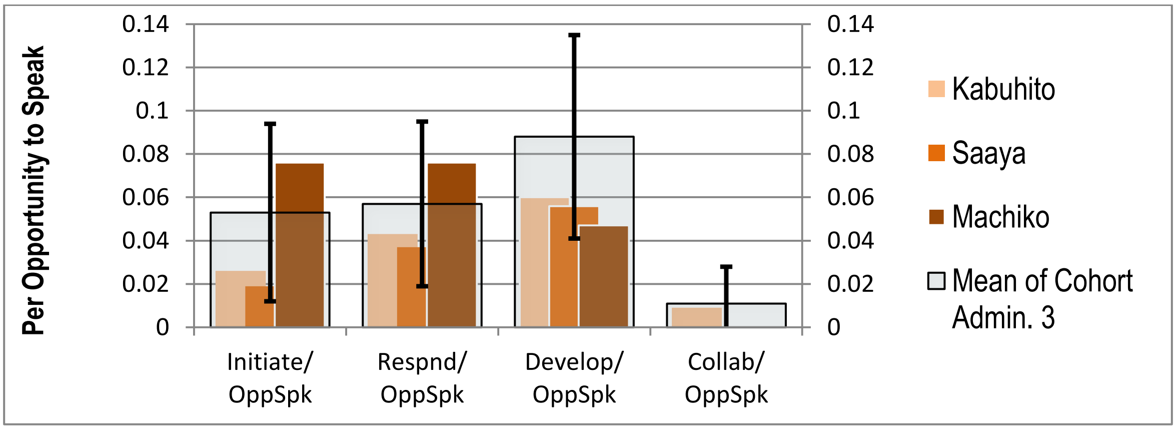


Figure 42: Interactive indices of low initial group in the third administration



involvement, and should be considered the minimum participation participation necessary for maintaining interaction.

The lack of Development functions in this administration shows that Kabuhito did not answer them at length. He did not initiate in the first two administrations, and just once in the final administration when his question to check his groupmates' understanding was counted as an 'initiation' even though, as noted in the qualitative section (Section 4.2.2.3, Excerpt 10), it seems more of a pre-planned move than 'initiation-in-conversation'.

Given the qualitative analysis, it is not surprising that the most impressive development is shown by Machiko, who shows improvements in Initiating features in every administration, as seen in Figures 41 and 42. Her responding and developing figures show that in the final test she developed less but responded more, which is consistent with the lack of time spent in turns over ten seconds in the final administration. The only negative note is that she did not take part in turns that would be counted in the Collaborating index. Like Machiko, Saaya also did not take part in a collaborative exchange over the three tests, and her figures also show improving ability to initiate, as noted in the qualitative analysis. Her responding figures in the final administration show that she took a greater role in the conversation by answering other's questions, which did not happen in her second GOT, in which she had a single long turn.

In summary, while the quantitative figures for the low initial scoring group can usefully illustrate trends in the figures, individual details about how they were produced may be overlooked. This is particularly so for lower ability students who speak few words and thus are more heavily influenced by odd events such as reading a prompt aloud or incorporating a single repair, as occurred in the first administration for Machiko and Kabuhito respectively. In the second administration, the influence of the prompt was clearly discernible on the figures through its impact on the pattern of discourse which resulted in startlingly different figures for Saaya and Kabuhito, particularly in the accuracy and error free counts. If only the figures were examined it would seem that Saaya speaks with considerably higher complexity, but we know from the qualitative analysis that this is more a result of her speaking in a single longer turn. The qualitative analysis makes it possible to interpret their statistics. The last subsection will continue this endeavour by investigating the extent to which their scores reflect their performance as described in the qualitative analysis and indices.

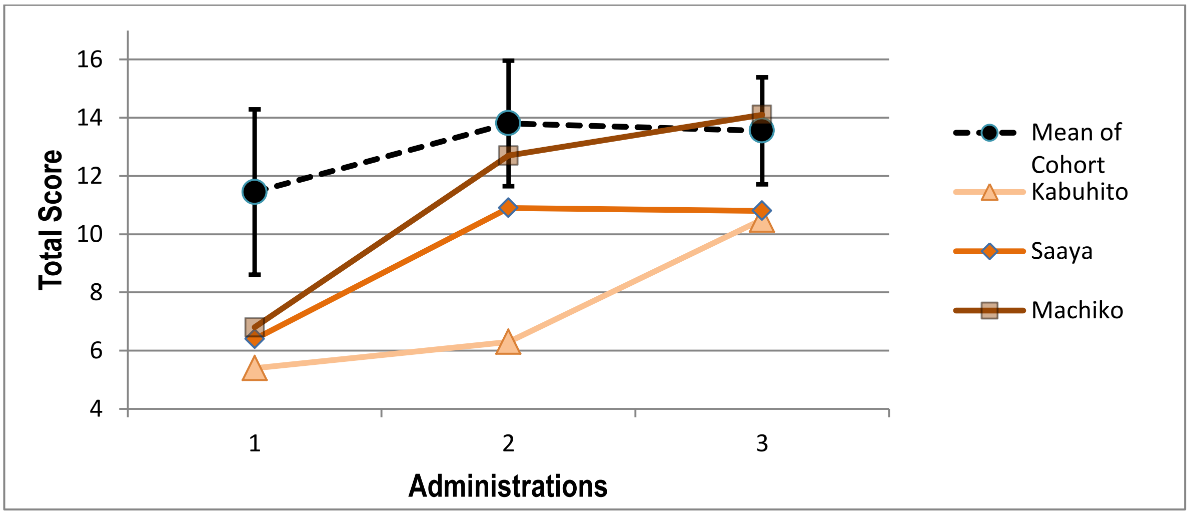
5.1.2 Comparison of low initial group with scoring

In this section, the performance of participants as shown by the indices will be compared with the scores they received. First a general picture will be provided of how their scoring unfolded by describing how their total score changed over the three GOTs they took part in. The following three subsections will compare their performance in the first, second and third administrations respectively. In these subsections the emphasis will be on comparing their performance per administration, not across administrations. It will not be until the summary of this section (5.1.3) that a developmental perspective will once more be adopted.

5.1.2.1 The overall pattern of scoring of the low initial group

The total score awarded to these students in the three administrations is displayed in the graph in Figure 43, which shows the pattern of their total scores over the three administrations. Here it can be seen that Kabuhito and Machiko achieve considerable gains in all administrations, with Kabuhito making his most impressive gain in the third test, while Machiko's score increases more in the second, than in the third. Saaya shows a different pattern by achieving a considerably improved score in the

Figure 43: Total scores of the low initial group over three administrations



second test, but then recording a slight drop in the third. Of these three test-takers, only Machiko manages to surpass the average score of the cohort, which she does in the final administration, though it can be seen that she comes close in the second. The other two members of this group fail to come within one standard deviation of the total average score of the cohort in any of the three tests, perhaps showing the difficulty of catching up from their low starting points.

To provide a reference for the discussions of their overall scores, Table 56 provides their scores over the three administrations in all bands, and their total scores. The next three sections discuss the relationship between the scores and performance as measured by the indices and from the qualitative analysis.

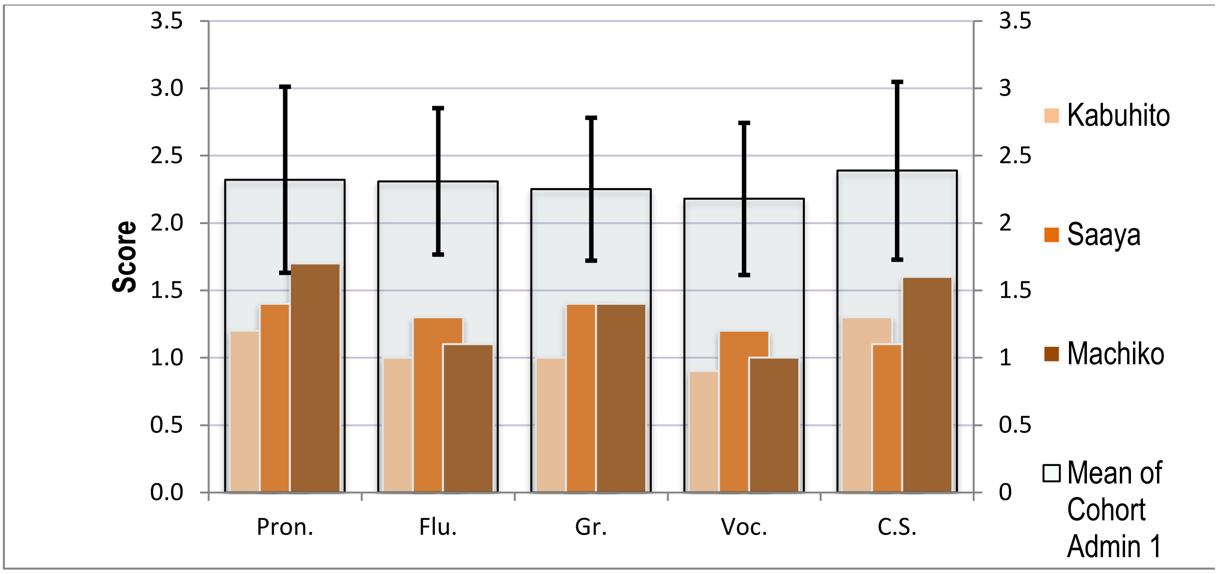
Table 56: Low initial group's scores over three administrations

		Pron.	Flu.	Gr.	Voc.	C.S.	Tot.
Kabuhito	Admin 1	1.2	1.0	1.0	0.9	1.3	5.4
	Admin 2	1.3	1.3	1.2	1.3	1.1	6.3
	Admin 3	2.1	2.5	2.1	2.0	1.8	10.5
Saaya	Admin 1	1.4	1.3	1.4	1.2	1.1	6.4
	Admin 2	2.1	2.2	2.1	2.0	2.5	10.9
	Admin 3	2.5	2.0	2.5	2.0	2.0	10.8
Machiko	Admin 1	1.7	1.1	1.4	1.0	1.6	6.8
	Admin 2	2.6	2.5	2.6	2.3	2.8	12.7
	Admin 3	3.1	2.4	3.1	2.8	3.1	14.1

5.1.2.2 The low initial group in the first administration

The scores in the bands of the low initial scoring group can be seen in Figure 44, in which a visual comparison is made with the average and standard deviation of their cohort. As seen here, the scores hover around the score of 1.0, with only Machiko breaking 1.5 in communicative skills and pronunciation. When considering the scoring bands (reproduced in Table 57 for convenience) from in conjunction with what is known through their performance in the indices and the qualitative analysis of their performance, apart from a couple of exceptions, they seem to be largely consistent. The wording in the 1.0 – 1.5 band is an appropriate description for these test-takers, and the scores in

Figure 44: Scores of low initial group in the first administration



general agree with the order of ability that is apparent from the qualitative analysis and indices, with the highest ability student being Machiko, followed by Saaya and Kabuhito.

The one clear discrepancy can be found in Machiko's score in the 'communicative skills/strategies' (CS) band. It is clearly stated in the 1.0 band that a candidate with a score between 1.0 and 2.0 "does not initiate interaction" whereas in the test she did, albeit by reading out directly from the prompt. Taking this into account a score of 2.0 seems more appropriate than the 1.6 she was awarded by the scorers at the time. Another possible mismatch in the first administration is between the evidence on the video and Kabuhito's pronunciation score of 1.2. The scale states that scores of up to should be awarded for "Very heavy accent… Japanese katakana-like phonology and rhythm; words are not blended together", and there is certainly evidence of this as pointed out in the qualitative analysis (see Excerpt 1). It could be conjectured that the raters, being familiar with Japanese accents, were overly forgiving when judging this scale, as can be inferred from a study by Major, Fitzmaurice, Bunta and Balasubramanian (2002). Since no objective measure was included in this dissertation, the doubts raised here cannot be verified.

Table 57: Scoring bands used in the GOT

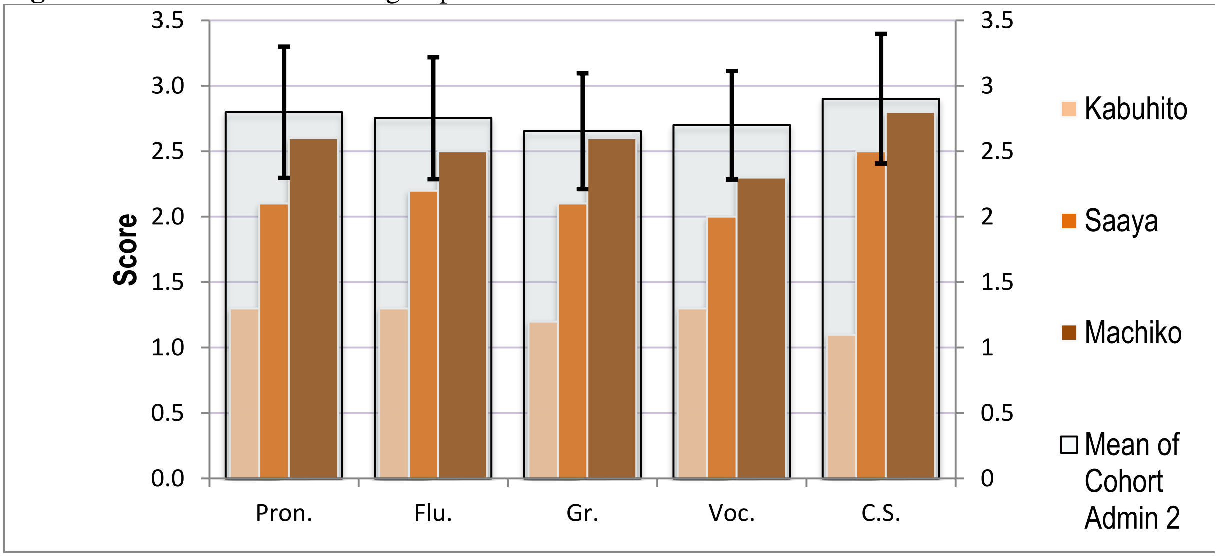
	Pronunciation Think about: • pronunciation • intonation • word blending	Fluency Think about: • automatization •fillers • speaking speed	Grammar Think about: • use of morphology • complexity of syntax (embedded clauses, parallel structures, connectors)	Vocabulary Think about: • range of vocab	Communicative skills/strategies Think about: • interaction • confidence • conversational awareness
0 .5	Very heavy accent, uses Japanese katakana-like phonology and rhythm; words are not blended together	Fragments of speech that are so halting that conversation is not really possible; nss would not think person had virtually no English	Does not use any discernable grammatical morphology	Shows knowledge of only the simplest words and phrases taught in junior high school or beginning high school	Shows no awareness of other speakers; may speak, but not in a conversation-like way
1.0 1.5	Somewhat katakana-like pronunciation; does not blend words together, they are pronounced in isolation	Slow strained speech, constant groping for words and long unnatural pauses; communication with a ns would be difficult	Doesn't have enough grammar to express an opinion clearly; makes frequent errors; no attempt at complex grammar	Lexis not adequate for task, cannot express opinion properly with the limited words used	Does not initiate interaction, produces monologue only; shows some turn-taking, may say, "i agree with you," but not relate ideas in explanation; too nervous to interact effectively
2.0 2.5	May not have mastered some difficult sounds of English, but would be mostly understandable to a naïve NS; makes some attempts to blend words	Speech is hesitant; some groping for words and unfilled spaces are present but generally don't impede communication completely	Relies mostly on simple (but appropriate) grammar, has enough morphosyntax to express meaning, complex grammar is attempted but may be inaccurate	Generally has enough lexis for expressing some opinion but does not demonstrate any particular knowledge of vocabulary	Responds to others without long pauses to maintain interaction; shows agreement or disagreement to others' opinions
3.0 3.5	Pronunciation is good but has still not mastered the sound system of English; accent does not interfere with comprehension; can blend words	May use some fillers, rarely gropes for words but speech may still not be quick	Shows ability to use some complex grammar, may make errors but they are only in late-acquired grammar	Shows some evidence of some advanced vocabulary	Generally confident, responds appropriately to others opinions, shows ability to negotiate meaning quickly and relatively naturally
4	Speaks with excellent pronunciation and intonation; has practically mastered the sound system of English	Excellent fluency, uses fillers effectively, shows ability to speak quickly in short bursts	Uses both simple and complex grammar effectively; may make occasional errors but they are only in late-acquired grammar	Shows evidence of a wide range of vocabulary knowledge	Confident and natural, asks others to expand on views, shows how own and others' ideas are related, interacts smoothly

5.1.2.3 The low initial group in the second administration

The low initial scoring group's performance in the second administration according to the raters can be seen in the graph in Figure 45. Kabuhito's scores now vary slightly between 1.0 and 1.5. According to the performance seen from the indices and the qualitative analysis, this seems an appropriate reward for his performance. While his role was mostly passive, he was involved quite frequently in the discussion by responding to questions, and this is in accordance with the descriptions in the bands of the scores received.

The same could be said for Machiko, for whom the qualitative analysis showed a leading performance, though the indices revealed that it was still below average compared to her peers. As such her scores vary around 2.5 in each band, and this appears to be appropriate according to the wording of the scales.

Figure 45: Scores of low initial group in the second administration

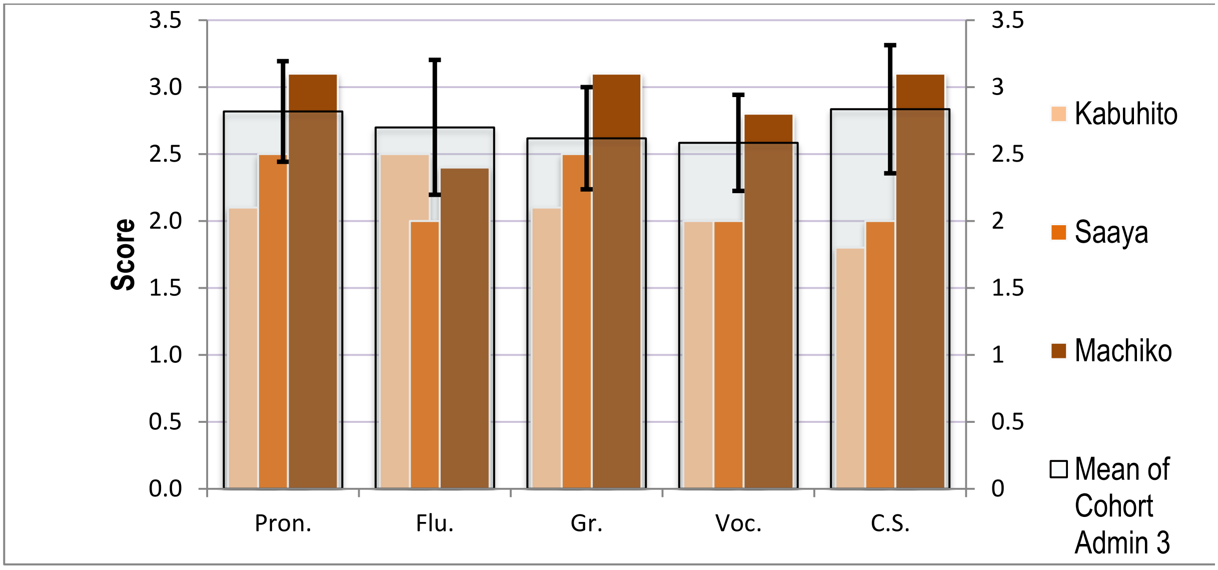


For Saaya the second administration sees her scores vary around the 2.0 level, except for her best score in the communicative skills (CS) band, which reaches 2.5. Apart from this score, Saaya's probably did enough to provide support for her other scores according to the indices and qualitative analysis. In the CS bands (Table 57) the relevant wording in the 1.0 – 1.5 band is that the speaker does "not initiate interaction", and in the 2.0 – 2.5 band "responds to others without long pauses to maintain interaction". In the qualitative analysis we saw that in the single turn that she took she was the last of the speakers to give her opinion and the long 7 second pause before it was evidence for it being a turn granted her by her fellow test-takers. Although it meets the bare minimum as an initiating statement according to the definitions used in the interactive analysis, it is difficult to justify according to the scoring bands. The evidence provided by her participation as noted in the qualitative analysis strongly suggests that a score between 1.0 and 1.5 in the CS band would have been appropriate.

5.1.2.4 The low initial group in the third administration

For the final administration, Figure 46 shows the scores they achieved. Kabuhito's scores are now just above the 2.0 band in every band except for communicative skills, which is at 1.8. The qualitative analysis showed that Kabuhito's performance still required much support from sympathetic listeners in order to be understood, and it is difficult to justify most of the 2.0 grades he was awarded.

Figure 46: Scores of low initial group in the third administration



For pronunciation, there are some contracted forms, but he still says "I am so relaxed" by annunciating each word separately, and katakana[10]-influenced pronunciation is still very evident when he says "ground". For fluency, he manages the first turns he takes in the group oral quite well, but this would have been the opinion he prepared in the minute before the test started, and quite likely, the initial turns of his group mates. In the turns he takes later in the GOT that he had much less time to prepare for there is certainly constant "groping for words", and that matches the criteria for the 1.0

[10] English influenced by native Japanese pronunciation

band. For the grammar bands there is no attempt at a complex form. The wording for the vocabulary 2.0 band is that the participant has "enough lexis for expressing some opinion", but the constant searching for words, the strain shown by his gesture (Gan & Davidson, 2011) and the amount of support he needed suggest that a score in the high 1.0 band might be more appropriate. Although he produced an impressive number of lower frequency words, these were nearly all concrete nouns rather than words suitable for academic discourse. It can be conjectured that the higher proficiency speakers who surrounded Kabuhito affected the raters' objective judgment, as one variety of the 'halo effect'. A less traditional explanation would be to view it from the perspective of IC, which Young describes as "locally constructed" (2009, 2011). Kabuhito appeared more skilful to the raters in this context because of the support he received from his groupmates, and the raters responded by giving him a higher score.

In general, Machiko's performance in the third administration appears to have been appropriately reflected in the scoring bands. In her third test she reaches 3.1 in the CS scale which is a fair reflection of her performance given that she fulfils the criteria of the 3.0 band of appearing "generally confident" and "responding appropriately", although her faltering ability to lead the group as was seen in the qualitative analysis casts some doubt on her ability to "negotiate meaning quickly and relatively naturally" as the band requires. Machiko's vocabulary score of above 2.5 looks supportable, as she only uses three words outside the first 1000 most frequent word band or the prompt, and she does not always use them accurately. Her lowest score is in fluency, which is the only score less than average for her cohort, and this seems to reflect her lower than average speed fluency indices, which are also well below the average of her cohort.

The score that seems the least justifiable is her grammar, which was given as 3.1, and is above the average of this cohort. According to the bands, for it to be in the 3.0 band she needs to be only making errors in "late acquired grammar", yet her transcript shows her still making errors like "go to shopping", and she also makes mistakes in the complex grammar she does attempt. In the indices for complexity and accuracy, although she uses an above average proportion of error-free clauses, the number of error-free clauses is average, and her complexity in terms of words per AS-unit is lower than average. Although these statistics are probably reflective of a lower than average number of clauses she spoke and the number of short turns she used respectively, a high score in the 2.0 band may have been more appropriate for her.

5.1.2.5 Summary of the low initial group

A summary of the scores awarded to the members of the low initial group in three administrations can be found in Table 58. Where there is evidence according to the above analysis that the score given should have been higher or lower an arrow can be seen, and if a question mark appears then it indicates that, as discussed in the analysis above, there are some grounds for accepting the original score.

Table 58: Aberrant scores of the low initial group over three administrations

		Pron	Flu	Gr	Voc	C.S.	Tot.
Kabuhito	Admin 1	1.2 ↓?	1.0	1.0	0.9	1.3	5.4
	Admin 2	1.3	1.3	1.2	1.3	1.1	6.3
	Admin 3	2.1↓?	2.5↓	2.1↓	2.0↓	1.8↓	10.5
Saaya	Admin 1	1.4	1.3	1.4	1.2	1.1	6.4
	Admin 2	2.1	2.2	2.1	2.0	2.5↓	10.9
	Admin 3	2.5	2.0	2.5	2.0	2.0	10.8
Machiko	Admin 1	1.7	1.1	1.4	1.0	1.6↑	6.8
	Admin 2	2.6	2.5	2.6	2.3	2.8↑	12.7
	Admin 3	3.1	2.4	3.1↓?	2.8	3.1	14.1

Key: ↑ - should have been scored higher, according to performance
↓ - should have been scored lower, according to performance
? - indicates possibility

Outside of the scoring of Kabuhito's performance in the third administration, the scores for this group of students seem to accord reasonably well with the performances as described in the qualitative analysis and indices, given the circumstances of the raters (it being an obligatory task for which they have just a couple of hours of training). These participants started low, but had the most room for improvement. The indices and qualitative analysis showed that their abilities did improve, and the raters recorded this in their scores relatively well for Machiko and Saaya over the three administrations, and Kabuhito in the first two. For Kabuhito, although the quantitative indices do show improvements in each succeeding administration, given Kabuhito's performance in the third administration scores in the mid to late 1.0 band would have been more appropriate, and still would have reflected his development in ability to participate in the GOT.

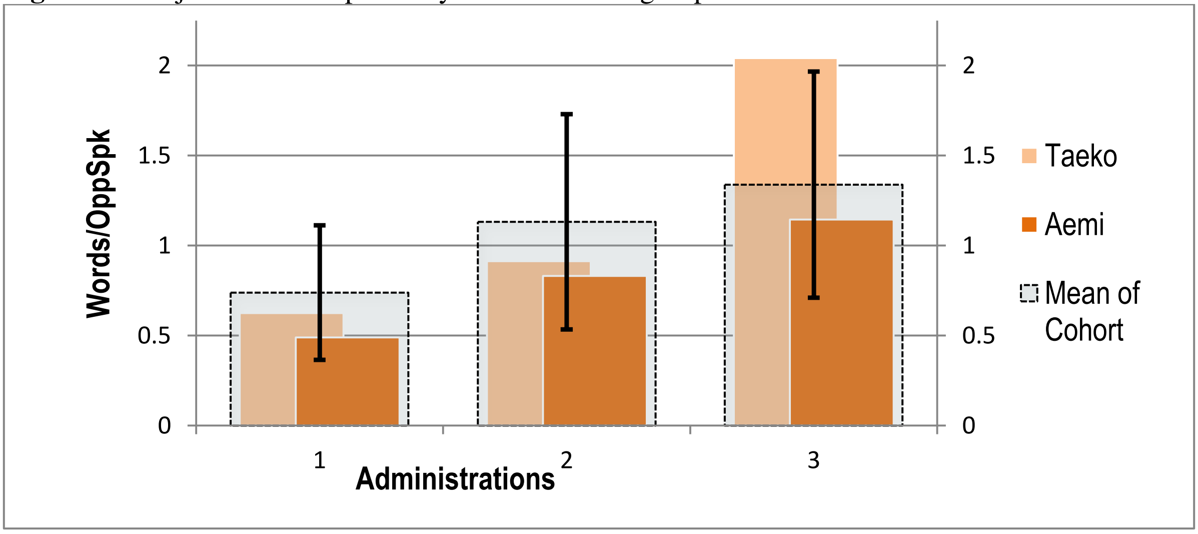
5.2 Comparison of medium initial group with quantitative analysis

5.2.1 The medium initial group's performance according to the indices

As for the previous section, the main findings will be displayed graphically in this section, and the corresponding descriptive statistics can be found in Tables 88 to 91 of Appendix H. The first indices to examine for the medium initial scoring group are the foundational ones of words spoken and length of turn in words in order to give an overview of medium initial test-takers' performances. The first of these can be seen in Figure 47, which shows the words spoken per opportunity to speak for the two members of this group, Taeko and Aemi. While both of these test-takers speak a below average number of words in the first two administrations, they are never outside of the range of the standard deviation. In particular, Taeko speaks a considerably greater number of words spoken in each administration, not only normalized as seen in this graph, but also the raw count.

There is a clear difference between Taeko's final administration and her first two: in her third test the number of words she produced was well above the average spoken, while in her first two she

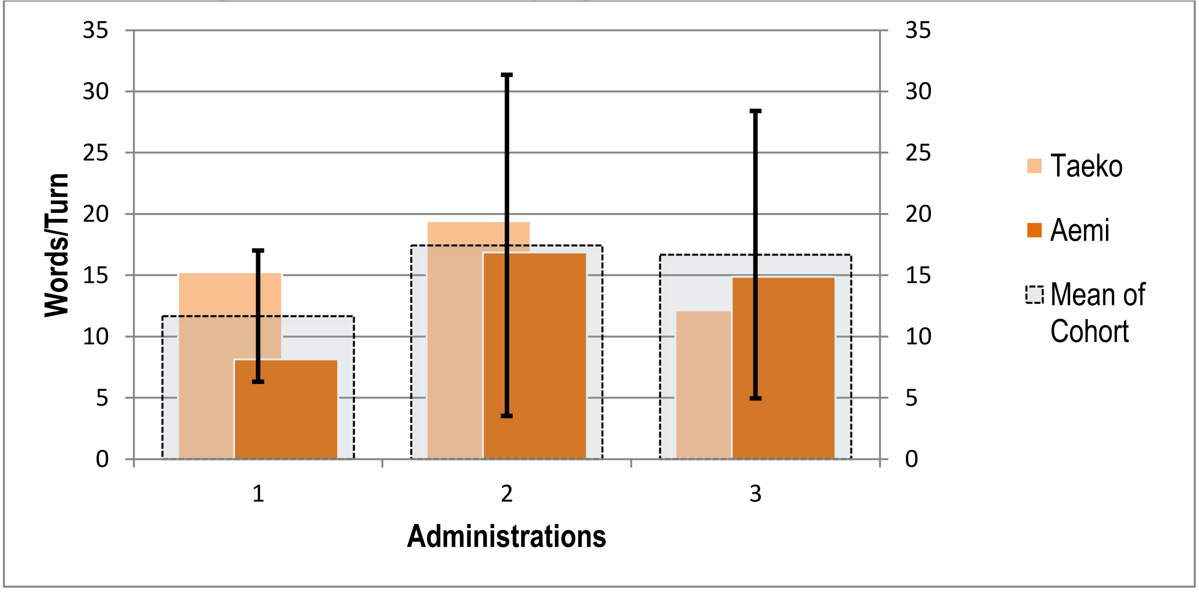
Figure 47: Adjusted words spoken by medium initial group over the three administrations



was just below. Aemi also produced more than double the number of words in her second administration, and although the count of words used do not increase by much in her third administration, the normalized figures, as shown in the graph, still show improvement. This may be reflective of Taeko taking on the identity of 'leader' in the final administration, while Aemi's effort saw her drop back from the leader role she displayed in the second GOT to be just an 'active participant'.

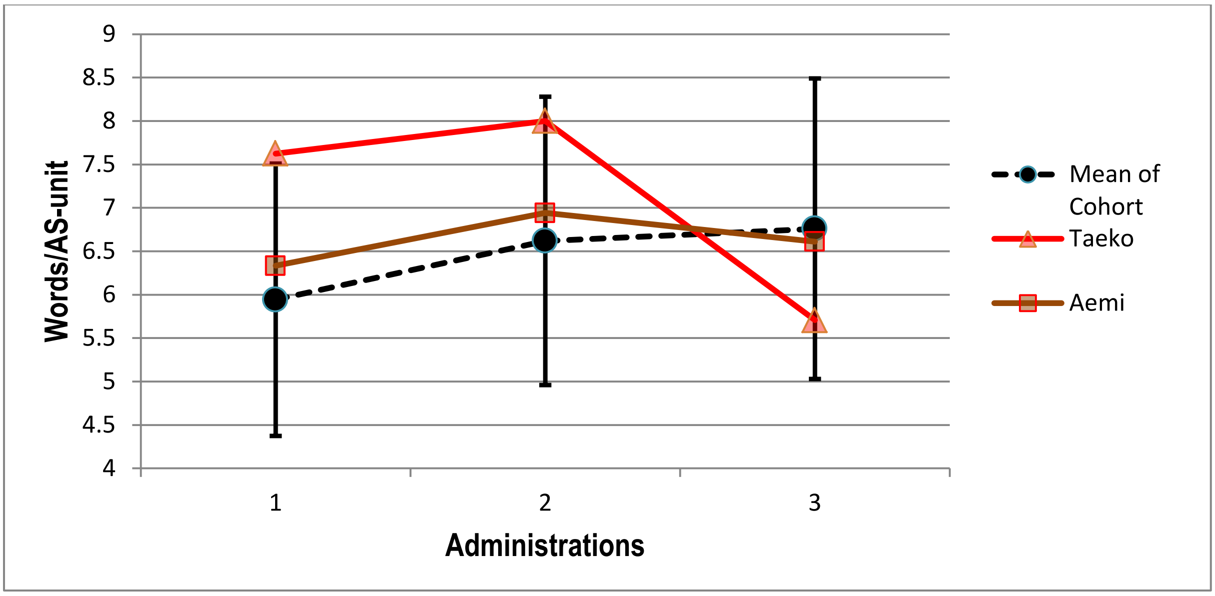
The figures for words per turn in Figure 48 shows a consistent trend with Saaya and Machiko of the low initial scoring group, by showing a bulge in the second administration before declining in the final administration. It is worthwhile reminding the reader that this is likely to be at least partly attributable to the prompt, which in this case tended to elicit longer turns. By chance they were both given the 'singles' prompt to which test-takers responded with significantly longer turns than other prompts (Leaper & Riazi, 2014). Taeko's statistics show a more pronounced dip in the third administration than Aemi's, probably due to the number of shorter terms she used to lead the discussion, as was seen in the qualitative analysis.

Figure 48: Words per turn of medium initial group over three administrations



The next graph (Figure 49) shows the complexity figures of words per AS-unit. As for the low initial scorers, the complexity index of words per AS-unit figures follow the general trend for figures for turn length, with both of these participant's highest figures also coming in the second administration. The plunge in the final administration of Taeko's figures is quite remarkable, as her statistic goes from being well above average to being well below average. By contrast Aemi's figures show much less variation over the three times she took the test, going from just above average in her first two administrations, to just under average in the final one.

Figure 49: Words per AS-unit of medium initial group over the three administrations



The graphs for the accuracy measures of proportion of error free clauses and normalized number of error-free clauses are displayed in Figures 50 and 51. Figure 50 shows both of these participants are similar in that they are higher in the first administration than in the second, suggesting that they were more careful to speak accurately in the first administration, but perhaps focused more on speaking fluently in the second at the expense of accuracy.

In the final administration, both improve to a similar level as their first, when they are speaking with a much greater number of clauses. This results in them speaking a greater number of error free clauses per opportunity to speak, as is shown in Figure 51. As is consistent with the much greater number of words spoken, as was seen in Figure 47, Taeko's figures can be seen rising dramatically in the third administration, and it is in this administration that she beats the average. Aemi's number of error-free clauses per opportunity to speak also follows her words spoken per opportunity to speak, and is likewise a gradual rise that is always just short of reaching the average for the cohort.

Figure 50: Proportion of error-free clauses by medium initial group over three administrations

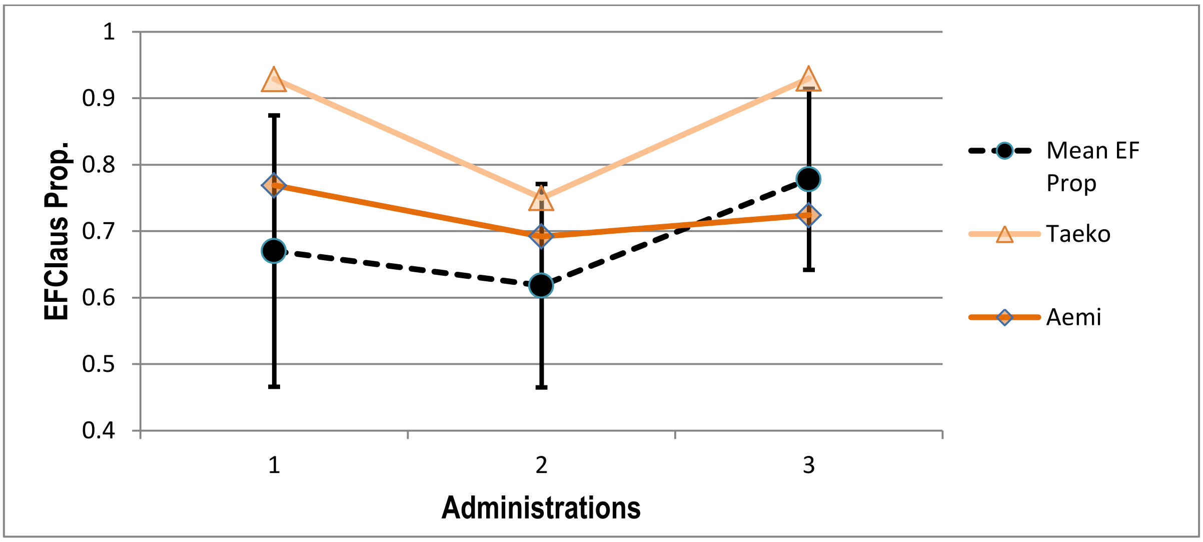
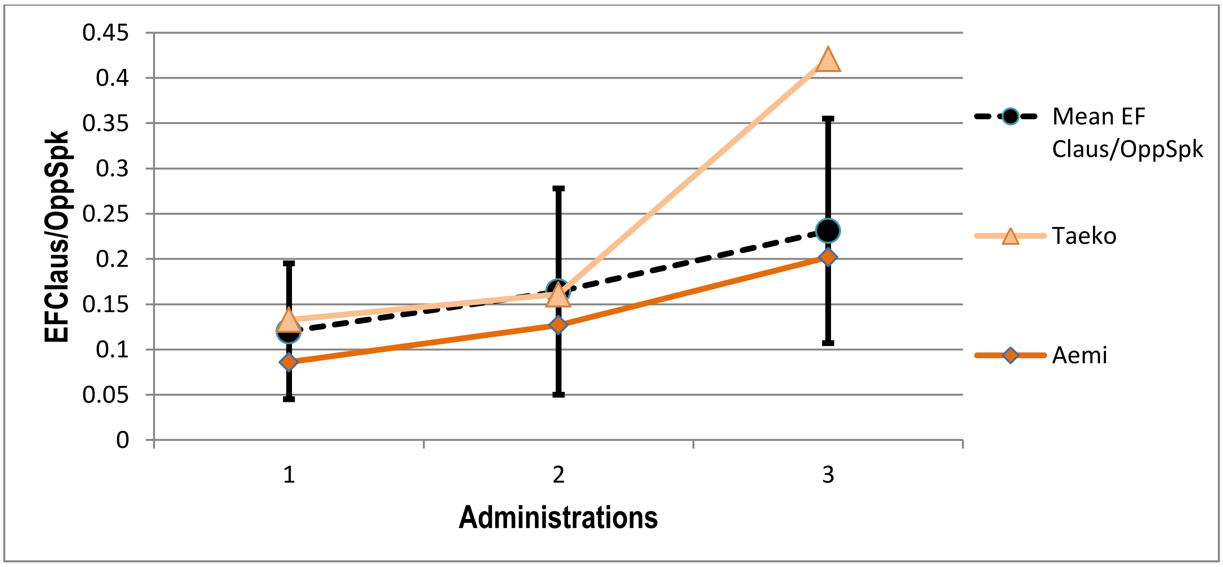


Figure 51: Adjusted error-free clauses of medium initial group over three administrations



For the fluency graphs, Figures 52 to 54 show how the medium initial scoring groups performed compared to the average of the cohort. Taeko and Aemi's articulation and speech rate figures provide interesting contrasts here, with Taeko showing continuous improvements over the three administrations, while Aemi's figures seem to affirm a degree of stagnation in the final administration.

