THE EFFECTS OF NOTICING TRAINING, MODEL INPUT AND TASK REPETITION ON L2 SPEECH PRODUCTION

By

Matthew Campbell

A thesis submitted to Macquarie University in fulfillment of the requirements for the degree of Doctor of Philosophy in Applied Linguistics

Macquarie University

Linguistics Department

October, 2019

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

(Signed) _____ Date 10 Oct 2019

Abstract

Task repetition (TR) has shown to be facilitative of oral fluency for L2 learners. However, the effects of TR on the grammatical accuracy of learners' speech performance remain unclear, with learners often carrying over errors from an initial delivery to the iteration(s). In order to target accuracy, it has been suggested (e.g. Ellis, 2009) that some kind of reflection is required by the learner on their initial performance before engaging in a repeat performance. The primary aim of this study was to examine whether L2 learners could be trained to a) notice linguistic gaps in their initial performance of a speaking task and mine model input to fill those gaps, and b) notice gaps in the way they used language and the way a model speaker used language to complete the same oral narrative task. A further aim was to investigate how noticing training impacts on L2 speech performance as measured by complexity, accuracy, and fluency (CAF).

Thirty-six ESL students took part in one of three groups: Guided Noticing (GN), Unguided Noticing (UN), and Control (C). Participants in all three groups completed a pre-test, a post-test, and a delayed post-test. Each test involved four stages: 1) participants' performance of an oral narrative task based on a picture sequence, 2) a stimulated recall session conducted by the researcher with each participant to identify what gaps they noticed in their interlanguage while performing the task (i.e. inter-language gaps [IL-gaps]), 3) a comparison stage where the participants listened and noted linguistic differences between a recording of their initial performance and a recording of a model speaker performing the same task (i.e. they noted interlanguage – target language gaps [IL-TL gaps]), and 4) a repeat performance by participants of the same narrative task. Stages 1, 2 and 4 were audio recorded and transcribed. Stage 1 was also video recorded for use in Stage 2 (stimulated recall).

Between the pre- and post-tests, each group took part in three training sessions. Training sessions involved three stages: 1) all participants performed an oral narrative task based on a picture sequence, 2) they then completed 7.5 minutes of training, and 3) they repeated the same oral narrative task. The training stage (Stage 2) differed according to group. Training for the GN Group involved the use of a guided noticing prompt designed to direct their attention to the formal features of their output and of model input. Training for the UN Group involved the use of an unguided noticing prompt designed *not* to direct their attention to any particular aspect of their output nor of model input, and training for the C Group involved 7.5 minutes of pronunciation practice unrelated to the narrative task. Participants' oral performances were analysed from all training and testing sessions for a range of measures of CAF. Stimulated recall transcripts from testing sessions were also analysed to determine the nature and number of the gaps they noticed in their output (IL-gaps), and participants' note-paper from Stage 3 of the tests was analysed for the number of gaps they noticed between their output and the model input (IL-TL gaps).

The findings revealed that following training (i.e. in post-tests) the GN Group noticed significantly more grammar-related IL-gaps compared to the UN and C Groups. Furthermore, this increased noticing by the GN Group resulted in significantly greater accuracy in their oral output compared to the UN and C Groups when given the chance to repeat a task. Importantly, this increased grammatical accuracy for the GN Group occurred while maintaining rates of fluency. Examination of participants' speech performance in training sessions suggested that the provision of model input in testing sessions mitigated gains in fluency that might otherwise have been made. The findings are explained in terms of the type of training provided to each group and how it impacted upon speech production. Theoretical and pedagogic implications are also discussed.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 BACKGROUND

Over three decades since task-based language teaching (TBLT) emerged as an approach to communicative language teaching (CLT), it remains "both an innovative language teaching method and a thriving area of investigation in the field of second language acquisition" (Ahmadian & Garcia Mayo, 2018, p. 1). Such is the volume of material currently being published on TBLT, keeping abreast of it all is "nigh on impossible" (Newton, 2016, p. 278).

TBLT has gone through significant development and been the subject of a number of criticisms since it appeared in the 1980s as a response to perceived weaknesses in other approaches at that time (e.g. Present Practise Produce). One point of contention has been the extent to which teaching of linguistic form should be included, with advocates of a 'strong form' of TBLT at one end rejecting any explicit focus on grammar, and proponents of a 'weak form' of TBLT at the other end believing it is necessary in order for second language (L2) learners to reach high levels of proficiency in the target language (TL). Currently, there is a general consensus in task-based language (TBL) literature that there needs to be some attention to form while maintaining a primary focus on meaning in order for acquisition to take place (Ellis & Shintani, 2013; Long, 2014).

During speaking tasks, however, finding ways to direct L2 learners' attention towards form while maintaining a primary focus on meaning has proved problematic. An explanation for this can be found in Skehan's (2009) Trade-off Hypothesis. According to the Trade-off Hypothesis (also known as the Limited Attention Capacity Hypothesis), L2 learners have limited processing capacity, and, owing to the communicative nature of speaking tasks, it is

natural for learners to devote this capacity to meaning; or as Skehan (2011) puts it, "when there is communicative pressure of any sort, meaning will be the priority, and form will be something of a luxury" (p. 398). A number of studies have found trade-off effects whereby a higher performance in one component of speech production, such as fluency (i.e. meaning) comes with a corresponding lower performance in another component, such as accuracy and/or complexity (i.e. form). It has, therefore, been the focus of one branch of TBL research to examine the extent to which L2 learners' speech performance can be influenced by the manipulation of the design and implementation variables of different tasks to bring about improved performance in the grammatical complexity, grammatical accuracy, and fluency (CAF) of speech production. This is of pedagogic importance to TBLT because:

> The extent to which TBLT is successful in promoting acquisition will depend on the skill of the task designer and the teacher in manipulating the design and implementation variables of different tasks to achieve a balance between complexity, accuracy and fluency. (Ellis & Shintani, 2013, p. 149)

Manipulation of task variables in order to bring about changes in L2 learners' speech performance can occur at the pre-task, during-task or post-task stage. At the pre-task stage, Yuan and Ellis (2003), for example, found that the provision of planning time led to improved grammatical complexity in L2 learners' speech during the performance of an oral narrative task. The addition of time pressure during a task has been shown to enhance oral fluency (Arevart & Nation, 1991; Boers, 2014; Nation, 1989; Thai & Boers, 2015), and a post-task reflection activity in a study by Lynch (2001) was found to encourage learners to focus on form when producing spoken output. Another implementation variable, and one of the focal points

of this study that has received a great deal of interest from TBL researchers (Ahmadian 2012) is task repetition.

1.2 TASK REPETITION

Task repetition (TR) requires a learner to repeat the same or similar task (Bygate, 2016) after a given interval. The interval between repetitions can broadly be categorised as either 'delayed' (e.g. one day, one week or one month after the initial task performance), or 'immediate' where the learner is asked to repeat a task with minimal (if any) delay.

The psycholinguistic underpinnings for TR and an explanation for how it can influence L2 speech performance can be found in Levelt's (1989) model of speech production. This model includes three main components: 1) the 'conceptualiser' where the speaker thinks about what it is they want to say, 2) the 'formulator' where the speaker encodes the message, and 3) the 'articulator' where the speaker turns the message into actual speech. In order for this process to operate smoothly (i.e. without interruption) the 'formulator' must process partially complete information as it is fed by the 'conceptualiser' (Level, 1989). For native speakers, this process of turning ideas into speech operates in parallel, thanks to the ability to draw upon automatised linguistic knowledge (Tavakoli & Foster, 2008). For L2 learners, however, the 'conceptualiser', 'formulator' and 'articulator' all compete for the same limited attentional resources resulting in slow and/or dysfluent and/or inaccurate speech.

During the repeat performance of a task, pressure is taken off the 'conceptualiser' as the speaker is already familiar with the content, and this in turn frees up attentional capacity which can then be directed to the 'formulator'. The 'formulator' includes a self-monitoring system that "enables a speaker to monitor his or her production prior to articulation and to reformulate his/her speech if/when necessary" (Ahmadian & Tavakoli, 2010, p. 36). It is believed that with increased attention to encoding and monitoring during the repeat performance of a task, an L2

speaker is better able to attend to the form of their speech, resulting in improved complexity and accuracy. Ultimately, as Bygate (2016) believes, task repetition can provide the conditions necessary to "bring together a focus on meaning and form" (p. 393).

A number of studies involving immediate TR have found that after an initial performance, learners perform with greater fluency during subsequent iterations (Arevart & Nation, 1991; Boers, 2014; Nation, 1989; Thai & Boers, 2015). However, despite the added attentional resources available during the encoding process in the repeat performance of a task, and the theoretical claims made in L2 literature that form is enhanced during repeat performances (e.g. Nguyen & Newton, 2019), the effects of TR on complexity and accuracy remain unclear. Boers (2014), for instance, found that errors made by participants in an initial delivery of a speaking task "were simply carried over" to repeat deliveries (p. 230). This has a potentially detrimental effect as it could lead to the consolidation of those errors in the learner's interlanguage. It has, therefore, been suggested that what is needed is some kind of intervention between deliveries to direct learners' attention to form in the iteration(s) (Ellis, 2009).

1.3 ENHANCED TASK REPETITION

Enhanced TR refers to "the second performance that a learner produces after having had the opportunity to engage in some sort of cognitive activity related to their first run" (Lynch, 2018, p. 196). Enhanced TR is synonymous with an output-input-output approach to TR (e.g. Adams, 2003, Uggen, 2012) where some kind of intervention is introduced at the input stage before the repeat performance is delivered. However, as Sheppard and Ellis (2018) note, very few TR studies have examined the effects of an intervention on the repeat performance(s) of a task, and fewer still have examined the effects of an intervention on the repetition of a *speaking* task using an output-input-output sequence (e.g. Izumi & Izumi, 2004; Lynch, 2001, 2007).

As the name suggests, an output-input-output (O-I-O) approach to TR (i.e. enhanced TR) asks a language learner to produce output (either spoken or written, e.g. a narrative task), following which some form of relevant input is provided (e.g. a model narrative) before the learner is then asked to produce output again (e.g. repeat the narrative task). The underpinnings of this approach stem from Swain's (1995) Comprehensible Output Hypothesis. Included in this hypothesis is the 'noticing function' (also called the 'triggering function'), that is, through the act of producing output, a learner becomes aware of gaps in their linguistic knowledge. Swain (2005) outlines the noticing function of output by stating that:

> While attempting to produce the target language . . . learners may notice that they do not know how to say (or write) precisely the meaning they wish to convey. In other words, under some circumstances, the activity of producing the target language may prompt second language learners to recognize consciously some of their linguistic problems: It may bring their attention to something they need to discover about their second language (possibly directing their attention to relevant input). This awareness triggers cognitive processes that have been implicated in second language learning. (p. 474)

After becoming aware of 'gaps' in their linguistic resources during the first output stage, if a learner is then provided with immediate exposure to relevant input, this may lead to deeper and more focused attention on the input (Uggen, 2012). This deeper processing of the input comes as the learner seeks language to fill the gaps they noticed in their prior output, which, according to Schmidt's (1990) original Noticing Hypothesis, is essential if input is to become uptake. If the learner is then given the opportunity to repeat their initial output, this time

utilising language provided in the input to fill the gaps, interlanguage development may be triggered.

Despite the popularity of Swain's Output Hypothesis, and Schmidt and Frota's (1986) similar concept of 'noticing the gap', in L2 literature, relatively few studies have sought to explore the impact of these different types of noticing on speech performance, and of the studies that have, the majority have focused on L2 learners' noticing in written production rather than oral (e.g. Hanaoka, 2007; Izumi, 2003; Izumi & Bigelow, 2000; Izumi & Izumi, 2004; Swain & Lapkin, 1995). Even fewer studies still (e.g. Lynch 2001, 2007; Lynch & Maclean, 2001) have examined the O-I-O cycle in a TBLT context with a view to informing L2 teaching practice, and as far as can be ascertained, no study to date has investigated whether L2 learners can be trained to develop noticing as a skill. In addition, a common criticism of many studies examining speech performance in a TBLT context is that they are one-shot performances (i.e. performance is measured on a single occasion only) (Fukuta, 2016), and this same criticism can be levelled against noticing studies. The aim of the present study, therefore, is to fill the existing gaps in current knowledge mentioned above while at the same time addressing the methodological shortcoming of previous studies that only examine TR and its impacts on speech performance in a single O-I-O sequence.

1.4 THE PRESENT STUDY

The present study was motivated by two main factors. Firstly, with the majority of O-I-O studies examining noticing in a single cycle only, the primary motivation was a desire to explore the idea of training L2 learners over time (i.e. on multiple occasions) to notice form-related problems in their oral output. The underlying rationale here is that L2 learners who are able to identify linguistic gaps in the complexity and/or accuracy of their output, and then mine

relevant input for language to fill those gaps, would be more likely to achieve a better balance between complexity, accuracy and fluency when asked to repeat the output.

Secondly, because so few studies have examined noticing in an O-I-O cycle in spoken modality, there is a need for a more in-depth understanding of what features of their oral output L2 learners notice as being problematic, and how this influences what they then notice upon subsequent exposure to relevant input. It would seem plausible that because L2 learners' attention is largely directed to meaning during speech production, it would primarily be meaning-related problems that they notice in their output, and, as a result, their attention would be orientated to meaning-related features of subsequently presented input. If true, this would likely enhance fluency in the second output stage, however, it would have little impact on complexity or accuracy. As mentioned above, however, achieving a balance between complexity, accuracy and fluency is essential if TBLT is to be successful in promoting L2 acquisition (Ellis & Shintani, 2013). This thesis will therefore explore whether learners can be trained to notice form-related problems in their L2 output, and then subsequently mine model input for language to solve their previously noticed problems, thus resulting in improved speech performance when asked to repeat initial output.