Figure 52: Fluency indices of medium initial group in the first administration

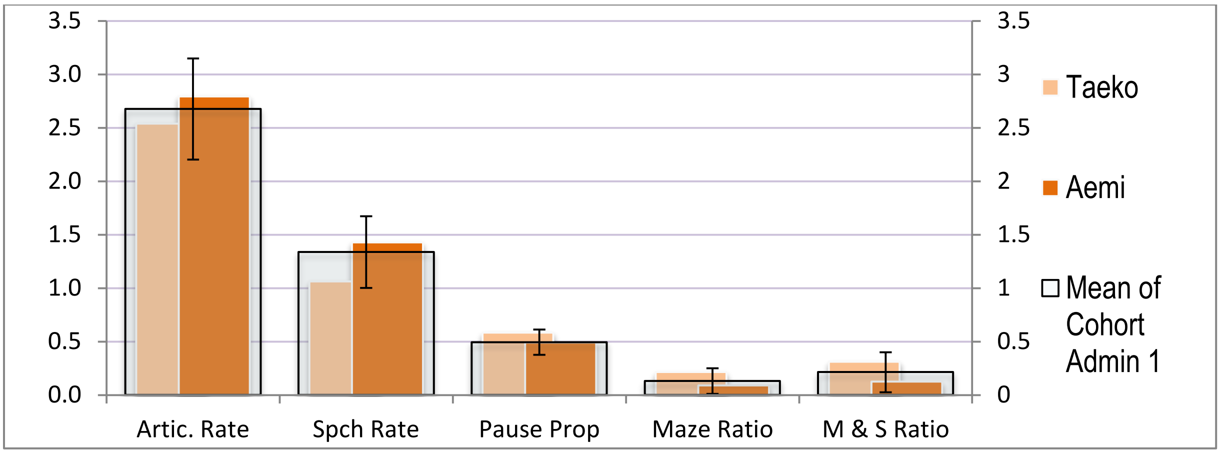


Figure 53: Fluency indices of medium initial group in the second administration

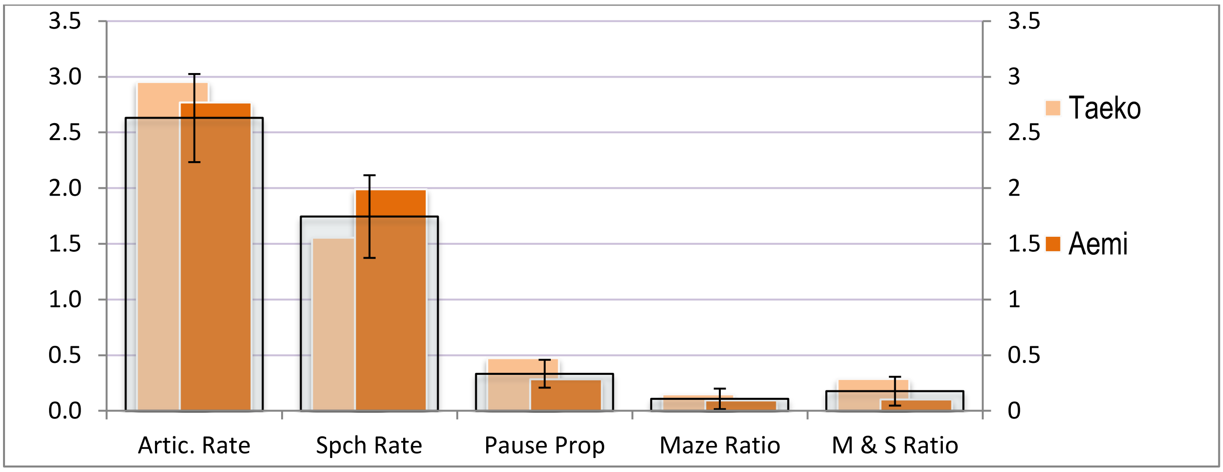
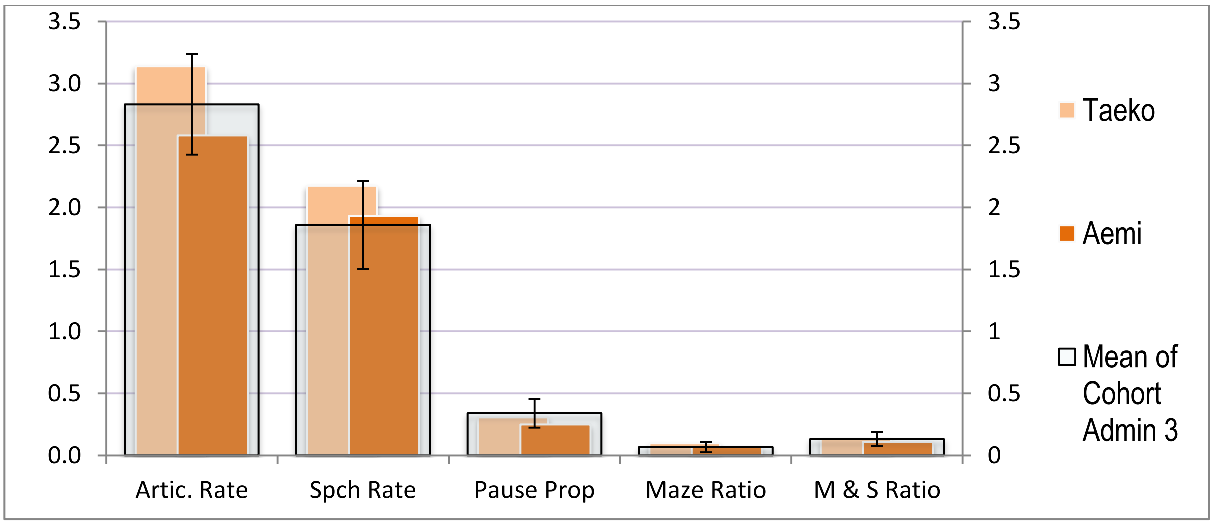


Figure 54: Fluency indices of medium initial group in the third administration



In the first two administrations, Aemi’s figures are slightly above average in both of these indices, but in the second administration she only improves her speech rate through reducing her pausing as her articulation rate does not change. At the same time the average articulation rate of the cohort improves so that in the final administration, Aemi’s articulation rate becomes below average. In the proportion of pauses, both Taeko and Aemi show continuous improvements, especially Taeko in her third administration who becomes better than the average. In the final two indices of repair fluency, Aemi’s performance in these indices varies within a narrow range over the three administrations, and Taeko improves from having slightly worse than the average, to about the level of the average.

When their development in lexis is examined, both Taeko and Aemi show strong growth in the variety of words that they use, as the type count can be seen to increase in successive administrations. They also succeed in improving the lower frequency words, though the sample size is small. However, only Taeko shows continuous improvement in the token count.

Table 59: Development of lexis of medium initial group over three administrations

	Administration 1			
	Taeko		Aemi	
	Type	Token	Type	Token
k1	30	60	34	59
k2 ≤	3	5	1	1
k2 ≤ words	prefer*, tour*, alone		overseas”	
AWL	0	0	1	1
AWL words	-			
	Administration 2			
k1	66	134	47	120
k2 ≤	3	3	3	4
k2 ≤ words	partner, communicate, quit		dream, boyfriend, girlfriend	
AWL	4	7	0	0
AWL words	job, jobs, partner, communicate		-	
	Administration 3			
k1	76	187	73	120
k2 ≤	6	6	5	5
k2 ≤ words	earn, fee, location, sea, ticket, weekdays		concentrate, freshmen, graduate, retired, focus*	
AWL	3	7		4
AWL words	fee, job*, location*		concentrate, focus*, job*, jobs*	
	*word from prompt			

The final category to be discussed is the interactive analysis, for which the graphs are displayed in Figures 24 to 26. Taeko's figures over the three administrations show remarkable consistency in the distribution of the functions and the consistent improvements succeeding administrations, with the most considerable improvement coming in the final one, in which she exceeds the average of every category, marginally in the Initiating features and comfortably in every other function. As was seen in the qualitative analysis, it was the third administration that she demonstrated an identity as leader, and it can be seen clearly in these statistics.

Aemi displays varying patterns over the three administrations: in initiating she reaches or comes near the average in the first two administrations; she does more responding in the first and last administrations, and does more developing in the final two administrations. The qualitative analysis found a section in the first administration which Aemi led, and described her 'leader' identity in the second administration, and it is reflected by her level of initiating in these two administrations. That she did not take on this role in the final administration is shown by her recording her lowest level of initiating and highest levels of responding and developing in that test.

As can be seen in this analysis, the qualitative analysis and performance as depicted by the indices of the medium initial scoring test-takers are largely in agreement. The qualitative analysis showed that Taeko's best performance was in the final administration, and Aemi's was in the second. The indices support this in Taeko's case by showing her most impressive improvements occurring in the third in almost every respect, except for the complexity figures, which probably reflect the greater number of shorter turns she took. For Aemi, for the most part the indices reflect a declining performance in the third administration, especially in the interactive and fluency indices which show either absolute or relative declines. Of the remaining indices, complexity as length of AS-unit showed also showed a decline, but it seems likely this followed a similar decline in the words per turn index, and her accuracy in terms of proportion and number of error-free clauses improved in the final administration. The next section will incorporate this analysis with the scores they were awarded in order to answer the question about how well they reflected the performance of the test-takers.

Figure 55: Interactive indices of medium initial group in the first administration

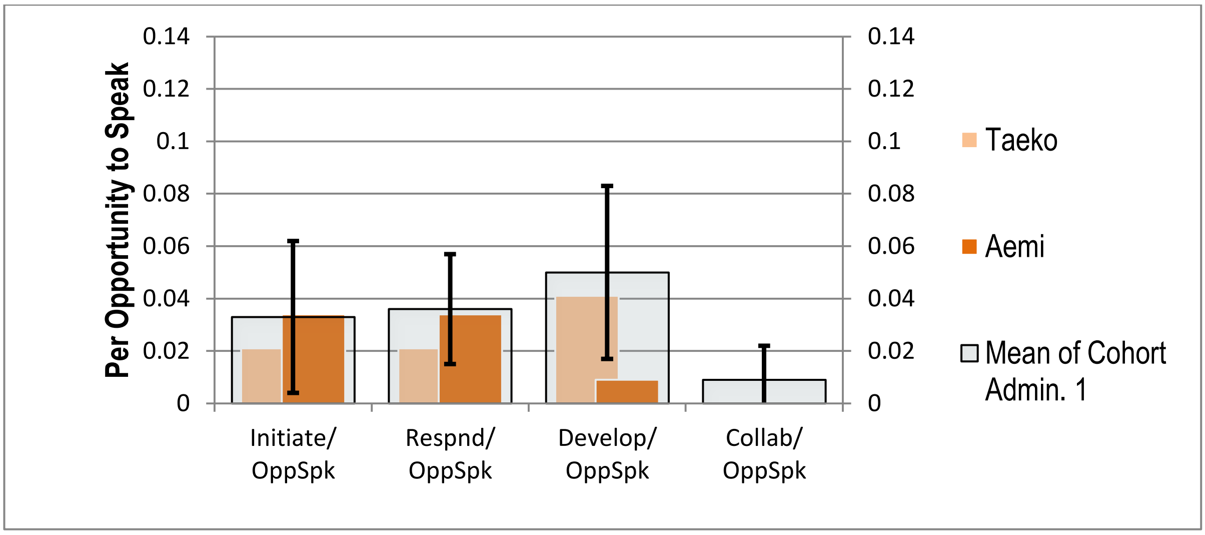


Figure 56: Interactive indices of medium initial group in the second administration

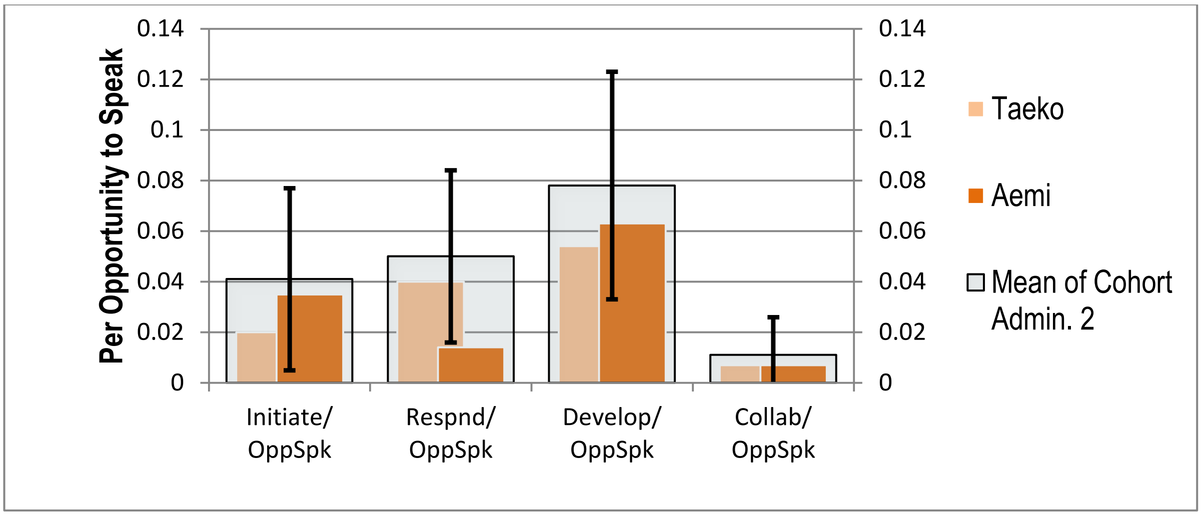
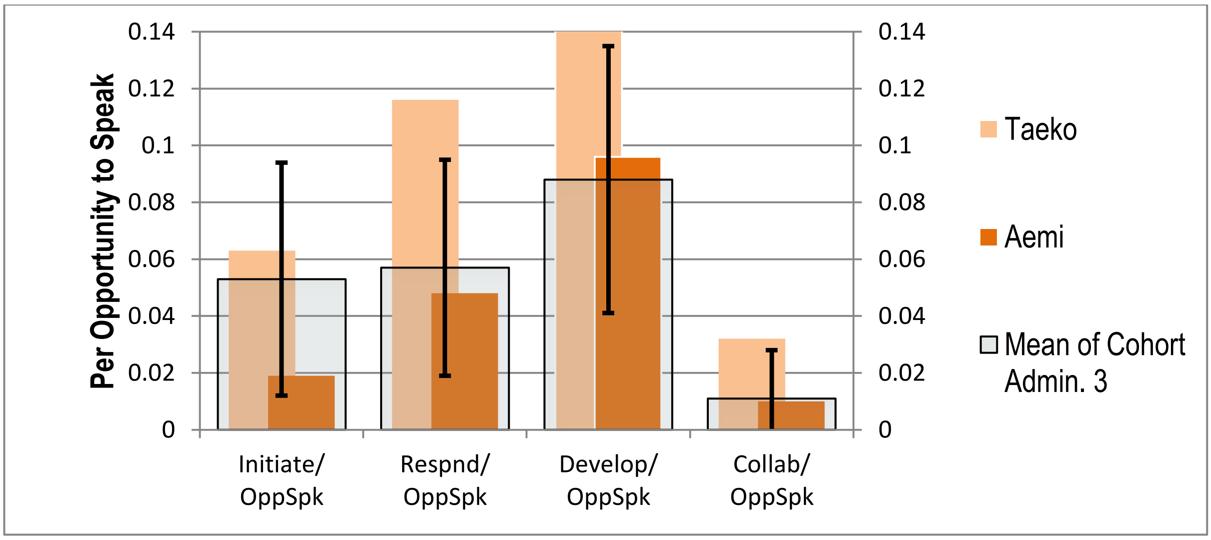


Figure 57: Interactive indices of medium initial group in the third administration

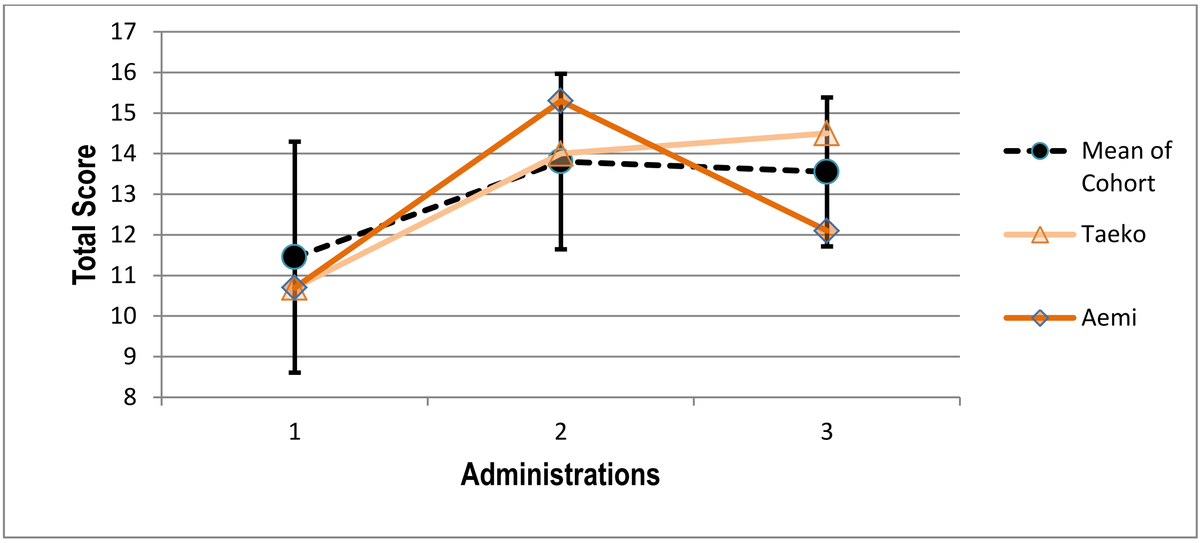


5.2.2 Comparison of medium initial group with scores

5.2.2.1 The overall pattern of scoring of the medium initial group

Before beginning the detailed discussion, the overall trend in scoring of medium initial scoring group is displayed in Figure 58, which shows the total scores of Taeko, Aemi and the cohort average.

Figure 58: Total scores of the medium initial group over three administrations



In Figure 58 it can be seen that although they both had considerable gains in the second administration, they experienced contrasting fortunes in the final one. Taeko went on to achieve a further but diminished gain, following the same pattern as Machiko in the low initial scoring group; and Aemi a suffered a considerable decline to a point that is just less than halfway between her initial and second administration scores.

Before launching into a detailed discussion of their scores in the bands, for reference Table 60 provides their performance as awarded by the raters in each band for each administration.

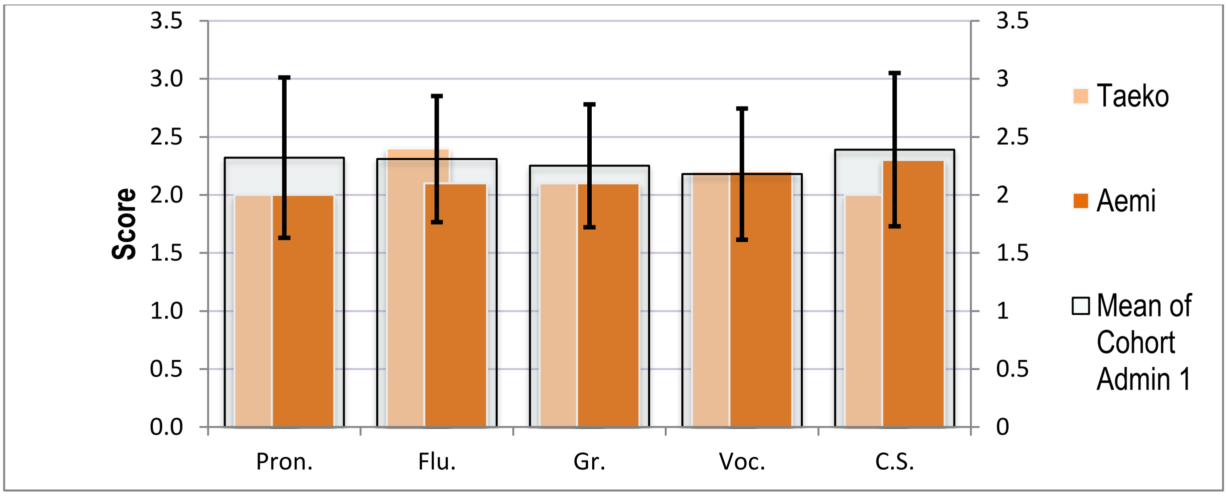
Table 60: Medium initial group's scores over the three administrations

		Pron	Flu	Gr	Voc	C.S.	Tot.
Taeko	Admin 1	2.0	2.4	2.1	2.2	2.0	10.7
	Admin 2	2.7	2.8	3.0	2.8	2.6	14.0
	Admin 3	3.2	2.5	3.2	2.6	3.1	14.5
Aemi	Admin 1	2.0	2.1	2.1	2.2	2.3	10.7
	Admin 2	2.8	3.3	3.0	3.0	3.3	15.3
	Admin 3	2.6	2.3	2.6	2.3	2.3	12.1

5.2.2.2 The medium initial group in the first administration

The performance of the medium initial scorers in the first administration can be seen in Figure 59, which shows that, not surprisingly, they scored very close to the average of the cohort, with their scores ranging from 2 to 2.4 in every band. When comparing their performance in relation to these scores in the communicative skills scale, Taeko's score of 2.0 seems generous. The single question that she asks in the entire test is "How about you?" which is used to transfer her turn after she responded to another person's question, and it appears that this was sufficient for the rater to move her beyond the level 1.0 band which states that a test-taker with this score "does not initiate interaction". It seems that the raters valued this minimal indication of initiative above the contrary evidence of pauses of almost 10 seconds before each of her long turns. This should rule her out of the 2.0 band, for which a test-taker "responds to others without long pauses to maintain interaction." This highlights a potential quandary for raters: if a test-taker does initiate, and yet also takes a long time between turns then the rater may feel torn between the 1.0 and 2.0 bands. In this case, since Taeko's transfer question barely counts as "initiating interaction", and there are consistent long pauses before her turns, a high score in the 1.0 band would seem to be more appropriate.

Figure 59: Scores of medium initial group in the first administration



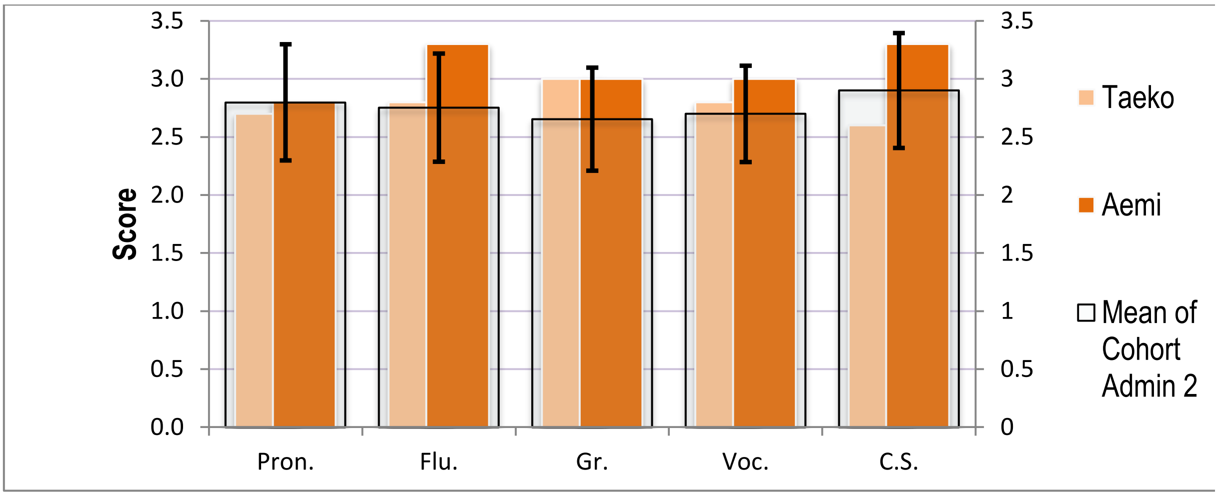
Taeko's scores in the remaining scales in the first administration could be justified if interpreted generously; pronunciation that is "mostly understandable", fluency that does not "impede communication completely", complex grammar that "is attempted but is inaccurate", and enough lexis for "expressing some opinion", though achieving her highest score in fluency is somewhat surprising.

For the other member in this group, as we saw in the qualitative analysis, Aemi shows involvement in her first GOT, even leading a part of the discussion. Her scores in the low 2.0 band are justifiable as it could be argued that she met the criteria for that band on each scale.

5.2.2.3 Medium initial group in the second administration

The graph in Figure 60 shows the scores awarded to Taeko and Aemi in the second administration where it can be seen that while Taeko was awarded scores that vary either side of the average of the cohort, Aemi exceeds the average in all bands except pronunciation, which is at the average.

Figure 60: Scores of medium initial group in the second administration



Taeko's performance in her second GOT is recorded in the indices as being around the average for her cohort, always within range of the standard deviation, and the qualitative analysis described her identity as an 'active participant'. This performance is reflected in her scores which are all above 2.5 with a 3.0 for grammar. Her high grammar score probably comes from such impressive constructions as "it has no meaning to come to university" and "it's difficult to communicate with each other if we have jobs" which in the raters eyes probably qualify her as showing "ability to use some complex grammar" as the 3.0 band in Table 57 has it. For CS, although she did not start a topic, in this administration she initiated several effective follow-up questions, and this would seem sufficient to justify a 2.6 score. Her indices and qualitative analysis point to her scores in fluency being acceptable, and for vocabulary she fits the description for the 2.0 band by having sufficient word usage for the opinions she expresses.

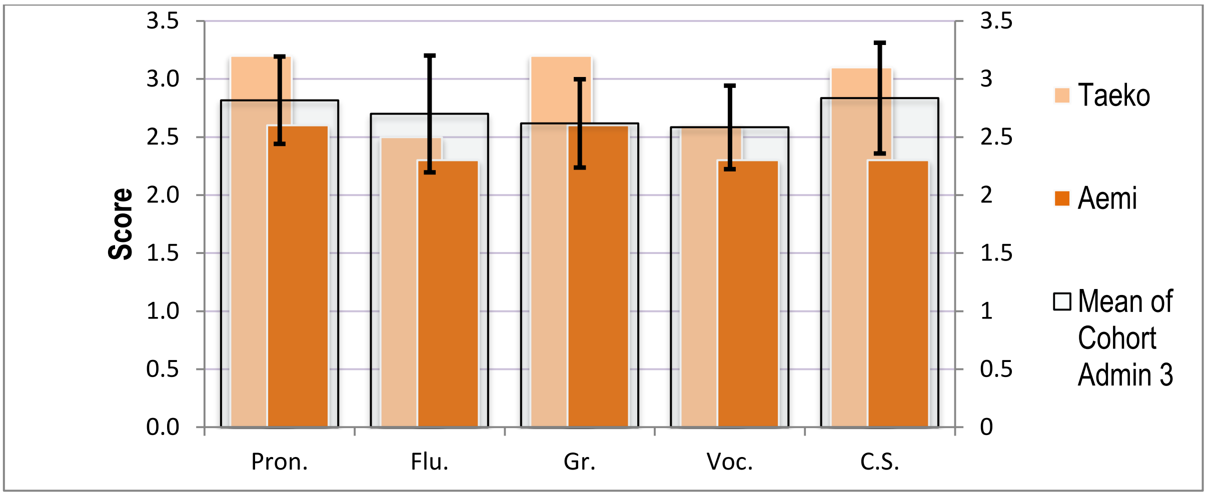
In the second test, Aemi's scores in the low 3.0s in all scales, except pronunciation, which is at 2.7, are sufficient to place her well above average among the other test-takers of her cohort. This contrasts with her indices, which point to a performance that was mostly at the average or slightly better than her cohort; however, the qualitative analysis showed that her performance was more than the sum of its collective parts. Her high score of 3.3 in the CS band was perhaps recognition that apart from her high involvement in initiating topics and asking follow up questions, she corrected another student's grammar, which is a rare event in these GOTs. Her grammar score of 3.0 is mostly like due to the raters being impressed by her attempt at complex grammar in one of her last turns of the test, when she asks the following question: "If you have a girlfriend or boyfriend, so you want you want to get married with her or him, so… but… he or she don't want to get married, how do you think?" Attempting to ask such a question in an assessment situation certainly shows courage, and as it demonstrates "ability to use some complex grammar" her score is justifiable. Her grammar related indices show a slightly higher than average words per AS-unit, and higher than average proportion of error free clauses, though the number of error free clauses was below average.

However, Aemi's score of 3.0 for vocabulary does seem generous, since "advanced vocabulary" is difficult to be justified in her transcript in terms of low frequency vocabulary: the three lower frequency words she uses are *dream, boyfriend,* and *girlfriend.* It could be contended that her score was influenced by her correct use of "get married" and "being married" which the raters would have witnessed many students getting wrong. Whether this should be considered "advanced" is debatable, and so her score here should be considered inflated. Her score for fluency also looks too high: she was awarded a score higher than the range of the standard deviation of her cohort, but her indices were just above the average. Perhaps the raters were influenced by the speed of her response times that were noted in the qualitative analysis.

5.2.2.4 Medium initial group in the third administration

In the final administration, the scores awarded to the medium initial scoring group can be seen in Figure 61. It can be seen in this graph that Taeko surpassed the average in three of the five bands, met the average in one of them, and was slightly less than average in one. Aemi, on the other hand only meets the average in the grammar band, and falls below in the remaining ones.

Figure 61: Scores of medium initial group in the third administration



In Taeko's last GOT, the fluency figures from the quantitative part of the study show that her speed fluency (articulation and speech rate), and breakdown fluency (pause proportion) and the maze and sound ratio for repair fluency are all better than average in her third administration, making it difficult to justify her lower than average score. Perhaps the raters were over-influenced by some examples of message abandonment and did not pay enough attention to the overall fluency with which she spoke that is recorded in the indices. For vocabulary, although she used four words outside the first 1000 most frequent word band, her score was average, and the words *earn, fee, sea, ticket, locate* all appear in the second band. If small things make a difference to the raters, then in the third administration, Taeko using a Japanese filler when searching for the appropriate lexis might have resulted in her not getting a higher score: "I need a lot of money, um, for for *nanyakore* for train, train fee". For the remainder of her scores, pronunciation and grammar seem to be justifiable, but her CS score of 3.1 seems to be low given her leading performance in this GOT. Unfortunately, some of the advanced things she does that should fall into this category are not mentioned in the scoring bands. For example, summarizing other's positions, referring to earlier elements in the conversation, and even her overall involvement in the group oral are absent from the criteria in the scoring bands.

Aemi's final test saw her scores fall below the 3.0 band in all the scales. Her CS score of 2.3 recognizes her lack of involvement in this test. Her score of 2.3 in vocabulary is justifiable given that there are no lower frequency lexical items in her transcript. For grammar, although there are certainly some complex sentences, the quantitative figures are very similar to her second test, and since there is no attempt to use complex grammar to create meaning this time, her score of 2.6 is justifiable. Her fluency score of 2.3 is well below the average of the cohort, which seems harsh when her fluency indices of speech rate, pause proportion and maze and sound ratios are better than average, though her articulation rate is lower. It is conceivable that her low score in fluency is a 'halo effect' from her other low scores, and probably influenced by her lack of involvement the third time she takes the test.

5.2.2.5 Summary of medium initial group

To summarize the ability of the scoring bands to account for the performance of the medium initial scoring group as recorded in the qualitative analysis and indices, some discrepancies could be detected, as presented in Table 61.

Table 61: Aberrant scores of the medium initial group over three administrations

		Pron	Flu	Gr	Voc	C.S.	Tot.
Taeko	Admin 1	2.0	2.4	2.1	2.2	2.0↓?	10.7
	Admin 2	2.7	2.8	3.0	2.8	2.6	14.0
	Admin 3	3.2	2.5↑	3.2	2.6	3.1↑	14.5
Aemi	Admin 1	2.0	2.1	2.1	2.2	2.3	10.7
	Admin 2	2.8	3.3↓?	3.0	3.0↓?	3.3	15.3
	Admin 3	2.6	2.3↑	2.6	2.3	2.3	12.1

Key: ↑ - should have been scored higher, according to performance
↓ - should have been scored lower, according to performance
? - indicates possibility

In the first two administrations the scoring was representative enough for all but two band scores that were overly generous: Taeko's communicative skills score seems to be more worthy of a high 1.0 band score, and Aemi's vocabulary score of 3.0 did not seem justified by the presence of "advanced" vocabulary as specified in the bands. By contrast, the final administration seems to have been marked more strictly, with Taeko's fluency score declining when the indices clearly show her improving, and the communicative skills score not recognizing her leadership of the group. This latter discrepancy is exacerbated by the wording of the band not including the functions that she used. If the scoring bands followed the progress of Taeko's development appropriately, she would have made a larger gain in the third administration.

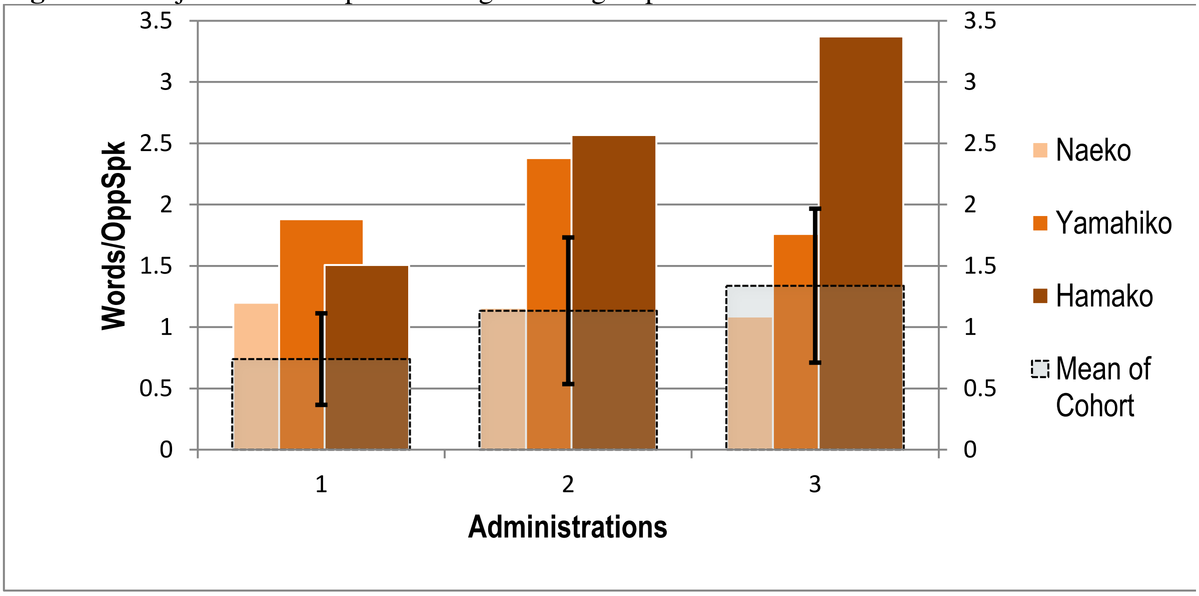
For Aemi, her lower scores in the final administration do recognize the lower level of performance for the most part, but it is difficult to justify a decline by an entire point in fluency, when the indices show her performance did not decline to that extent from the previous administration.

5.3 Comparison of the high initial group with quantitative analysis

5.3.1 Comparison of high initial group with indices

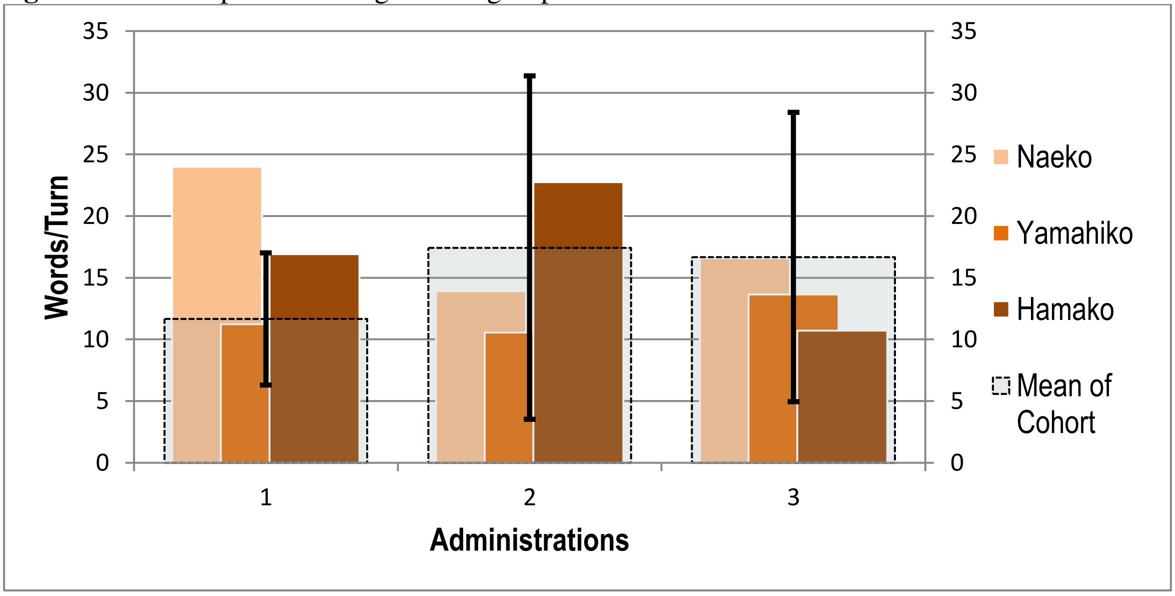
As for the previous two sections, graphs will be used to display the findings here, and the full descriptive statistics are available in Tables 92 to 94 in Appendix H. The first figures to consider for the high initial group are the core statistics of words spoken per opportunity to speak and words per turn. Figure 62 shows the graph for words spoken per opportunity to speak, and it can be seen that the three participants display a different pattern. Naeko and Hamako both showed consistent trends, but where Naeko's number of words constantly declined, Hamako used considerably more words relative to the time of test and participants in each successive administration. By contrast, Yamahiko used more words in the second administration than the first, but then declined precipitously in the third administration to a lower level than the first. Despite this decline, he still spoke considerably more words than the average in his last GOT.

Figure 62: Adjusted words spoken of high initial group over the three administrations



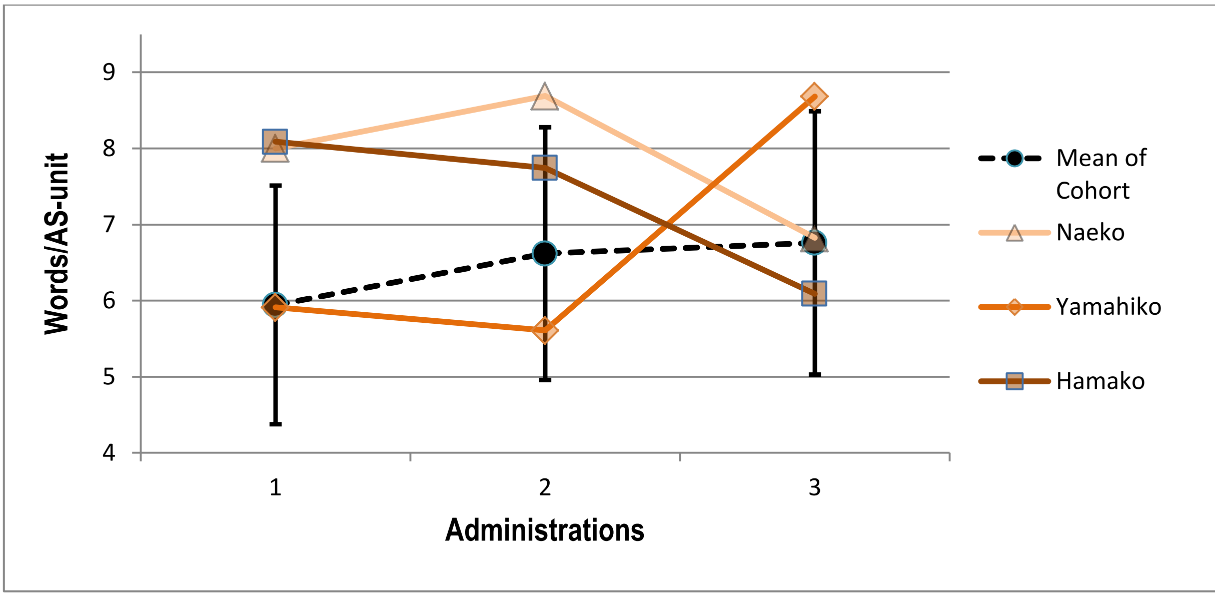
The words per turn statistics vary widely, as can be seen in Figure 63, and this variance is illustrated by the administration in which the three participants recorded their longest average words per turn: Naeko's highest figure is in her first administration, Yamahiko's in his third and Hamako's in her second. The number of turns Naeko took peaked in the second administration, before decreasing in the third, though still at a higher level than her first administration, and coupled with the number of words spoken per administration meant that her longest average turn length was in the first administration, the shortest in her second with the third administration in between. The number of turns Yamahiko took followed the same trend as his words spoken in the first two administrations, and this is reflected in his words per turn statistics being similar. In his third administration he used fewer words in even fewer turns, leading to a longer average length of turn. Hamako's figures show an increase in turn length in the first two administrations, and although she spoke her highest number of words in the final administration, she also increased the number of turns used, bringing down the mean turn length in this administration.

Figure 63: Words per turn of high initial group over the three administrations



The complexity statistics in Figure 64 depict an interesting range of patterns. Naeko's figures show an opposite trend to her words per turn by increasing in her second administration and then declining steeply in her final. By contrast, the longer turns that Yamahiko took in the third administration are reflected in the highest complexity figures coming in that administration. For Hamako, the figures show her highest complexity in the first administration, a small decline in the second administration, and a steeper one in the final.

Figure 64: Words per AS-unit of high initial group over three administrations



For accuracy, Figure 65 shows the proportion of error free clauses and Figure 66 the adjusted number of them. Naeko's accuracy figures in both of these graphs show consistent increases in every administration: a considerable rise in the proportion and a small rise in the number. Yamahiko and Hamako both experience a decline in the proportion of error free clauses in the second administration, with almost exactly the same figures as each other.

Figure 65: Proportion of error free clauses by high initial group over the three administrations

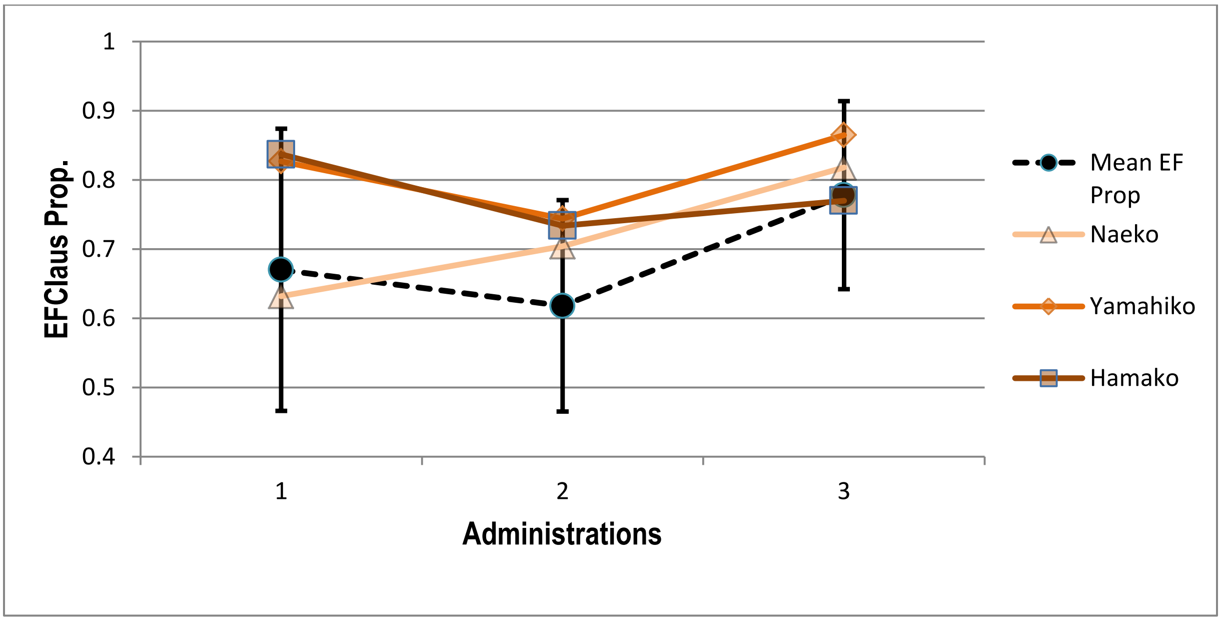
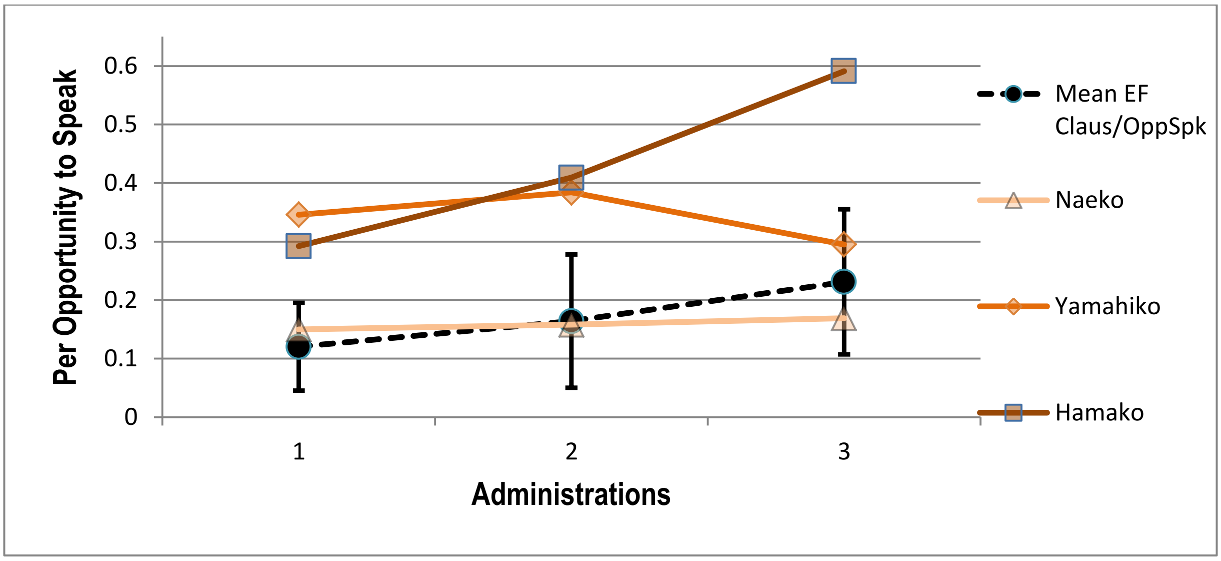


Figure 66: Adjusted error-free clauses spoken by high initial scorers over the three administrations



In the third administration they part ways, with Yamahiko's improving more sharply than Hamako's. Figure 66 shows that, at the same time as their decreasing proportions, they both used an increasing number of error free clauses in their second administration that came with the increased number of words and clauses they spoke. Due to the lower number of words he spoke in the third administration, Yamahiko's number of error free clauses dropped, and in contrast, Hamako's number of error free clauses spoken improved in this administration in line with her increased amount of talk. The high scoring group's fluency statistics are shown in Figures 67 to 69. In these three graphs Naeko's and Hamako's fluency figures are similar in some respects: Over the three administrations they show a constantly improving speech rate, along with constantly declining pause proportion and maze related ratios. However, their articulation rate shows more variance, with it declining slightly in the second administration before improving in the third to a greater level than their first administration. Yamahiko seems to have produced quite varying performances over the three administrations. His articulation rate declines in each administration, while his speech rate also declines in the second administration before improving markedly in the third administration. This may be related to the increased competition for the floor which he experience in his final GOT compared to the first two where there was little competition. He also improves his pause proportion in the third administration, which had not changed much in the first two administrations. After the first administration, his maze ratio improves in the following administration, but the final administration does not see an improvement.

Figure 67: Fluency indices of high initial group in the first administration

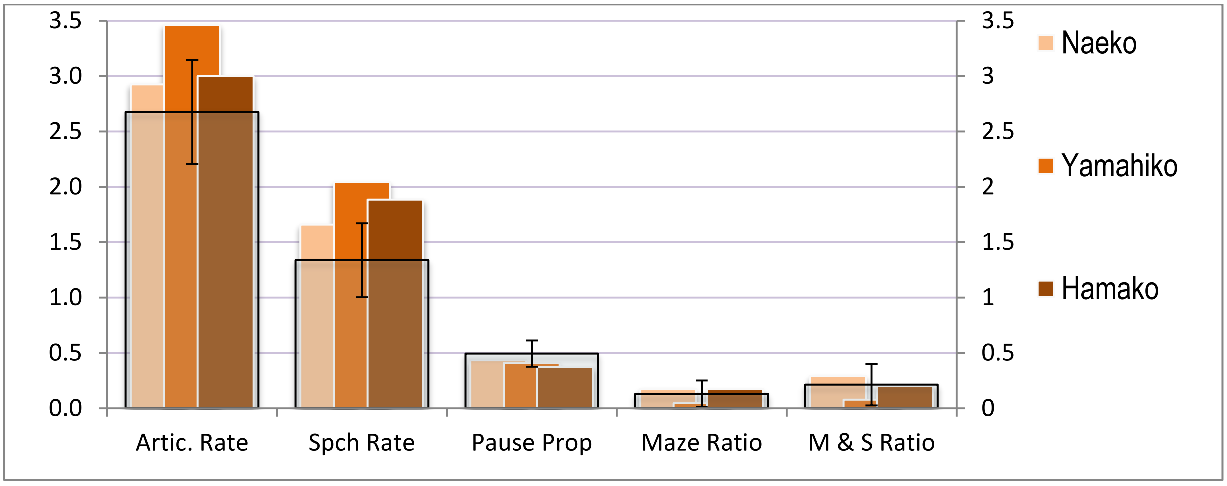


Figure 68: Fluency indices of high initial group in the second administration

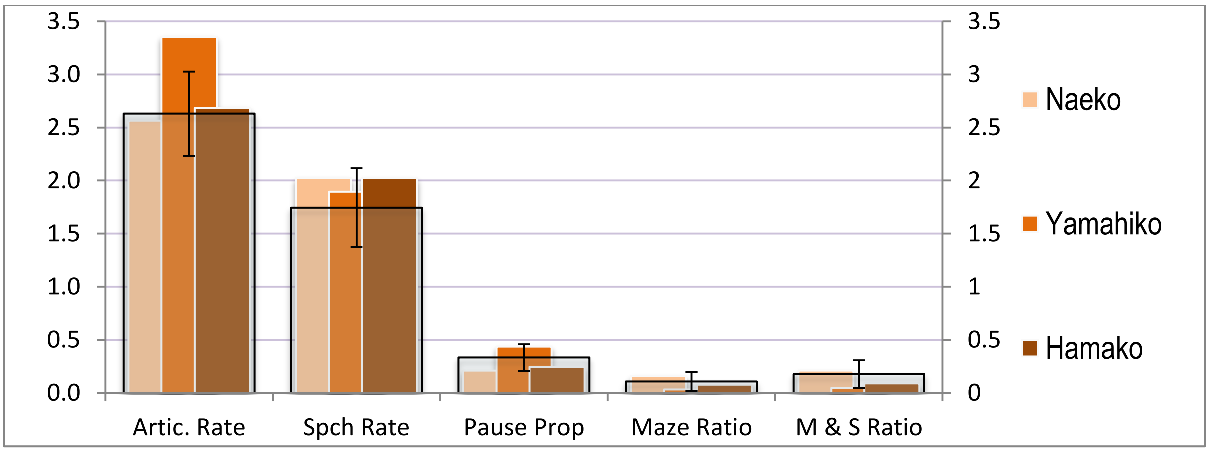
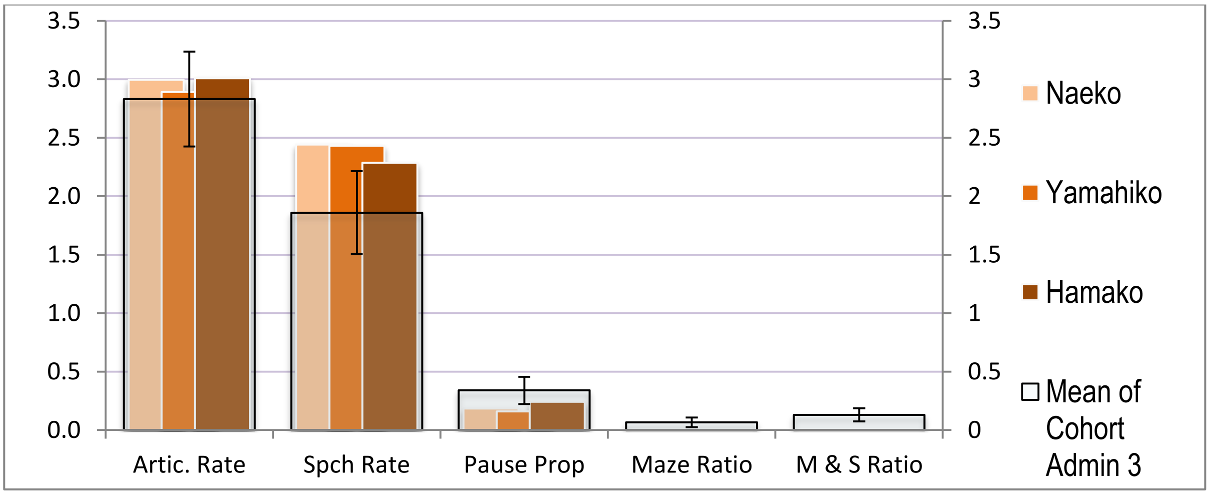


Figure 69: Fluency indices of high initial group in the third administration



When examining the vocabulary counts of the high initial scorers, similarities can be seen in the usage of lower frequency words between Naeko and Yamahiko. For both of them, their first two administrations show their highest usage, before a decline in the final administration. For Naeko and Yamahiko this decline probably results from the smaller sample as they spoke fewer words. After Hamako's improvement in the number of lower frequency words in her second administration, her development can be seen in the way she maintains the number of lower frequency types in her final administration despite also speaking fewer words.

Table 62: Development of lexis of high initial group over the three administrations

	Administration 1					
	Naeko		Yamahiko		Hamako	
	Type	Token	Type	Token	Type	Token
k1	58	91	117	311	85	190
k2 ≤	8	11	12	23	4	5
k2 ≤ words	Asia, Caribbean, culture, exchange, foreign, tropical, Japanese, abroad*,		ancient, Australia, China, Chinese, cool, foreign, Ireland, Italy Japan, Japanese, modern, abroad*		Australia, Chinese, comfortable, tour*	
AWL	1	2	1	1	1	1
AWL words	culture		area		major	
	Administration 2					
k1	78	137	122	336	108	355
k2 ≤	6	6	11	28	11	15
k2 ≤ words	crises, elementary, Japan, Japanese, kindergarten, lonely		alone, addicted, annoying, bathroom, complicated, display, elementary, junior, mobile, tons, yen		ballerina, ballet, cute, dance, divorced, lonely, partner, sad, someday, advantage*, prefer*,	
AWL	1	1	1	1	1	1
AWL words	economical		display		partner	
	Administration 3					
k1	59	118	85	191	104	259
k2 ≤	2	2	2	3	11	13
k2 ≤ words	boring, former		countryside, entire		convenience, convenient, fat, Japan, Japanese, rice, store, taste, variety, vegetables, vitamin	
AWL	1	3	0	0	1	1
AWL words	job		-		affected*	
	*word from prompt					

The final set of graphs in this section (Figures 70 to 72), display the high initial group's interactive statistics over the three administrations, and show contrasting patterns.

Figure 70: Interactive indices of high initial group in the first administration

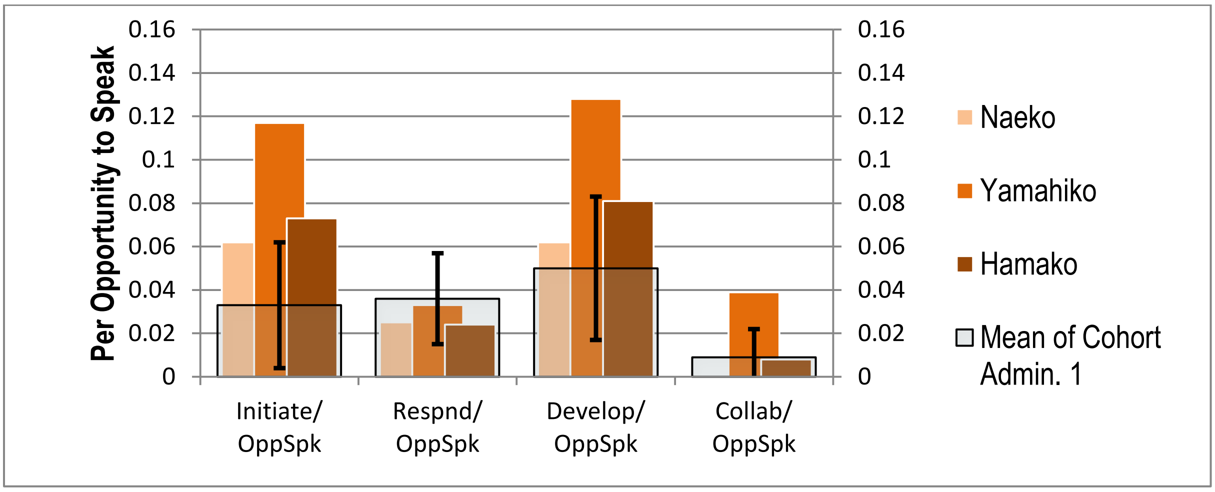


Figure 71: Interactive indices of high initial group in the second administration

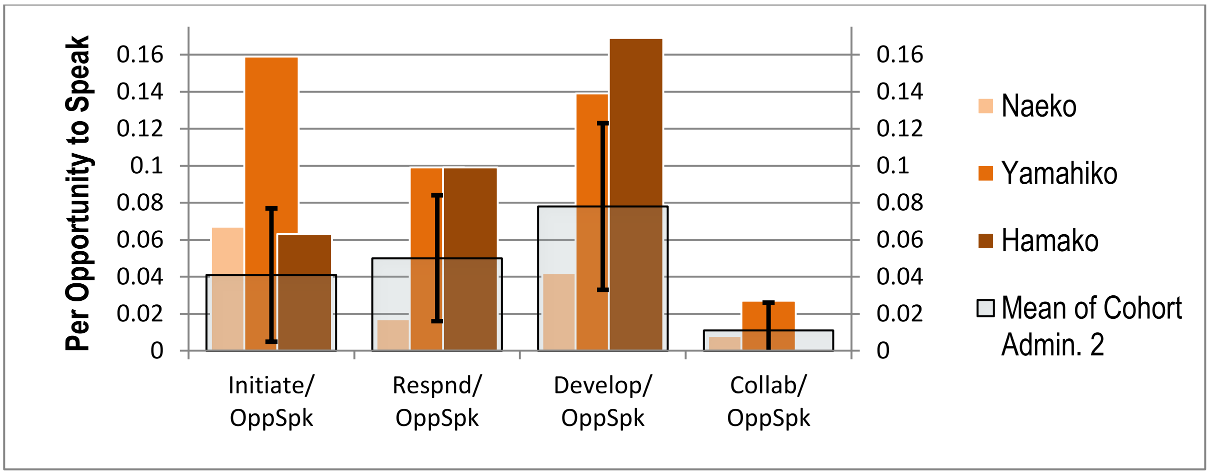
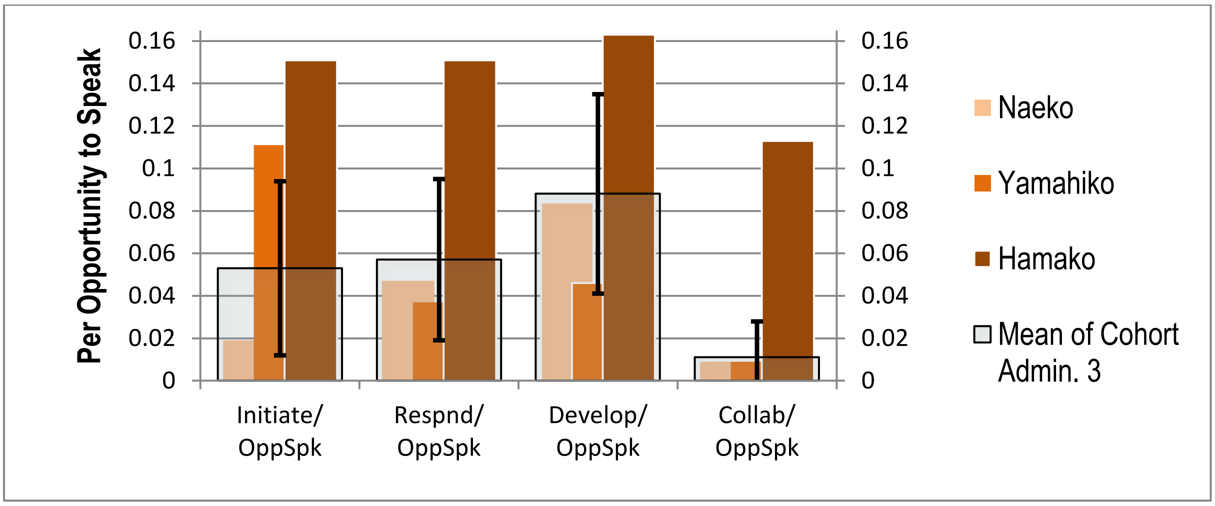


Figure 72: Interactive indices of high initial group in the third administration



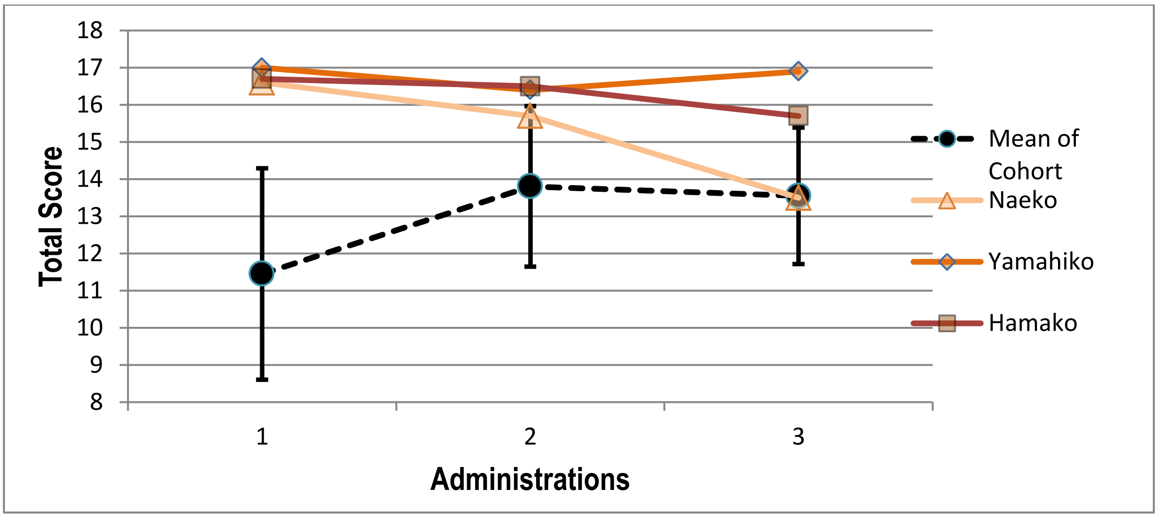
Naeko had a similar level of initiating in her first two administrations, but dropped off in her final one, in which she achieved her highest responding and developing figures, an accurate reflection of the role in the third GOT as analysed in the qualitative section. Yamahiko's figures show that he peaked in the use of Initiating, Responding and Developing features in the second administration, but had more Collaborating functions in his first administration. His final GOT, when he was grouped with higher level learners shows a decline in the number of these functions, but as the qualitative analysis shows, it was probably the discussion in which he showed the most communicative ability. Hamako's figures also show that her strongest performance was in her final administration, with only the Developing features being higher in the second administration where she spent more time explaining her own stance.

5.3.2 Comparison of high initial group with scores

5.3.2.1 The overall pattern of scoring of the high initial group

The overall picture of the scoring for test-takers in this group is depicted in Figure 73 where a different scoring pattern to the participants in the other groups can be seen. The highest score in this test for all three of these participants comes in their first test, and they never quite manage to reach that point again.

Figure 73: Total scores of the high initial group over three administrations



For Naeko, the story of her progress seems to be one of gradual decline, as her total score drops considerably year on year, from a total of 16.6 to a mere 13.5. Yamahiko shows the least variance of the members of this group, with never more than 0.6 points between his totals over the three tests. Hamako's scores also drop, but not as dramatically, through there is a difference of an entire point between the total scores of her first and her last test. A summary of their scoring in the various bands over the three administrations can be found in Table 63.

It seems likely that the scoring of students in this group is an aberration not only on a prima facie level, where it might be considered unlikely that the students' conversational skills would decline over two years of exposure to English only instruction on a fairly intensive program, but also on the basis of the qualitative and quantitative analyses which has already made it clear that this did not happen to all members of this group. As such, the justification for their scores need to be examined with particular care, as will be done in the next section for the first administration.

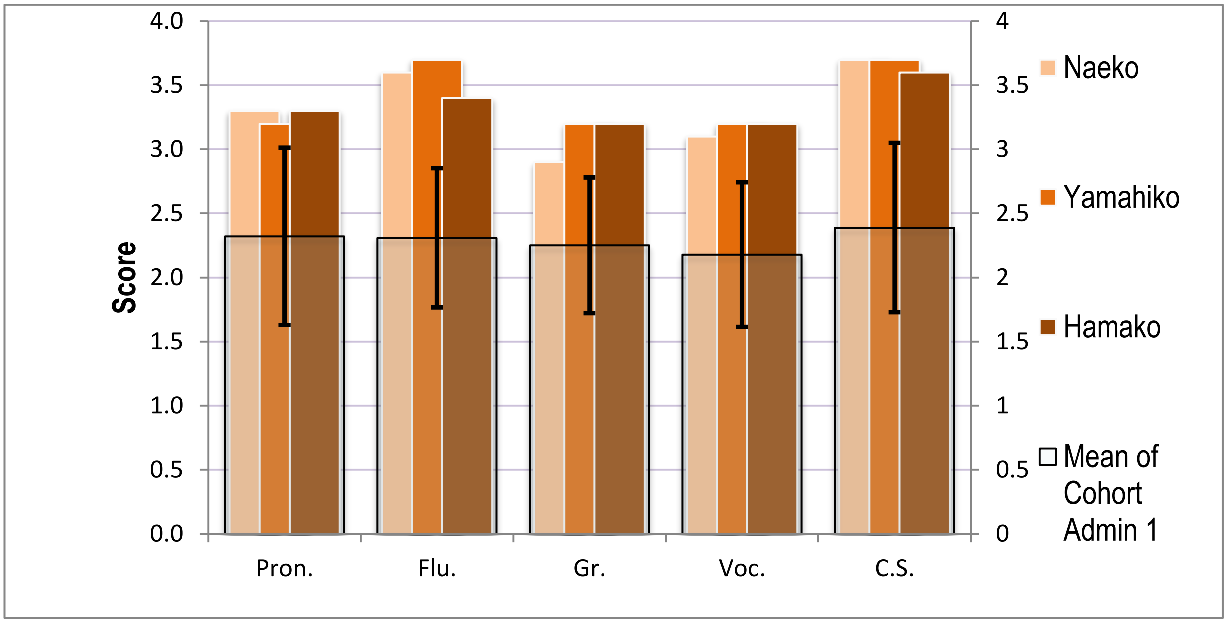
Table 63: High initial group's scores over the three administrations

		Pron	Flu	Gr	Voc	C.S.	Tot.
Naeko	Admin 1	3.3	3.6	2.9	3.1	3.7	16.6
	Admin 2	3.1	3.0	3.1	3.1	3.3	15.7
	Admin 3	2.9	3.0	2.9	2.3	2.5	13.5
Hamako	Admin 1	3.3	3.4	3.2	3.2	3.6	16.7
	Admin 2	3.3	3.0	3.4	3.1	3.7	16.5
	Admin 3	3.0	3.1	3.0	2.9	3.8	15.7
Yamahiko	Admin 1	3.2	3.7	3.2	3.2	3.7	17.0
	Admin 2	3.2	3.4	3.2	3.3	3.4	16.4
	Admin 3	3.4	3.0	3.4	3.1	3.5	16.9

5.3.2.2 High initial group in the first administration

As can be seen in Figure 74, the first candidate, Naeko scored in the 3.0 band in all scales except grammar, which was just below at 2.9. For her score of 3.3 in pronunciation, the crucial distinguishing point in the scale appears to be whether there is evidence of her "blending words". She

Figure 74: Scores of high initial group in the first administration



does this by quite clearly using contractions “I’m” and “I’d” and when she speaks she has runs like “as an exchange student from my high school” without pausing. In this utterance there is clear blending between the first three words, and provides evidence that her pronunciation score is justified. In the fluency scale she scored 3.6, which puts her beyond the range of the standard deviation of the average for her cohort, and so is an extraordinarily high score. The description in the scoring bands require that she “rarely gropes” for words though her speech “still may not be quick”, and since her score is over halfway to the fourth band, she should show some of the qualities of the highest band, which includes the description of using “fillers effectively” and speaking “quickly in short bursts”. In support of this high score, an examination of her performance does not reveal overly long groping for words in the long turns that she takes, and the fact that she did speak in long turns probably counted in her favour. Additionally, the passages quoted above in support for her pronunciation score could also help account for the score she was awarded. However, she does not use fillers, and in her indices, as we saw in Figure 67, her speed fluency was only above average, perhaps not to the extent that her high score warranted since both her articulation and speech rate fell well within the range of the standard deviation. For breakdown fluency, her proportion for pausing is better than average, but the repair fluency measures of are both worse than average. This makes the awarding of her high fluency score dubious; a score between 2.5 and 3.0 may have been more appropriate.

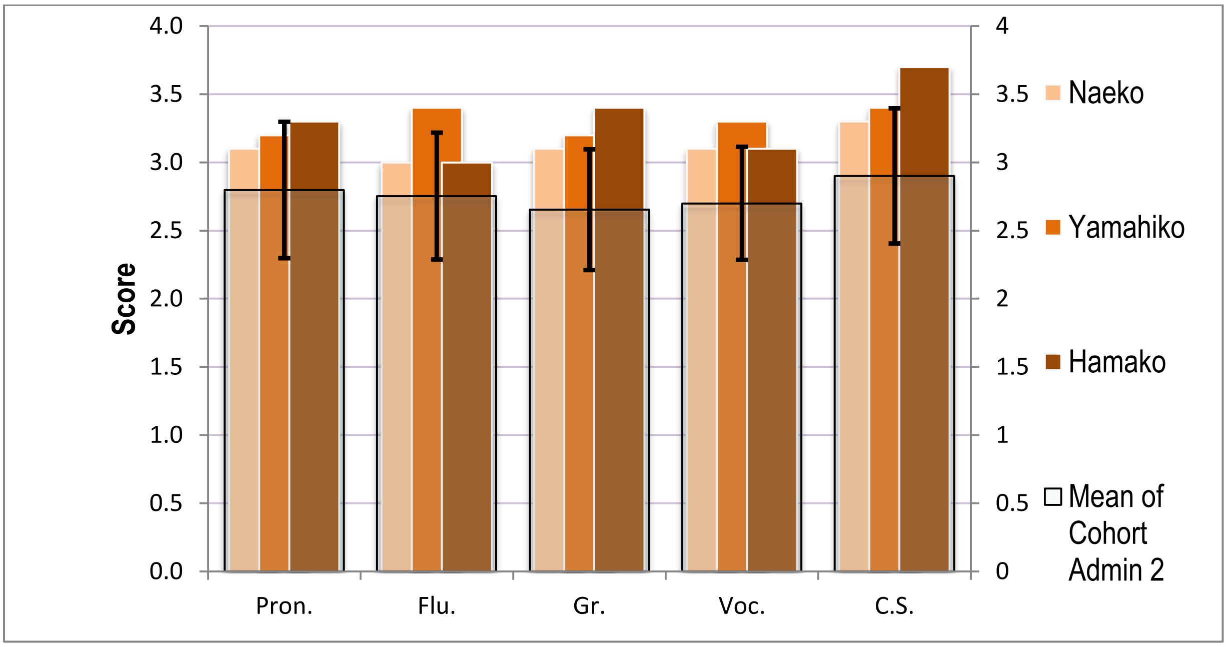
The raters scored her grammar and vocabulary as the weakest of Naeko's speaking skills. For her grammar score of 2.9, complex grammar needs to be "attempted", even if "inaccurate". An examination of her transcript shows she can use 'verb + to infinitive', as she demonstrates in a question she answers: "I'm interested in tropical country… ah so I I'd like to work in… Caribbean or… ah… Southeast Asia." but mostly she uses simple sentences strung together with co-ordinating conjunctions, so a score of 2.9 is probably on the generous side of the 2.0 range. Her vocabulary score of 3.1 was probably boosted by her accurate use of lexical items like "culture" and "tropical country" (in the second and fourth 1000 most frequent word lists respectively of the British National Corpus [BNC]); on the other hand she can be seen substituting high frequency words instead of more appropriate lower frequency words you might expect students to be familiar with, for example, she says "little bad point" instead of "small disadvantage". For generous raters, you could say this 'shows evidence' of some advanced vocabulary, but a high score in the 2.0 band would probably be more appropriate. Finally, to earn her 3.7 CS band, Naeko starts the test and then restarts when the conversation ran down, but a person scoring this should be able to "negotiate meaning quickly and relatively naturally", and as noted in the qualitative analysis there were long pauses before her initiations, and this makes her almost perfect score in this category unreasonably generous. Indeed a score of 3.7 means that it should contain some elements of the highest grade of 4.0 in it, and this includes the criteria "asks others to expand on their views, shows how own and others' ideas are related, interacts smoothly", and it is implausible to find traces of this in her first GOT. It seems then, that of the five scales, she may have been over-generously scored in four of them.

The scores for pronunciation and fluency of the other two participants in the first administration, Yamahiko and Hamako, ranged from 3.2 to 3.7, and for confirmatory evidence there is no shortage of examples of word blending. Additionally, using fillers "like" and "right, let's say" would have contributed to Yamahiko's score of 3.7 for fluency. Moreover, according to the indices, both of Yamahiko's speed fluency indices are well beyond the range of the standard deviation of the cohort, as is Hamako's speech rate. For grammar and vocabulary Yamahiko and Hamako had the same score of 3.2, and there are sufficient examples of complex grammar to support a 3.0 band score. Yamahiko can put together grammatical constructions like "You don't want to live there or anything like that?" and Hamako, "It will be good for us to go with many people because we can help each other". To earn their vocabulary score of 3.2 there is evidence for less frequent vocabulary according to the BNC: for Yamahiko, "ancient", "cool" and expressions like "it's hard to do"; and for Hamako, and "wherever you want to". Their scores of 3.6 and 3.7 for communicative skills, while generous, can also be justified since they adequately fulfil the criteria of being "generally confident", they respond 'appropriately to others opinions' and there is evidence of their ability to 'negotiate meaning quickly and relatively naturally' in their quick reaction times, as noted in the qualitative analysis. To fulfil this criteria, Hamako puts forward her opinion, asks others what they think, and then puts forward an opposing opinion, negotiation both sides of the argument in a relatively efficient way. Yamahiko dominates his discussion by means of his questions, which he tailors according to information that he learnt about his interlocutors earlier in the conversation, and in this way it can be said that he "negotiates meaning". Thus, the scores for Hamako and Yamahiko in the first administration are justifiable, unlike those for Naeko.

5.3.2.3 High initial group in the second administration

In the next administration, the scores in the scales of these participants fall in a narrow range between 3 and 3.7, as can be seen in Figure 75. Naeko's scores range from 3.0 in fluency to 3.3 in CS. As justification for her CS score, Naeko led the discussion by initiating topics and asked a clarification question, which counted as a Collaborating feature. However, as noted in the qualitative section, Naeko found herself in a group in which there was a distinct lack of competition for floor time, and thus took the identity as 'leader by default'. To justify a score in the 3.0 band the stipulation in the criteria is to show "ability to negotiate meaning quickly and relatively naturally" and it is questionable whether she does this. As such, a score high in the 2.0 band might have been more appropriate here.

Figure 75: Scores of high initial group in the second administration



Naoko's fluency score in the second administration was 3.0, which is an above the average of her cohort. Supporting this grade was her speech rate and pause proportion, which were better than average, and she makes some use of fillers (for example, when she says "… so yeah I didn't feel lonely"). Counting against this was an articulation rate that was slower than average and slightly worse than average repair fluency indices. Overall, given evidence from appropriately used fillers and positive fluency indices, it could be said that her score in the 3.0 band for fluency is justifiable.