1.5 LAYOUT OF THE THESIS

This thesis has eight chapters. In this chapter (Chapter 1), the overall issue addressed in this research and some key terms and concepts are introduced. In Chapter 2, the relationship between output, noticing and second language acquisition is explored drawing largely on Schmidt's (1990) Noticing Hypothesis and the variations that have followed. Chapter 3 reviews empirical studies that have investigated output, noticing, and L2 performance and development. In Chapter 4 'tasks' and TBLT are discussed, and the various ways in which TBL researchers have attempted to manipulate task implementation variables to bring about

changes in L2 learners' speech performance with a specific focus on TR are reviewed. Also in Chapter 4 is an outline of how speech performance is measured in TBLT, the psycholinguistic underpinnings of TR and a review of relevant TR literature to date. Chapter 5 presents the methodological procedure used in this study, including information about the context, participants, data collection, coding and analysis. Results are presented in Chapter 6, before interpretation and discussion of the results in Chapter 7. Chapter 8 concludes with an outline of the limitations of this study, along with theoretical and pedagogical implications and recommendations for future research.

2 Output, Input and Noticing in SLA

2.1 INTRODUCTION

Almost three decades since Schmidt (1990) first proposed the Noticing Hypothesis, it remains highly influential in SLA research (Leow, 2018). Noticing is closely related to both input and output, and it is these three constructs that form the basis of this chapter. Following this introduction, in Section 2.2, is an overview of the role of input and output in SLA along with their relationship to noticing. Next, in Section 2.3, is a brief background of the Noticing Hypothesis, followed, in Section 2.4, by an examination of the different types of noticing with a particular focus on the two types under investigation in the current study. Section 2.5 looks at common methods researchers have used to measure noticing, and the chapter concludes with an overall summary.

2.2 THE ROLE OF INPUT AND OUTPUT IN SLA

Until the 1980s it was largely thought that input (i.e. the language that learners are exposed to) was the driving force behind SLA. In one influential study, Krashen (1980) introduced the idea of *comprehensible input*, which he defined as language that is slightly more advanced than the learner's current level of language. He used the formula $i + 1$ to explain what is meant by comprehensible input, where i = the learner's current level of L2 proficiency and 1 is the level just beyond. Comprehensible input formed one of the five elements of Krashen's (1985) Monitor Hypothesis of language learning which included the claim that comprehensible input is the "essential ingredient for language acquisition" (p. 4). Swain (1985), however, countered Krashen's claim by stating that while input is an important part of the language acquisition process, it is only part of the picture that also includes *output*.

Swain (1985) arrived at the conclusion that output plays a vital role in language learning after examining L2 French immersion students in a Canadian school. She observed that although the L2 French students performed on par with the L1 French students in terms of level of reading and listening proficiency, they performed much lower in the productive skills of speaking and writing, even after seven years of study. Furthermore, she observed that the L2 French students were less actively engaged in the productive use of French language during the 'French' part of the day compared to their level of active engagement in English language during the 'English' part of the day. This led Swain (1985) to believe that input alone was insufficient to account for language learning, and as a result, she suggested that:

> Comprehensible output . . . is a necessary mechanism of acquisition independent of the role of comprehensible input. Its role is, at minimum, to provide opportunities for contextualized, meaningful use, to test out hypotheses about the target language, and to move the learner from a purely semantic analysis of the language to a syntactic analysis of it. (p. 252)

Later, Swain (1995) proposed the Comprehensible Output Hypothesis which includes the notion that the productive use of language is a necessary element for acquisition to take place.

It's important to note that the role of output being discussed here is not simply the practicing of language that has been received in input. Instead, it is the *act* of producing output that Swain (1995) argues forces the language user to process language syntactically. This is in contrast to the semantic processing required when receiving input, or in Gass's (2015) words, "when producing language, one is forced to use syntax, whereas, in comprehending language there are circumstances in which one need only understand lexical items and not necessarily

the details of syntax" (p. 184). It is the use of syntax when producing language that Swain (1995) argues can lead a learner to *notice* that they are unable to form their message as intended due to limitations in their L2, and this not only makes them aware of what they do not know in their L2, it also forces them to modify their output in order to communicate their message.

Similar to Swain's (1995) hypothesis that the act of producing language prompts an L2 learner to notice what they cannot say (or write) in their L2, Schmidt and Frota (1986) suggested that an L2 learner has the opportunity to notice differences (e.g. errors) in their output when given the chance to compare it to relevant input from a more proficient speaker. It was this notion that becoming consciously aware of (i.e. noticing) gaps when comparing one's L2 output to relevant input from a more proficient speaker that formed the basis of the Noticing Hypothesis (Schmidt 1990, 1994, 2001).

It should be stated at this point that the term *noticing* is being used here (and throughout this thesis) in the same way it was used by Schmidt and Frota (1986) and Swain and Lapkin (1995) who state the term is used "in the normal sense of the word, that is consciously" (Schmidt & Frota, 1986, p. 311).

2.3 BACKGROUND OF THE NOTICING HYPOTHESIS

A large number of empirical and theoretical papers have explored the noticing hypothesis with the underlying assumption that it plays an integral role in driving L2 development forward (Izumi, 2013). According to Schmidt's (1990, 1994, 2001) Noticing Hypothesis, "input does not become intake for language learning unless it is noticed, that is, consciously registered" (Schmidt, 2010, p.1). The Noticing Hypothesis has been labelled, "one of the most influential theoretical underpinnings in SLA over the last two decades" (Leow, 2018, p.1), and while first proposed in 1990, the roots of the Noticing Hypothesis can be traced back to two landmark studies: Schmidt (1983) and Schmidt and Frota (1986).

Schmidt's (1983) study spanned five years and followed the English language progress of a Japanese male who had immigrated to the United States. Schmidt gave the man the pseudonym 'Wes' and documented his English acquisition beginning in 1978. Schmidt (2010) recalls that:

> Wes was a remarkably good learner of English in many ways. His pronunciation was good from the beginning, and he developed quickly along the dimensions of fluency, lexical development, listening comprehension, conversational ability, pragmatic appropriateness, and especially strategic competence, the ability to get his message across in spite of the limitations of his interlanguage . . . His development in the area of grammar—morphology and syntax—was very limited, however. One possible explanation may be that he didn't care much for the small grammatical details of language. Or perhaps he just didn't notice them. For example, after several years of exposure he continued to say things like *Yesterday I'm go beach* and *Tomorrow I'm go beach* (with no articles, no prepositions, and no tense marking), even though he surely heard people say things like *I went to the beach yesterday*, but apparently without registering the forms. (p. 2)

At the time, the dominant view was that second language development was an unconscious and implicit process (Rebuschat & Williams, 2013). However, after analysing 30 hours of recordings of Wes' English, and noting his persistent problems with grammar, Schmidt (1983) concluded that unconscious learning of grammar is not possible for adult L2 learners and that conscious attention to form is necessary.

The second landmark study that lead directly to the Noticing Hypothesis involved Schmidt's own learning of Portuguese while living in Brazil for a period of five months. Schmidt and Frota (1986) analysed notes that Schmidt had kept in a journal about his L2 Portuguese learning, records of what he was taught in class, and tape-recordings taken monthly of his L2 Portuguese production, and they found that it was not until Schmidt had consciously noticed certain grammatical forms in input that he then acquired them. This led Schmidt (1990) to investigate the role of consciousness in language learning, and to propose the Noticing Hypothesis for the first time stating, "…subliminal language learning is impossible, and noticing is the necessary and sufficient condition for converting input to intake" (p. 129).