For her grammar Naeko was rated in her second administration at 3.1, which is higher than the average for her cohort, and is supported by her higher than average words per AS-unit and proportion of error indices. However, the 3.0 band for grammar stipulates that the errors that she makes be "only in late-acquired grammar". As such, the rater may consider some well-constructed sentences like "when I was in elementary school my mother didn't work so… so yeah I didn't feel lonely" as being above the 2.0 band which states that she "relies mostly on simple… grammar". At the same time, other utterances have a multitude of ungrammatical elements and could hardly be described as late acquired grammar, like "Age of ma-marriage, um more older… becoming becoming more older". If the rater is sympathetic to a test-taker operating under communicative pressure to get their message out, then her score could be considered justified. Naeko's vocabulary score also received a well above average rating of 3.0. This seems justifiable given her use of such words as *kindergarten*, *lonely* and the phrase *economical crises* (sic.), which must have been seen as signs as "some evidence of some advanced vocabulary", according to the scoring bands.

The second time Yamahiko takes the test his scores are well above average in all scales and vary in a narrow range from 3.2 to 3.4. His CS score of 3.4 was clearly well earned from his almost complete domination of the test as was seen in his qualitative analysis, and supported by a far above average initiating rate. When comparing his performance to the wording in the 3.0 band, it is difficult to understand why he was not given a higher score. At the one point it was seen in the qualitative analysis that another participant ignored his question and effectively hijacked the conversation, but it was not long before Yamahiko interceded and resumed the previously established pattern. This seems like an example of "ability to negotiate meaning quickly and relatively naturally", and if he met this criteria, the rater should have looked at his performance in terms of the 4.0 band. As such, his performance perhaps deserved a higher score in CS.

For vocabulary he scored 3.3, which is 'evidence of some advanced vocabulary', but among the 11 lower frequency words he spoke were an interesting array of formal (*addicted*, *annoying*, *complicated*) and casual language (*tons of*), arguably showing "a wide range of vocabulary knowledge", which is the criteria for a 4.0 band score. It is certainly enough evidence to raise the question that Yamahiko was scored too low in this test, and that a score in the middle to higher regions of the 3.0 band would have been more appropriate. His grammar and fluency scores, at 3.2 and 3.4 respectively, seem to be mostly in accordance with his performance in the indices and qualitative performance. Both of these scores are well above average, as are the relevant indices for the most part.

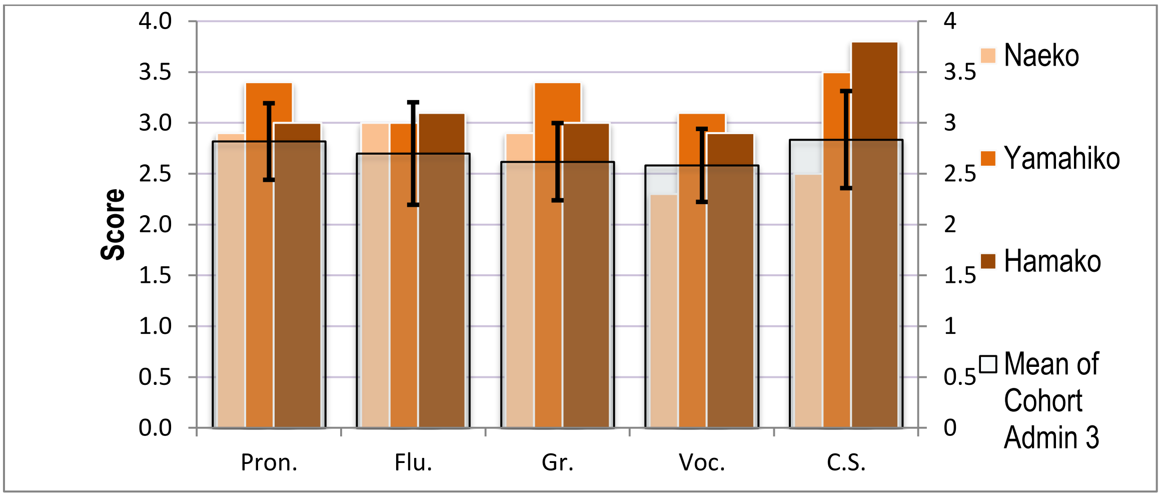
With scores between 3.0 and 3.7, Hamako's, second administration scores are all comfortably above the average of her cohort. For communicative skills, her high score of 3.7 was well justified as she led the group by initiating topics and managing the interaction. Her vocabulary score of 3.4 seems to be in congruence with her performance. The lexis she used included 11 items in word frequency bands higher than the first 1000 word band, sufficient for "some evidence" of advanced vocabulary. With a grammar score of 3.2, her score is beyond the range of a standard deviation from her cohort's average, and this seems in congruence with her complexity and accuracy indices also being well-above average. Fluency also appears to have been scored appropriately, with both indices and score being above the average of her cohort.

5.3.2.4 The high initial group in the third administration

As can be seen in the graph in Figure 76, by the time of the third administration, for the first time one of the members has scored below the average of the cohort in a couple of the bands. Naeko's vocabulary and communicative skills scores are now lower than the average of her cohort. For her CS score, as we saw in the qualitative and quantitative analyses, Naeko plays a minor role in the GOT in this third administration. The qualitative analysis showed that in the face of two group members who dominated the conversational floor, she was passive in the conversation, not initiating any topics and only asking one transfer question, and her Initiating features index reflects this by being well below average. Unless the transfer question counts as initiating interaction in the CS scoring band, her score of 2.5 should be in the 1.0 to 2.0 range. Naeko's lowest score is 2.3 for vocabulary, which is also well below the average of her cohort. Although she only uses two words from the less frequent word lists (*boring* and *former*, as seen in Table 62), even these could have provided "some evidence of some advanced vocabulary". In addition there are phrases such as "the important point for me" and "most important thing" that point towards a score at least substantially higher in the 2.0 band, if not the 3.0 band, as being more appropriate for her.

Naeko's highest score is in fluency which at 3.0, is above average for her cohort. In support of this score, her speech rate and pause proportion indices being substantially superior to the average of her cohort. She might have been held back from an even higher score in this band by having average repair fluency figures, and perhaps influence from low scores in the other scales, which were all in the 2.0 band. Figure 76 shows Naoko's grammar score at 2.9 to be above average for her cohort. For this score, her transcript should show her being reliant on simple grammar, with inaccurate attempts at

Figure 76: Scores of high initial group in the third administration



more complicated grammar, and as it is close to the 3.0 band, the errors should mostly be in late acquired grammar. The following utterance of her longest turn provides an example of her speech that the raters might have paid attention to:

> The important point for me… is whether I can enjoy... um uh my my former part time job it was so boring… and I didn't like it but… now my job is so fun… um I can… I could I could meet… many people in… different ages or… different type of students. So it's very… um good for me.

The use of subordination introduced with 'whether' is impressive, but Naoko makes a mistake with the intransitive use of 'enjoy'. She also erroneously self-corrects the modal 'can' to 'could' to mean a frequent occurrence. The remaining clauses are mostly independent, strung together with co-ordinating conjunctions, which seems consistent with the wording for the 2.0 band. Further support for the score awarded to her comes from the indices, where it can be seen that her proportion of error free clauses is slightly higher than average, while the error free clauses per opportunity to speak falls below, probably a reflection of her lower than average quantity of words spoken. Also, her complexity index of words per AS-unit was lower than average, though as pointed out in Section 5.2.1 it could have been partly due to the shorter turns she was taking, as well as her habitual use of co-ordinating conjunctions, as noted in the example above. When this information it considered, it can be said that her grammar score was justified by her overall performance.

Naoko's pronunciation score is also in the 2.0 band, and at 2.9 it is sitting just above the average of her cohort. According to the bands, she is "mostly understandable" while making "some attempts to blend words". It seems likely that this low score was influenced by most of her other scales now falling into the 2.0 band.

One possible factor in Naoko's falling scores in the final administrations may well be related to who she tested with. In her first two tests, her fellow test-takers seem to be more introverted than two assertive test-takers in her third test, and so it appears she impressed the raters for being relatively more willing to talk than her group mates. This did not apply in the third test, in which Naeko shows either a lack of ability or will to more actively engage when her group mates were more competitive.

The next person in this group, Yamahiko, scores very consistently between 3.0 and 3.5, with his scores being higher than the range of the standard deviation above average in all the bands except fluency. His highest score of 3.5 comes in the CS band, and this seems to reflect his participation in this test, in which the qualitative analysis showed he was under competitive pressure to have his turn, but still managed to put his turns in. His score is consistent with his high incidence of initiating in the graph in Figure 72. For vocabulary he scored 3.1, which means that he showed "some evidence of advanced vocabulary". He does this by using two words outside the first 1000 most frequent words list (*countryside*, *entire*), and perhaps by using such phrases as "I think pretty much the same" and "What do you think about that?" when other students would just say "I agree" or "How about you?" What counts against Yamahiko achieving a higher score in vocabulary in this GOT may well be the reduced number of words that he spoke compared to his previous ones: he only spoke 191 this time as opposed to more than 300 in the previous two. In this restricted number, it seems he did not have enough opportunity to use any more lower frequency words, or perhaps as a result of the communicative pressure he ws under.

Yamahiko's score of 3.4 for grammar, as seen in Figure 76, was his second highest score after CS, and was well above the average of the cohort. This score can be justified by his high complexity as indicated by the words per AS-unit index and above average accuracy in terms of proportion and number of error free clauses. The wording of this band, that he makes errors “only in late-acquired grammar” seems particularly appropriate. Finally, his fluency score of 3.0 is his lowest score, and seems to be too low compared to a speech rate, pause proportion and repair fluency indices that are higher than the range of the standard deviation from the average of the cohort. In addition, the raters seem to ignore Yamahiko’s effective use of fillers, which in the scoring band is included in the highest 4.0 band. One example is when he is talking about the quality of tap water: “It’s got, like, white things in there.” As such a score at least in the middle of the range would be more appropriate for Yamahiko’s fluency score.

In Hamako's third test her scores in the scales span a considerable range from 2.9 in vocabulary to 3.8 in communicative skills, with the latter being vastly superior to the average of her cohort, and the remaining scores being comfortably ahead. For CS, as seen in the qualitative and quantitative analysis, her leading performance amply justifies the 3.8 that she was awarded. Her fluency score was rated at 3.1, which seems low given that her speech rate is beyond the range of the standard deviation for the average. Examining the difference in the wording of the 3.0 and 4.0 bands, to get a higher score she needs to use fillers “effectively” and speak “quickly in short bursts”, whereas in the 3.0 band she “may” use fillers, “rarely” searches for the appropriate word, and speed that “may still not be quick”. In favour of a higher score, Hamako does use fillers when she says “I have to cook and eat (.7) but (2.0) y’know I have to (1.2) I have to take time to cook (.3) and.. (1.0) and.. (ah, ha) (1.0) so (3.1) once a week”. It is difficult to argue that this is not an ineffective use of a filler, and it is not the only example. At the same time, this utterance shows examples of a disfluency when she abandons an utterance, as well as a short, fluent burst in the run of “I have to take time to cook”. Hamako’s speech shows traits described in both sets of wording, and so her score would be better placed in the middle of the band than in its lower reaches.

Her grammar score was rated as 3.0, for which the indices give mixed support. A score in the 3.0 band is supported by a lower than average words per AS-unit index, though this is more a reflection of the higher interaction which is shown by the low number of words per turn she used (see

Figure 63) and an average proportion of error free clauses. On the other hand the number of error free clauses spoken is high, though likely results from the high number of words she spoke in the third administration. When examined qualitatively, she did use complex grammar such as conditionals and subordinating clauses, for example in her utterance: "If… we if if I eat every day and every meal… I think it's really bad for health but… we have to consider… how much we eat". It is easily argued here that this complex grammar is used effectively, albeit with false starts and repetitions. As such a score in the middle of the 3.0 range should have been the lowest score her performance deserved. Similarly, Hamako's vocabulary of 2.9 slips into the 2.0 band, meaning that she was considered by the raters to "not demonstrate any particular knowledge of vocabulary", as it is worded in the scoring bands. This seems contradictory to her use of nine words outside the 1000 most frequent word band of the BNC, which in the limited time available in the GOT is a considerable number, and thus a higher score in the 3.0 range would have been warranted.

For fluency, her score of 3.1 places her above the average of her cohort, though according to her indices her speech rate and pause proportion were beyond the range of the standard deviation of the average of her cohort, and she was above average in all her other indices. Again, this appears to be a score that does not reflect the extent to which she was in advance of her cohort in the indices, and a score in the middle of the 3.0 band would have been more appropriate.

5.3.2.5 Summary of high initial group

A summary of the analysis in the previous three sub-sections is presented in Table 64. It seems that the raters had the most difficulty in consistently awarding scores to the students in the high group, as among the 45 scores awarded, an examination of their performance in the indices and compared to a qualitative analysis shows that up to 17 of them, or 37.7%, may not have been appropriate. One of the problems seen here is the difficulty the raters had judging the ability of Naeko in the first two administrations, and that her score became inflated through being the best in her group, rather on her actual performance. This led to a skewed appearance of her development. According to her indices,

Table 64: Aberrant scores of high initial group over the three administrations

		Pron	Flu	Gr	Voc	C.S.	Tot.
Naeko	Admin 1	3.3	3.6↓	2.9↓?	3.1↓	3.7↓	16.6
	Admin 2	3.1	3.0	3.1↓?	3.1	3.3↓?	15.7
	Admin 3	2.9↑?	3.0	2.9	2.3↑	2.5↓?	13.5
Yamahiko	Admin 1	3.2	3.7	3.2	3.2	3.7	17.0
	Admin 2	3.2	3.4↑	3.2	3.3↑	3.4↑?	16.4
	Admin 3	3.4	3.0↑	3.4	3.1	3.5↑?	16.9
Hamako	Admin 1	3.3	3.4	3.2	3.2	3.6	16.7
	Admin 2	3.3	3.0	3.4	3.1	3.7	16.5
	Admin 3	3.0↑?	3.1↑	3.0↑	2.9↑	3.8	15.7

Key: ↑ - should have been scored higher, according to performance
↓ - should have been scored lower, according to performance
? - indicates possibility

aspects of her fluency and accuracy improved in every administration, but her scores in fluency and grammar fail to reflect this.

In the CS scale, Naeko seems to have produced a weaker performance in the final test than in previous tests, and this could be related to either her own will or lack of ability in the face of more highly skilled group members than she had previously encountered. Yet, Naeko showed she was able to initiate topics in the first two tests without the pressure of competition for the floor, and was rewarded for it. This raises an important issue: how many more test-takers with a low CS score would have been good enough to initiate topics if their groupmates had not been so good? It is a question that reveals unfairness in this format, and a threat to its validity.

Finally, for Naeko's pronunciation, even though this dissertation has not included an objective measure for it, it is possible to comment on the scores she received. The scores in Table 64 reveal a continual decline, even dropping into a lower band in the final administration even though it seems unlikely that her pronunciation could actually get worse. Indeed from the evidence available, it improved in some respects. For example, Naeko's pronunciation score was highest in the first administration, where she can be distinctly heard saying "want to", but in the second came in at 3.1. Yet, improvements can be seen: in the first test she was using "want to", but this has become a more conversational “wanna” in the second, showing improvement in word blending. This shows the subjective nature of the pronunciation scale as interpreted by human judges.

As shown in Table 64, Yamahiko’s scores show the least fluctuations. For pronunciation, grammar, vocabulary and CS his score does not vary by more than 0.3 points over the three administrations, with the variance sometimes being higher sometimes lower. His fluency varies by 0.7, but as the analysis revealed this was scored aberrantly low for his final performance. His quantitative statistics also fluctuate in a narrow band: for grammar, his complexity statistic for words per AS-unit declines slightly in his second test and improves dramatically in the third, while the proportion of error free clauses again drops in the second administration before recovering to be just above his starting figure in the third administration. If based purely on the indices, the development shown in his grammar scores can be seen to be justified. For vocabulary, although his first test registers the most words outside the most frequent word frequency band, most of these are names of countries. As such his best performance is in his second test and as it happens his highest score is consistent with this. However, CS scores do not reflect the performance in the final test when he was in a group of more competitive speakers, and where he perhaps showed a greater level of skill to claim floor time for himself. This was in contrast to the first two GOTs in which he consistently dominated the group by initiating many topics. However, the clearest trend in his scores is a declining fluency score in every administration. The quantitative figures that show a small decline in his speech rate and a small increase in his pause proportion support the decrease in score in the second administration. However his figures show marked improvements in his third administration in these two indices, which show that the final decline in his score was not justified. Overall though, given the wording of the scoring bands, it seems it would be difficult for Yamahiko to escape the third band in any of the scales, despite the development that he shows in the indices.

Hamako, like Yamahiko, shows little variance within each scale: the most her scores vary is 0.4. Although all the margins are small and likely within the standard error of measurement, some general trends can be commented on. First, the scores for her CS are consistent with her performance in that it also shows improvements throughout. However, according to the qualitative analysis, her second and in particular her final GOTs showed agreater levels of performance than her first, and yet the scores only saw an incremental increase of 0.1 in each administration. At the same time her vocabulary scores see continual small declines over the three administrations, contrary to the trend in the number of lower frequency words she uses. In the first administration she only used three words outside the first 1000 word band, compared to eleven in the second and third administrations respectively. Her grammar score shows her peak performance in the second administration and her lowest in the final, which seems quite unrelated to the trends in the indices. Her scores also seem out of sync with her performance indices for fluency, which show her highest scores coming in her final administration, whereas her highest score came in the first. Hamako's score for pronunciation is lowest in the final administration which sees it drop to 3.0, which seems unlikely though. However, as for Yamahiko, while many of her scores are possible to judge as justified as being within the 3.0 band on the basis of a single administration, when seen in a developmental context, the scores fail to match the trends shown by her indices and qualitative performance.

For Yamahiko and Hamakos' even in the first administration, these students had well developed communicative competence, allowing them to contribute more to the discussion and gain high scores. These high scores were boosted, it seems, by seeming especially good compared to the low level of their fellow test-takers. In the following administrations, the average of the cohort gained ground on them and they do not seem quite as good by comparison, even if the indices show that their performance does not necessarily decline. This is seen most clearly in their fluency scores. Both of these test-takers spoke with lower repair and breakdown fluency (as shown by their maze and maze and sound ratios and pause proportion indices) and higher speed fluency (speech rate) in the final than the first administrations, and both of them saw their fluency scores decline.

With communication skills, the Yamahiko and Hamako encounter the problem identified earlier in the discussion of Taeko's score (Section 5.2.3) that the communicative skills scale fail to identify higher level communicative skills. Both of them displayed greater skill in communicative skills in their final two administrations, but this was barely reflected by their scores, as Hamako's registers a slight increase from first to last, and Yamahiko's declines slightly. It is in the final administration where examples of the really collaborative aspects of communication can be found. In such discussions it seems that determination to have your say becomes an important factor, and Yamahiko and Hamako demonstrate that they are sufficiently assertive and capable of achieving this. The issue this raises is whether test-takers such as Yamahiko or Hamako could have performed with the same skill in the first administration in communicative skills, but were prevented from doing so because their groupmates level was insufficient for them to deploy their skills.

5.4 Summary of the qualitative-quantitative synthesis

The qualitative section of the study provides insights into two issues. The first was the extent the various indices represent the performances of the test-takers. It can be seen from this analysis that while the information that the indices provide may be more useful on a group level, at an individual level it is necessary to complement the information with a qualitative analysis, as in a mixed methods study. Among the indices related to words and turns, the most consistent was the words per opportunity to speak. When adjusted for the time of the test and number of participants, this index clearly showed that most students were using more words in each successive test, with the exceptions of the high group's Naeko (high group), whose figures declined in each successive GOT, and Yamahiko, whose figure declined in the final one. The turn related indices vary considerably since they are affected by discourse patterns, as noted above, and as such may be more useful indicators of the kind of interaction that was elicited. Besides, turn length does not necessarily increase with increasing ability in discussion (see Galaczi, 2013): although it seems likely that the test-takers are certainly more capable of speaking in long turns as their ability improves, the discussions in the final administration tended to have more shorter turns than previous administrations. Being related to the length of turn, the complexity figure of words per AS-unit also showed considerable variance at an individual level, and whenever it was referred to, its relationship with turn length has to be acknowledged. The accuracy index of proportion of error free clauses is of some value, but care should be taken lest it reward those who are overly cautious in their usage of language. It needs to be used in conjunction with the measure of error free clauses per opportunity, which gives information about the quantity of correct clauses used by the speaker, but this needs to be discussed in relation to the quantity spoken. Amongst the fluency indices, perhaps the most consistent indicator was the pause proportion, which clearly showed improvement for all of the speakers except Saaya in the third administration. The articulation rate and speech rates both varied according to what the test-takers were trying to say, and so may also be affected by the length of the turn. The maze and maze and sound indices also varied somewhat, perhaps being influenced by some of the prompts in the second administration (Leaper & Riazi, 2014). Finally, in the interactive statistics the best indicator of how the participant took part in the discussion is the Initiating features index. The Responding features index is a useful indication of their passive engagement, and the Developing index indicated the length of turns. The Collaborating features index was less useful, firstly because examples of collaboration were relatively rare in the corpus, and also because it incorporated both active and passive collaboration. Since lower level learners are more likely to receive help than give it, it should be divided into ‘collaborate-initiate’ and ‘collaborate-respond’.

Given the utility of these indices, a qualitative examination was necessary in order to explain what happened in the discussion. This is particularly so in the case of the lowest performing test-takers who fail to produce enough words to gain an accurate picture of their ability. This was also true for the higher scoring students where, for example, the number of words Yamahiko speaks shows a dramatic plunge in his final performance, despite what is arguably a more impressive performance.

The second issue concerned how the scores represented the performances of the test-takers. Although a majority of the scores given were justifiable, the most significant problems occurred because the raters seem to be prone to being influenced by the test-taker’s groupmates. This effect was most obvious with Naeko and Kabuhito in the first and third administrations respectively. In the first administration, Naeko’s scores were amplified by being the least unwilling to participate in the test: she was overly rewarded for being the best in the group. On the other hand, Kabuhito found himself amongst a supportive group of higher level test-takers in the third administration, and his scores were dragged upwards as a result. Given that the theory of IC states that performances are context dependent, it seems contradictory to criticize the raters for being influenced by one of the most salient context related elements in the GOT. Nonetheless, in this case they are being assessed as individuals, and according to a careful comparison of their performance with the criteria presented in the scoring bands, it is clear that training procedures need to be improved to account for this.

While the most significant examples involved a test-taker being graded incorrectly across the scales, there are further instances of lack of independence in the rating. For example, it seems Aemi's vocabulary score in the second administration was rated more highly perhaps because most of her other scales fell into the 3rd band, and Naeko's score in pronunciation likely was rated lower in her third administration due to most of her scores being in the 2nd band. Finally, there are examples of features in the GOT participants' speech that were not being paid due attention by the raters, leading to an incorrect score being awarded. A prime example is Machiko (low initial group) initiating a topic in her first administration that should have garnered a higher score in the CS band, but must have been missed by the raters.

Amongst the three groups, the students who receive the low grades seem to be graded fairly, according to the scales. The students with low initial scores can improve their communicative skills and be rewarded for them, as Machiko was. This is more difficult for other students, such as Saaya and Kabuhito, who are improving, but at a slower rate, and perhaps less evenly. However, it is worrisome for the validity of this format that even low performers like Kabuhito can get a higher score by virtue of being with a group of supportive higher level students, as happened in the third administration. Amongst those who scored in the middle range in the initial administration, we can see that it is possible for them to continue to improve, but that it may depend on their will. Although Aemi made great improvements in the second administration, her body language suggested that she was not making the same effort in the third. The other student in this group showed that she could keep improving, though it is unfortunate for her that the scoring bands are too blunt to recognize her abilities, and she was not rewarded for the skills she showed as she should have been.

However, this study also shows that the students who are treated the most harshly by this scoring system are those who enter the university with already strong conversational skills. Along with the double advantage noted above of being both more capable than their peers and having less competition for floor space, they suffer a double disadvantage when their scores are viewed over a longer perspective. Firstly, they may be rated more generously than they should have been for their domination, as Naeko seemed to be. While this gives them a short term boost, it might also mean that even if they continue to improve they may never score as highly in succeeding tests, which must seem very unfair to them on a personal level. The second issue is that the scoring bands do not adequately describe advanced level performance. It is quite possible to meet the criteria of every scale in the third band, and keep improving as Hamako and Yamahiko did, and yet never escape the third band.

Finally, in Section 3.2.4.1 where the scoring of the cohort in the administrations was investigated, in the final administration it was noted that the scores did not significantly gain from the second administration, despite small but significant gains being made in some of the indices. One contributing factor suggested by the analysis in this chapter is that the narrowing distance in ability between the students means that those who were over-rewarded in previous administrations for being relatively superior to their groupmates no longer stand out in this final administration. Thus, they no longer attracted high grades, and this influenced the overall pressure downwards. Moreover, an impression that may have been a contributing cause to the lack of progress in the scoring bands is that the scoring in general seems to be stricter in the third administration than preceding ones. Supporting this impression, a comparison of the indices for Taeko (medium initial group) in her second administration and Machiko (low initial group) in her third administration shows despite Machiko having superior indices in words per opportunity, error free proportion and per opportunity, speech rate, pause proportion, maze ratio, sound and maze ratio, Initiating and Responding features per opportunity, they had the almost the same total score. Any variance in the strictness of scoring may to some extent be the result of the practice at this institution of creating new training materials for each administration rather than standardising them. Of course, to influence the scoring of the cohort, such effects as noted above should be collectively stronger than the effect of excpetions such as Kabuhito (low initial group) who was over-rewarded in the third administration. Further research would be necessary to clarify this matter.

In summary, this investigation has revealed two issues that need to be addressed. The foremost need is to adjust the upper levels of the scoring bands so that they adequately capture the high level discussion skills that these students are capable of. The other key issue that needs to be addressed in the scoring bands are the conflicts that make it difficult to assign a band. In the first administration, Taeko did initiate interaction, which means that she is higher than the first band of CS, but she also paused for a long time when maintaining interaction, which should rule her out of being in the second band of CS. These contradictory statements in the first and second bands mean that the rater is wrong no matter what grade is given. After such adjustments as these have been made, it is necessary to work on training so that raters can learn to recognize potential situations in which they might feel influenced to score higher or lower than they should. These are issues related to the scoring, and highlighting them here allows administrators to recognize and avoid problems in other tests in which peers interact.

CHAPTER SIX

Summary, Discussion, and Conclusions

6.0 Introduction

This chapter commences with a summary of the study, including quantitative and qualitative phases and a synthesis of both as presented in Chapters 3 to 5. Subsequently, the research questions that were introduced in Chapter 1 will be addressed. The implications for this format of test will follow and it will all be concluded with some limitations and suggestions for future research.

6.1 Summary of the study

An institution that has the purpose of teaching foreign languages has certain curricular needs in order to fulfil its function. To graduate students who are capable communicators in a foreign language, the institution must set a curriculum that provides opportunities for its students to learn communicative skills in the target language with an appropriate assessment system to monitor the students' development while encouraging their language learning. For the institution where the research project reported in this dissertation took place, the group oral test (GOT) was the chosen method of assessment, and the purpose of this dissertation was to evaluate how well it performs these functions. To this end a complementarity mixed methods research design (Riazi & Candlin, 2014) was devised in which a sample of 53 students were captured on video the three times they took the GOT over their first two years of study at this university. For the quantitative part of the study, the transcripts were analysed through counts of words, turns and turns over ten seconds in length, and indices were created to measure the syntactic complexity and accuracy of their grammar, the repair, breakdown and speed of their fluency, their vocabulary and their interactive functions. This enabled the study to answer questions about the consistency of their performance and their scores were used to evaluate the GOT's ability to measure it. To complement the quantitative section of the test, eight students who were classified according to low, medium, and high scores on their initial test were chosen and analysed from the perspective of interactional competence (IC) (Young, 2009, 2011). Their progress was analysed qualitatively and in terms of their indices and then compared to the scores they received. This allowed a micro-perspective that focussed on individual progress that was lacking in the quantitative section. The results of these various elements of this dissertation are summarized in the following sections under each research question.

6.1.1 Summary of results for research question 1

RQ1. How does the language elicited by the GOT show the development of syntactic complexity, accuracy, fluency, and range of lexis in test takers' performance in the three GOTs taken over the first two years of study at university?

The results for this question are crucial for the GOT's ability to display the test-takers' development as a cohort, and so is of particular interest to administrators and other stakeholders. For syntactic complexity two measures were used, clauses per AS-unit, and words per AS-unit. It was found that while the former did not significantly change over the three administrations, which is consistent with the results of Serrano et al., (2012), whose participants were on a one year study abroad program. The measure of words per AS-unit saw significant improvement only in the second administration. The lack of development shown by students in the measure of clauses per AS-units might be explained by it being more sensitive to such differences within an administration as the prompt. Evidence supporting this explanation were the findings from a study that used data from the second administration of this dissertation that showed that different prompts could elicit significantly different clause per AS-unit ratios (Leaper & Riazi, 2014). The other measure, words per AS-unit, seems to be a more stable measure of development, as it showed that students could improve by using longer AS-units in terms of words in the second administration, although the test-takers did not continue to improve significantly in this index in the final administration. This finding should be tempered by the recognition that on an individual level, the words per AS-unit index is influenced by the average length of turn, as was seen in Sections 5.1.1, 5.2.1, and 5.3.1.

The complexity figures in long turns were also calculated, and again the clause per AS-unit ratio was insignificant over the three administrations, and this time the words per AS-units showed significant growth only in the third administration from the first, suggesting that students' improvements in speaking in long turns is more gradual than their ability overall. The long turn figures were consistently higher than the overall figures in both of these measures, showing that speaking in shorter turns uses shorter AS-units that are less complex in terms of subordination, something that may have been overlooked in previous studies (see Nitta & Nakatsuhara, 2014).

The measures for global accuracy were the number of error-free clauses per opportunity to speak, the proportion of error-free clauses in all turns and the proportion in long turns. When comparing the figures for long turns verses overall, the proportion was consistently higher in their overall turns, showing that speaking in long turns was generally less accurate than in long turns. In terms of development, the proportion of error free clauses overall and in long turns saw significant improvements in the third administration from the second, which is again consistent with Serrano et al., (2012), whereas the number of error-free clauses per opportunity significantly improved in successive administrations. The picture of development provided by these two figures is that a test-taker's accuracy improves first in terms of quantity, by speaking more and thus producing a greater number of error-free clauses as they did in the second administration; and then in terms of quality by improving the proportion of error-free clauses, which accordingly drove up the number of error-free clauses again.

The measures for fluency revealed a complex mix of different elements that developed at varying rates in this cohort. Speed fluency, as measured by the articulation rate, showed no significant improvement in the second administration, but improved in the third administration. The measure of breakdown fluency was the proportion of pauses, and this showed improvement in the second administration, but not the third. The speech rate, which incorporates elements of both speed fluency and breakdown fluency, improved significantly in both the second and third administrations. Repair fluency showed different characteristics depending on whether it included voiced fillers or not. The maze ratio is arguably a purer measure of repair fluency since it is the ratio of just such disfluencies as false starts, corrections and rephrasing to the total words spoken. It showed improvement in the third from the second administration, but not in the second from the first. The maze and sound ratio is related since it adds voiced fillers to the same phenomena as in the maze ratio to arrive at a more general index of disfluency, and it showed significant improvements in successive administrations. It seems that these test-takers improved their fluency by first reducing the unfilled and filled pauses, before then improving the speed with which they talked and the number of maze words.

Since the timed data for the fluency indices were only collected for longer turns, the only indices that allowed comparison with their overall figures were for repair fluency (speed and breakdown fluency indices were only calculated for longer turns of over ten seconds). The figures for these indices were consistently higher for longer turns than for all turns, suggesting that test-takers lose fluency when constructing longer turns as they focus on giving the message, and conversely, when taking shorter turns they may appear to be more fluent.

Finally, the measurement of lexis using lexical diversity indices proved problematic due to many of the test-takers not succeeding in reaching the minimum threshold for valid measurement. Amongst those who did, no significant differences could be detected across the three administrations. One explanation is that the language produced in discussion, being a mixture of multiple short and longer turns with various communicative functions, is simply not amenable to these kinds of measures, which are more applicable to single texts such as oral speeches or presentations or writing. To get an overall picture of their lexical development, the test-takers' lexis was analysed according to their frequency and by means of *Vocabprofile* (Cobb, 2002). The frequency analysis revealed that the raw number of tokens grew remarkably from 4215 in the first administration, to 7456 in the second, but declined slightly to 7439 in the final administration. The number of types showed the same pattern of a large increase in the second administration followed by small decline in the third. Although this made it seem like the test-takers' ability to use vocabulary did not improve in their final year, the percentage of types by frequency band reveal an ever decreasing dependence on lexis in the first 1000 word band. The pattern shows a continual decrease from 83.46% (first administration) to 80.72% (second administration) to 77.39% (third administration), and this corresponds with an increasing proportion of words in the second 1000 word band from 9.85% to 10.44% to 13.59%, showing a greater use of less frequent words to express themselves. Although these figures might seem like incremental gains, if they were normalized according to the participants' opportunity to speak (as the findings from the first research question show), they would seem more substantial. The *Vocabprofile* analysis therefore shows that students improved in each successive administration by being less dependent on lexical items from the most frequent band, increasingly using words in the second most frequent band, and in more frequent use of lexis from the academic word list (AWL) which were not from the prompt.

When comparing the indices for longer turns of over ten seconds with all turns taken, it appears that the candidates consistently used more complex language that was less accurate and fluent, which may have significantly impacted on other studies that did not distinguish between the language of shorter and longer turns. For example, Nitta and Nakatsuhara (2014) found turns to be significantly longer among students who had planning time, but did not separate data from shorter and longer turns. They found that the unplanned condition was significantly more fluent according to the number of words per second[11], but according to the findings of this dissertation this is not surprising because short turns can be spoken more fluently than long turns.

Overall, the development shown by the indices for complexity, accuracy, fluency and lexical diversity is one of large gains in the second administration, with incremental and more subtle improvements being made in the third administration. How representative the findings from this dissertation are in general is an interesting question, since examples of longitudinal studies are quite rare and usually are case studies with few participants, or did not use comparable indices to those used in this dissertation. This investigation into the indices allowed the test takers' performances to be quantified without taking into consideration how they were interacting in the GOTs. An investigation

[11] Although surprisingly, Nitta and Nakatsuhara (2014) found pause time to be higher in the unplanned condition. This may be because they included the time taken between turns in addition to the pauses within the speech.

into the words and turns that they produced in their discussions sheds more light on this in quantitative terms, which is the subject for the next research question.

6.1.2 Summary of results for research question 2

RQ2. To what extent do the test-takers' number of words, frequency and length of turns show a consistent pattern of development in their performances in the three GOTs they take over two years?

This question focuses on the pattern of discourse that is prevalent in the GOT as seen through its word and turn statistics. For the number of words, as was seen in the answer to the first research question, the raw counts showed a spectacular rise in the second administration, but the slightest of declines in the third, as was the case for the number of words in long turns. When the words spoken were normalized for number of participants and time of test, in fact a significant increase was registered in the third administration from the second as well. It seems that this was the result of the increase in speed fluency that was noted in the results for research question one. The raters are allowed to call for the test to end when they have a rateable sample, and they did this earlier than in the previous administrations mostly likely because the test-takers could speak more quickly.

The number of turns taken followed the same pattern as the unadjusted number of words by having a large increase in the second administration, and then the slightest of declines for the final one. When adjusted for the time of test and number of participants, it showed a much more gradual increase, so that the only significant gain was between the third administration and the first. The raw number of long turns also had a significant increase in the second administration, and then a significant decline in the third from the second administration, with no difference between the first and third administrations. When normalized, this index saw no significant increases or decreases.

The final set of statistics were for the words per turn, words per long turn, long turn words per words and long turns per turns. Of these ratios, only words per turn and words per long turn showed significant differences, and these occurred in a gain between the first and third administration, with the word per turn showing a small effect size, and the words per long turn showing a medium effect size.

For both of these figures there is no significant difference between the second and the third, but the gain is maintained in the third so that they are also significantly more than the first administration.

Collectively, the statistics devised to answer this research question show that in the second administration test-takers spoke a significantly greater number of words overall in longer turns, as well as a non-significant rise in the number of turns. In the third administration, a further increase in words spoken is made, and a further non-significant increase in the number of turns used, which is sufficient to be significantly more than the number spoken in the first administration. However, their long turns do not continue to get longer, suggesting that test-takers' path to improvement does not lie in the direction of speaking at greater length. In support of this, throughout the three tests, the proportions of words in long turns to all words and the long turns to all turns does not change significantly, showing that in the second administration the improvement was in general toward improving everything, and their ability to take long turns along with it. In the second period up to the third administration, the number of long turns decreases significantly, and the number of turns they take is significantly greater than their first administration, showing a gain in their ability to take short turns. These findings are consistent with Galaczi (2010, 2013) who suggests that as test-takers' abilities advance towards a tendency to have a greater number of short turns in discussion.

The first two research questions were focussed on the GOT as an assessment that elicits a performance from the test-takers. The next question investigates the role of the GOT as a tool within a curriculum created to encourage language learning by investigating its ability to elicit relevant features of conversation.

6.1.3 Summary of results for research question 3

RQ3. To what extent is the GOT able to elicit features of conversation that have been found to be beneficial to language learning?

This question required an analysis of the interactive function of the transcripts of the test-takers over the three administrations. This analysis resulted in four categories of interactive functions: Initiating, Responding, Developing and Collaborating. Initiating functions are the kind of utterances that test-takers use to commence an interaction or seek further information or to add a new point of view (when not in response to a question). The use of initiating functions can be seen as an index of the willingness of the test-taker to be involved in the conversation, and research has found that those who involve themselves are more likely to benefit in terms of learning (McDonough, 2004). Responding functions are passive in that they can only be used after another person has initiated, and are considered the minimum level of participation required in the GOT. Developing functions serve to extend a turn, and may be indirectly related to language learning by being indicative of the speaker having attentional resources (Schneider & Shiffrin, 1977) available to organize and refer to other speaker's utterances. Finally, the Collaborating features include moves such as negotiation of meaning and co-construction, which studies have found to be related to language learning (Foster & Ohta, 2005; Mackey, 2012).

The counts of these features revealed that, consistent with He and Dai (2006) and Van Moere (2007), Collaborating features were quite rarely used. Moreover, their use was relatively constant and did not change significantly over the three administrations. It seems quite likely that this may be at least partly related to developmental issues. In the first administration most of the test-takers may be largely focussed on delivering and comprehending the message, like the low level learners in Galaczi (2010, 2013). In later administrations it may be that the test-takers do not see an opportunity to use them, or, as the qualitative analysis suggests, it may depend on having equally high level learners in the group in order to effectively deploy them.

Nonetheless, the remaining functions (Initiating, Responding and Developming) can also be related to language learning, and the analysis found evidence of a staged pattern of development. In the first period until the second administration, the test-takers as a whole improved their Responding and Developing features significantly. In the second period (from the second to the third administration) these showed no significant difference while the Initiating features gained significantly. This shows that the test-takers improve by developing their ability to answer questions and develop their responses before they improve their ability to engage more actively in the discussion. It seems then, that the GOT can elicit a wide variety of functions, although not all test-takers are able to deploy them in the test. It may be partly a matter of development, with many students being at too low a level to find opportunities to produce initiating or collaborating functions until the final administration.

The questions answered thus far have shown how the GOT elicits test-takers' skills through the CAF measures, how their interaction develops through indices related to words and turns, and what interactive functions they use. The next question aimed to improve our understanding of the extent to which these relatively objective indices are measured by the raters.

6.1.4 Summary of results for research question 4

RQ4. How well do the speaking performance indices of test-takers who take the GOT over the three administrations in two years predict the scores awarded to them?

In the first part of the answer to this question the test takers' scores were analysed through repeated measures ANOVA (RM ANOVA), which found that the scores increased significantly in the second administration, but for the most part there was no significant difference between the third administration and the second, except for vocabulary which saw a significant decrease. This suggests that the impressive improvements seen in the second administration were faithfully recorded by their scores, but the more subtle gains in the third administration were not.

The second stage of the answer to this question was to use single regressions, standard and sequential multiple regressions (MR) to determine the extent the indices used in this study to analyse GOT interactions could predict the scores assigned by the raters. For the scales of the scoring bands for which there were relevant indices (fluency, grammar, vocabulary and communicative skills) the procedure was to include both specific indices relevant to the scale, where possible, and three general ones that might be thought to have an overall effect (words and turns per opportunity to speak, words per turn). First a single regression was performed for each index to determine their order of importance, and then a standard multiple regression and finally a sequential regression was run. It was decided that in the multiple regressions the specific indices would be input before the general ones, since it is the role of the specific indices that is of central concern to this dissertation.

For fluency, the single regressions showed that the most important specific indices over the three administrations in declining order were the speech rate, maze and sound ratio and articulation rate, but the standard indices of words and turns per opportunity played an important role. In the standard regression, only the number of words predicted the assigned score in this band. The sequential MR showed that the speech rate and the number of turns could be relevant factors if the number of words spoken were not part of the model. It seems for the fluency band, the raters were more influenced by the number of words spoken than the speed at which they spoke. The likely explanation for this is that the more fluent speakers generally speak the most, as shown by the consistently high correlation between the words spoken and the speech rate (as shown in Table 76 of Appendix F). It might be inevitable that for raters one of the most obvious factors relevant to fluency would be the quantity spoken by the test-takers.

The analysis of grammar in the single regressions showed that over the three administrations the most important predictors were the indices for error-free clauses per opportunity to speak and as a proportion, with the number of words and turns also of importance. The standard MR showed that the number of error-free clauses was a significant predictor in the first administration, and the number of words per clause was significant in the first and second administration, but there were no significant predictors in the final GOT. The sequential MR confirmed that the number of error-free clauses was the most important predictor, followed by the number of words per clause in the first and second administrations, but not the third. The error-free proportion was important in some models in the second and third administrations, but not the first. Overall, it seems that when raters adjudicate the grammar scale, the error count seems to be more of a focus than the test-takers' ability to produce more complex grammar, as the words per AS-unit fail to reach significance. This is despite the wording in the scoring bands including both what the test-takers do with their grammar as well as their mistakes. Complexity is taken into account to some degree by the raters, as is shown by the occasional significance of the words per clause measure. However, this only includes subordination indirectly. For accuracy, in general, the raters seem to have found the number of error-free clauses more pertinent than the proportion of them, as this proves to be a better predictor. The importance of this index probably accounts for the lack of significance of the words spoken index in these models, as they are both measures of quantity.

The regression analysis of the scores for vocabulary was hampered by the lack of specific indices, as explained in Section 3.2.1.5, and so the analysis was run with just the general indices of normalized words spoken, normalized turns taken, and words per turn. This could mean a contradiction between the indices, which are a pure measure of the quantity of words, and the scales, which focus on what the test-takers do with the words to make meaning. The standard MR showed that in the first administration none of the vocabuluary indices were significant predictors, and only the words spoken per opportunity were significant in the second and third administrations. This may have been due to the few words that were spoken in the first administration: the problems they had communicating received relatively more attention from the raters. By contrast, in the second and third administrations the figures suggest that speaking more words became relatively more important. The figures also confirmed that turn taking factors were not considered relevant to the raters when assigning a vocabulary score, as these failed to be significant in any administration of the standard or sequential MR.

The final scale to be addressed is that of the communicative skills (CS) band. According to the results of the single regressions, the index that had the most weight in determining the assigned score over the three administrations were the Initiating features, followed by the Developing, Responding and Collaborating features. One interesting discrepancy in these results was that the Initiating features had the strongest showing in the first and third administrations, but in the second it was the Developing index that was the most important predictor of the CS score. The second administration was the subject of the paper by Leaper and Riazi (2014) in which it was found that two of the four prompts elicited significantly longer words per turns than the other two. The high showing of Developing in the second administration may mean that those who took long turns may have been rewarded for it in the communicative skills band. Over the three administrations though, Developing was the second strongest after Initiating, suggesting that there is a tendency for raters to use this scale to reward those who spoke in longer turns. This is consistent with studies conducted at the same institution (Kobayashi & Van Moere, 2003), and in a different context (Nitta & Nakatsuhara, 2014). Nitta and Nakatushara (2014) found that the planned condition participants had significantly longer turns and were rewarded with higher scores despite no significant difference being found in their indices (see Section 2.2.4.4).

In the standard MR only the Initiating features in the test-takers' first administration were significantly related to the CS score, and among the general indices, only the number of words spoken in the first and last administrations were significant, leaving no significant indices in the second administration. In the sequential MR, Initiating features were significant in all models of the first administration, and Developing features in all but the last model in which words per opportunity to speak was included, supporting the inference that Developing features are related to the raters rewarding the quantity of speaking in the CS band. The second administration sees a quite different pattern in which initiating is only significant when it is the only index in the model, and then the developing function is significant in all models it is included in until the final one, which adds words spoken and in which nothing is significant. In the third administration, Initiating features become relevant again, being significant in all models except the final one, in which words per opportunity is included, and is the only significant index.

The answer to this research question is that most indices do not appear to play a statistically important role in the raters' awarding of scores. What little role they do play seems to diminish with each succeeding administration, perhaps showing that raters are confronted with more complex, multifaceted performances as students' skills develop. In fluency, it is perhaps not surprising that the quantity that the test-takers speak appears to be the most important determiner of the score since fluent speakers may be more likely to do more speaking. The sequential MR showed that speech rate also played a role, and when it is considered that this combines speed and repair fluency elements (De Jong, 2013) perhaps it is not surprising that it is the most important of the fluency indices. The number of words spoken was also the most important element of their CS score, which was also represented by the occasional significance of the Developing features, especially in the second administration. In the CS scale, words spoken were the least important in the second administration, but the sequential MR showed that the quantity of words was included, albeit indirectly in the form of Developing, which was significant in four of the five models it was included in. However, Initiating features are more relevant to the descriptions of the bands in the CS scale (see Table 57), and were significant in most models in the first and third administrations.

It seems then that the indices from research questions 1 to 3 (Sections 3.2.1 to 3.2.3) only represent the awarded scores to a limited degree. For grammar, fluency and CS the raters were paying attention to at least some relevant features that are represented by the indices included in this study, even if the general index of words spoken was perhaps the weightiest factor. The question remains open as to what the other elements were in the test takers' performances that the raters were paying attention to, and in the literature Orr (2002) and May (2011) show that raters may draw on elements not within the bands in order to describe the performances they see, and may even compare the performance to other participants, contrary to their training (Orr, 2002). Ducasse and Brown (2009) show the importance of non-verbal communication such as body language may have on raters' impressions of a communicative interaction. Insight into such matters may be provided by a qualitative examination of the test takers' performances in relation to their scores, as the next research question purports to do.

6.1.5 Summary of results for research question 5

RQ5. How well do the test-takers speaking performance indices and scores awarded to test-takers of the GOT over the three administrations of the test taken in two years represent their performance compared to a qualitative analysis of their performance?

In order to answer this question a sub-sample of eight test-takers were chosen. These participants were divided into three groups: three test-takers with low initial scores, two with medium initial scores, and three with high initial scores. The first stage in answering this question was to use an IC framework to qualitatively analyse their performance (Chapter 4). Secondly, their progress according to the indices from the quantitative phase was described; and then in the final stage these elements were synthesized in an analysis of how well they met the criteria given in the scoring bands (Chapter 5). This summary will focus on describing the findings in terms of problems that were identified in the scoring.

The first scoring problem that was revealed was one that occurs when a rater misses an element of the test-taker's behaviour, and results in the test-taker being awarded a lower score than was justified by the performance. For example, in the first administration, the test-taker Machiko (low initial group) initiated a topic by reading it out directly from the prompt, but was still awarded a score in the 1.0 band in contradiction to the band's dictum that the test-taker in this band "does not initiate interaction".

Another kind of problem occurs when the rater seems to miss, ignore or misinterpret some element in the scoring bands, and rewards the test-taker more highly than deserved according to the performance. Although this dissertation does not contain an independent index for pronunciation, it is apparent that Kabuhito's pronunciation in his first administration is so heavily influenced by his native pronunciation that it should fall in the 0-1.0 band. The description is "uses Japanese katakana-like phonology and rhythm; words are not blended together", and there is strong evidence of this in the recording of his performance.

In another example from the low initial group, Saaya's score in the Communicative Skills (CS) band of 2.5 in the second administration seems to be a clear example of the scoring bands being misinterpreted. The scoring bands for the 1.0 band state that such a test-taker "does not initiate interaction" and "produces monologue only". In the qualitative analysis it was obvious that since the other group members had given their response it was Saaya's turn, and they waited patiently through a long pause for her to take it, as such it was not initiating. It seems that rater training needs a clearer definition of what counts as initiation. Moreover, in this test Saaya contributes a single long turn which can only be described as a monologue, which is decribed. Taeko in the medium group also failed to initiate a topic in her first administration, though did use a question to transfer a turn, which may have been enough to move her beyond the 1.0 band. This reveals a potential problem in the CS scales: In the 2.0 band appears the description that the test-takers' responses be "without long pauses to maintain interaction" and there is no mention of initiating, leaving the rater in a quandary if a test-taker's performance includes both long pauses and initiation, as the rater may have felt to be the case here.

Sometimes the score given disagrees with the data shown by the indices, or seems inconsistent with past performances, something that seems particularly prone to happen with the fluency indices. Aemi (medium initial group) in the second administration was awarded a fluency score of 3.3, well above the average for her cohort, yet her indices placed her only above average. In the next administration the opposite problem occurs as her fluency score plummets an entire point to be well below average, and yet her fluency indices are slightly above average. This may well be a case of her fluency scores being influenced by the scores in her other scales: in the second administration, Aemi's fluency was one of four scales with scores in the 3.0 band, while in the third, every scale was in the 2.0 band. This suggests that raters may be influenced by the halo cast by scores awarded in other scales rather than the evidence in the performance itself.

There are also examples of the influence of the surrounding test-takers seeming to exert an influence on the scores. This seems to have occurred with Naeko (high initial group) particularly in the first, but also in the second administration. The qualitative analysis showed that in the first two she found herself in groups with low competition, in groups with similar characteristics to the 'Parallel Interaction' that Galaczi (2008) noted in some paired tests. Since she was the least unwilling to initiate topics and nominate other speakers to talk she found herself to be the leader by default, which was recognized by the raters giving her scores which were well above the average. This was in contrast to her indices that only placed her at above average in the first administration. This effect seems to work in a similar way to that found in the experimental studies of Berry (2004) and Ockey (2009). Both of these researchers found that extravert test-takers could have their scores elevated by being in groups with a low level of extraversion. In a similar way, Naoko stood out to the raters as speaking when her group members did not, and was rewarded for it.

The data for Kabuhito shows the same effect working in the opposite direction. Berry (2004) found that introverts could have their scores boosted when in a group of extraverts, and a similar effect can be observed in Kabuhito's (low initial group) third administration. In this test he is made to seem a higher level than he deserves by the support of his group members, who are at higher level than him. These higher level groupmates asked questions to draw him into the conversation, supported him by suggesting words to help him express himself and did so on two occasions. Although his improving indices support a higher score in the final administration than his previous GOTs, achieving scores in the 2.0 band in all scales except CS was higher than his performance warranted.

As well as identifying problems in the interpretation and application of the scoring bands as they stand, this dissertation identifies an inadequate description of what constitutes higher level ability in the CS scale. This was particularly evident in the performance of Taeko (medium initial group) whose performance in the third administration included such features as summarizing, referring back to what was said earlier, rephrasing to overcome a misunderstanding and completing others' sentences. None of these are included in the scoring bands at all, but they are all features of advanced talk among these GOT test-takers. This lack of adequate description of the highest levels also was damaging for Yamahiko and Hamako (high initial group) in the final administration when they showed they could take a leadership role among other advanced learners by competing successfully for the floor, which is not mentioned in the bands.

6.2 Conclusion

The impetus that gave rise to the study reported in this dissertation was the question regarding the effectiveness with which the GOT can play its role in a curriculum intended to prepare students for communication in a target language. The institution, where data for the current study was collected, requires a method of assessment that allows the progress of the students to be monitored and to provide washback that encourages communicative practice in the classroom. This can be broken down into two issues. The first is whether students taking the GOT show development in their language and communicative skills over the period of the study. If it can be shown that students display development in their skills in this test, then it suggests that the test has the potential to generate positive washback. The results of the quantitative phase of this study show that this question can be answered in the affirmative. The students in this study showed a substantial gain in most facets related to speaking in their first period of study, and although the gains made in the second period were more subtle they could be detected too. This study thus demonstrates that at a group level, students can show progress in their speaking skills on the GOT.

The second issue is whether the rating system employed for this particular GOT can be used to faithfully record students' progress in their speaking skills, and here the answer is qualified. The scores successfully tracked the significant progress made in the second administration of the test, but recorded insignificant losses in the third administration in all scales except for vocabulary, where the decline was significant. This suggests that something is not functioning as it should be in this format of the GOT. One of the reasons inferred by the multiple regression analysis was that the indices contribute little to the scores the test-takers received. As the test-takers' speaking skills, as recorded by the indices, develop, the raters seem to pay more attention to other factors than those represented by the indices. This point may not be as serious an issue if the rating bands are not closely related to the indices. However, the wording in two of the scoring bands align closely the indices used in this study. The fluency scale descriptors mention aspects of breakdown, speed and repair fluency, which are all represented in the indices. The grammar scale describes the construct in terms of accuaracy and complexity, which are both represented in the indieces. On the other hand, the vocabulary indices did not function due to the nature of the source data, and the communicative skills descriptors were somewhat more general than the feature counts that the indices were based on. Even after considering this caveat, this study points to the need for a greater understanding of the elements that the broad categories represented in the scoring bands are based on. It can be suggested that basing the scoring bands more closely on their foundational elements will allow a greater accountability of the raters and improve the alignment of the scores to the performance.

There are some possible explanations for the GOTs inability to record any improvement in the third administration from the second that may operate at a global level. Firstly, there was an impression that the scores awarded in the final administration were slightly but noticeably less generous than those assigned in the preceding two administrations, and some support could be found by comparing performances in the second and third administrations (see Section 5.4) but further research is necessary for confirmation. If the third administration was rated more strictly, it may be related the training sessions that use a different set of materials for each administration, and may have oriented the raters to be more severe than usual.

Another factor which cannot be quantified is that some students may have been less motivated to do well in their third administration than their preceding ones. This is because the results of the first and second administrations are used not only as a portion of their grades, but also to place the students into their next classes, whereas the final administration marks the end of their core courses and so is not used for placement purposes. It can be conjectured that students no longer had the motivation to do as well as possible in the third GOT so that they could stay in the same class or move to a higher class the following year. Among the eight test-takers in the qualitative, there was at least one who displayed less engaging body language and performed at a lower level in her final administration and so may have been affected in this way (Aemi, medium initial group). If this feeling was more pervasive amongst other test-takers it could have the effect of depressing the overall scores in the final administration. Further research is however necessary to ascertain the extent of these anecdotal observations.

Moving from these global considerations, the qualitative phase of this dissertation illuminated this issue on a more personal scale by choosing eight participants from the original 53 student sample. When the data were examined qualitatively, it seems that those who start with low or medium scores in the GOT are more likely to have their development tracked by the scores awarded in subsequent

GOTs, albeit with the odd score out of place. A major exception occurred when one student (Kabuhito, low initial group) had scores in each scale that failed to represent his speaking ability. This was likely to be due to supportive higher level groupmates who helped him by asking questions to draw him into the conversation and collaboratively built meaning with him. In terms of Interactional Competence, it can be said that his boosted scores represented what he could do under favourable circumstances, but the scores awarded did not represent his individual speaking ability in terms of the scoring bands. This situation is reminiscent of Berry (2004) who found that introverts scores could be boosted if they were in groups that were otherwise composed of extraverts. The raters in Berry (2004) may have been influenced by the more extravert groupmates to award the sole introvert a higher score. In a similar way, Kabuhito in the third administration seemed more able when surrounded by higher ability groupmates.

Nonetheless, aside from this major exception, the scores awarded in the GOT track the development of low and middle ranking students reasonably well. It could be expected that the majority of the students who entered the institution would fall into these ability ranges, and would be able to see their scores rise alongside their abilities. Unfortunately this would not be the case for those who entered the university with speaking skills that are already more advanced than their peers. These students suffer a double disadvantage in the GOT as it is run at this institution. In the first administration their ability stands out relative to their groupmates who had little opportunity to build up their “architecture of the practice” (Young, 2000, p. 6) of communicative acts, and the raters over-compensate these skilled students’ scores for their greater relative speaking ability. It is telling that the highest score for all three members of the high initial group came in the first administration. This effect of the raters being overgenerous to an individual relative to his or her groupmates is a parallel situation to that described by Ockey (2009) and Berry (2004) who found that a single extravert may receive a higher score if they are in a group of introverts.

The second part of this double disadvantage makes itself apparent in subsequent administrations, when their cohort has improved their English speaking skills. Not only do those who come in with well-developed speaking skills lose the relative advantage of being considerably better than their groupmates, but they also have the disadvantage of a failure of the scoring bands to adequately describe advanced interactional skills. Of the three members of the high initial group, Yamahiko and Hamako are both trapped in this situation. They dominate their first two administrations with their advanced speaking ability, though it is noticeable that their groupmates are considerably better in the second administration than the first. It is only in the third administration that they experience competition for the floor, and they both show that even in this situation they can lead the conversation. Yet the skills they display to do this are not described in the scoring bands, and so they find themselves forever trapped in the 3.0 band of the scales, which have effectively created a ceiling effect. As such, the scoring bands need to be subjected to further research to explore this issue in more depth.

These high level students illustrate another problem for the GOT: the level of their interlocutor. For Yamahiko and Hamako, the problem of the GOT is that their full range of abilities is only displayed when they are among similar level groupmates. This raises the troublesome issue of the ability of this format to test the upper limits of the test-takers' ability: how much earlier could they have shown these skills if they had had similar level groupmates? Van Moere (2007, p. 6) pointed out that in an interview led test, the interviewer uses 'probes' and 'level checks' (ETS, 1982, p. 64-5) in order to ascertain the candidate's level, but this deliberate exploration of the upper reaches of a test-takers' ability cannot occur in a peer-interaction speaking test like the GOT. It is both a strength and a weakness; it is the absence of such features that make the GOT potentially more like a conversation, and at the same time a weakness of its use as a tool of assessment.

For the other member of high initial group, Naeko, the effect of her groupmates is slightly different, but perhaps even more damaging. Like the other two members of this group, she is over-rewarded for being the leader of her group in the first two administrations. Yet the qualitative examination of her performance in Section 4.2.4, reveals that in the first two administrations she was with particularly unassertive groupmates with no competition for the floor. In this context, Naeko seemed to become the leader because she was the least unwilling to take this role, and the raters boosted her score accordingly. However, in the final administration two of her four groupmates compete heavily for the floor, and it seems that Naeko does not have the ability or will to assert herself in the same way that Hamako and Yamahiko could. As such, some of the scores become somewhat depressed, and when compared to her previous performances, and her overall score declines. From this it seems that making a distinction between leading a discussion in a non-competitive group and a highly competitive group would be desirable, though it raises further issues about the degree to which a GOT test-takers scores should depend on who they are to be tested with.

For the administrator then, the final answer is that the use of the GOT has some positive points in terms of how students can display their speaking ability in it. At the same time the findings of points to the need for improving the rating and the rating scales so that these positive points may come to the fore. In particular, the rating scales need to be improved so that they can represent the abilities of those with more advanced communicative skills, and raters needs to be sensitized to the situations in which they may be inclined to overcompensate the test-taker who stands out for either being comparatively better, or made to look better by the support of his or her groupmates. The minimum proposed solution is to reform the rating bands as suggested above, and improve training not only by standardising interpretations as much as possible, but also to recognize that the test-takers' interactional competence is dependent on the context of the discussion, and that this may inevitably impose limits on the skills they are able to display.

6.3 Theoretical and practical implications

The results of the study as presented in this dissertation point to a number of methodological issues as well as the theoretical and practical implications of using the GOT as a format for testing speaking ability.

6.3.1 Methodological implications

The theoretical foundation of this dissertation is the notion of Interactional Competence (IC), upon which the qualitative analysis was conducted using Young's (2009, 2011) framework. At the same time, the basic assumption under which the assessment program in the curriculum operates is that scores should be awarded on an individual basis. This study was designed to straddle the line between these disparate perspectives, and thus has used an MMR design to collect data and provide inferences that acknowledge both sides. Where the quantitative phase of the study relied on indices that were produced from the transcripts of the individual interactions in GOTs, the qualitative phase examined the performance of the test takers as embedded in the mutually constructed discussions that took place in GOTs. In the synthesis phase these two elements were brought together, and it was here that conflicts between the test takers' collaborative performance in the particular GOT and as represented by a score awarded by the raters were clearly delineated. This occurred because it seems that the rater is also prone to viewing the individuals in the GOT as part of a collaborative co-constructed performance that is consistent with IC.

Accordingly, MMR design has proved to provide applied linguist researchers with unique opportunities to tackle more complex problems and provide more comprehensive inferences about the phenomenon under study (Riazi & Candlin, 2014) as was the case in the current study. The use of MMR design made it possible to investigate issues related to GOT not only as a means of language learning, by focusing more collaborative communication, but also as an assessment tool, attempting to assess and record individual performance within an IC framework, as discussed in the next section.

6.3.2 Implications for using Interactional Competence as the theoretical background

The most substantial differences between the test-takers' performances in terms of the scoring bands and the scores they were awarded coincided with the GOTs when there was the most considerable difference in ability between the test-takers. Amongst the low initial group this could be seen with Kabuhito in final administration, where his ability to express himself is at its highest with the support of the more advanced test-takers, and this performance was rewarded more highly than it should have been according to the criteria in the scoring bands. In terms of IC, it could be argued that his success in communicating his thoughts with support from his group members justified the higher score he received.

By contrast, Saaya (low initial group) in the final administration attempted to put forward a topic, but it was not taken up by her apparently unsympathetic groupmates. If it had been taken up as a topic by the group, it cannot be said that she would have garnered a higher score than she did, but the fact that she put forward a topic was recognized by her communicative skills score breaking into the 2.0 band (while Kabuhito who never initiated a topic earned a CS score of 1.8). In this we can see an accommodation between the scores awarded to the individual for actions taken, and scores earned through the act of collaboration within the context of a group discussion. An implication is that the meaning of their individual scores could be made clearer by contextualizing notes from the scorer giving information about the specific nature of the discussion. In this case, the note would state that the candidate performed as such in a group of higher level groupmates who supported his attempts to communicate. Such a note would allow outside users a clearer idea how to interpret the individually awarded scores given, and pay due respects to the nature of communication according to IC. The work of Galaczi (2008) in characterizing paired format interactions would be foundational in this regard, and some preliminary work on the GOT can be seen in O'Sullivan and Nakatsuhara (2011). Further research building on these studies would enable raters to identify the interaction of the GOT in such a contextualized note.

The other situation that was captured in the IC analysis was where a high ability student achieved a high grade because she (Naeko, initial high group) was the only one in the group seemingly willing to take a leadership role. Under the IC perspective, again it can be argued that her scores were deserved because she did take on the mantle of leader, and it was this role within her group that raters responded to by awarding her higher scores than she might otherwise have had. Here a contextualizing note might state that she is capable of leading a group of lower level learners in a low competition discussion.

Introducing contextualized notes would certainly help to explain scores, but of course it would also result in an extra layer of complexity. For the learners, it would increase the stakes regarding who they are grouped with. At present placement into groups is randomized within certain conditions (see

Section 3.1.2.3) but if similar ability students could be grouped together, perhaps on the basis of self-assessment surveys or other background records, then it may lessen the interlocutor effect and result in less variance within groups. For raters, having to write contextualizing notes is an extra burden that may not be popular among staff for whom rating is an obligation (see Section 3.1.2.2). The load on the rater might be eased if GOTs can be characterized in categorisable ways (Galaczi, 2008; O'Sullivan & Nakatsuhara, 2011) that can be selected from drop down menus in a computerized grading system. Such a system would have advantages in terms of the implications related to rater training as is explained in the next section.

6.3.3 Implications for the use of the GOT to elicit collaborative language

Consistent with He and Dai (2006) and Van Moere (2007), the interactive analysis in this dissertation found that the GOT does not succeed in eliciting a high frequency of Collaborative features that are most closely related to language learning, and this finding has implications that need to be considered. Perhaps the most obvious one is that by changing the task of the GOT to a goal oriented one would result in more collaboration, as Van Moere's (2007) research suggests. However, this outcome needs to be put into the context of the development of the test-takers ability in interaction in general. This dissertation showed that students significantly improve in their use Responding and Developing features before they go on to improve their Initiating of topics by the end of the period of study. The possibility needs to be entertained that with further improvements they would significantly improve in their Collaborating features, as is inferable from Galaczi's (2010, 2013) research into paired interactions at different levels. The qualitative analysis of the performance of the high level group in Section 4.2.4 is encouraging in this regard, as it suggests that more collaboration will result when higher level test-takers interact with each other. Moreover, as pointed out in Leaper and Riazi (2014, p. 180), Van Moere (2007) only used a single prompt that may have influenced a particular pattern of longer turns and less collaboration. As such, the implication is that further research is necessary to refine expectations of how much collaboration can be expected in GOTs depending on the level of learner and how this interacts with other such facets as the influence of the prompt.

6.3.4 Implications for training raters for the GOT

This dissertation has found examples of the threats to scoring validity analogous to those described in experimental designs by Berry (2004) and Ockey (2009). By manipulating the number of introvert and extravert students in groups, these studies identified situations in which raters are inclined to over or under-reward the scores of individuals. While this dissertation did not include measures of personality, it did replicate the situation in which one person's score was boosted relative to his or her groupmates as found by Berry (2004) and Ockey (2009). Observing how two of the members of the high initial group assertively took turns (Yamahiko and Hamako) it seems highly likely they would score highly on an extravert scale, and Kabuhito's behaviour (low initial group) seemed consistently introverted. Although this cannot be confirmed empirically, it is sufficient to say that Berry (2004) and Ockey (2009) show that candidates who stand out may receive a higher grade than expected given the performance according to the scoring bands. Of course, the GOT at this institution is supposed to be criterion-based: raters should only compare the test-takers to the scoring bands, but the literature shows that making comparisons to other test-takers is something that raters may do regardless of their training (Orr, 2002). It may be that in a GOT where there are up to four test-takers to be graded concurrently it may be even more tempting to compare them with each other than to the criteria in the scoring bands. Nonetheless, the implication from this study is that rater training needs to highlight situations in which raters may be more likely to give test-takers higher scores than they warrant. A system in which similar ability students are grouped together, as noted above, would reduce the frequency of situations in which the test-takers' abilities differ to such a wide degree.

Also, this study clearly demonstrates the need for clear definitions and examplars of key words in the scoring bands. For example, the phrase "initiate interaction" in the scoring bands is crucial for determining a score in the 1.0 band or the 2.0 band of the CS scale. Yet when Machiko (low initial group) in the first administration read out a prompt to restart the conversation, it was not recognized in her score as "initiating" and she did not get credit for it. By contrast, in the same administration, Taeko (medium initial group) asked a single transfer question ("How about you?"), which came after her own response to another group member's question. As such it seems to function more like "continuing" a conversation rather than "initiating", and yet she got credit for it by receiving a score in the 2.0 band of the CS scale. Another candidate in the low initial group, Saaya, in the second administration received a score in the 2.0 band without initiating a topic, which seems even less justified. Making the definitions clear would enable raters to more quickly recognize what counts and what does not, and improve the scoring of this test. Additionally, creating standardized sets of training examples which operationalize such definitions should be made a priority to reduce the chance of overall levels of strictness and leniency varying from administration to administration.

6.3.5 Implications for the scoring bands

The results presented in this dissertation have pointed to some areas in which the scoring bands of the GOT need to be improved. One of the most important is the finding that advanced communication skills are not adequately described in the higher bands of the communicative skills scale. Such moves as referring to or summarizing what other speakers previously said, completing others sentences, speaking without a gap or in a turn overlapping the preceding speaker and collaboratively building turns are all features of more advanced speakers and if included would differentiate the higher levels from lower ones more clearly. This could have an additional effect of improving their scores in the later administrations, and thus more accurately reflect the achievement shown by this cohort as discussed earlier.

Also in the communicative skills band is the problematic language that is used from what seems to be incomplete descriptions in the bands. As noted above, it is clearly stated that test-takers in the 1.0 band "do not initiate interaction", but the "initiation" is absent from the 2.0 band, which instead states that the test-taker "responds to others without long pauses to maintain interaction", a feature that is not mutually exclusive with "initiating interaction". In the case where a test-taker does not "initiate an interaction", but does "respond without long pauses" the rater is caught in a contradiction. According to the former criteria the score should be in the 1.0 band, but according to the latter it should be at least 2.0. The remaining criteria in the 2.0 band are of little use, since a test-taker who “shows agreement or disagreement to others’ opinions” does not resolve the contradiction. Indeed, from the very first administration most test-takers seem more than capable of showing agreement, which makes this criterion a questionable fit for the 2.0 band. A revision of the communicative skills bands should see a clearer demarcation between the more passive functions below the 2.0 band of agreeing and responding, and the more active ones of initiating and disagreeing above that mark. The co-constructed functions that more advanced speakers are capable of should be more clearly described above the 3.0 mark, allowing those who display such features to be rewarded for them. However, such revisions would be rendered ineffectual if the co-constructed nature of the interaction is not taken into account. High level uses may not be able to employ such techniques if there is no competition from the floor by other high level users, as may have been the case with those in the high initial scoring group in this study. While contextualising notes would go some way to addressing this issue, grouping similarly skilled test-takers together would be the ideal solution. This could be done by means of appropriately detailed ‘can do’ self-assessments, if other measures of their speaking ability are lacking.

Another aspect that needs to be improved is the recognition of the test-takers’ use of language in an assessment situation, which the grammar and vocabulary bands fail to do. If the GOT represents a real act of communication, then the priority for the group members should be on being comprehensible to their fellow group members. In order to be comprehensible it is important to use grammar and vocabulary that fellow conversants understand. Using language that is advanced is risky because even if it is used correctly, the group members may not be able to understand it. A misunderstanding might be embarrassing and lead to loss of face (Luk, 2010), and it is unpredictable how the negotiation for meaning might be resolved. For a test-taker, such unquantifiable outcomes would surely result in a conservative strategy for using vocabulary and grammar. Yet the bands for grammar and vocabulary in bands above 2.0 demand that ‘complex’ and ‘late-acquired’ grammar to be at least attempted, and looks for evidence of vocabulary that is ‘advanced’ and in ‘a wide range’ without specifying what exactly this means. If a test-taker wants a high score in these bands, it seems that test-takers must be prepared to be misunderstood by their peers. For the language user who finds him or herself in a group of apparently lower level speakers, it might result in simplifying the language they would otherwise use, and ending up with a lower grade than they would otherwise have. The problem is exacerbated by grammar and vocabulary being in two scales, making 40% of their score subject to this conflict between language as an act of communication, and language as a tool of assessment. The likelihood of the former issue occurring can be reduced by taking the measure suggested in 6.3.2 above: by having similar level test-takers in each GOT this situation is less likely to occur. A further improvement would be to combine the vocabulary and grammar scales, so that it reduces the weight of these bands. [12]

The fluency bands also could be improved. At the highest level it includes the description "shows the ability to speak quickly in short bursts", but without adequate definition it is almost meaningless. As was noted, Hamako (high initial group) had a burst of nine words between pauses in her second administration, and this was not enough to give her a higher score than 3.0 in fluency. If it were possible to benchmark such statements as these, it would make the scoring bands easier to use. An implication of this study is the need for a sustained investigation of fluency at various levels in order to rewrite the rating bands so that they can be clearly differentiated. This would be a major study since it would need to be conducted in conjunction with a study on raters and the extent to which the features of fluency could be identified.

6.3.6 Implications for the use of indices to track performance in GOT

This study is one of the few in the literature to use Complexity, Accuracy, and Fluency (CAF) and vocabulary indices to track development in conversation rather than in monologic or one-way formats, and one of even fewer to mix a qualitative analysis with it. As such, this dissertation makes a valuable contribution to the already considerable literature on the use of these indices. Moreover, this study has made explicit the difference between using these indices for the language in short and long turns. It was found that the average CAF statistics derived from turns of over ten seconds in length are more

[12] In the time between the collection of the data for this dissertation and its completion, a revision of the scoring bands saw the grammar and lexis scales being combined.

complex, less accurate and less fluent than the overall figures, and future studies seeking to use them to analyse conversational language needs to take this into account (for example, see Nitta and Nakatsuhara, 2014).

One of the startling findings in this dissertation was the contrast between the significant differences in the indices between the second and third adminstrations, and the lack of recognition for this in the scores awarded. An implication that may be entertained is that the improvements in the indices in the final administration were too subtle for the raters to notice. It has been pointed out in this dissertation that the gains made in the final administration were less overt than those made in second administration, but if it is not possible to detect such differences, then the GOT can hardly be at fault. It is simply an assumption that a significant difference in an index should be noticeable to a human judge, but it is an important issue to clarify for future research in this field.

An inference that can be made from the apparent ability of some of the indices to track development at a group level over time is its independent of IC. While it is necessary to be cautious since the only source of data available is from the test itself, it seems that the core figures of words, clauses and AS-units per opportunity to speak, words per AS-unit, error-free proportion and per opportunity to speak, and the fluency indices of articulation rate, speech rate, pause proportion and the maze and sound ratio have the potential to be used to guage user development. Since research that used data from the second administration of this dissertation found that fluency and complexity indices were significantly affected by the prompt (Leaper & Riazi, 2014), it is important to qualify this further. If all due measures were taken to ensure consistency of the prompt and conditions, and the rating is of sufficient quality, then the GOT can plausibly measure an individuals development in speaking ability.

In this dissetation, using the indices to track development on a cohort scale was remarkably successful: most of the indices served to show how the test-takers developed in various stages. There were two notable exceptions: the failure of the vocabulary indices and the index of syntactical complexity (the ratio of clauses to AS-units) to track development. These are discussed in the sections below.

6.3.6.1 Implications for the use of vocabulary indices to track performance in GOTS

For the vocabulary indices, the major handicap was that many test-takers did not reach the threshold for stable measurement (Koizumi, 2012; Koizumi & In'nami, 2012) and amongst those that did no significant difference could be detected. Several explanations may be put forward for this failure. The first is that the number of participants reaching the minimum number of words in more than one administration may have been too small for significant differences to be detected. With the number of qualifying test-takers varying between 10 and 26 it may not be surprising that no significant differences could be detected. Also, it should be considered that these indices were usually developed for the analysis of writing, not speaking. Unlike a written text which may consist of a single cohesive unit, a person's part in a conversation typically consists of a number of turns collected together, which these indices were usually not designed to cope with. This factor is probably multiplied in this study in which one conversation is compared to another in which the pattern of long and short turns may differ markedly, depending on such factors as the prompt (Leaper & Riazi, 2014). Finally, there is a real possibility that in fact their lexical diversity did not change significantly from administration to administration. As it was argued above (Section 6.3.4) since the test-takers were responding to the communicative situation in which they used vocabulary that they knew their groupmates would understand, the GOT may not be an appropriate arena to measure the use of advanced vocabulary. Given these factors, it may not be surprising that the vocabulary indices failed to show development over the three administrations.

6.3.6.2 Implications for the use of complexity indices to track performance in GOTS

This study used two measures of complexity: words per AS-unit and clauses per AS-unit. While the former followed the familiar pattern of development in this dissertation of a rise in the second administration and no change in the third, it was also notable that the latter showed no significant change over the three administrations. A reason for this was suggested in Leaper and Riazi's (2014) study which used data from the second administration of this study that shared many of the same participants. They found that syntactical complexity differed significantly depending on the specific prompt used, suggesting that it is more subject to variance in response to more immediate test context factors, such as the prompt.

Instead it seemed that the measure of words per AS-unit provided a measure of complexity that allowed the development of grammar to be measured. However, when used to track progress of the individuals in the qualitative study, it was found to vary markedly depending on the length of the turn: it declined when the candidates produced a greater quantity of short turns, and rose if longer turns were predominantly (or exclusively in the case of Saaya) used, and this phenomenon could be observed in the members of all three groups. Given the conclusion from research question 2 that the path of development trends towards having a greater number of shorter turns in more advanced learners, as inferred from Galaczi (2013), this would reduce its value as a measure to track development of conversational ability. The implication of this, when considered with the arguments about the student's conservative use of grammar, as discussed in Section 6.3.4, is that complexity indices should be interpreted with caution in the GOT.

6.4 Limitations and suggestions for future research

Along with the implications, there are limitations that need to be considered, and from the limitations ideas arise for future research. The first limitation is that the only source of data was the video recordings of the GOTs, and transcripts produced from them. Adding an insider view through interviewing and surveying participants, both test takers and raters, would have enabled a broader foundation of information from which to draw conclusions. For example, while I have pointed out the similarities between the conditions described by Ockey (2009) and Berry (2004) under which raters can be induced to overcompensate test-takers, without surveying their assertiveness it cannot be conclusively asserted that it is the result of the same phenomenon that is driving the effect for Kabuhito (low initial group) and Yamahiko and Hamako (high initial group), but not to the same extent as Naeko (high initial group), whose leadership seems to be borne out of being the least reluctant to lead the group. Also, an insider view would be invaluable in uncovering the reason Aemi's (medium initial group) performance seemed to decline in the last administration. Her body language seemed to indicate that she lacked the will to be more engaged than she was, but only an introspective account from the participant could be definitive. Further research that adds such sources of information would add richness to the description of the learners' progress in the GOT.