Although in early versions of the Noticing Hypothesis Schmidt (1990, 1994) stated that noticing was a necessary condition for learning, this claim was somewhat softened when Schmidt (2001) acknowledged that subliminal learning may be possible, but the more learners notice, the more they learn (Ellis & Mifka-Profozic, 2013), or in Schmidt's (2001) words, "attended learning is far superior" (p. 3).

It should be noted, however, that the Noticing Hypothesis is not without its critics. Truscott (2008), for instance, argues that it is not clear what details of input need to be noticed, and that L2 researchers investigating the Noticing Hypothesis need to be clearer about what it is specifically that learners are becoming aware of. Truscott and Sharwood Smith (2011) also criticise the Noticing Hypothesis by claiming that there does not seem to be agreement in the literature about what exactly *noticing* is. Furthermore, they argue that it is not simply a language learners' awareness of input, which is central to the Noticing Hypothesis, but more specifically it is awareness of forms in the input, as can be seen in the following quote:

> Noticing is more than just awareness of input; it involves awareness specifically of forms in the input. However, it is much less than full awareness of form, as conscious understanding is excluded. Thus, noticing necessarily has a lower boundary that distinguishes it from simple awareness of input and an upper boundary that distinguishes it from awareness at the level of understanding. (Truscott & Sharwood Smith, 2011, p. 501).

Despite these criticisms, the Noticing Hypothesis continues to be used in L2 research to better understand the relationship between input and L2 output as part of the process of learning a second language.

As a result of around three decades of empirical investigation, the concept of noticing has evolved and, as Izumi (2013) notes, the "idea of noticing has been interpreted in various ways and given rise to discussion of different types of noticing in the SLA literature" (p. 26).

2.4 DIFFERENT TYPES OF NOTICING

More than three decades have passed since *noticing* first appeared in L2 literature, and even a brief review of studies to date reveals a number of different types of noticing along with various noticing-related terms which, due to inconsistency in use, have the potential to lead to confusion. For instance, Schmidt and Frota (1986) proposed *noticing a gap*, while Swain (1998) proposed *noticing a hole* and *noticing a form* which are part of the *noticing function* - one of the three functions of Swain's (1995) Comprehensible Output Hypothesis which also refers to learners *noticing gaps* in their interlanguage. More recently, Izumi (2013) has suggested four types of noticing: 1) *Noticing a form(-meaning-function) relationship*, 2) *noticing a gap between IL and TL*, 3) *noticing holes in IL*, and 4) *noticing a gap in one's ability*.

While there is similarity in terminology, the terms above have been used to refer to slightly different types of noticing. To complicate things further, *noticing a gap* and *noticing a hole* have been used interchangeably by some authors while others have used the same terms to refer to subtly different kinds of noticing. Izumi (2013), for example, treats *noticing a gap* and *noticing a hole* as separate principles. While both occur during the speech production process, in his definition, the noticing of a *hole* refers to an L2 learner who becomes consciously aware that they have a complete lack of knowledge of the form in question, while the noticing of a *gap* refers to the partial absence of a form. Izumi (2013) illustrates the difference with an example where a learner was aware of the term *traffic jam* but was unsure what verb to use with it, leading the learner to use a common verb such as *is* or *has*. Izumi believes a learner would likely feel that this verb usage is incorrect, therefore, leading them to notice a 'gap' between their IL and the precise meaning intended. Conversely, Izumi (2013) states a *hole* is the complete absence of a certain form, in which case, he believes, "production problems do not become manifest in the learners' output (as any relevant knowledge of the form is missing)." (p. 28).

Due to the potential for confusion in terminology, before going any further it is important to label and define the types of noticing under investigation in the present study, namely, *noticing an IL gap*, and *noticing an IL-TL gap.*

2.4.1 Noticing an IL gap

The first type of noticing investigated in the present study is what Swain (1995) called the *noticing function* of the Output Hypothesis. It has also been referred to as *noticing a hole* (Doughty & Williams, 1998; Swain, 1998) and *noticing the gap in one's ability* (Izumi, 2013). However, for the purposes of this study, this type of noticing will be referred to as *noticing an IL-gap*. While it could be argued that adding the term *noticing an IL gap* to the list of terms already used to date might complicate matters further, it is felt that *noticing an IL gap* is a

clearer term in that it identifies *where* the gap is noticed (i.e. in IL), whereas, terms such as *noticing a gap* and *noticing a hole* do not. Furthermore, the term *noticing an IL gap* is more easily contrasted with the second type of noticing under investigation in the present study, namely, *noticing an IL-TL gap* (see section 2.4.2).

Noticing an IL-gap occurs when a learner attempts to produce their L2 and notices (i.e. becomes consciously aware) that they cannot form their message as intended because of a lack of L2 linguistic resources. Swain (1995) argues that the act of producing one's L2 raises awareness of limitations in one's IL, and this consequently pushes the user to modify their output in order to overcome these limitations to communicate their message. It is this modification that results in the learner engaging in cognitive processes that are involved in the process of L2 learning. Swain (1995, 2000) further claims that noticing limitations in one's L2 primes the learner to attend more carefully to language forms subsequently presented in relevant input.

2.4.2 Noticing an IL-TL gap

The second type of noticing under examination in the present study is what Schmidt and Frota (1986) termed *noticing the gap,* and what Izumi (2013) calls *noticing the gap between IL and TL* (and will hereafter be referred to as *noticing an IL-TL gap*). Unlike noticing an IL gap which happens *during* the speech production process, noticing an IL-TL gap happens *after* the language production process while a learner has a chance to compare how he used language to convey meaning to the way a more proficient speaker (or writer) used language to convey the same meaning. In other words, when an L2 learner is exposed to relevant input from a more proficient speaker, the learner may notice a difference between how they use their L2 to express an idea, and how the more proficient speaker uses language to express the same idea. Schmidt and Frota (1986) believe it is through conscious comparison between one's L2 output and target language input that the learner can notice errors in their IL and seek to correct them, thus

driving development of their IL forward. The two types of Noticing used in the present study are defined in Table 2.1.

Table 2.1 - *Definitions of Noticing used in the present study*

Type of Noticing	Definition
Noticing an IL gap	During L2 speech production when a learner attempts to produce their L2 and notices (i.e. becomes consciously aware) that they cannot form their message as intended because of a lack of L2 linguistic resources.
Noticing an IL – TL gap	After producing L2 speech, when a learner has the opportunity to compare the language they used to that of a model speaker, the learner notices a difference between the way they used language to express an idea and the way the model speaker used language to express the same idea.

Despite the prominence of Swain's (1995) *noticing function* and Schmidt and Frota's *noticing the gap*, few studies have sought to investigate their role in L2 learning (Hanaoka, 2007; Hanaoka & Izumi, 2012; Izumi & Bigelow. 2000; Izumi, Bigelow, Fujiwara & Fearnow, 1999; Swain & Lapkin, 1995, see Chapter 3 for a review). One reason for this may be due to difficulties in measuring noticing.

2.5 MEASURING NOTICING

Having outlined different types of noticing, and provided their definitions in the previous section, the current section examines how noticing can be measured. This is important because in order to examine the relationship between noticing and SLA, it is essential that there exist valid and reliable measures of what it is learners notice and when (Smith, 2012). Measuring

noticing presents a number of challenges as it involves the examination of learner-internal processes (Izumi & Bigelow, 2000). However, with careful planning and preparation, these challenges can be overcome. Noticing as it relates to the present study can be measured at two different times: firstly, what learners notice during the act of L2 production (i.e. what IL gaps they notice), and secondly, what learners notice when they compare their output to relevant input (i.e. what IL-TL gaps they notice).

2.5.1 Measuring IL-gaps noticed

Methods of measuring IL-gaps noticed can be broadly separated into 'offline' and 'online' measures. 'Offline' measurement takes place *after* the language has been produced and commonly used measures include post-task questionnaires (e.g. Robinson, 1997) and stimulated recall sessions (e.g. Mackey, 2006). Conversely, 'online' measurement takes places *while* the language is being produced with the most commonly employed method being a think aloud protocol (e.g. Bowles, 2010; Hanaoka, 2007; Hama & Leow, 2010; Leow, 1997, 2000). Although online measures are generally acknowledged to be more reliable due to the immediate reporting of what is noticed, the decision of which method to employ is largely contingent upon the modality of the language output that is under investigation. Researchers investigating noticing in written output are generally able to choose either offline or online measures, whereas, those examining noticing in oral output generally only have offline measures available as any online measure will interfere with the speech production process (i.e. an L2 learner cannot be asked to complete a speaking task and simultaneously talk about their thought processes without interruptions to their task performance). The most commonly used offline method to measure what L2 learners notice in oral output, and the method used in the present study, is a stimulated recall (SR) approach (Egi, 2008; Gass & Mackey, 2000; Mackey, 2006; Mackey, Philip, Egi, Fujii & Tatsumi, 2002; Philip & Iwashita, 2012; Uggen, 2012).

2.5.1.1 Stimulated recall.

SR is an introspective technique in which "participants are asked to recall thoughts they had while performing a prior task" (Gass & Mackey, 2017, p.22). As a method for investigating what learners' were thinking during a past event, SR has been gaining increasing prominence in L2 literature (King, 2016) with numerous studies employing it as a research methodology (Dörnyei & Kormos 1998; Mackey, 2002; Mackey, 2006; Mackey, Gass, & McDonough, 2000; Sato, 2007; Watanabe, 2008; Watanabe & Swain, 2007). The 'stimulus' in SR generally comes from exposure to an extract of the written or spoken language a learner produces during the prior task (Ellis, 2008). In the case of oral production, the stimulus is often a video recording of the learner's performance.

SR as a technique for eliciting data on participants' thought processes is not without potential pitfalls. The main criticism relates to veridicality (Bowles, 2010). Veridicality refers to the accuracy of the retrospective verbal report given by the participant during the SR as memory decay potentially impacts a participant's ability to accurately recall what they were thinking during the prior event. With memory decay becoming increasingly likely as the time between the SR and the original event grows (Gass & Mackey, 2017), the problem of veridicality can be mitigated by conducting the SR as soon as possible after the original event and, where possible, by providing an audio *and* visual stimulus to help prompt recollection (as opposed to audio only).

A second potential issue that can arise relates to the possible influence of how the SR session is run by the researcher leading to variability between sessions. Variability can come from the timing of the SR session (i.e. how soon after the original event), the wording of instructions given to participants and the kinds of questions asked in order to prompt recall. To overcome these potential pitfalls, Gass and Mackey (2013) suggest that instructions be read from a script, and careful consideration be given beforehand to the types of questions to be

asked and the way in which they intend to be asked. For instance, they warn against "fishing for recall comments" (Gass & Mackey, 2013, p. 44) which can happen when a participant says that they do not recall their thought process at a particular time, yet the researcher pursues a response with further questioning. Fishing for recall comments increases the likelihood that a participant will feel the need to provide a response, resulting in the participant reporting what they are thinking now (rather than at the time they were undertaking the task), thus supplying an inaccurate comment.

2.5.2 Measuring IL-TL gaps noticed

Because the noticing of IL-TL gaps occurs when a learner is exposed to input, not when they are producing output, online measures are possible regardless of output modality. Online measures are the preferred choice as they enable the immediate reporting of what is noticed and thus reduce the chance of memory decay. The vast majority of studies that have investigated what IL-TL gaps learners notice when presented with input have used underlining as an indication of noticing (e.g. Izumi, Bigelow, Fujiwara, & Fearnow, 1999). However, underlining has come under criticism for its validity as a measure for noticing. Uggen (2012), for instance, used two methods to measure noticing of language presented in written input: the online measure of underlining and the offline measure of stimulated recall. She found that some of what learners reported noticing in the stimulated recall session was not underlined, therefore indicating that not everything that is noticed is underlined.

With regard to examination of IL-TL gaps noticed in oral input, as far as can be ascertained, the only empirical investigation of what L2 learners notice when provided with the opportunity to compare their output to model input comes from Sheppard's (2006) unpublished doctoral dissertation. In that study, the linguistic features learners noticed when comparing their performance of an oral narrative task to that of a model speaker were measured

by examination of notes learners were asked to take while comparing. This same method is used in the present study (see Chapter 5 for more detail on methods used for data collection).

2.6 CHAPTER SUMMARY

In this chapter, an overview of the role of input and output in SLA was given while highlighting the importance that both play in L2 development. Also presented was an outline of Swain's (1995) noticing function of the Comprehensible Output Hypothesis which includes the claim that the act of producing output can lead a learner to become consciously aware of (i.e. notice) what it is they do not know in their L2. This led to a discussion of noticing that began with a brief background of Schmidt's (1990, 1994, 2001) Noticing Hypothesis before clarifying noticing-related terminology used in the literature to date. Following this was a description of different types of noticing, including an explanation and definition of the two types that are under investigation in the present study, namely, noticing an IL gap, and noticing an IL-TL gap. The final section of this chapter explored both offline and online methods used to measure noticing, while highlighting potential problems associated with different methods and how they might be overcome.

3 Noticing Research

3.1 INTRODUCTION

As discussed in the previous chapter, the concept of *noticing* has evolved since it first appeared in SLA literature, and it has been interpreted in various ways (Izumi, 2013). Unsurprisingly, there have been numerous studies that have investigated *noticing* from a range of perspectives. However, pertinent to the research presented in this thesis are studies that have investigated Swain's (1995) noticing function of the Comprehensible Output Hypothesis, and Schmidt and Frota's (1986) noticing the gap. Therefore, it is empirical investigation of these two constructs that are the focus of the review of literature presented in this chapter.

The chapter begins in Section 3.2.1 with an overview of relevant noticing research while highlighting methodological approaches and a common limitation of studies to date. The remainder of the chapter looks firstly, in Section 3.2.2, at studies investigating noticing in L2 writing, followed, in Section 3.2.3 by a review of studies investigating noticing in L2 speaking. The chapter ends with an overall summary.