A limitation that would have important consequences for the GOTs validity is the lack of evidence of the test-takers ability outside the performances they show on the GOT. Having sources of their speaking ability from classroom tasks or conversations in other contexts would enable a study such as this to ascertain the degree to which the GOT in a relatively narrow assessment context can tap their ability. Such information would have widened the scope of this investigation and added valuable insights into the GOT's role as a measure of student development. Additionally, as the reader will have noticed, the amount that some speakers participate in the GOT is vanishingly little. Having such outside data would allow the extent that the few words spoken are representative of their overall ability to speak to be ascertained. Indeed, when calculating the indices, having data of their speaking outside the assessment would have allowed the data obtained from the test to be verified as representative of their overall speaking ability. Additionally, since the fluency indices were calculated only from their longer turns, the pool of data from which to draw upon was usually further diminished – especially in the case of the few participants in the first and third GOTs who did not have a single turn longer than 10 seconds in length. Being able to state the extent to which their performance on the GOT represents their ability would enable a more realistic perspective of the validity of this format. Future research that collects conversational data from test-takers outside the test and compares it to what they can do in the test is a research gap begging to be filled.

Another limitation stems from the inability to include an objective measure of pronunciation in this dissertation. The literature on pronunciation is problematic in that it necessarily involves interpretation by the listener. Studies that have included pronunciation use judges opinions about specific features such as target-like pronunciation, intonation, and rythm (Brown, Iwashita & McNamara, 2005; Iwashita, Brown, McNamara, O'Hagan, 2008). However, a frequent issue raised among raters is how familiarity with the accent affects comprehension (Brown, 2006), and other research has demonstrated that such concerns are justified, as a significant number of raters who had prolonged exposure to a specific accent have been found award higher scores to its speakers (Carey, Mannell & Dunn, 2011). Since the auther has moved countries since collecting the data for this dissertation and had limited access to judges sufficiently familiar with Japanese accented English, it was impractical to include pronunciation in this study. As such this is an area for future research to investigate.

The qualitative section of this dissertation described the non-verbal behaviour of the candidates where it was relevant to describing their attempts to communicate. While this was revealing, a more systematic description of non-verbal communicative behaviour within a framework such as that described by Gan & Davidson (2011) would have allowed a richer analysis of the test-takers' performances. Besides the above mentioned study, there has been little attention paid to non-verbal communication in speaking assessment contexts, and future research in this area could be rewarding.

This dissertation has made a contribution to the field by clarifying the extent to which it can elicit various features of interactive language over time at the level of a cohort, but was limited in its investigation of the extent to which it generates washback in the curriculum. Although the features more directly related to language learning – the collaborative functions – where found to be used at a low level, it is a question for future research to identify whether that is sufficient for the role of the test in the administration. That is, knowing that these kinds of skills may be used in the test, how and to what extent they drive teachers to practice them in their classes and students to practice them when studying for the GOT remains unclear. Future research could clarify these matters and lead the way for improved teaching of speaking and a closer realignment of assessment and teaching at such institutions as these.

Related to the above notion are areas for future research that focus on interactivity in the GOT. This dissertation has shown certain aspects of students' development of skills in the GOT over time, and some consistencies can be seen with the cross-sectional study of Galaczi (2013), but how students develop in these skills is an area that little research has been conducted on. For example, a question that has yet to be answered is if teaching interactivity empowers students to develop such conversational skills earlier, or if students simply cannot use them until they have developed sufficiently to allow attentional resources to devote to them. If so, at what point does this occur in the cycle of development? By answering such questions as these, appropriate steps can be taken to align the curriculum to assessment to the benefit of learning to communicate in the foreign language.

The ultimate objective of a language teaching institution is to prepare students to use the target language in communicative contexts in the real world. To do this it must provide the most effective curriculum, teaching and learning activities and assessment tools to allow its students to build their “architecture of the practice” (Young, 2000, p. 6). To the author’s knowledge, this dissertation is the first in the literature on peer assessed speaking formats to track the progress of the same learners over the course of years and administrations to illustrate its development. In doing this, if this dissertation has provided an insight into the effectiveness of a key part of an educational system, then it has performed a valuable service and thereby contributed to the field of foreign language education.

REFERENCES

Adams, R. (2007). Do second language learners benefit from interacting with each other. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies*, (pp. 29-51). Oxford: Oxford University Press.

Agard, F. B., & Dunkel, H. B. (1948). *An investigation of second-language teaching*. Chicago, IL: Ginn.

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: does it transfer to a new task? *Language Learning Journal*, 39(3), 269-280.

American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines*. Yonkers, NY: ACTFL.

American Council on the Teaching of Foreign Language (1999). *ACTFL Oral Proficiency Interview Tester Training Manual.* Yonkers, NY: ACTFL.

Austin, J. L. (1962). *How to do things with words*. Oxford: Oxford University Press.

Bachman, L. F. (1988). Problems in examining the validity of the actfl oral proficiency interview. *Studies in Second Language Acquisition,* 10(2), 149-164.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task based language performance assessment. *Language Testing,* 79 (4), 453-476.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. D. Fox (Ed.) *Language testing reconsidered* (pp. 41-71). Ottawa, Ontario: University of Ottawa Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal,* 70(4), 380-390.

Bennett, R. (2012). Is linguistic ability variation in paired oral language testing problematic? *ELT Journal*, 66(3), 337-346.

Berkoff, N. A. (1985). Testing oral proficiency: A new approach. In Y. P. Lee (Ed.) *New directions in language testing*, (pp. 93-100). Oxford: Pergamon Press.

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets* (Unpublished doctoral dissertation). King's College, University of London, U.K.

Bonk, W.J. (2003), *KEPT Senegal 2003 administration: Report and analysis*. Unpublished report. Kanda University of International Studies, Chiba, Japan.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.

Bonk, W. J., & Van Moere, A. (2004, March). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores.* Paper presented at the Language Testing Research Colloquium, Temecula, California.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2), 177-220.

Bradac, J. J., & Street Jr, R. L. (1989). Powerful and powerless styles of talk: A theoretical analysis of language and impression formation. *Research on Language & Social Interaction*, 23(1-4), 195-241.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing,* 26(3), 341-366.

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. *IELTS research reports,* 6, 71-90.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1-25.

Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Woods (Ed.), *Research Reports 1997* Vol. 1, (pp. 173-191). Sydney: ELICOS.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1), i-157.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. NY: Longman.

Bruton, A., & Samuda, V. (1980). Learner and teacher roles in the treatment of oral error in group work. *RELC Journal,* 11(2), 49-63.

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research*, 3(3), 185-214.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), 201-219.

Carr, T. H., & Curran, T. (1994). Cognitive factors in learning about structured sequences. *Studies in Second Language Acquisition,* 16(02), 205-230.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369-383.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.) *Handbook of research in second language teaching and learning* (pp. 815-832). Mahwah, NJ: Erlbaum.

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, 36(4), 498-506.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L.F. Bachman and A.D. Cohen (Eds.) *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19: 254-272.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, 25(1), 39-62.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly*, 3 (3), 295–306

Clancy, P. M., Thompson, S. A., Suzuki, R., & Tao, H. (1996). The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, *26*(3), 355-387.

Coates, J. (1994). No Gap, Lots of Overlap; Turn-taking Patterns in the Talk of Women Friends. In D. Graddol, J. Maybin, & B. Stierer (Eds.). *Researching language and literacy in social context: A reader*, (pp. 177-192). Avon, Clevedon: Multilingual Matters.

Cobb, T. (2002). VocabProfile program [software]. Available from http://www.lextutor.ca/vp/

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*(4), 161-170.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3), 573-605.

Csépes, I. (2002). Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores. *novELTy*, 9(1), 22-45.

Curran, T., & Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,19(1), 189-202.

D'agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4), 316-321.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26(3), 367-396.

De Jong, N. (2013, August). *How do I choose which measure of fluency to use?* Paper presented at EUROSLA Conference, Amsterdam.

Dings, A. (2007). *Developing interactional competence in a second language: A case study of a Spanish language learner* (Unpublished doctoral dissertation). The University Of Texas at Austin: Texas, TX.

Doughty, C., & Pica, T. (1986). “Information gap” tasks: do they facilitate second language acquisition? *TESOL Quarterly*, 20 (2), 305-325.

Drew, P., & Heritage, J. (1992). *Analyzing talk at work: An introduction*. In P. Drew, & J. Heritage, (Eds.), *Talk at work: Interaction in institutional settings* (Vol. 8) (pp. 3-65). Cambridge: Cambridge University Press.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423-443.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, 5 (2), 160-167.

Eckman, F. R., Bell, L., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of english as a second language. *Applied Linguistics*, 9(1), 1-20.

Eggins, S., & Slade, D. (1997). *Analysing casual conversation.* London: Cassell.

Egyud, G., & Glover, P. (2001). Readers respond. Oral testing in pairs-secondary school perspective. *ELT Journal*, 55(1), 70-76.

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *IELTS Research Reports*, 4, 207-254.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics,* 30(4), 474-509.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Ellis, R., Tanaka, Y., & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of l2 word meanings. *Language Learning*, 44(3), 449-491.

ETS. (1982). *Oral proficiency training manual.* Princeton: Educational Testing Service.

ETS. (1996). *TOEIC language proficiency interview manual.* Princeton: Educational Testing Service.

Field, A. (2005). *Discovering statistics with SPSS*. London: Sage.

Folland, D., & Robertson, D. (1976). Towards Objectivity in Group Oral Testing. *ELT Journal,* 30(2), 156-167.

Foot, M. C. (1999a). Relaxing in pairs. *ELT Journal*, 53(1), 36-41.

Foot, M. C. (1999b). Reply to Saville and Hargreaves. *ELT Journal*, 53(1), 52-53.

Foster, P. (1998). A Classroom Perspective on the Negotiation of Meaning. *Applied Linguistics,* 19(1), 1-23.

Foster, P., & Ohta, A. S. (2005). Negotiation for Meaning and Peer Assistance in Second Language Classrooms. *Applied Linguistics,* 26(3), 402-430.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3), 299-323.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354-375.

Fujii, A., & Mackey, A. (2009). Interactional feedback in learner-learner interactions in a task-based EFL classroom, *International Review of Applied Linguistics in Language Teaching,* 47(3/4), 267-301.

Fulcher, G. (1987). Tests of Oral Performance: the need for data-based criteria. *English Language Teaching Journal,* 41, 4, 287 - 291.

Fulcher, G. (1996a). Testing tasks: issues in task design and the group oral. *Language Testing*, 13(1), 23-51.

Fulcher, G. (1996b). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing,* 13, 2, 208 - 238.

Fulcher, G. (2003). *Testing second language speaking*. London: Longman/Pearson Education.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, 20(3), 321-344.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the first certificate in english examination. *Language Assessment Quarterly*, 5 (2), 89-119.

Galaczi, E. D. (2010, September). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? In R. McColl, M. Durham & M. Durham (Eds.), *Applied linguistics, global and local.* Paper presented at BAAL Annual Conference, University of Aberdeen, Scotland, 9-11 September (pp. 91-103). London: Scitsiugnil Press.

Galaczi, E. D. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics.* Advanced online publication. doi:10.1093/applin/amt017

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher-and lower-scoring students. *Language Testing*, 27(4), 585-602.

Gan, Z. (2011). An investigation of personality and L2 oral performance. *Journal of Language Teaching and Research*, 2(6), 1259-1267.

Gan, Z. (2013). Task type and linguistic performance in school-based assessment situation. *Linguistics and Education*, 24(4), 535-544.

Gan, Z., & Davison, C. (2011). Gestural behavior in group oral assessment: a case study of higher-and lower-scoring students. *International Journal of Applied Linguistics*, 21(1), 94-120.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2009). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics*, 30(3), 315-334.

García Mayo, M. D. P., & Pica, T. (2000). Interaction among proficient learners: are input, feedback and output needs addressed in a foreign language context? *Studia Linguistica,* 54(2), 272-279.

Gass, S. M. (1997). *Input, interaction, and the second language learner.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Gass, S. M., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA Review*, 19(1), 3-17.

Gass, S. M., Mackey, A., & Ross-Feldman, L. (2005). Task-Based Interactions in Classroom and Laboratory Settings. *Language Learning*, 55(4), 575-611.

Gass, S. M. & Selinker, L. (2008). *Second Language Acquisition: An Introductory Course (3rd Edition).* New York: Routledge/Taylor Francis.

Geluykens, R. (1993). Topic introduction in English conversation. *Transactions of the Philological Society*, 91(2), 181-214.

Goffman, E. (1959). *The presentation of self in everyday life.* Garden City, NY: Double Day.

Goffman, E. (1967). *Interaction ritual: essays on face-to-face interaction.* Garden City, NY: Double Day.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Massachusetts: Harvard University Press.

Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.

Greer, T., & Potter, H. (2008). Turn-taking practices in multi-party EFL oral proficiency tests. *Journal of Applied Linguistics*, 5(3), 297-320.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. Morgan (Eds.) *Syntax and Semantics Volume 3: Speech Acts*. New York: Academic Press, 41-58.

Hall, J. K. (1993). The role of oral practices in the accomplishment of our everyday lives: The sociocultural dimension of interaction with implications for the learning of another language. *Applied Linguistics*, 14(2), 145-166.

Hall, J. K. (1995). (Re) creating our worlds with words: A sociohistorical perspective of face-to-face interaction. *Applied Linguistics*, 16(2), 206-232.

Halliday, M.A.K. (1994). *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.

Harris, R. J., Lee, D. J., Hensley, D. L., & Schoen, L. M. (1988). The effect of cultural script knowledge on memory for stories over time. *Discourse Processes*, *11*(4), 413-431.

Harris Wright, H., Silverman, S., & Newhoff, M. (2003). Measures of lexical diversity in aphasia. *Aphasiology*, 17(5), 443-452.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 402-435). Rowley, MA: Newbury

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, 23(3), 370—401.

He, A. W., & Young, R. (1998). *Language proficiency interviews: A discourse approach*. In R. Young & A. W. He (Eds.), Talking and testing: Discourse approaches to the assessment of oral proficiency (pp. 1–24). Amsterdam: John Benjamins.

Heatley, A., & Nation, P. (1996). Range [software]. Wellington, New Zealand: Victoria University of Wellington.

Hilsdon, J. (1991). The group oral exam: Advantages and limitations. In J. C. Alderson & B. North (Eds.), *Language testing in the1990's* (pp. 189-197). London: Modern English Publications and the British Council.

Horst, M., & Collins, L. (2006). From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, 63(1), 83-106.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461-473.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Ishida, M. (2009). Development of interactional competence: Changes in the use of ne in L2 Japanese during study abroad. In H. Nguyen & G. Kasper (Eds.), *Talk-in-interaction: Multilingual perspectives,* (pp. 351-385).Hawaii: National Foreign Language Resource Centre.

Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics*, 33(12), 1859-1880.

Iwashita, N. (1996). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing*, 5(2), 51–65.

Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative–nonnative interaction in Japanese as a foreign language. *System,* 29(2), 267-287.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24-49.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on language and social interaction*, 28(3), 171-183.

Johnson, M. (2001). *The art of non-conversation: A reexamination of the validity of the oral proficiency interview.* New Haven, CT: Yale University Press.

Johnson, M., & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). Amsterdam: John Benjamins.

Kaulfers, W. V. (1944). Wartime Development in Modern-Language Achievement Testing. *Modern Language Journal*, 28(2), 136-150.

Kobayashi, M. & Van Moere, A., (2003, September). *Group oral testing: Amount of floor time and score variance*. Paper presented at British Association of Applied Linguistics Conference, Leeds, UK.

Koizumi, R. (2012). Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens? *Vocabulary Learning and Instruction*, 1(1), 60-69.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40(4), 554-564.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal*, 70(4), 366-372.

Krashen, S. D. (1981). *Second language acquisition and second language learning*. Oxford: Oxford University Press.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Krashen, S. (1998). Comprehensible output? *System,* 26(2), 175-182.

Lado, R. (1961) *Language testing: the construction and use of foreign language test*. London: Longman.

Lantolf, J.P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal*, 69, 337-345.

Lantolf, J. P., & Frawley, W. (1988). Proficiency - understanding the construct. *Studies in Second Language Acquisition*, 10(02), 181-195.

Larsen-Freeman, D. (1991). Second Language Acquisition Research: Staking out the Territory. *TESOL Quarterly*, 25 (2), 315-350.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16 (3), 307-322.

Lazaraton, A. (1992). The Structural Organization Of A Language Interview: A Conversation Analytic Perspective. *System*, 20(3), 373-386.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests,* (Vol. 14). Cambridge: Cambridge University Press.

Lazaraton, A. (2006). Process and outcome in paired oral assessment. *ELT Journal*, 60(3), 287-289.

Lazaraton, A., & Davis, L. (2008). A Microanalytic Perspective on Discourse, Proficiency, and Identity in Paired Oral Assessment. *Language Assessment Quarterly*, 5(4), 313-335.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing*, 31(2).

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning,* 40(3), 387-417.

Li, Y., & Qian, D. D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System*, 38(3), 402-411.

Liddicoat, A. J. (2011). *An introduction to conversation analysis*. New York: Continuum.

Liski, E., & Puntanen, S. (1983). A study of the statistical foundations of group conversation tests in spoken English. *Language Learning*, 33(2), 225-246.

Liski, E. P., & Puntanen, S. (1985). A percentile regression model with an application to error frequency in group conversation tests. *Canadian Journal of Statistics*, 13(1), 71-78.

Loban, W. (1976). *Language development: Kindergarten through grade twelve* (Vol. 18). Urbana, Illinois: National Council of Teachers of English.

Loewen, S., & Philp, J. (2006). Recasts in the adult english l2 classroom: Characteristics, explicitness, and effectiveness. *Modern Language Journal*, 90(4), 536-556.

Lombardo, L. (1984). Oral testing: getting a sample of real language. *English Teaching Forum* (Vol. 22, (1), 2-6.

Long, M. H. (1981). Input, interaction, and second language acquisition. *Annals of the New York Academy of Sciences,* 379(1), 259-278.

Long, M. H. (1983a). Linguistic and conversational adjustments to non-native speakers. *Studies in Second Language Acquisition*, 5(2), 177-193.

Long, M. H. (1983b). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, 4(2), 126-141.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition: Vol. 2. Second language acquisition* (pp. 413–468). New York: Academic Press.

Long, M. H., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *Modern Language Journal*, 82(3), 357-371.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, 96(2), 190-208.

Luk, J. (2010). Talking to score: impression management in L2 Oral assessment and the co-construction of a test discourse genre. *Language Assessment Quarterly*, 7(1), 25-53.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, 26(03), 399-432.

Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford: Oxford University Press.

McCarthy, P. M. (2011). Gramulator (Version 6.0) [software]. Available from https://umdrive.memphis.edu/pmmccrth/public/software/software_index.htm

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459-488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392.

McDonough, K. (2004). Learner-learner interaction during pair and small group activities in a Thai EFL context. *System*, 32(2), 207-224.

McDonough, K. (2005). Identifying the impact of negative feedback and learners' responses on esl question development. *Studies in Second Language Acquisition*, 27(1), 79-104.

McLaughlin, B. (1987). *Theories of Second Language Learning*. London: Edward Arnold.

McNamara, T. F. (1997). ‘Interaction’ in second language performance assessment: Whose performance? *Applied Linguistics*, 18(4), 446-466.

Mackey, A., Abbuhl, R., & Gass, S. M. (2012). Interactionist approach. In S. M. Gass, & A. Mackey, (Eds.). *The routledge handbook of second language acquisition,* (pp. 7-23). Oxon: Routledge.

Mackey, A., Oliver, R., & Leeman, J. (2003). Interactional input and the incorporation of feedback: an exploration of ns–nns and nns–nns adult and child dyads. *Language Learning*, 53(1), 32.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *Modern Language Journal*, 82(3), 338-356.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36(2), 173-190.

Malone, M. E., & Brooks, R. L. (2013). Language testing in the government and military. In C. A. Chapelle (Ed.) *The encyclopedia of applied linguistics*. Retrieved from http://onlinelibrary.wiley.com/book/10.1002/9781405198431/homepage/6_Language_Testing_in_the_Government_and_the_Military.pdf

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, 15,197-237.

Markee, N. (2000). *Conversation analysis*. New Jersey: Routledge.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26(3), 397-421.

May, L. (2011). Interactional competence in a paired speaking test: features salient to raters. *Language Assessment Quarterly*, 8(2), 127-145.

Mercer, N. (1995). *The guided construction of knowledge: Talk amongst teachers and learners*. Cleavedon: Multilingual matters.

Messick, S. (1989) Validity. In R. L.Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-103). Washington , DC : American Council on Education & National Council on Measurement in Education.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, *45*(3), 241-259.

Moll, LC (Ed.). (1990). *Vygotsky and education*. Cambridge, England: Cambridge University Press.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*,46(4), 610-641.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, 28(4), 483-508.

Negishi, J. (2010). Characteristics of interactional management functions in group oral by japanese learners of english. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(1) 57-79.

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System*, 39(4), 533-538.

Nguyen, H. T. (2008). Sequence organization as local and longitudinal achievement. *Text & Talk-An Interdisciplinary Journal of Language, Discourse Communication Studies*, 28(4), 501-528.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, 31(2), 147-175.

Nobuyoshi, J., & Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT journal*, 47(3), 203-210.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555-578.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT journal*, 59(4), 287-297.

Norton, B. (2006). Identity as a sociocultural construct in second language research. *TESOL in context [special issue]*, 22-33.

Norton, J. (2013). Performing identities in speaking tests: co-construction revisited. *Language Assessment Quarterly*, 10(3), 309-330.

Ockey, G. J. (2006). *Making a case for the group oral discussion test: The effects of personality on the group oral's score-based inferences* (Unpublished doctoral dissertation). University of California, Los Angeles.

Ockey, G. J. (2009). The effects of group members' personalities on a test-taker's L2 group oral discussion test scores. *Language Testing,* 26(2), 161-186.

Ockey, G. (2011). Self-consciousness and Assertiveness as Explanatory Variables of L2 Oral Ability: A Latent Variable Approach. *Language Learning*, 61(3), 968-989.

Our Philosophy, (n.d.), In *Kanda University of Foreign Studies.* Retrieved August 1, 2014, from http://www.kandagaigo.ac.jp/kuis/english/overview/philosophy/

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, 30(2), 143-154.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277-295.

O'Sullivan, B., & Nakatsuhara, F. (2011). *Quantifying conversational styles in group oral test discourse*. In B. O'Sullivan (Ed.), Language testing: Theories and practices. London: Palgrave.

Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairmentapplication of D. *Journal of Speech, Language, and Hearing Research*, *45*(5), 927-937.

Pallant, J. F. (2011). *SPSS survival manual: A step-by-step guide to data analysis with SPSS (4th Ed.).* Crows Nest: Allen & Unwin.

Palmer, M. T. (1989). Controlling conversations: Turns, topics and interpersonal control. *Communications Monographs*, 56(1), 1-18.

Pavesi, M. (1986). Markedness, discoursal modes, and relative clause formation in a formal and an informal context. *Studies in Second Language Acquisition*, 8(01), 38-55.

Pavlou, P. (1997). Do different speech interactions in an oral proficiency test yield different kinds of language? In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment - proceedings of LTRC 1996* (pp. 185-201). Jyväskylä: University of Jyväskylä and University of Tampere.

Pica, T. (1994). Research on negotiation: what does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning,* 44(3), 493-527.

Pica, T., Lincoln-Porter, F., Paninos, D., & Linnell, J. (1996). Language learners' interaction: How does it address the input, output, and feedback needs of l2 learners? *TESOL Quarterly,* 30(1), 59-84.

Pica, T., & Doughty, C. (1985). The role of group work in classroom second language acquisition. *Studies in Second Language Acquisition*, 7(02), 233-248.

Piehl, K. (2011). Can Adults learn a second language? research findings and personal experience. *Delta Kappa Gamma Bulletin,* 78 (1), p33-38.

Pine, J. M., & Lieven, E. V. (1990). Referential style at thirteen months: Why age-defined cross-sectional measures are inappropriate for the study of strategy differences in early language development. *Journal of Child Language*, 17(03), 625-631.

Plough, I. C., MacMillan, F., & O'Connell, S. P. (2011). Changing tasks… changing evidence: A comparative study of two speaking proficiency tests. In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. Prieto Botana, and E. Rhoades (Eds.) S*elected proceedings of the 2010 second language research forum,* (pp. 91-104). Massachusetts, University of Maryland

Porter, P. (1986). How learners talk to each other: Input and interaction in task-centered discussions. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition,* (pp. 200-222). Rowley, MA: Newbury House.

Raffaldini, T. (1988). The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition*, 10(2), 197-216.

Ranney, S. (1992). Learning a new script: An exploration of sociolinguistic competence. *Applied Linguistics*, 13(1), 25-50.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1-32.

Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching*, *47*(02), 135-173.

Richards, B. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language*, 14(02), 201-209.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes,* 14(4), 423-441.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, 45(2), 283-331.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction*, (pp. 287-318). Cambridge: Cambridge University Press.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching*, 45(3), 193-213.

Robinson, P. (2011). Task-Based Language Learning: A Review of Issues. *Language Learning*, 61(1), 1-36.

Ross, S. J. (2007). A comparative task-in-interaction analysis of OPI backsliding. *Journal of Pragmatics*, 39 (11), 2017-2044.

Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14(02), 159-176.

Ross-Feldman, L. (2007). Interaction in the L2 classroom: Does gender influence learning opportunities? A. Mackey. *Conversational interaction in second language acquisition: A series of empirical studies,* (pp. 53-77). Oxford: Oxford University Press.

Sacks, H. (1992). *Lectures on Conversation (Vol. 1).* Oxford: Basil Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 696-735.

Sandlund, E., & Sundqvist, P. (2011). Managing task-related trouble in L2 oral proficiency tests: contrasting interaction data and rater assessment. *Novitas-ROYAL (Research on Youth and Language)*, 5(1), 91-120.

Sato, M. (2007). Social relationships in conversational interaction: A comparison of learner-learner and learner-NS dyads. *JALT Journal*, 29(2), 183.

Sato, M., & Lyster, R. (2007). Modified output of Japanese EFL learners: Variable effects of interlocutor verses feedback types. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies,* (pp. 123-142). Oxford: Oxford University Press.

Savignon, S. J. (1985). Evaluation of Communicative Competence: The ACTFL Provisional Proficiency Guidelines. *Modern Language Journal*, 69(2), 129-134.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT journal*, 53(1), 42-51.

Saville-Troike, M., & Kleifgen, J. A. (1986). Scripts for school: Cross-cultural communication in elementary classrooms. *Text - Interdisciplinary Journal for the Study of Discourse*, 6(2), 207-222.

Schank, R. C. & Abelson. R.P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* New York, NY: Halsted.

Schegloff, E. A. (1987). Analyzing single episodes of interaction: An exercise in conversation analysis. *Social Psychology Quarterly*, 50(2), 101-114.

Schegloff, E. A. (2007). *Sequence organization in interaction*. Cambridge: University Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53, 361-382.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, 8(4), 289-327.

Schmidt, R. (1983). Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. In N. Wolfson, & E. Judd (1983). *Sociolinguistics and language acquisition* (pp. 137-174), Rowley, MA: Newbury House.

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, 11(2), 129-158.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition,* (pp. 237-326). Rowley, MA: Newbury House.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, 48(2), 281-317.

Schneider, W. & R. M. Shiffrin. (1977). Controlled and automatic human information processing: 1. Detection, search, and attention. *Psychological Review*, 84(1), 1-66.

Serrano, R., Tragant, E., & Llanes, A. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, 68(2), 138-163.

Shehadeh, A. (2003). Learner output, hypothesis testing, and internalizing linguistic knowledge. *System*, 31(2), 155-171.

Shohamy, E., Reves, T., & Bejarano, T. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal*, 40(3), 212-220.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics,* 30(4), 510-532.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93-120.

Stang, D. J. (1973). Effect of interaction rate on ratings of leadership and liking. *Journal of Personality and Social Psychology*, 27(3), 405.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (eds.) *Input in second language acquisition,* 165-179. Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (eds.) *Principle and practice in applied linguistics: Studies in honour of HG Widdowson*, (pp. 125-144). Oxford: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.) *Handbook of research in second language teaching and learning* (pp. 471-483) .Mahwah, NJ: Erlbaum.

Swain, M. (2008, May). *The output hypothesis: its history and its future.* Paper presented at the 5th International Conference on ELT in China, Beijing, China. Retrieved from http://www.celea.org.cn/2007/keynote/ppt/merrill%20swain.pdf

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *Modern Language Journal*, 82(3), 320-337.

Taguchi, N. (2007). Task difficulty in oral speech act production. *Applied Linguistics*, 28(1), 113-135.

Tannen, D. (1993). What's in a frame? Surface evidence for underlying expectations. In D. Tannen (Ed.) *Framing in discourse,* (pp. 14-56). New York, NY: Oxford University Press.

Taylor, L. (2000). Investigating the paired speaking test format. *Research Notes* 2, 14-15.

Taylor, L. (2003). The paired speaking test format: Recent studies. *Perspective*, 55(1), 70-76.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58(2), 439–473.

ten Have, P. (1999). *Doing conversation analysis*. London: Sage.

ten Have, P. (2007). *Doing conversation analysis* (2nd ed). London: Sage.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, 16 (02), 183-203.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics,* 17(1), 84-119.

Trahey, M., & White, L. (1993). Positive evidence and preemption in the second language classroom. *Studies in Second Language Acquisition*, 15(02), 181-204.

Truscott, J. (1998). Noticing in Second Language Acquisition: A Critical Review. *Second Language Research,* 14(2), 103-135.

Truscott, J. (2006). Optionality in second language acquisition: A generative, processing-oriented account. *International Review of Applied Linguistics in Language Teaching*, 44(4), 311-330.

Truscott, J., & Smith, M. S. (2011). Input, intake, and consciousness. *Studies in Second Language Acquisition,* 33(4), 497-528.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, 49(1), 3-12.

Valls Ferrer, M. (2011). *The development of oral fluency and rhythm during a study abroad period* (Unpublished doctoral dissertation). Universitat Pompeu Fabra: Barcelona, Spain.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly,* 23(3), 489–508.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing,* 23(4), 411–440.

Van Moere, A. (2007). *Group oral test: How does task affect candidate performance and test score?* (Unpublished PhD thesis). University of Lancaster, UK.

Van Moere, A. & Johnson, F.C. (2002, May). *Communicative assessment in a personal curriculum at Kanda University Of International Studies*. In A. Mackenzie, & T. Newfields, (Eds.), *Curriculum innovation, testing and evaluation.* Paper presented at the 1st Annual JALT Pan-SIG Conference. Retrieved from http://jalt.org/pansig/2002/HTML/VanJoh.htm

Varonis, E. M., & Gass, S. M. (1985). Miscommunication in native/nonnative conversation. *Language in Society*, 14(3), 327-343.

Vercellotti, M. L. (2012). *Complexity, Accuracy, and Fluency as Properties of Language Performance: The Development of the Multiple Subsystems over Time and in Relation to Each Other* (Unpublished doctoral dissertation). University of Pittsburgh, PA.

Wells, G. (Ed.). (1981). *Learning through interaction: The study of language development* (Vol. 1). Cambridge: Cambridge University Press.

Williams, J. (1999). Learner-generated attention to form. *Language Learning*, 49(4), 583-625.

Yagi, K. (2007). The development of interactional competence in a situated practice by Japanese learners of English as a second language [Electronic Version]. *Hawaii Pacific University TESL Working Paper Series, 5 (1).* Retrieved February 2, 2014 from http://www.hpu.edu/ CHSS/LangLing/TESOL/ProfessionalDevelopment/TESOL_WPS/Archives_TESOLWPS/

Vol_5_1Spring07.html

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning*, 45(1), 3-42.

Young, R. F. (2000, March). *Interactional competence: Challenges for validity*. Paper presented at Language Testing Research Colloquium, Vancouver, British Columbia, Canada.

Young, R. F. (2003). Learning to talk the talk and walk the walk: Interactional competence in academic spoken English. *North Eastern Illinois University Working Papers in Linguistics*, 2, 26-44.

Young, R. F. (2009). *Discursive practice in language learning and teaching*. Malden, MA: Wiley-Blackwell.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel, (Ed.). *Handbook of research in second language teaching and learning* (Vol. 2), (pp. 426-443). New York, NY: Routledge.

Young, R. (2013). Learning to talk the talk and walk the walk: Interactional competence in academic spoken English. *Ibérica*, 25, 15-38.

Young, R., & Milanovic, M. (1992). Discourse variation is oral proficiency interviews. *Studies in Second Language Acquisition,* 14, 403–424.

Young, R. F., & Miller, E. R. (2004). Learning as changing participation: Negotiating discourse roles in the ESL writing conference. *Modern Language Journal*, 88(4), 519-535.

Yule, G., & Macdonald, D. (1990). Resolving referential conflicts in l2 interaction: the effect of proficiency and interactive role. *Language Learning,* 40(4), 539-556.

APPENDICES

Appendix A: Prompts

Administration 1 (2004)

Prompt 1
Travelling in a group has been a popular way to travel for Japanese, but many younger travelers prefer to go by themselves or with their friends. Do you prefer to on group tour or as an independent tourist? Do you prefer to go to new places by yourself or with a group of friends? Give reasons for your answers.

Prompt 2
These days more and more people are going abroad not just for a holiday but also to live and work. Is this something that you want to do? What do you think are the good and bad points of living and working abroad?

Prompt 3
If you travel you can go to regular places, where many tourists go, or more unusual places where fewer people may visit. Do you want to go to unusual places or places that are more regular? What are the good and bad points of going to more unusual places?
2006

Administration 2 (2005)

Prompt 1
In the traditional Japanese family, men earned the money and women did the housework. Is your family traditional or not? Why do you think so? What are the advantages and disadvantages of the traditional family? Do you think the situation in Japan is changing? Why?

Prompt 2
Around the world, people are marrying later and later. What are the advantages of being single? Why? What are the advantages of being married? Why? Do you want to get married or would you prefer to be single? In future, do you think people will still want to get married?

Prompt 3
More and more young people are spending their free time inside the house watching TV, using the Internet and playing computer games. Do you like to spend your free time inside or do you prefer to do outdoor activities? Should we all try to spend some time doing outdoors activities? Why or why not? Is it healthy to spend all our time indoors?

Prompt 4
These days, lots of people have mobile phones and they are becoming very important in our lives. Do you have a mobile phone? Why? Do you often use it? How do you feel about

mobile phones? What are some good points and bad points about them? Why? Could you live without your mobile phone?

Administration 3 (2006)

Prompt 1
These days there are many types of fast food available. How often do you eat fast food? Do you prefer to eat Western fast food or Japanese fast food? Should young people be more careful about what they eat? Should we eat more home cooked food? How do you think fast food has affected Japanese society?

Prompt 2
Living in the country and living in the city can be quite different. Would you prefer to live in the city or in the country? Why? What are the benefits of living in the city? What are the benefits of living in the country? Where would you prefer to live in the future?

Prompt 3
More and more young people are spending a lot of their free time shopping. How often do you go shopping? What kinds of products do you buy? Do you like to shop alone or with your friends? Why? Do young people spend too much time shopping? Why or why not? Should young people spend more time doing other things? For example?

Prompt 4
Nowadays, more and more young people have part-time jobs. Do you have a part-time job? Why or why not? What factors are important when you select a part-time job? (e.g. when you work, your pay, the location, etc) What is the best kind of part-time job to have? Should students have part-time jobs or focus more on their studies?

<u>Appendix B:</u> Initial Transcription Coding System (Eggins & Slade, 1997, p. 5)

Symbol	Meaning
.	- certainty, completion (typically falling tone)
no end of turn punctuation	- implies non-termination
,	- parcelling of talk; breathing time
?	- uncertainty (rising tone, or wh-interrogative)
!	- "surpised" intonation (rising – falling tone)
WORDS IN CAPITALS	- emphatic stress and/or increased volume
" "	- change in voice quality in reported speech
( )	- untranscribable speech
(words within parentheses)	- transcribers guess
[words within square brackets]	- non-verbal information
= =	- overlap
…	- short hesitation within a turn (less than three seconds)
[pause – 4 secs]	- indication of inter-turn pause length
Dash – then talk	- false start/restart

Appendix C: Statistical Tables for Research Question 1

Table 65: Results of Friedman's ANOVA for core indices

	1	2	3	4	5	6
	AS-units	Clauses	Words	AS-unit/ OppSpk	Clause/ OppSpk	Words/ OppSpk
N	53	53	53	53	53	53
χ^2	34.124	37.206	34.872	29.925	32.725	39.736
P	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
In all cases, degrees of freedom = 2						

Table 66: Results of Friedman's ANOVA for complexity and accuracy

	1	2	3	4
	Complexity		Accuracy	
	Clauses/ AS-unit	Words/ AS-unit	E-Free Prop	E-Free/ OppSpk
N	53	53	53	53
χ^2	5.610	7.396	26.190	45.962
P	0.061	0.025*	0.000***	0.000***
In all cases, degrees of freedom = 2				

Table 67: Results of Friedman's ANOVA for fluency

	1	2	3	4	5	6
	Long Turns	Artic. Rate	Speech Rate	Pause Prop	Maze ratio	M & S Ratio
N	48	48	48	48	53	53
χ^2	9.000	6.125	41.292	38.292	17.128	13.736
P	0.011*	0.047*	0.000***	0.000***	0.000***	0.001**
In all cases, degrees of freedom = 2						

Table 68: Results of Freidman's ANOVA for core statistics, complexity and accuracy in long turns

	1	2	3	4	5	6	7	8	9
N=48	AS-units	Clauses	Words	LT Words Prop.	Clauses/ AS-unit	Words/ AS-unit	E-Free Prop	Maze Ratio	M & S Ratio
χ^2	19.181	26.138	19.542	2.021	1.380	8.576	22.340	15.906	26.167
P	0.000***	0.000***	0.000***	0.364	0.502	0.014*	0.000***	0.000***	0.000***
In all cases, degrees of freedom = 2, n = 53									

Table 69: Wilson Signed Ranks results for long turns

		1	2	3	4
		No. Long Turns	AS-units (LT)	Clauses (LT)	Words (LT)
Admin 1	Z	-3.910[b]	-4.243[b]	-4.767[b]	-4.979[b]
Vs	p	0.000***	0.000***	0.000***	0.000***
Admin 2	*r*	0.329	0.420	0.472	0.493
Admin 2	Z	-2.094[a]	-.807[a]	-.797[a]	-.309[a]
Vs	p	0.036*	0.420	0.426	0.757
Admin 3	*r*	-0.198	-0.079	-0.078	-0.030
Admin 1	Z	-1.219[b]	-3.260[b]	-3.557[b]	-4.072[b]
Vs	p	0.088	0.001**	0.000***	0.000***
Admin 3	*r*	0.122	0.326	0.356	0.407

a – based on positive ranks   b – based on negative ranks

<u>Appendix D:</u> Statistical Tables for Research Question 2

Table 70: Results of Freidman's ANOVA for turn taking

	Words	Turns	Long Turns	Words/ OppSpk	Turns/ OppSpk	L.Turns/ OppSpk	Words/ Turn	Turns /L.Turns
χ^2	35.010	17.949	12.409	39.094	6.491	4.829	9.223	2.490
p	0.000***	0.000***	0.002**	0.000***	0.039*	0.089	0.010**	0.288

In all cases, degrees of freedom = 2, n=53

Table 71: Initiating features elicited in the three administrations

		Q-O	Q-T	Q-Fu	Opin-O	Opin	Opin-Con	Opin-sug	App	Sum	Fun-I	Man-I	Jpn-I
Administration 1	Total	30	52	43	14	43	4	0	7	0	0	6	5
	Proportion	14.71%	25.49%	21.08%	6.86%	21.08%	1.96%	0.00%	3.43%	0.00%	0.00%	2.94%	2.45%
	Min.	0	0	0	0	0	0	0	0	0	0	0	0
	Median	0	1	0	0	0	0	0	0	0	0	0	0
	Max.	5	5	6	4	5	1	0	2	0	0	2	2
	Mean	0.57	0.98	0.81	0.26	0.81	0.08	0.00	0.13	0.00	0.00	0.11	5
	StdDev	0.99	1.29	1.18	0.74	1.26	0.27	0.00	0.39	0.00	0.00	0.38	0.35
Administration 2	Total	36	56	45	14	66	18	4	11	3	7	8	1
	Proportion	13.38%	20.82%	16.73%	5.20%	24.54%	6.69%	1.49%	4.09%	1.12%	2.60%	2.97%	0.37%
	Min.	0	0	0	0	0	0	0	0	0	0	0	0
	Median	0	1	0	0	1	0	0	0	0	0	0	0
	Max.	4	4	10	2	5	4	1	3	2	2	2	1
	Mean	0.68	1.06	0.85	0.26	1.25	0.34	0.08	0.21	0.06	0.13	0.15	0.02
	StdDev	1.03	1.13	1.60	0.59	1.27	0.71	0.27	0.57	0.30	0.44	0.46	0.14
Administration 3	Total	49	50	51	10	84	25	0	4	1	9	12	0
	Proportion	16.61%	16.95%	17.29%	3.39%	28.47%	8.47%	0.00%	1.36%	0.34%	3.05%	4.07%	0
	Min.	0	0	0	0	0	0	0	0	0	0	0	0
	Median	1	1	1	0	1	0	0	0	0	0	0	0
	Max.	4	6	5	2	7	3	0	2	1	2	3	0
	Mean	0.92	0.94	0.96	0.19	1.58	0.47	0.00	0.08	0.02	0.17	0.23	0.00
	StdDev	1.07	1.23	1.18	0.48	1.55	0.77	0.00	0.33	0.14	0.43	0.54	0.00

Table 72: Responding features elicited in the three administrations

Administration 1						
	Ans	Ag	React	Fun-R	Man-R	Jpn-R
Total	175	21	15	4	2	1
Proportion	80.65%	9.68%	6.91%	1.84%	0.92%	0.46%
Min.	0	0	0	0	0	0
Median	3	0	0	0	0	0
Max.	12	3	2	2	1	1
Mean	3.30	0.40	0.28	0.08	0.04	0.02
StdDev	2.55	0.72	0.57	0.33	0.19	0.14
Administration 2						
Total	198	52	48	23	12	0
Proportion	59.46%	15.62%	14.41%	6.91%	3.60%	0
Min.	0	0	0	0	0	0
Median	4	1	0	0	0	0
Max.	9	8	8	4	2	0
Mean	3.74	0.98	0.91	0.43	0.23	0
StdDev	2.24	1.50	1.56	0.95	0.54	0
Administration 3						
Total	205	46	46	16	3	0
Proportion	64.87%	14.56%	14.56%	5.06%	0.95%	0
Min.	0	0	0	0	0	0
Median	4	0	0	0	0	0
Max.	11	4	5	2	1	0
Mean	3.87	0.87	0.87	0.30	0.06	0
StdDev	2.55	1.16	1.21	0.50	0.23	0

Table 73: Developing features elicited in the three administrations

Administration 1					
	Dev	Ref	Fini-sum	Fini-yea	Fini-trail
Total	260	1	28	17	6
Proportion	83.33%	0.32%	8.97%	5.45%	1.92%
Min.	0	0	0	0	0
Median	4	0	0	0	0
Max.	20	1	3	3	1
Mean	4.91	0.02	0.53	0.32	0.11
StdDev	4.34	0.14	0.80	0.64	0.32

Continued on next page

Administration 2					
Total	427	13	61	15	14
Proportion	80.57%	2.45%	11.51%	2.83%	2.64%
Min.	0	0	0	0	0
Median	6	0	1	0	0
Max.	28	2	5	3	2
Mean	8.06	0.25	1.15	0.28	0.26
StdDev	5.56	0.52	0.99	0.63	0.56
Administration 3					
Total	391	16	69	24	12
Proportion	76.37%	3.13%	13.48%	4.69%	2.34%
Min.	0	0	0	0	0
Median	7	0	1	0	0
Max.	18	2	5	3	2
Mean	7.38	0.30	1.30	0.45	0.23
StdDev	4.54	0.54	1.07	0.75	0.47

Table 74: Collaborating functions elicited in the three administrations

Administration 1								
	Q-Cl	Q-con	Cor	Compl	SW	Incom	RespH	Enc
Total	18	8	0	1	4	6	16	2
Proportion	32.73%	14.55%	0.00%	1.82%	7.27%	10.91%	29.09%	3.64%
Mean	0.34	0.15	0.00	0.02	0.08	0.11	0.30	0.04
StdDev	0.65	0.63	0.00	0.14	0.33	0.32	0.57	0.19
Max.	3	4	0	1	2	1	2	1
Median	0	0	0	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0
Administration 2								
Total	19	24	2	7	6	4	10	0
Proportion	26.39%	33.33%	2.78%	9.72%	8.33%	5.56%	13.89%	0.00%
Mean	0.36	0.45	0.04	0.13	0.11	0.08	0.19	0.00
StdDev	0.65	0.85	0.19	0.39	0.42	0.27	0.52	0.00
Max.	3	3	1	2	2	1	2	0
Median	0	0	0	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0

Continued on next page

Administration 3								
Total	6	27	1	2	12	10	3	0
Proportion	9.84%	44.26%	1.64%	3.28%	19.67%	16.39%	4.92%	0.00%
Mean	0.11	0.51	0.02	0.04	0.23	0.19	0.06	0.00
StdDev	0.32	0.78	0.14	0.19	0.75	0.44	0.23	0.00
Max.	1	3	1	1	5	2	1	0
Median	0	0	0	0	0	0	0	0
Min.	0	0	0	0	0	0	0	0

Table 75: Results of Freidman's ANOVA on interactive categories

Normalized Interactive Features by Category						
	Init.	Resp.	Dev.	Collab.	Jpn	Total
N	53	53	53	53	53	53
χ^2	10.106	6.577	21.981	0.146	5.2	31.358
P	0.006**	0.037*	0.000***	0.93	0.074	0.000***
In all cases, degrees of freedom = 2						

Appendix F: Statistical Tables for Research Question 4

Table 76: Fluency correlations in administration 1-3

Administration 1							
N = 49		SRpermin	MazeSnd Ratio	ArtiRate	PauseProp	TurnsOpp*	WordsOpp*
Fluency Score	*r*	0.696	-0.387	0.306	-0.455	-0.472	0.704
	p	0.000***	0.004**	0.018*	0.001**	0.000***	0.000***
SRpermin	*r*		-0.419	0.426	-0.693	-0.569	.658
	p		0.002**	0.001**	0.000***	0.000***	0.000***
MazeSnd Ratio	*r*	-0.419		-0.316	0.206	0.486	-0.366
	p	0.002**		0.015*	0.082	0.000***	0.006**
ArtiRate	*r*	0.426	-0.316		0.345	-0.187	0.328
	p	0.001**	0.015*		0.009**	0.104	0.012*
PauseProp	*r*	-0.693	0.206	0.345		0.448	-0.400
	p	0.000***	0.082	0.009**		0.001**	0.003**
TurnsOpp*	*r*	-0.569	0.486	-0.187	0.448		-0.646
	p	0.000***	0.000***	0.104	0.001**		0.000***
WordsOpp*	r	0.658	-0.366	0.328	-0.400	-0.646	
	p	0.000***	0.006**	0.012*	0.003**	0.000***	

Continued on next page

Administration 2

N = 50		SRpermin	MazeSnd Ratio	ArtiRate	PauseProp	TurnsOpp*	WordsOpp *
Fluency Score	*r*	0.397	-0.195	0.215	-0.290	-0.374	0.596
	p	0.002**	0.087	0.067	0.021*	0.004**	0.000***
SRpermin	*r*		-0.536	0.540	-0.723	-0.445	0.413
	p		0.000***	0.000***	0.000***	0.001**	0.001**
MazeSnd Ratio	*r*	-0.536		-0.359	0.354	0.625	-0.230
	p	0.000***		0.005**	0.006**	0.000***	0.054
ArtiRate	*r*	0.540	-0.359		0.182	-0.251	0.308
	p	0.000***	0.005**		0.103	0.039*	0.015*
PauseProp	*r*	-0.723	0.354	0.182		0.328	-0.223
	p	0.000***	0.006**	0.103		0.010*	0.060
TurnsOpp*	*r*	-0.445	0.625	-0.251	0.328		-0.459
	p	0.001**	0.000***	0.039*	0.010*		0.000***
WordsOpp*	r	0.413	-0.230	0.308	-0.223	-0.459	
	p	0.001**	0.054	0.015*	0.060	0.000***	

Administration 3

N = 50		SRpermin	MazeSnd Ratio	ArtiRate	PauseProp	TurnsOpp*	WordsOpp *
Fluency Score	*r*	0.563	-0.244	0.270	-0.368	-0.436	0.625
	p	0.000***	0.044*	0.029*	0.004**	0.001**	0.000***
SRpermin	*r*		-0.159	0.445	-0.692	-0.397	0.482
	p		0.136	0.001**	0.000***	0.002**	0.000***
MazeSnd Ratio	*r*	-0.159		-0.035	0.134	0.287	0.037
	p	0.136		0.403	0.176	0.022*	0.399
ArtiRate	*r*	0.445	-0.035		0.334	-0.020	0.361
	p	0.001**	0.403		0.009**	0.446	0.005**
PauseProp	*r*	-0.692	0.134	0.334		0.407	-0.204
	p	0.000***	0.176	0.009**		0.004**	0.047*
TurnsOpp*	*r*	-0.397	0.287	-0.020	0.407		-0.397
	p	0.002**	0.022*	0.446	0.004**		0.002**
WordsOpp*	r	0.482	0.037	0.361	-0.204	-0.397	
	p	0.000***	0.399	0.005**	0.047*	0.002**	

Table 77: Grammar correlations in administrations 1-3

Administration 1							
N = 52		EFOpp	EFProp	WordsClaus	Words ASUnit	TurnsOpp	WordsOpp
Grammar Score	*r*	0.714	0.300	0.249	0.232	-0.513	0.694
	p	0.000***	0.015*	0.037*	0.049*	0.000***	0.000***
EFOpp	*r*		0.559	-0.118	0.054	-0.695	0.768
	p		0.000***	0.203	0.351	0.000***	0.000***
EFProp	*r*	0.559		-0.379	-0.215	-0.355	0.019
	p	0.000***		0.008**	0.063	0.000***	0.446
Words Clause	*r*	-0.118	-0.379		0.694	0.102	0.387
	p	0.203	0.008**		0.000***	0.237	0.002**
Words AS-unit	*r*	0.054	-0.215	0.694		0.202	0.398
	p	0.351	0.063	0.000***		0.075	0.002**
TurnsOpp	*r*	-0.695	-0.355	0.102	0.202		-0.568
	p	0.000***	0.000***	0.237	0.075		0.000***
WordsOpp	r	0.768	0.019	0.387	0.398	-0.568	
	p	0.000***	0.446	0.002**	0.002**	0.000***	

Administration 2							
N = 50		EFOpp	EFProp	WordsClaus	Words ASUnit	TurnsOpp	WordsOpp
Grammar Score	*r*	0.639	0.497	0.364	0.329	-0.176	0.551
	p	0.000***	0.000***	0.005**	0.010*	0.110	0.000***
EFOpp	*r*		0.559	0.077	0.068	-0.461	0.845
	p		0.000***	0.297	0.318	0.000***	0.000***
EFProp	*r*	0.559		-0.201	-0.036	-0.163	0.108
	p	0.000***		0.081	0.403	0.129	0.228
Words Clause	*r*	0.077	-0.201		0.734	0.295	0.425
	p	0.297	0.081		0.000***	0.019*	0.001**
Words AS-unit	*r*	0.068	-0.036	0.734		0.519	0.285
	p	0.318	0.403	0.000***		0.000***	0.023*
TurnsOpp	*r*	-0.461	-0.163	0.295	0.519		-0.396
	p	0.000***	0.129	0.019*	0.000***		0.002**
WordsOpp	r	0.845	0.108	0.425	0.285	-0.396	
	p	0.000***	0.228	0.001**	0.023*	0.002**	

Continued on next page

Administration 3

N = 53		EFOpp	EFProp	WordsClaus	Words ASUnit	TurnsOpp	WordsOpp
Grammar Score	*r*	0.651	0.516	0.059	-0.060	-0.279	0.575
	p	0.000***	0.000***	0.337	0.335	0.021*	0.000***
EFOpp	*r*		0.485	-0.016	-0.016	-0.510	0.925
	p		0.000***	0.453	0.455	0.000***	0.000***
EFProp	*r*	0.485		-0.001	-0.193	-0.301	0.266
	p	0.000***		0.496	0.083	0.014*	0.027*
Words Clause	*r*	-0.016	-0.001		0.741	0.269	0.249
	p	0.453	0.496		0.000***	0.026*	0.036*
Words AS-unit	*r*	-0.016	-0.193	0.741		0.533	0.232
	p	0.455	0.083	0.000***		0.000***	0.047*
TurnsOpp	*r*	-0.510	-0.301	0.269	0.533		-0.393
	p	0.000***	0.014*	0.026*	0.000***		0.002**
WordsOpp	r	0.925	0.266	0.249	0.232	-0.393	
	p	0.000***	0.027*	0.036*	0.047*	0.002**	

Table 78: Vocabulary correlations in administrations 1-3

Administration 1

N = 49		WordsOpp	TurnsOpp	WordsTurn
Vocabulary Score	*r*	0.612	-0.285	0.402
	p	0.000***	0.024*	0.002**
Words Opp	*r*		-0.477	0.627
	p		0.000***	0.000***
Turns Opp	*r*	-0.477		0.314
	p	0.000***		0.014*
Words Turn	r	0.627	0.314	
	p	0.000***	0.014*	

Administration 2

N = 51		WordsOpp	TurnsOpp	WordsTurn
Vocabulary Score	*r*	0.602	-0.435	-0.033
	p	0.000***	0.001**	0.410
Words Opp	*r*		-0.428	0.219
	p		0.001**	0.061
Turns Opp	*r*	-0.428		0.690
	p	0.001**		0.000***
Words Turn	r	0.219	0.690	
	p	0.061	0.000***	

Continued on next page

Administration 3				
N = 51		WordsOpp	TurnsOpp	WordsTurn
Vocabulary	*r*	0.607	-0.327	0.040
Score	p	0.000***	0.010*	0.390
Words	*r*		-0.313	0.302
Opp	p		0.002**	0.059
Turns	*r*	-0.313		0.749
Opp	p	0.002**		0.000***
Words	r	0.302	0.749	
Turn	p	0.059	0.000***	

Table 79: Communicative skills correlations in administrations 1-3

Administration 1							
N = 52		Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
Comm. Skills	*r*	0.641	0.542	0.151	0.153	-0.511	0.766
Score	p	0.000***	0.000***	0.143	0.139	0.000***	0.000***
Initiating	*r*		0.344	-0.066	0.318	-0.603	0.663
features	p		0.006**	0.321	0.011*	0.000***	0.000***
Developing	*r*	0.344		0.073	0.118	-0.348	0.791
features	p	0.006**		0.305	0.203	0.006**	0.000***
Responding	*r*	-0.066	0.073		-0.014	-0.577	0.088
features	p	0.321	0.305		0.461	0.000***	0.268
Collaborating	*r*	0.318	0.118	-0.014		-0.372	0.248
Features	p	0.011*	0.203	0.461		0.003**	0.038*
Turns	*r*	-0.603	-0.348	-0.577	-0.372		-0.568
Opp	p	0.000***	0.006**	0.000***	0.003**		0.000***
Words	r	0.663	0.791	0.088	0.248	-0.568	
Opp	p	0.000***	0.000***	0.268	0.038*	0.000***	

Continued on next page

Administration 2

N = 49		Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
Comm. Skills Score	*r*	0.455	0.575	0.228	0.192	0.651	-0.337
	p	0.001**	0.000***	0.058	0.093	0.000***	0.009**
Initiating features	*r*		0.454	0.366	0.407	0.674	-0.613
	p		0.001**	0.005**	0.002**	0.000***	0.000***
Developing features	*r*	0.454		0.217	0.212	0.877	-0.196
	p	0.001**		0.068	0.072	0.000***	0.088
Responding features	*r*	0.366	0.217		0.512	0.364	-0.636
	p	0.005**	0.068		0.000***	0.005**	0.000***
Collaborating Features	*r*	0.407	0.212	0.512		0.389	-0.488
	p	0.002**	0.072	0.000***		0.003**	0.000***
Turns Opp	*r*	0.674	0.877	0.364	0.389		-0.409
	p	0.000***	0.000***	0.005**	0.003**		0.002**
Words Opp	r	-0.613	-0.196	-0.636	-0.488	-0.409	
	p	0.000***	0.088	0.000***	0.000***	0.002**	

Administration 3

N=51		Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
Comm. Skills Score	*r*	0.592	0.379	0.445	0.350	-0.573	0.614
	p	0.000***	0.003**	0.001**	0.006**	0.000***	0.000***
Initiating features	*r*		0.333	0.323	0.365	-0.561	0.586
	p		0.008**	0.010**	0.004**	0.000***	0.000***
Developing features	*r*	0.333		0.185	0.128	-0.157	0.872
	p	0.008**		0.097	0.186	0.135	0.000***
Responding features	*r*	0.323	0.185		0.392	-0.685	0.290
	p	0.010**	0.097		0.002**	0.000***	0.019*
Collaborating Features	*r*	0.365	0.128	0.392		-0.360	0.252
	p	0.004**	0.186	0.002**		0.005**	0.037*
Turns Opp	*r*	-0.561	-0.157	-0.685	-0.360		-0.348
	p	0.000***	0.135	0.000***	0.005**		0.006**
Words Opp	r	0.586	0.872	0.290	0.252	-0.348	
	p	0.000***	0.000***	0.019*	0.037*	0.006**	

Table 80: Results of the sequential MR for fluency in administrations 1-3

Administration 1

Model	Total R2	△R2	SRpermin *B*	MazeSnd Ratio *B*	ArtiRate *B*	PauseProp *B*	TurnsOpp* *B*	WordsOpp* *B*
1	0.485	0.485***	0.024*** (0.02,0.03)					
2	0.496	0.011	0.022*** (-0.01,0.03)	-0.115 (-2.39,0.83)				
3	0.496	0.000	0.022*** (-0.01,0.03)	-0.791 (-2.44,0.86)	-0.012 (-0.36,0.34)			
4	0.527	0.031	0.070* (0.01,0.013)	-1.062 (-2.71,0.59)	-1.517 (-3.34,0.34)	7.737 (-1.65,17.12)		
5	0.527	0.000	0.069* (0.01,0.13)	-1.055 (-2.89,0.78)	-1.514 (-3.43,0.40)	7.719 (-1.99,17.43)	0.054 (-5.86,5.97)	
6	0.623	0.096**	0.049 (-0.01,0.11)	-1.091 (-2.75,0.57)	-1.106 (-2.85,0.64)	5.605 (-3.27,14.48)	-3.665 (-9.50,2.17)	0.848** (0.31,1.38)

Administration 2

Model	Total R2	△R2	SRpermin *B*	MazeSnd Ratio *B*	ArtiRate *B*	PauseProp *B*	TurnsOpp* *B*	WordsOpp* *B*
1	0.157	0.157**	0.397** (0.00,0.01)					
2	0.158	0.000	0.009* (0.00,0.02)	0.106 (-1.30,1.51)				
3	0.158	0.000	0.009* (0.00,0.02)	0.107 (-1.32,1.54)	0.004 (-0.39,0.39)			
4	0.159	0.001	0.001 (-0.06,0.06)	0.150 (-1.33,1.63)	0.296 (-2.01,2.60)	-1.159 (-10.17,7.85)		
5	0.231	0.072*	0.001 (-0.05,0.06)	0.979 (-0.68,2.63)	0.258 (-1.97,2.49)	-0.943 (-9.67,7.78)	-0.017* (-0.03,0.00)	
6	0.397	0.166**	-0.008 (-0.06,0.04)	0.471 (-1.04,1.98)	0.444 (-1.56,2.45)	-2.004 (-9.85,5.84)	-0.006 (-0.02,0.01)	0.386** (0.16,0.61)

Administration 3

Model	Total R2	△R2	SRpermin *B*	MazeSnd Ratio *B*	ArtiRate *B*	PauseProp *B*	TurnsOpp* *B*	WordsOpp* *B*
1	0.317	0.317***	0.011*** (0.01,0.015)					
2	0.342	0.025	0.010*** (0.01,0.02)	-1.133 (-2.85,0.59)				
3	0.343	0.001	0.010*** (0.01,0.02)	-1.142 (-2.88,0.60)	0.030 (-0.22,0.28)			
4	0.347	0.004	0.024 (-0.03,0.08)	-1.124 (-2.88,0.63)	-0.495 (-2.53,1.54)	2.363 (-6.71,11.44)		
5	0.393	0.046	0.026 (-0.03,0.08)	-0.727 (-2.50,1.04)	-0.609 (-2.60,1.38)	3.070 (-5.82,11.96)	-0.009 (-0.02,0.00)	
6	0.531	0.138**	0.007 (-0.04,0.06)	-1.401 (-3.02,0.22)	-0.077 (-1.87,1.72)	0.177 (-7.90,8.26)	-0.003 (-0.01,0.01)	0.294** (0.13,0.46)

Table 81: Results of the sequential MR for Grammar in administrations 1-3

Administration 1

Model	Total R2	△R2	EFOpp	EFProp	WordsClaus	Words ASUnit	TurnsOpp	WordsOpp
1	0.510	0.510***	5.048*** (3.64,6.45)					
2	0.525	0.015	5.623*** (3.94,7.31)	-0.390 (-1.03,0.25)				
3	0.623	0.098**	5.292*** (3.76,6.82)	0.026 (-0.60,0.65)	0.234** (0.10,0.37)			
4	0.627	0.004	5.390*** (3.82,6.96)	0.017 (-0.61,0.64)	0.276** (0.10,0.46)	-0.029 (-0.11,0.06)		
5	0.627	0.000	5.347*** (3.09,7.60)	0.019 (-0.62,0.66)	0.275** (0.08,0.47)	-0.028 (-0.12,0.07)	-0.001 (-0.02,0.02	
6	0.629	0.003	6.817* (1.19,12.49)	-0.224 (-1.29,0.84)	0.320* (0.07,0.57)	-0.024 (-0.12,0.07)	-0.002 (-0.02,0.02)	-0.296 (-1.33,0.74)

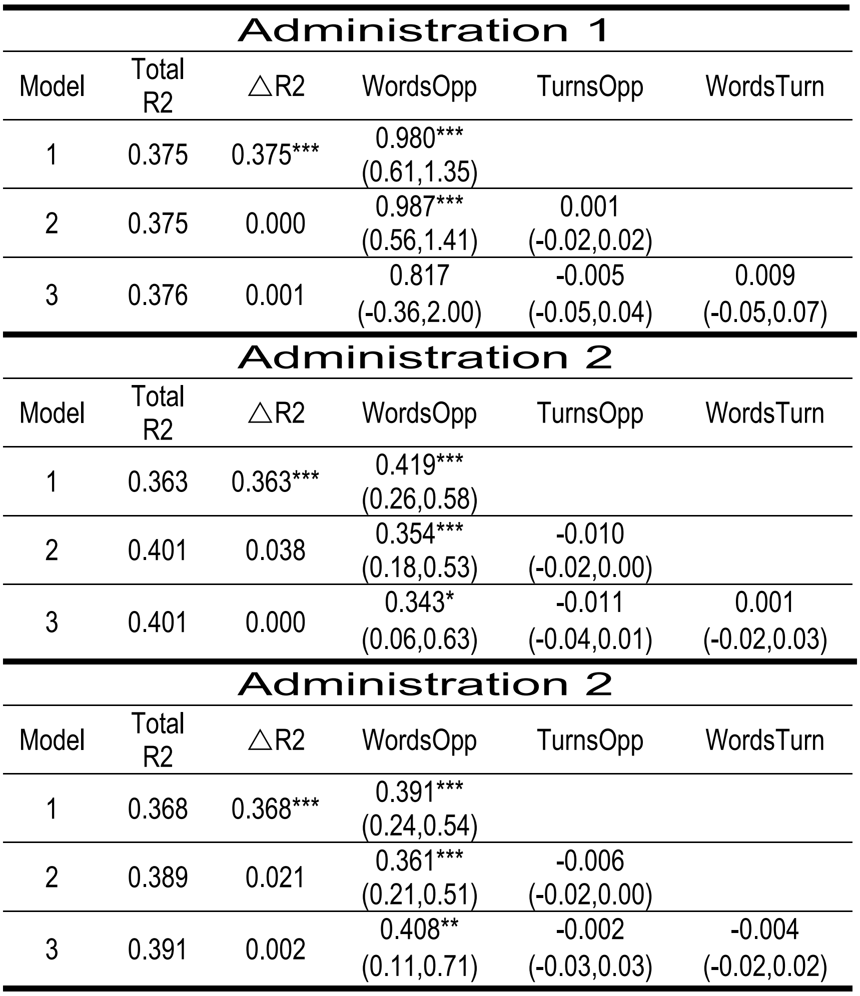
Administration 2

Model	Total R2	△R2	EFOpp	EFProp	WordsClaus	Words ASUnit	TurnsOpp	WordsOpp
1	0.408	0.408***	3.293*** (2.14,4.44)					
2	0.437	0.029	2.706*** (1.34,4.08)	0.674 (-0.21,1.55)				
3	0.583	0.146***	2.135** (0.91,3.36)	1.145** (0.35,1.95	0.265*** (0.13,0.40)			
4	0.584	0.001	2.153** (0.91,3.40)	1.122** (0.30,1.95)	0.244* (0.04,0.44)	0.012 (-0.07,0.09)		
5	0.592	0.008	1.754* (0.22,3.29)	1.197** (0.35,2.04)	0.236* (0.04,0.44)	0.034 (-0.06,0.13)	-0.006 (-0.02,0.01)	
6	0.599	0.007	4.439 (-1.83,10.71)	0.469 (-1.39,2.32)	0.314* (0.05,0.58)	0.048 (-0.05,0.15)	-0.009 (-0.02,0.01)	-0.429 (-1.40,0.54)

Administration 3

Model	Total R2	△R2	EFOpp	EFProp	WordsClaus	Words ASUnit	TurnsOpp	WordsOpp
1	0.424	0.424***	2.021*** (1.36,2.68)					
2	0.477	0.052*	1.627*** (0.90,2.36)	0.740* (0.08,1.40)				
3	0.481	0.005	1.631*** (0.90,2.37)	0.738* (0.07,1.41)	0.042 (-0.09,0.17)			
4	0.488	0.007	1.678*** (0.93,2.43)	0.648 (-0.06,1.36)	0.101 (-0.10,0.30)	-0.028 (-0.10,0.04)		
5	0.507	0.019	2.052*** (1.13,2.98)	0.573 (-0.14,1.29)	0.144 (-0.06,0.35)	-0.066 (-0.16,0.03)	0.008 (0.00,0.02)	
6	0.508	0.001	1.572 (-2.5,5.64)	0.673 (-0.42,1.77)	0.121 (-0.16,0.40)	-0.067 (-0.16,0.03)	0.008 (0.00,0.02)	0.091 (-0.66,0.85)

Table 82: Results of the sequential MR for vocabulary



Administration 1					
Model	Total R2	△R2	WordsOpp	TurnsOpp	WordsTurn
1	0.375	0.375***	0.980*** (0.61,1.35)		
2	0.375	0.000	0.987*** (0.56,1.41)	0.001 (-0.02,0.02)	
3	0.376	0.001	0.817 (-0.36,2.00)	-0.005 (-0.05,0.04)	0.009 (-0.05,0.07)
Administration 2					
Model	Total R2	△R2	WordsOpp	TurnsOpp	WordsTurn
1	0.363	0.363***	0.419*** (0.26,0.58)		
2	0.401	0.038	0.354*** (0.18,0.53)	-0.010 (-0.02,0.00)	
3	0.401	0.000	0.343* (0.06,0.63)	-0.011 (-0.04,0.01)	0.001 (-0.02,0.03)
Administration 2					
Model	Total R2	△R2	WordsOpp	TurnsOpp	WordsTurn
1	0.368	0.368***	0.391*** (0.24,0.54)		
2	0.389	0.021	0.361*** (0.21,0.51)	-0.006 (-0.02,0.00)	
3	0.391	0.002	0.408** (0.11,0.71)	-0.002 (-0.03,0.03)	-0.004 (-0.02,0.02)

Table 83: Results of the sequential MR for Communicative skills

Administration 1

Model	TotalR2	△R2	Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
1	0.411	0.411***	14.300*** (9.44,19.16)					
2	0.528	0.117**	11.506*** (6.82,16.19)	7.002** (2.96,11.04)				
3	0.553	0.025	11.875*** (7.25,16.50)	6.667** (2.67,10.66)	4.865 (-1.05,10.78)			
4	0.557	0.003	12.301*** (7.42,17.18)	6.678** (2.65,10.70)	4.876 (-1.08,10.83)	-2.992 (-13.16,7.18)		
5	0.558	0.001	13.075** (5.87,20.29)	6.789** (2.65,10.92)	6.041 (-3.90,15.98)	-2.383 (-13.46,8.70)	0.005 (-0.03,0.04)	
6	0.644	0.087**	7.657* (0.33,14.98)	-1.204 (-7.35,4.94)	6.342 (-2.67,15.36)	-2.651 (-12.71,7.40)	0.013 (-0.02,0.05)	1.171** (0.46,1.88)

Administration 2

Model	TotalR2	△R2	Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
1	0.207	0.207**	7.212** (3.07,11.35)					
2	0.378	0.171**	3.878 (-0.29,8.04)	6.040** (2.62,9.46)				
3	0.379	0.002	3.643 (-0.78,8.07)	6.004** (2.54,9.47)	0.702 (-3.36,4.77)			
4	0.380	0.001	3.767 (-0.85,8.38)	6.007** (2.50,9.51)	0.912 (-3.63,5.46)	-1.421 (-14.53,11.69)		
5	0.396	0.016	2.306 (-3.09,7.70)	6.350** (2.79,9.92)	-0.466 (-5.72,4.79)	-2.287 (-15.49,10.92)	-0.010 (-0.03,0.01)	
6	0.443	0.047	-0.480 (-6.51,5.55)	0.522 (-6.60,7.65)	-0.922 (-6.05,4.21)	-4.746 (-17.85,8.36)	-0.010 (-0.03,0.01)	0.582 (-0.04,1.20)

Administration 3

Model	TotalR2	△R2	Initiating F.	Develop F.	Respond. F.	Collab. F.	Turns Opp	Words Opp
1	0.351	0.351***	8.773*** (5.35,12.20)					
2	0.388	0.037	7.765*** (4.20,11.33)	2.067 (-0.37,4.51)				
3	0.452	0.064*	1.769** (3.04,10.16)	1.165 (-0.51,4.17)	1.563* (0.51,6.80)			
4	0.456	0.004	6.292** (2.57,10.01)	1.846 (-0.52,4.21)	3.339* (0.01,6.67)	3.756 (-8.51,16.02)		
5	0.487	0.031	4.610* (0.42,8.80)	2.081 (-0.26,4.42)	1.164 (-0.05,5.38)	3.763 (-8.29,15.81)	-0.013 (-0.03,0.00)	
6	0.576	0.089**	1.130 (-3.37,5.63)	-4.087 (-8.71,0.54)	1.007 (-2.87,4.89)	2.752 (-8.36,13.86)	-0.012 (-0.03,0.00)	0.680** (0.23,1.13)

Appendix G: Interactive coding and examples

Initiating features

1. Opening question (Q-O) These are questions that either comes as the opening line of the conversation and is related to the topic, or can be identified as a new topic in the conversation.

These correspond to the *open-initiate-demand* moves in Eggins and Slade (1997, p.193). A typical example of a starting conversation can be found in Excerpt 1 below. This was taken from the first turn of the conversation, so it was indisputably the opening question of the discussion.

Excerpt 1: 1-4308-2*

Code	Turn	Person	
Q-O	1	B:	**Have you ever been to a foreign country?**
	2	C:	Yes I have./ I have been England and Korea

More problematic to identify were opening questions that came after the group oral had commenced. Here the concept of the 'bounded' topic (Sacks, 1992) could be employed. According to Sacks (1992), these occur when a new topic is introduced into the conversation that comes after a pause, or by use an introductory phrase such as 'By the way…" or simply by starting a new topic that just occurred to the speaker. Bounded topics bring in new propositional content to the conversation. Additionally, Galaczi's (2008) study was useful here. Since the interactions in her study used images as a prompt, she could operationalize the topic of conversation as "a spate of talk that referred to a specific visual" (p.96). This analysis follows her by defining a sufficient but not necessary condition of 'topic' as a sequence of talk that refers to a specific question from the prompt, with the other conditions being those pointed out by Sacks (1992) as using an introductory phrase, coming after a pause in the conversation or bringing in new prepositional content. An example is given in excerpt 2 below.

Excerpt 2: 2-4309-6

Code	Turn	Person	
	12	A:	Have a baby./Yes
Q-O	13	C:	**So ...[B: um] {so what do you think} ...ah what do you think are the advantages of being single?/** [A: nnn] {I think + being single is uh hmmm how} I don't know + {how can I I how I how can} how do I say English/ ur I think + being

			single is ur {less than due [AB : hnnn] less due} / mm I don't know[1]/ maybe {less than} less due than get married/ mm ...ah ...{I'd like} uh if I'm... single + mm I can spend some money [A: hm m] + what I'd like to do [AB: hmm] / or I'd spend the time [A: hm] + what I want to do [A: hm]
	14	A:	It's free
	15	C:	Yes
	16	ABC:	Huh huh huh huh
			…
Q-O	17	B:	**Hmm.. If you have a girlfriend or boyfriend + [A: mm] so {you want} you want + to get married with her or him**/ [A;nn] {so} But he or she don't want to get married/ [C: hmm] {How do you think} [C: hmm haha] what do you think?
	18	A:	It's difficult problem.

[1] Interjecting AS-unit

In Excerpt two, two Opening questions can be identified using this criteria. The first occurs in line 13, after A accepted a correction, C signals a new topic by means of the 'So', gives the question, then answers it herself. The next participant gives her answer, C agrees and there is a spate of joint laughter. The opening question in line 17 can be identified by a short pause and by introducing new propositional content. The vocalized 'hmm' is another signal that the speaker is ready to speak, but it does not suggest in the same way as 'so' in line 13 that a new topic will be injected into the interaction.

2. ***Transfer Question (Q-T)*** The specific purpose of these is to bring in another test-takers participation, not to set a new topic or bring any content into the interaction, similar to the *eliciting opinion* of Brooks (2009) and the *monitor* move of Eggins and Slade (1997, p.195). In the interactions in this corpus, with non-native speakers of almost certainly lower proficiency than Brooks (2009) or the native speakers of Eggins and Slade (1997) they seem to take on a more deliberate role with a limited range of expressions being employed – almost invariably "How about you?". In the corpus these features could appear as an independent turn, with the purpose of bringing in a participant who had not yet voiced his or her opinion, as in line 13 of Excerpt 3 below.

Excerpt 3: 3-4305II3

Code	Turn	Person	
Q-T	11	D:	I working at Disneyland/ [ABC: ah wow] and {it's} it's so fun./ **How about you?**
	12	A:	My {tim -} part time job is um drug store, + to sell the drugs (B:yeah A:Ahem[?])
Q-T	13	C:	**How about you?** Huh-huh [D: hm]
	14	B:	I work at McDonalds

Often they came at the end of a statement, or answer to a question, as in line 11 of excerpt 3, and so served the additional function of signalling that the speaker's turn had ended.

3. Follow-up Question (Q-Fu) These were questions that developed the conversation stepwise (Sacks, 1992) by seeking further information about a topic that is currently being discussed. These appear in Brooks (2009) as *prompting elaboration* and are close to the *track-probe* move in Eggins and Slade (1997, p.210).

Excerpt 4: 1-4308-10

Code	Turn	Person	
	5	B:	Have you ever + been any.. world
	6	C:	Ah I went to Las Angeles last month.
Q-Fu	7	B:	**{Wh} why?**
	8	C:	Ah it kind of sightseeing um + also homestaying [B: homestay]./ But just one week./ Too short../ Have you ever + been to overseas?

In excerpt 4, the question in line 7is coded as a follow-up question by virtue of B using it to get more information from C about his trip to China. Though it is difficult to be sure from the brief description and example, this seems close to the 'prompting elaboration' feature in Brooks (2009).

4. Opening Opinion (Opin-O) Instead of beginning a conversation about a topic by asking a question, a participant could perform the same function by stating an opinion about it, as in Eggins and Slade's *initiate-information* move (1997, p.193). Similar to opening questions, identifying these were uncontroversial when they occurred as the first line of conversation about a topic, as in turn 1 in Excerpt 5 below.

Excerpt 5: 2-4308-5

Code	Turn	Person	
Opin-O	1	B:	**In my family both of my parents are working {because of [A: yeh] yeah} because of economical crisis.**[C: oh oh]
	2	C:	Ah my family is traditional family/ So father earn money /and mother stayed at home all day.
	3	A:	Yeah/ My family is not traditional family + because my father {work in the} work in the Funibashi city office / and my mother also do so.
	4	D:	So {my family} my mother don't have the job/ and ah she stay the home all day + and do housework/ but I don't think + it traditional one + because ah sometimes my father help her with dishes or something like that.

Opin-O	5	B:	**Ah when I get married + I don't want + to stay home + because I wanna + enjoy my life**./ [A: yh] ah what do you think about it?
	6	A:	Yeah/ I want + to {work} work out + because I want + to be a English teacher/ [B: aha] so I want + to work hard./ [B: okay yeah C:hmmm] How about you?

They require judgment to be exercised when they occur after the opening exchange. Using the same conditions as described in *opening questions*, an *opening opinion* can be seen in line 5 of Excerpt 5 which opens a new topic by virtue of presenting new propositional content and being on a different topic to the previous turns. Lines 1-4 answered the question in the prompt that asked if the test-taker's family could be considered 'traditional'. After the four participants have given their opinion on this topic, participant B states in line 5 that she would not cease working if she got married, which is sufficiently different to qualify as a new topic of conversation. That this is a new topic is supported by B adding a transfer question to give other participants an opportunity to voice their opinions on it.