3.2 RESEARCHING NOTICING

3.2.1 Overview of noticing research

The two types of noticing being examined in the present study include three main claims, two from Swain's (1995) noticing function of the Comprehensible Output Hypothesis, and one from Schmidt and Frota's (1986) noticing a gap. These claims are:

1. The act of producing output leads an L2 learner to become aware of limitations in their IL (Swain, 1995). In other words, producing output is the trigger for an L2 learner to notice IL gaps.
2. After noticing IL gaps, learners attend more closely to subsequently presented relevant input in order to find language to fill those gaps (Swain, 1995).
3. When given the opportunity to compare their L2 output to that of a more proficient speaker, a learner has the opportunity to identify differences in how they used language to express meaning, and how the more proficient speaker uses language to express the same meaning, that is, they have the potential to notice IL-TL gaps (Schmidt & Frota, 1986).

To date, researchers have investigated a number of aspects related to noticing. For example, studies have examined whether the act of producing output leads L2 learners to notice IL gaps (Cumming 1990; Swain & Lapkin, 1995), whether producing output leads to more noticing of target forms in subsequently presented input (Izumi & Bigelow, 2000; Izumi, Bigelow, Fujiwara, & Fearnow, 1999; Uggen, 2012), and whether, when given the chance to repeat output, learners fill IL gaps noticed in initial output using language presented in relevant input (Hanaoka, 2007). Only one study to date has examined learner-generated noticing (Hanaoka & Izumi, 2012), and very few have examined noticing of both IL gaps and IL-TL gaps in speech (Izumi & Izumi, 2004; Sheppard, 2006). Finally, all the aforementioned studies have investigated noticing in a single output-input-output cycle, or in the case of Cumming (1990) and Swain & Lapkin (1995) a one-shot output performance. No studies to date have looked at noticing in an output-input-output cycle on multiple occasions, and this is one limitation the current research aims to address.

Additionally, the majority of studies have examined researcher-generated noticing rather than learner-generated noticing. In researcher-generated noticing, the researcher has a pre-selected target form which, in the case of written input, is usually presented using either input enhancement or input flooding. Input enhancement refers to target forms that have been highlighted, underlined or boldened in a text, while input flooding relates to input that contains a high frequency of target form use. Comparatively fewer studies have examined learner-generated noticing, which is when the learner is left undirected to notice (or not notice) whichever linguistic aspects of input they wish (see Section 3.2.3 for a review).

3.2.2 Noticing in writing

A number of studies have sought to investigate the nature of IL gaps noticed when producing L2 written output. One of the earlier studies with this aim came from Cumming (1990) who conducted think-aloud protocols with 23 adult Francophone learners of English. Each participant took part in two separate L2 writing tasks (an informal letter writing task and an argument essay writing task). Cumming (1990) analysed think aloud transcripts to find instances of when the learners were making decisions about their writing, and these instances were further coded to determine what each learner was attending to while making the decisions. Results showed that while writing, learners spent almost a third of their decision-making time simultaneously attending to 'gist' (i.e. the content of their composition) and 'metalinguistic' concerns, while the majority of the remaining decisions involved cross linguistic comparisons (Cumming, 1990).

Similarly, Swain and Lapkin (1995) aimed to test whether the act of producing output makes learners aware of (i.e. notice) linguistic problems in their IL, and whether upon noticing learners are pushed into modifying their output. Nine adolescent French immersion students in Canada took part in the study. Each student was asked to write one to two paragraphs about a given topic in French (their L2). A think aloud protocol was used to identify instances of the

students noticing a language-related problem, or *language related episode (LRE)* as Swain and Lapkin term it. After analysing transcripts of think-aloud protocols, Swain and Lapkin categorised each LRE into one of seven ways in which the students solved each problem they encountered (e.g. by applying a grammatical rule, through a lexical search, through translation). Findings showed that the students became aware of linguistic problems as they were writing and, furthermore, they engaged in thought processes to solve those problems which required them to syntactically process the language which, the authors conclude, can play a role in L2 learning.

Izumi et al (1999) took the testing of the noticing function of the Comprehensible Output Hypothesis one step further when they set out to investigate Swain's (1995) claim that the act of producing output can lead a learner to discover what they do not know in their L2. In addition to identifying whether learners notice IL gaps, Izumi et al (1999) also examined whether the noticing of IL gaps pushes the learner to attend more closely to language in subsequently presented input. Their study involved 22 college students enrolled in an ESL course at a community college in the U.S. At the outset, all participants completed a pre-test that involved a grammaticality judgement test and a test of written output based on a picture prompt that included the need for the use of the past-hypothetical conditional (which was the target form in their study).

Next, in phase one of the study, both those in the experimental group (EG) and those in the control group (CG) were presented with a reading that contained the target form (past hypothetical). Participants were asked to "underline the word, words, or parts of words" (Izumi et al., 1999, p. 427) that they thought would be necessary for the next stage of the experiment (participants were made aware of the requirements of each stage at the outset). In the next stage, those in the EG were asked to reconstruct the passage they had been given in the prior input, while those in the CG were required to answer comprehension questions. This was

repeated a second time (presentation of input and reconstruction of the passage for the EG, and comprehension questions for the CG). In that way, both groups received the same input, however, only the EG produced output, while the CG answered comprehension questions related to the input.

In phase two of the study, the EG group was asked to complete a writing task that required the participant to imagine they were on a school committee deciding how a designated amount of money should be spent. This task called for the use of the target structure (past hypothetical). The CG group, on the other hand, was asked to write about a topic unrelated to that of the EG which did not call for the use of the past hypothetical. Next, a model essay was presented to the EG for which they were required to read and underline the parts they felt would be necessary for the following stage where they were asked to once again write on the same topic previously given. Rather than underline, those in the CG were required to answer comprehension questions about the model essay. Contrary to what was hypothesised, both groups (i.e. the EG who produced output and the CG who produced no output) improved significantly in their noticing of the target form in the input presented. However, when it came to incorporation of the target form in the post-test (a grammatical judgement test and a test of written output), the EG outperformed the CG by incorporating more of the target form in their output. Finally, partial confirmation was found for the hypothesis that the EG would show greater gains in accurate use of the target form with significant improvements in accuracy of the past hypothetical shown by the EG in phase two of the study (but not phase one), while the CG did not improve in accurate use of the target form in phase two. Izumi et al (1999) concluded that their study provides partial support for the noticing function of the Output Hypothesis in that the production of initial output leads to noticing of target forms in input and subsequent incorporation of target forms in repeated output. The authors argue that their results

provide support for the idea that output prompts noticing and learning and is therefore useful for SLA.

Izumi et al (1999) conceded that there was the possibility of a cumulative effect which led to the phase 2 results. In other words, phase one of the study had a priming effect on phase 2. Thus, Izumi & Bigelow (2000) sought to address this issue in their study which replicated Izumi et al (1999) but reversed the order of tasks. Although their findings showed that output production did not always result in learners attending to the target form, they did find that opportunities to produce output and exposure to relevant input were vital for learners to improve their use of the target form.