5. Opinion (Opin) This is a statement that a test-taker makes on the topic that is being discussed. These show initiative because these utterances are volunteered, and not a response to a question being asked. As Eggins and Slade point out for their *respond-develop* moves, they indicate a high degree of acceptance of the current topic as they expand upon it (1997,p.202). Excerpt 6 below shows an example of this.

Excerpt 6: 2-4308-5

Code	Turn	Person	
	1	B:	In my family both of my parents are working {because of [A: yeah] yeah} because of economical crisis.[C: oh oh]
Opin	2	C:	**Ah my family is traditional family**/ so father earn money /and mother stayed at home all day.

In Excerpt 6 person C is adding their opinion to what person B has said about the extent to which their family could be described as traditional. Person B gave no verbal signal that C should be the next person to answer the question, and in the video there is no gesture or body language that anoints person C as the next speaker, thus C takes the initiative by stating his opinion.

6. Controversial Opinion or Question (Opin-con) As for *Opinion*, these are not in response to a question, but make a statement that is controversial, because it contradicts or challenges what somebody else has said, or is different in a surprising way. The most similar move to this in Eggins and Slade (1997) is their *confront-challenge* move. In this corpus this coding was rare, so questions that were asked with the seeming intention of challenging were also included in this category.

Excerpt 7: 1-4308-2

Code	Turn	Person	
	14	D:	Uhh {I think mmm} I think + a group tour is better for me + {because} um because I don't have not so much knowledge of other countries/ So mm group tour is maybe safety
	15	B:	I think so
Opin-con	16	C:	But I think + uh myself is best / Because if I travel with friends {I mu} + I must be depend on friends

In Excerpt 7 two of the participants have agreed that traveling in a group is best, and then C puts forward an opposing opinion, signalled clearly by the contrasting 'But', which is the typical signal that the test-takers used for this kind of move.

7. Suggesting (Opin-sug) These interactional features are used to propose that another participant take some form of action, as perhaps Eggins and Slades' notion of *initiate-offer* encapsulates (1997, p.194). Not surprisingly they are rare in the corpus because the discussion prompts did not allow much scope for their use. Excerpt 8 is an example.

Excerpt 8: 1-4307-5

Code	Turn	Person	
	59	B:	{What} what kind of movie?
	60	A:	Ah it's a Korean soldier [B: yes C: ahum]and North Korean soldier + make friendship/ [C: um] but it is not allowed [B: oh] for each government./[BC um]
	62	D:	You like war movie?
	63	A:	Yeah/ [BCD: hahah] I like action movie. [D: yeah]
	65	B:	I've never watched Korean movie.
Opin-sug	66	D:	**You should go**

Here the prompt on outdoor and indoor activities has wandered onto the topic of movies, and this presents an opportunity for a test-taker to recommend that another participant go and watch a movie that he enjoyed using the *suggest* interactive feature.

8. Opinion – Appraise (Opin-Ap) An opinion that passes judgment on some aspect of performance or level of difficulty was coded as opinion-appraise. To pass judgment in a conversation is to step outside the conversation and assess it on its own terms, and so was thought worthy of an independent code. In excerpt 9, speaker C has the confidence to pass judgment on A's ability to speak in English, despite never having studied abroad.

Excerpt 9: 1-4308-10

Code	Turn	Person	
	41	A:	I have never been abroad.
	42	C:	Never?
	43	A:	Never.
Opin-Ap	44	C:	**English pretty good though**./ [AB ahaha] So you went + to study abroad?

9. Summarizing – (Opin-Sum) In this rare function, the participant synthesizes other speaker's opinions or the current status of conversation. The phrase is often introduced with an initial 'so', as in the excerpt below.

Excerpt 10: 2-4309-12

Code	Turn	Person	
	25	D:	I'm from Chiba
	26	A:	Aahhh how about you
	27	C:	I came from Fukushima/ [[AB: ahh] so it's too far from here/ so I'm living alone/ [A: Ahhh] I don't feel any homesick or kind of that [A: ahh] + 'cos my mom [D: ahh your mum] sometimes calling me/ [A: ahh] and I hate that (A: eh hmmm)
	28	A:	How about you
	29	B:	I'm from Hiroshima
	30	A:	Yeah
Opin-Sum	31	C:	**So it's the farthest**

In excerpt 10 it is important to know that this conversation takes place in Chiba Prefecture, and that Fukushima is not as far away as Hiroshima. Instead of giving new information or moving onto a new topic, speaker C synthesizes information from the other participants to determine whose home town is the most distant, and for this reason it is classified as *opinion-summarize*.

10. Function – Initiate (Fun-I) The utterances that were given this code were not part of the discussion on the prompt, but fulfilled a function such as, most commonly, starting by greeting and self-introducing at the beginning. Those who began such an exchange were credited with the *function-initiate* feature, and those who responded were coded in the responding category as *function-response,* as in Excerpt XX below .

Excerpt 11: 2-4308-2

Code	Turn	Person	
Fun-I	1	B:	**How do you do?** / Nice to meet you,
Fun-R	2	A:	**Nice to meet you**
Fun-R	3	C:	**Nice to meet you**

11. Manage topic – initiate (Man-I) In this category were utterances that had the sole purpose of managing the topic by talking about what question should be answered next. This feature was also found in Brooks (2009), though for the purposes of this study I divide this category into initiating and responding management turns, as can be seen in Excerpt 12.

Excerpt 12: 3-4305-II5

Code	Turn	Person	
Man-I	1	C:	**Mm Can I go first?**
Man-R	2	B:	**Yeah sure**

12. Japanese-initiate (Jpn-I) Although it is an assessment of spoken English, students sometimes resorted to using their native language, and when they did it was often spoken inaudibly. From what little that was audible (if they used Japanese for communicative purposes it was usually whispered), it seemed that most of what they say is about the task itself and what they are supposed to do. Whatever they say, if the participant could be identified as initiating talk in Japanese, they were coded for this interactive feature, as in the example below. This coding was only necessary to be used in the first administration: following that there was no whole turns in Japanese. Although this feature was counted, it was omitted from the statistical analysis since the objective of the test is to measure the candidate's use of English.

Responding features

These features are elicited from test-takers by initiating or repair features, and as such will tend to dominate the features used by more passive participants.

13. Answer (Ans) If somebody asks any question, the initial AS-unit of the response will be classified as an answer. In Eggins and Slade (1997) this is the main category of respond. Though their analysis breaks this down into many smaller categories, there was no need for such a finely detailed breakdown for the purposes of this dissertation. Excerpt 10 shows an example of an *answer* in line 28 that is a response to a *transfer question* that B has asked. As it might be expected, this was one of the most often used codes in the corpus.

Excerpt 14: 2-4307-5

Code	Turn	Person	
	26	B:	{What do} what do you do?
Ans	28	A:	**Yn if I go outside + I go to Starbucks (C: Starbucks) + and drink coffee.**
	29	D:	It's kind of inside
Ag	30	A:	**Yeah**

14. Agreement (Ag) Quite common in the corpus are turns in which a test-taker adds agreement to a statement, even if not directly asked. This is classified as a responding feature because it adds no new proposition to the discussion, and usually means the turn passes to a different participant. It is more than a backchannel since it represents a positive response to the proposition, rather than be merely a signal that the speaker is listening. In Eggins and Slade, this is *reply-agree*, which is a subcategory of respond.

In Excerpt 14, the agreement in line 30 is in response to the prompt on outdoors and indoors activities in the second administration. Here, A acquiesces to D's categorization of the activity of drinking coffee as an inside activity, despite proffering it as an example of an outdoor activity.

15. Reacting (React) This interactive function occurs in the initial AS-unit of a response to an initiating move and is typically a short response that repeats elements from the previous speaker's utterance, or shows surprise like an initial 'Really?' before continuing a response, and has similarities to Eggins and Slades' *respond-register* category, except in this corpus often the speaker who uttered it continued speaking. Like *Agreement*, this interactive function runs close to being considered backchannel. Two differences distinguish its interactive function: firstly, while backchannel is not responded to, a reaction might be acknowledged by another participant; and secondly, it has the potential of being expanded on if the speaker chooses to. Excerpt 15 shows an example of this happening.

Excerpt 15: 2-4308-10

Code	Turn	Person	
58		B:	Do you think + mobile phone is kinda troublesome {at} in the train?
59	React	A:	**In the train yeah**/ some guys mobile phone is ringing in trains/ that's very annoying,/[B: nn] I hate it./ When somebody call me + when I'm on train, + I don't like that.

The *reacting* move is the first AS-unit in line 59. As well as showing agreement with the previous speaker, this move allows test-taker A more time to compose his reply.

16. Function – response (Fun-R) These were the responses to the Function-initiate category, as seen in Excerpt 11 above. As well as responses to greetings, also included in this feature were apologies and any other turns in the exchange that were uttered when, for example, two participants tried to speak at once.

Excerpt 16: 3-4305-II3

Code	Turn	Person		
	1	B:	Do you + have a part time job?	<simultaneous speech>
	2	C:	Do you…	
Fun-I	3	B:	**I'm sorry**	
Fun-R	4	C:	**It's okay**	

17. Manage topic – response (Man-R) As described above, these were in responses to *manage topic-initiate*, as described and exemplified in excerpt 12 above.

18. Japanese – response (Jpn-R) This is the responding counterpart to ***Japanese–initiate.*** Additionally, if a test-taker asked answered a question in Japanese that was asked in English, it was also counted as ***Japanese-response,*** as even though they commenced the use of Japanese, it was in response to the question they were asked.

Excerpt 17: 1-4308-4

Code	Turn	Person	
	5	B:	Mm I want + to go to South Africa + because ... {in} in our high school mm {I had} I had opportuni ties + to take part in the {volun teer} volunteer activities and student activities./ {I} I have been interested in NGO / so I wanna + {go} go there / and you?
Jpn-R	6	C:	Uh hh **<Japanese whispering to D>**
	7	D:	Oh… ok uh../ What do you think + are the good and bad points of + going to … that kind of place./ { What what what} do you think?
	8	B:	Uhh in South Africa there are many people + who {who} live under bad condition/ and ...{annn ...} if I go there + {?m} .. there are a lot of desires like war ...(D: yes)/ so mm I think + that it is a bad point/ {.hmm bad point...} and you?/

In this example, it seems that person C has told D in Japanese that she does not want to respond to the question, so D has to restart the conversation with an opening question. As for its counterpart in the initiating category, it was not counted in the statistical analysis.

Developing features

The initial AS-unit of an utterance may be coded as an *opening opinion*, *answer*, or a *question* (amongst others), but the speaker may decide not to relinquish the turn and instead continue talking. These subsequent AS-units may be used to expand the reply in some way. The AS-units that do this were counted as 'developing features'. Within these developing features, various categories could be distinguished, and they are described below.

19. Develop (Dev) This is the default code for any continuation of an initial AS-unit. These most often came following *answer* or *opinion* moves, which are the equivalent of Slade and Eggin's *prolong* moves (1997, p.196), as in the example below.

Excerpt 18: 1-4308-2

Code	Turn	Person	
	1	B:	Have you ever been to a foreign country?
Dev	2	C:	Yes I have./ **I have been England and Korea**

In Excerpt 18, the first AS-unit in the response to B's question is classified as an answer, and although it would have been possible to finish the turn at that point, C chose to continue by mentioning which countries she had been to in the second AS-unit, making this a *develop* feature.

This feature was also used to ensure that interactive features were not counted twice. For example, sometimes in the corpus a test-taker would expand on a question, as in Excerpt 19 below.

Excerpt 19: 2-4308-7

Code	Turn	Person	
Q-T Dev	32	C:	How about you,/ {you} your parents um works or anything + {yo-}you were little children?

The topic of conversation previous to this question was whether the parents worked or whether one of them stayed at home and looked after them as children, and B has not given an opinion on the topic. In Excerpt 13 participant C asks a transfer question to find out B's opinion, but repeats the issue being discussed as a question. It really is only one question, not two questions, and by treating the first AS-unit as *transfer question* and the second as *development*, is credited to the speaker as a single question.

20. Reference (Ref) On occasion within a multi-AS-unit turn, a test-taker would refer to something that another participant had said, and if they did so it was classified as reference. An example can be found in Excerpt 20 below.

Excerpt 20: 3-4309-III1

Code	Turn	Person	
	26	C:	But mm I like shopping / but um I can't go because of too much homework/ (B: yeah) **and as Momoko said + um I don't have much money** (A: yeah

B: hmm)

Referring to another person’s speech adds cohesion to the conversation and has the effect of personalizing it. It was not considered a Collaborating feature since these were usually built into longer turns and did not invite a response from another participant.

It appears to be a more advanced function as it was very rare in the first administration , and although it was used more often in the second and third administrations it was still not a common feature in the corpus. This is the *referring to partner’s ideas* feature of Brooks (2009). However, her classification system also had a feature called *paraphrasing* which seems to be similar in that they both refer directly to another’s words. Due to this and their rarity in this corpus, the *reference* function here incorporates them both.

21. Finish-summary (Fini-S) Upon examining the transcripts it was noted that when some participants began talking on a topic, they often felt the need to provide a signal to their fellow test-takers that they had finished speaking, and that somebody else should speak. One of the ways they appeared to be doing this was by the use of ‘so’. Although in general, these Japanese test-takers made heavy use of this connector, in particular they seemed to be using it as a final statement to indicate that their turn had finished. The participant in Excerpt 22 is a clear example of how this was typically done.

Excerpt 21: 2-4308-5

Code	Turn	Person	
	12	A:	{me too but} me too/ and {a} sometimes I missed my mother/ and ah when
Fini-S			I got home + {nobody} nobody didn't welcome**,/ so I miss my family.**

Here the test-taker starts by *agreeing* with the previous speaker, then *develops* her position before rounding it off with a ‘so’ statement to summarize and indicate turn completion.

22. Finish-yeah (Fini-Y) When students felt the need to signal the end of their turn, they could use a final single word or short phrase, in most cases ‘yeah’, as can be seen in Excerpt 23 below. Another typical finishing statement was “yeah, I think so”. Here ‘so’ is part of the phrase, and so such statements were not categorized as *finish-trail* (see below). As long as it was a deliberately used word or short phrase that did not add or summarize content, it would count as *finish-yeah.*

Excerpt 23: 1-4310-8

Code	Turn	Person	
Fini-Y	58	D:	I want + to go to Mexico too + because um I want + to study Spanish also/ Spanish in this college, **/ yeah /**

23. Finish-trail (Fini-T) Instead of finishing in a definitive way as for *finish-yeah* and *finish-summary*, the speaker could finish by trailing off, as in excerpt 24 below.

Excerpt 24: 3-4306III3

Code	Turn	Person	
Fini-T	26	B:	So I can't go shopping/ (AB haha) yeah but now I want+ to buy my suitcase/ [A: ohh] **I don't have it so.. um**

Often a conjunction was used with the speaker elongating vowel sound, followed by a non-word vocalization. Sometimes though, it was done in the middle of a sentence and was clearly a result of the participant searching for the right word. The pause gave other participants the opportunity to collaborate by suggesting a word in a repair move (see below), as in excerpt 25 below.

Excerpt 25: 2-4308-2

Code	Turn	Person	
Fini-T	16	B:	It is good + when I get home + {someone} (A: yeah C: hmmm) someone..
	17	C:	waiting for you
	18	B:	yeah [C: hmm]

Collaborating features

The Collaborating features are those in which a participant interacts more directly with the meaning of what was said by another participant, as happens in negotiation or repair. Repair features take place when there is some uncertainty over the meaning, brought about because one of the participants does not give full information, makes a mistake, or when a correction or suggestion for improved vocabulary, pronunciation or grammar is given. Sometimes they are brought about by the speaker's uncertainty, other times by the listener asking a question to clarify or confirm the speaker's meaning. Other forms of collaboration occur when there is not a breakdown, but another participant involves him or herself in what the other person said anyway, perhaps in order to invoke humour.

24. Question-clarify Q-Cl These are questions that aim to clarify the meaning of what the previous speaker had just said, similar to Eggins and Slades *rejoinder-clarify* (1997, p.210) though here they were perhaps used in a narrower sense, closer to Brooks (2009*) requesting clarification.*

Excerpt 26: 2-4308-4

Code	Turn	Person	
	32	B:	So, do you want to have a mobile phone which have so many functions in your future?
Q-Cl	33	A:	**future**?
	34	B:	Now [A: nn] you have not
	35	A:	Maybe no.

In this conversation in turn 32, person B starts a new topic about mobile phones, but A is not quite sure about what B means when he says 'future' and so turns it into a clarification question. B's explanation is clear and allows A to give an answer.

25. Question-confirmation Q-con These are questions that have the effect of confirming meaning. The answer to them is nearly always 'yes', and the question also often includes elements from the previous utterance it is asking about. These are not necessarily asked as a result of a breakdown but may function to forestall a breakdown, express surprise and they often allow the asker extra time to compose an answer. These are considered questions and not backchannel since they evoke a direct response, while backchannel does not.

Excerpt 27: 3-4305II3

Code	Turn	Person	
	33	B:	Shopping,/ maybe same things / but er I have a car
Q-con	34	D:	**You have a car?**
	35	B:	Yeah/ so {I have} I have to go + you know gas(A: nn D: un)

This discussion is on shopping, and the topic is about how the participants spend their money. Test-taker B also spends her money on shopping, and then admits to having a car, to which D uses a *question-confirmation.*

These appear in Eggins and Slades *rejoinder-confirm* (1997, p.209) and Brooks (2009) *seeking confirmation*. Often there is an element of paraphrasing about them, as a skilled speaker puts the previous speaker's words together in a different way when asking the question, as seen in excerpt XX. This seems like a more advanced way of speaking, but were difficult to differentiate from other *question-confirmation* features.

Excerpt 28: 2-4309-12

Code	Turn	Person	
	17	D:	Do you guys live in alone
	18	A:	Yes
	19	B:	Yes
	20	C:	Yeah
	21	A:	From ah {last}… last month
	22	D:	Last month
	23	A:	Yes
Q-Con	24	C:	**So really recently you came here**
	25	A:	Yes …

26. Correction (Cor) this occurs when one test-taker gives a correction of another's improper use of English. This is the equivalent of Brooks (2009) *other correcting.* In excerpt XX below, participant B corrects A's usage of 'baby' as a verb with the correctly collocated verb 'have'. This was a rare feature in the corpus, possibly due to participants being unwilling to threaten the face of fellow test-takers in this assessment situation, or not being confident of usage.

Excerpt 29: 2-4309-6

Code	Turn	Person	
	10	A:	I want to baby too
Cor	11	B:	**Have a baby**
RespH	12	A:	**Have a baby / yes**

27. Completing Sentences (Compl) When a participant supplies the end of a sentence that another test-taker started it was assigned this code. This is the same as Brooks (2009) *finishing sentences*. In Excerpt XX below, C's pause before finishing her sentence is opportunity enough for A to complete the sentence by suggesting an appropriate ending.

Excerpt 30: 1-4310-3

Code	Turn	Person	
	34	C:	Foreign people are very…
Compl	35	A:	**friendly**
	36	C:	friendly yeah/… my host family brings with me to ah shopping and sea./ It was {very} very beautiful

28. Suggest Words (SW) In the corpus there were some occasions when one test-taker suggests some words to another test-taker even if there was no obvious opportunity such as a speaker leaving a trailing sentence, as occurs in excerpt XX. This is the same as Brooks (2009) *suggesting words*.

Excerpt 31: 3-4305II4

Code	Turn	Person	
	31	B:	In my opinion, um if I want + to buy some clotheses + if {my friend with} I {sh} go to shopping with my friend + I can get some opinion [D: mm] or something [C: yea-yea]
SW	32	D:	**Some kind advice**
	33	B:	Yeah [D: nn]

This also occurred in lines of dialogue in which students collaboratively build meaning by adding examples of words, sometimes as a result of a participant not being able to supply words, but also in a lighter way with the apparent objective of providing amusement by means of shared humour.

29. Incomprehension (Incom) Occasionally test-takers admit that they do not know how to answer or what to say, inviting others to supply what was missing or explain, or they openly state they do not understand what another person said, as happens in excerpt 32 below. This replicates Brooks (2009) feature, *expressing incomprehension*.

Excerpt 32: 1-4309-04

Code	Turn	Person	
	26	D:	Parents is not
	27	C:	Not?
Incomp	29	A:	Not?/ **What do you mean?/** Do you mean + do we + prefer friends or family?

In this, the test-takers have been discussing who they like to travel with, friends or parents, and D offers an ambiguous statement which appears to be a complete sentence. Both C and A react to this, A by expressing *incomprehension* as part of her response. This example parallel’s the expressing incomprehension feature in Brooks (2009), however, also included in this category were incidents were the incomprehension was embedded in the turn,

30. Respond to help (RespH) The test-takers who received help or had another participant ask for clarification or suggest words often acknowledged it by repairing, incorporating or thanking the participant. This is similar to Brooks (2009) *responding to help,* but in this corpus a broader definition was used, since it incorporates the response categories of *resolve* and *repair* in Eggins and Slade (1997, p.209), and as such would include the *incorporate words* category of Brooks (2009). Excerpt 29 above shows an example of the speaker repeating the correction that was made to her incorrect usage of *baby* as a verb. If instead of showing signs of uptake they merely acquiesced, then it would be classed as *agreement*, as happens in excerpt 31 line 30, when B's "yeah" merely acknowledges D's suggestions.

Appendix H: Statistics of Test-takers chosen for qualitative analysis

Table 84: Low initial group's words and turns statistics

	Words	Words/ OppSpk	Turns/ OppSpk	Words/ Turn
Administration 1				
Kabuhito	27	0.221	0.025	9.000
Saaya	19	0.178	0.019	9.500
Machiko	43	0.374	0.035	10.750
$\bar{x}$of 53	85.6	0.739	0.069	11.660
std of 53	58.0	0.373	0.036	5.356
Administration 2				
Kabuhito	31	0.241	0.101	2.38
Saaya	37	0.325	0.009	37.00
Machiko	85	0.794	0.047	17.00
$\bar{x}$of 53	144.6	1.133	0.085	17.443
std of 53	82.3	0.598	0.054	13.919
Administration 3				
Kabuhito	70	0.605	0.069	8.750
Saaya	84	0.787	0.047	16.800
Machiko	116	1.098	0.152	7.250
$\bar{x}$of 53	144.6	1.338	0.102	16.688
std of 53	64.2	0.628	0.059	11.726

Table 85: Low initial group's complexity and accuracy statistics

	Admin. 1			Admin. 2			Admin. 3		
	Word/ AS-unit	EFClaus Prop.	EFClaus / Opp.	Word/ AS-unit	EFClaus Prop.	EFClaus Opp.	Word/ AS-unit	EFClaus Prop.	EFClaus Opp.
Kabuhito	4.500	0.500	0.041	2.818	0.538	0.054	4.375	0.765	0.112
Saaya	6.333	0.200	0.009	9.250	0.143	0.009	7.000	0.381	0.075
Machiko	8.600	0.778	0.061	6.071	0.667	0.131	5.524	0.828	0.227
std of 53	1.568	0.204	0.075	1.660	0.153	0.114	1.729	0.136	0.124
$\bar{x}$ of 53	5.943	0.670	0.120	6.619	0.618	0.164	6.759	0.778	0.231

Table 86: Low initial group's fluency statistics

	Administration 1					Administration 2					Administration 3				
	Artic. Rate	Spch Rate	Pause Prop	Maze Ratio	M & S Ratio	Artic. Rate	Spch Rate	Pause Prop	Maze Ratio	M & S Ratio	Artic. Rate	Spch Rate	Pause Prop	Maze Ratio	M & S Ratio
Kabuhito	2.612	29.440	0.812	0.222	0.259	2.098	62.342	0.505	0.387	0.613	2.315	74.887	0.461	0.100	0.157
Saaya	2.785	62.749	0.625	0.211	0.263	3.157	98.801	0.478	0.270	0.486	2.946	80.507	0.545	0.083	0.262
Machiko	2.814	50.871	0.699	0.279	0.302	2.694	88.176	0.454	0.141	0.282	2.285	100.246	0.269	0.086	0.129
std of 53	0.472	20.051	0.119	0.120	0.187	0.396	22.263	0.125	0.091	0.129	0.405	21.279	0.116	0.042	0.057
$\bar{x}$ of 53	2.676	80.235	0.494	0.131	0.213	2.630	104.624	0.333	0.108	0.177	2.831	111.540	0.340	0.067	0.131

Table 87: Low initial group's interactive statistics

	Administration 1				Administration 2				Administration 3			
	Initiate	Respnd	Develop	Collab	Initiate	Respnd	Develop	Collab	Initiate	Respnd	Develop	Collab
Kabuhito	0.000	0.008	0.025	0.016	0.000	0.062	0.008	0.023	0.026	0.043	0.060	0.009
Saaya	0.000	0.019	0.009	0.000	0.009	0.000	0.026	0.000	0.019	0.037	0.056	0.000
Machiko	0.009	0.026	0.009	0.000	0.047	0.019	0.075	0.000	0.076	0.076	0.047	0.000
std of 53	0.029	0.021	0.033	0.013	0.036	0.034	0.045	0.015	0.041	0.038	0.047	0.017
$\bar{x}$ of 53	0.033	0.036	0.050	0.009	0.041	0.050	0.078	0.011	0.053	0.057	0.088	0.011

All figures adjusted for time of test and number of speakers

Table 88: Medium initial group's group's words and turns statistics

	Words	Words/ OppSpk	Turns/ OppSpk	Words/ Turn
Administration 1				
Taeko	61	0.625	0.041	15.250
Aemi	57	0.490	0.060	8.143
$\bar{x}$of 53	85.6	0.739	0.069	11.660
std of 53	58.0	0.373	0.036	5.356
Administration 2				
Taeko	136	0.914	0.047	19.43
Aemi	118	0.831	0.049	16.86
$\bar{x}$of 53	144.6	1.133	0.085	17.443
std of 53	82.3	0.598	0.054	13.919
Administration 3				
Taeko	194	2.042	0.168	12.13
Aemi	119	1.145	0.077	14.88
$\bar{x}$of 53	144.6	1.338	0.102	16.688
std of 53	64.2	0.628	0.059	11.726

Table 89: Medium initial group's complexity and accuracy statistics

	Admin. 1			Admin. 2			Admin. 3		
	Word/ AS-U	EFClaus Prop.	EFClaus/ Opp.	Word/ AS-U	EFClaus Prop.	EFClaus Opp.	Word/ AS-U	EFClaus Prop.	EFClausOpp.
Taeko	7.625	0.929	0.133	8.000	0.750	0.161	5.706	0.930	0.421
Aemi	6.333	0.769	0.086	6.941	0.692	0.127	6.611	0.724	0.202
std of 53	1.568	0.204	0.075	1.660	0.153	0.114	1.729	0.136	0.124
$\bar{x}$ of 53	5.943	0.670	0.120	6.619	0.618	0.164	6.759	0.778	0.231

Table 90: Medium initial group's fluency statistics

	Administration 1					Administration 2					Administration 3				
	Artic.	Spch	Pause	Maze	M & S	Artic.	Spch	Pause	Maze	M & S	Artic.	Spch	Pause	Maze	M & S
	Rate	Rate	Prop	Ratio	Ratio	Rate	Rate	Prop	Ratio	Ratio	Rate	Rate	Prop	Ratio	Ratio
Taeko	2.538	63.699	0.582	0.213	0.311	2.952	93.218	0.474	0.147	0.287	3.139	130.474	0.307	0.098	0.134
Aemi	2.792	85.544	0.489	0.088	0.123	2.769	119.389	0.282	0.093	0.102	2.580	116.006	0.251	0.067	0.109
std of 53	0.472	20.051	0.119	0.120	0.187	0.396	22.263	0.125	0.091	0.129	0.405	21.279	0.116	0.042	0.057
$\bar{x}$ of 53	2.676	80.235	0.494	0.131	0.213	2.630	104.624	0.333	0.108	0.177	2.831	111.540	0.340	0.067	0.131

Table 91: Medium initial group's interactive statistics

	Administration 1				Administration 2				Administration 3			
	Initiate	Respnd	Develop	Collab	Initiate	Respnd	Develop	Collab	Initiate	Respnd	Develop	Collab
Taeko	0.021	0.021	0.041	0.000	0.020	0.040	0.054	0.007	0.063	0.116	0.158	0.032
Aemi	0.034	0.034	0.009	0.000	0.035	0.014	0.063	0.007	0.019	0.048	0.096	0.010
std of 53	0.029	0.021	0.033	0.013	0.036	0.034	0.045	0.015	0.041	0.038	0.047	0.017
$\bar{x}$ of 53	0.033	0.036	0.050	0.009	0.041	0.050	0.078	0.011	0.053	0.057	0.088	0.011

Table 92: High initial group's fluency statistics

	Administration 1					Administration 2					Administration 3				
	Artic.	Spch	Pause	Maze	M & S	Artic.	Spch	Pause	Maze	M & S	Artic.	Spch	Pause	Maze	M & S
	Rate	Rate	Prop	Ratio	Ratio	Rate	Rate	Prop	Ratio	Ratio	Rate	Rate	Prop	Ratio	Ratio
Yamahiko	3.462	122.551	0.410	0.047	0.077	3.354	113.667	0.435	0.031	0.047	2.892	145.812	0.160	0.037	0.058
Hamako	3.000	113.080	0.372	0.172	0.199	2.685	121.309	0.247	0.077	0.088	3.009	137.095	0.241	0.049	0.075
Naeko	2.924	99.440	0.433	0.177	0.292	2.564	121.433	0.211	0.158	0.209	2.995	146.452	0.185	0.078	0.121
std of 53	0.472	20.051	0.119	0.120	0.187	0.396	22.263	0.125	0.091	0.129	0.405	21.279	0.116	0.042	0.057
of 53	2.676	80.235	0.494	0.131	0.213	2.630	104.624	0.333	0.108	0.177	2.831	111.540	0.340	0.067	0.131

Table 93: High initial group's interactive statistics

	Administration 1				Administration 2				Administration 3			
	Initiate	Respnd	Develop	Collab	Initiate	Respnd	Develop	Collab	Initiate	Respnd	Develop	Collab
Yamahiko	0.117	0.033	0.128	0.039	0.159	0.099	0.139	0.027	0.111	0.037	0.046	0.009
Hamako	0.073	0.024	0.081	0.008	0.063	0.099	0.169	0.000	0.151	0.151	0.163	0.113
Naeko	0.062	0.025	0.062	0.000	0.067	0.017	0.042	0.008	0.019	0.047	0.084	0.009
std of 53	0.029	0.021	0.033	0.013	0.036	0.034	0.045	0.015	0.041	0.038	0.047	0.017
of 53	0.033	0.036	0.050	0.009	0.041	0.050	0.078	0.011	0.053	0.057	0.088	0.011

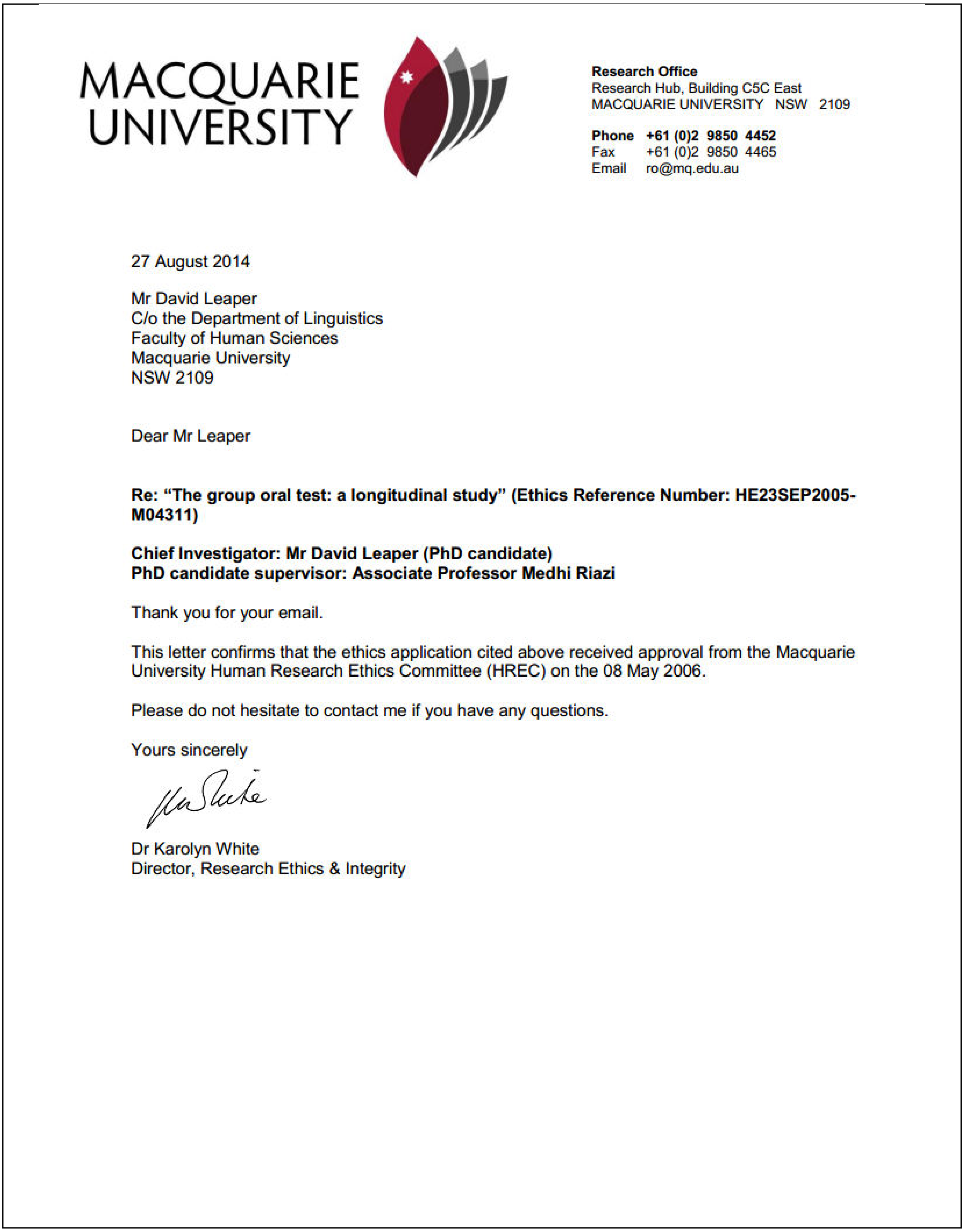
Table 94: High Initial group's words and turns statistics

	Admin. 1				Admin. 2				Admin. 3			
	Words	Words/ OppSpk	Turns/ OppSpk	Words/ Turn	Words	Words/ OppSpk	Turns/ OppSpk	Words/ Turn	Words	Words/ OppSpk	Turns/ OppSpk	Words/ Turn
Yamah'o	337	1.880	0.167	11.233	359	2.379	0.225	10.56	191	1.761	0.129	13.64
Hamako	186	1.508	0.089	16.909	364	2.567	0.113	22.75	268	3.371	0.314	10.72
Naeko	96	1.199	0.050	24.000	139	1.158	0.083	13.90	116	1.087	0.066	16.57
std of 53	58.0	0.373	0.036	5.356	82.3	0.598	0.054	13.919	64.2	0.628	0.059	11.726
of 53	85.6	0.739	0.069	11.660	144.6	1.133	0.085	17.443	144.6	1.338	0.102	16.688

Appendix I: Scoring bands used in the GOT

	Pronunciation Think about: • pronunciation • intonation • word blending	Fluency Think about: • automatization •fillers • speaking speed	Grammar Think about: • use of morphology • complexity of syntax (embedded clauses, parallel structures, connectors)	Vocabulary Think about: • range of vocab	Communicative skills/strategies Think about: • interaction • confidence • conversational awareness
0 .5	Very heavy accent, uses Japanese katakana-like phonology and rhythm; words are not blended together	Fragments of speech that are so halting that conversation is not really possible; nss would not think person had virtually no English	Does not use any discernable grammatical morphology	Shows knowledge of only the simplest words and phrases taught in junior high school or beginning high school	Shows no awareness of other speakers; may speak, but not in a conversation-like way
1.0 1.5	Somewhat katakana-like pronunciation; does not blend words together, they are pronounced in isolation	Slow strained speech, constant groping for words and long unnatural pauses; communication with a ns would be difficult	Doesn't have enough grammar to express an opinion clearly; makes frequent errors; no attempt at complex grammar	Lexis not adequate for task, cannot express opinion properly with the limited words used	Does not initiate interaction, produces monologue only; shows some turn-taking, may say, "i agree with you," but not relate ideas in explanation; too nervous to interact effectively
2.0 2.5	May not have mastered some difficult sounds of English, but would be mostly understandable to a naïve NS; makes some attempts to blend words	Speech is hesitant; some groping for words and unfilled spaces are present but generally don't impede communication completely	Relies mostly on simple (but appropriate) grammar, has enough morphosyntax to express meaning, complex grammar is attempted but may be inaccurate	Generally has enough lexis for expressing some opinion but does not demonstrate any particular knowledge of vocabulary	Responds to others without long pauses to maintain interaction; shows agreement or disagreement to others' opinions
3.0 3.5	Pronunciation is good but has still not mastered the sound system of English; accent does not interfere with comprehension; can blend words	May use some fillers, rarely gropes for words but speech may still not be quick	Shows ability to use some complex grammar, may make errors but they are only in late-acquired grammar	Shows some evidence of some advanced vocabulary	Generally confident, responds appropriately to others opinions, shows ability to negotiate meaning quickly and relatively naturally
4	Speaks with excellent pronunciation and intonation; has practically mastered the sound system of English	Excellent fluency, uses fillers effectively, shows ability to speak quickly in short bursts	Uses both simple and complex grammar effectively; may make occasional errors but they are only in late-acquired grammar	Shows evidence of a wide range of vocabulary knowledge	Confident and natural, asks others to expand on views, shows how own and others' ideas are related, interacts smoothly

Appendix J: Ethics Approval



MACQUARIE UNIVERSITY

Research Office
Research Hub, Building C5C East
MACQUARIE UNIVERSITY NSW 2109

Phone +61 (0)2 9850 4452
Fax +61 (0)2 9850 4465
Email ro@mq.edu.au

27 August 2014

Mr David Leaper
C/o the Department of Linguistics
Faculty of Human Sciences
Macquarie University
NSW 2109

Dear Mr Leaper

Re: "The group oral test: a longitudinal study" (Ethics Reference Number: HE23SEP2005-M04311)

Chief Investigator: Mr David Leaper (PhD candidate)
PhD candidate supervisor: Associate Professor Medhi Riazi

Thank you for your email.

This letter confirms that the ethics application cited above received approval from the Macquarie University Human Research Ethics Committee (HREC) on the 08 May 2006.

Please do not hesitate to contact me if you have any questions.

Yours sincerely

Dr Karolyn White
Director, Research Ethics & Integrity