More recently, utilising an output-input-output TR cycle, Uggen (2012) conducted a conceptual replication of Izumi and Bigelow's (2000) hypothesis that producing the TL directs L2 learners' attention to language presented in subsequent input. Furthermore, she investigated whether level of complexity of the target structure influences attentional processes. Her study involved 30 ESL learners divided into two experimental groups (EGs); EGpast, received the more complex target structure (past hypothetical) presented in written input during the treatment phase, and EGpres, who received the less complex (present hypothetical) target structure presented in written input in the treatment phase.

Treatment conditions included an initial output stage, followed but an input stage, and then a second output stage. The initial output stage elicited either the past or present hypothetical in an initial written output stage for the EGs, while the CG was not required to produce any output. During the following stage, all groups received model input that contained the relevant target structure (for the EGs), and they took part in reading and underlining activities designed to draw their attention to those structures. In a second output stage, participants were required to produce a second piece of writing using the relevant target structure. Analysis of participants' underlining from the input stage revealed no quantitative

difference between the EGs and the CG in terms of what was noticed in model input. However, qualitative analysis of participants' comments from a stimulated recall session conducted after the second output stage showed that learners exposed to the more complex target structure were more aware of vocabulary and limitations in their L2 (i.e. they had more problems accurately using the target structure) than those exposed to the less complex structure and those in the CG. This suggests that requirement to process more complex target structures induces more noticing, whereas less complex structures, as Uggen (2012) explains, may not attract noticing because they are more familiar.

Izumi (2002) also included an external device, namely, visual input enhancement which used underlining, bolding, highlighting and shadowing of target forms in an attempt to draw learners' attention to target forms. His study, involving 61 students enrolled in ESL programs at two U.S. universities, had two main aims: firstly, to investigate whether the act of producing output promotes noticing of target language forms presented in input and how that affects learning of form, and secondly, whether visual input enhancement designed to draw learners' attention to problematic forms in input has an impact on what is noticed and learned. Results showed that participants who produced output and were then exposed to input outperformed (in terms of amount of target form noticed) those who were exposed to the same input but were required to answer comprehension questions (i.e. not required to produce output). No significant advantage was found for those who received input enhancement compared to those who did not.

With several studies finding support for noticing, Hanaoka (2007) sought to investigate the nature of what learners notice when they produce written output (i.e. what IL gaps are noticed), what they notice when they compare their output to model input (i.e. what IL-TL gaps are noticed), whether there is an impact of noticing on subsequent incorporation of input, and finally whether level of L2 English proficiency impacts on what is noticed and incorporated.

Hanaoka's (2007) study began, in stage one, by asking 37 Japanese college students at two distinct levels of L2 English proficiency to write a story based on a picture prompt, then, in stage two, participants were required to compare their writing to two native-speaker models before an immediate rewrite (stage three) and a delayed rewrite (stage four).

While composing their initial written narrative in English, participants took notes (in their L1) on a separate sheet of paper about the problems they encountered while writing their story, that is, they were asked to write down IL gaps noticed. The note-taking sheet was collected when students entered stage two which required them to compare their story with two native-speaker models. During this stage, participants were asked to note (in their L1) the linguistic differences they noticed between their narration and that of the models, in other words, they were asked to note IL-TL gaps. In both an immediate post-test and a delayed post-test two months later participants rewrote their original story (i.e. the story they wrote in the pre-test).

IL gaps were measured through examination of participants' note-taking during stage one of the study. They were analysed to see what aspects of language in the native-speaker models the learners noticed. Instances of noticing fell into one of four categories: lexis, grammar, content, and other. Results showed that the nature of IL gaps noticed was overwhelmingly dominated by lexis with more than 90% of instances of noticing falling into this category. When comparing their writing to the native-speaker models, the IL–TL gaps noticed were dominated by lexis (63%) and content (29%). Results also showed that when comparing their writing to the models, participants noticed around two-thirds of solvable lexical problems (i.e. lexical problems they noticed in their output for which lexical solutions were available in the model input) and incorporated 92% of them in their immediate rewrite (i.e. post-test). When examining delayed post-test writing, results showed that 40% of solutions found and incorporated in the post-test rewrite had been retained. Finally, when

comparing the groups in terms of proficiency, no difference was found in the amount participants noticed while writing. In other words, level of proficiency had no impact on the number of IL gaps noticed. However, the higher-proficiency group noticed more when comparing their writing to the models indicating that level of proficiency does impact on number of IL-TL gaps noticed.

Hanaoka and Izumi (2012) also used an output-input-output design to examine the effects of what they termed 'overt' and 'covert' noticing on the uptake of target-language input. 'Overt' noticing was defined as problematic features that learners noticed while producing written output which were also addressed by using the form learners believed to be appropriate. 'Covert' noticing, on the other hand, included problematic features that learners noticed while writing but did not address or only partially addressed. For instance, a problematic feature that was noticed but subsequently avoided in output, or a problematic feature that was dealt with by the use of L1 were considered 'covert'. Following Hanaoka (2007) noticing was measured through examination of a note-taking sheet participants wrote on while composing their narration. One difference between Hanaoka and Izumi's (2012) study and previous studies came in the manner in which model input was provided to participants. Following their initial output, participants' compositions were collected and reformulated (i.e. rewritten with errors corrected but meaning unchanged) by a native speaker. One week later, participants were asked to rewrite their original composition after being exposed to input. The input included two texts: 1) a native speaker model of the original writing task, and 2) the reformulated version of each participants' initial output. The texts were presented one-at-a-time and were counterbalanced to control for ordering effects (i.e. half the participants received the model text first and then reformulated text second, while the other half of participants received the reformulated text followed by the model text). While participants were presented with the first input text (model essay for half the participants, and a reformulated essay for the other half),

they were asked to make notes (in their L1) of any language they noticed as they compared their original output with the input. After a time limit of seven minutes had passed, participants notes and the first written input were collected before the second written input was presented and participants were again asked to make notes on a new sheet of paper of language they noticed as they compared their output to the second input.

Results showed that model input was more useful in helping learners overcome problems noticed than reformulated input, and that 'overt' and 'covert' problems noticed were solved roughly equally during the final output stage. Another interesting finding revealed that 80% of the problems noticed by participants were lexis related, reinforcing similar findings from Hanaoka (2007). Furthermore, the authors concluded that their results lend support to the noticing function of the Output Hypothesis in that the act of producing output triggers learners to notice limitations in their IL.

With the majority of studies finding that the opportunity to produce output leads to greater noticing of language presented in input, Song and Suh (2008) set out to investigate whether the type of output task used impacts on the noticing and uptake of a target form (past counterfactual conditional). Instead of an output-input-output design, this study began with input, followed by an output stage then a second input stage and finally a second output stage. The second input and output stages were repetitions of the first. After being presented with the first input that contained text with the target form in around 70% of sentences, the L2 learners of English in the study were then given either a reconstruction task, which required them to rewrite the narration presented in input with the aid of picture prompts (but not the original input text) to lessen the burden on memory, or a picture-cued writing task, which required participants to rewrite the narration with scaffolding provided. A third group acted as a control and were given comprehension questions to answer during this stage (i.e. they did not have the opportunity to produce output).

Song and Suh (2008) found that participants who were given the opportunity to produce output noticed more instances of the target language form presented in input than those who did not produce output. Furthermore, those who had the opportunity to produce output in the treatment phase of the study performed significantly better in terms of acquisition of the target form in the post-test than those who had no opportunity for output during treatment, however, no significant difference was found between the two treatment groups (reconstruction task vs picture-cued writing task).

3.2.3 Noticing in oral output

Compared to investigation into noticing in written input and output, far fewer studies have examined noticing in speaking. One reason for this may be the methodological challenges involved in measuring what is noticed by learners during speech production.

Following similar studies focusing on L2 writing, Izumi and Izumi (2004) examined the effects of the opportunity for oral output on the acquisition of relative clauses by L2 learners of English. Twenty-seven students enrolled in an ESL program at a U.S. university were assigned to either a control group, an output group, or a non-output group. Around the time of their study, there were claims (e.g. Swain & Lapkin, 1995; Izumi, 2003) that the act of producing output forces learners to process language syntactically compared to the largely semantic processing required to comprehend input. As a result, in addition to a control group who took part in separate and unrelated treatment, Izumi and Izumi (2004) included two experimental groups in their study; one that was asked to produce output (and therefore hypothesised to require syntactic processing of language), and another that was not asked to produce output (therefore requiring only semantic processing of language). Both groups were exposed to the same aural and visual input, afterwards the output group took part in a picture description task whereas the non-output group completed a picture sequencing task for which they were not required to produce any output. Unexpectedly, the non-output group

outperformed both other groups in the learning of the target form. This led the authors to conclude that although they believe output has an important role to play in SLA, any output task must be evaluated to ensure that "relevant psycholinguistic processes are really engaged in the learners" (Izumi & Izumi, 2004, p. 606), rather than simply repeating target language input.

While many O-I-O noticing studies have employed a researcher-generated approach, that is, there has been a preselected target form, some studies have examined learner-generated noticing. Mennim (2007), for example, gave participants 'noticing exercises' over the period of one academic year. These exercises included: a) language development awareness sheets which required students to write down any new L2 English language they had noticed over the previous week, b) post-presentation questionnaires which were designed to direct the presenter's attention to the form used in their speech, and c) transcription exercises where students transcribed five minutes of their own speech and later corrected errors. Mennim (2007) noted several developments in participants' L2 English over than time, but with no control group in his study, improvements cannot be ascribed to the 'noticing exercises' used.

Lynch (2001) examined the effects on speech performance of asking L2 learners to transcribe and revise (in pairs) their initial performance of a speaking task. Eight students in an oral communication skills class in an ESL context took part in an activity which required pairs of L2 learners to role-play various 'academic scenarios' in which one learner acted as a university student and the other acted as a university teacher. In the scenarios, the 'student' had to make a request to the 'teacher', for instance a request for an extension of a deadline. Performances of scenarios were audio recorded and transcribed by participants in pairs. Once a verbatim transcription had been finalised, participants were then required to revise the transcript by correcting errors and changing any language they thought needed to be improved. The process of transcription required negotiation between participants in order to arrive at the

final transcription. When finished, transcripts were taken away and reformulated by the researcher. Reformulations involved changing certain parts of the transcripts that were linguistically incorrect, or wording that was deemed to be unnatural. When later examining the transcripts himself, Lynch (2001) identified further errors that had not already been noticed by the learner-learner dyads. Results showed that the learner pairs noticed and corrected, on average, 40 grammatical errors in their transcript, while the researcher found a further 34 grammatical errors. Furthermore, the learner pairs made an average of eight lexical corrections, whereas the researcher identified another 28 instances where lexical corrections were needed. Other areas where smaller numbers of errors were identified by the learners and the researcher were related to editing (i.e. removal of redundancies, repetitions and false starts) and reformulation.

Lynch (2007) conducted a similar study with the aim of investigating whether learner-produced transcripts or teacher-produced transcripts were more effective in improving learners' output. Sixteen students enrolled in an English for academic purposes course at a British university took part in the study. Participants were asked to complete a dyadic speaking task which was audio recorded and transcribed either by the learners or by the researcher. Learners then took part in 'reprocessing' activities which required the learners to examine their transcripts. Analysis of a performance of the same task two days after the first (and after learners had had the opportunity to examine their transcript), and a second repeat performance of the same task four weeks after the first (i.e. a delayed post-test) showed that more gains were made in accuracy in post-tests for participants who transcribed their own performance than for those whose performance was transcribed by the researcher.

More recently, Sheppard and Ellis (2018) aimed to test the impacts on speech performance of conducting a stimulated recall session between an initial and a repeat performance of the same speaking task. Furthermore, the authors sought to test whether any

improvements made in speech performance post-intervention would transfer to a new task. Forty EFL students in Japan were randomly assigned to either a task-repetition group or a stimulated recall group. Both groups performed a monologic oral narrative task based on a picture sequence at time one (i.e. the pre-test). After this initial performance, those in the task repetition group took part in a general conversation with the researcher before repeating the same narrative task a second time. Participants in the experimental group took part in a stimulated recall (SR) with the researcher, the purpose of which was to raise participants' awareness of their performance through identification of IL gaps. Immediately following treatment (conversation for the TR group, and SR for the SR group) all participants repeated the same oral narrative task that was used at Time 1. Two weeks later, a delayed post-test was conducted which required all participants to perform the same oral narrative task again. Finally, immediately after the delayed post-test, all participants were asked to perform a new oral narrative task. All four performances from each participant were audio recorded, transcribed and analysed for measures of CAF.

Findings revealed that structural complexity increased from pre- to post-test for both groups and was maintained in both the delayed post-test and in the performance of the new task at Time 4. Grammatical accuracy was found to remain the same from the initial performance to the repeat performances of the same task for both groups, however, accuracy dropped for both groups when asked to perform the new task. In terms of fluency, although gains were made from the pre- to the post-tests for both groups, gains were significantly greater for the SR group. Level of fluency for both groups declined in the performance of the new task, however, the authors point out that despite the drop in fluency in the performance of the new task, levels were still higher than in the pre-test. Sheppard and Ellis (2018) concluded that SR as a procedure to raise L2 learners' awareness of their speech performance is facilitative of fluency development in the short term, but has no impact on accuracy or complexity.

Furthermore, when looking at speech performance in the new task, the authors note that there was some transfer effect for fluency for the SR group (i.e. gains made in fluency during repeat performances of the same task were carried over to the performance of the new task), but not for the task repetition group, and not for complexity or accuracy for either group. The authors conclude that SR is effective in raising L2 learners' awareness of their speaking performance and that this supports fluency development in terms of number of pruned words (speech with dysfluencies such as pauses and hesitation removed) uttered per minute.

3.2.4 Noticing and uptake

Uptake refers to "a student's response to the provision of information about a linguistic form that the student has produced incorrectly" (Loewen, 2004, p. 154). The provision of linguistic information often comes by way of corrective feedback from the teacher. However, it can also come from other sources, such as model input, as is the case in the present study. Some researchers have investigated the relationship between noticing and uptake in task repetition studies (see section 4.6 for more on task repetition). In such studies, for example, L2 learners are first given the opportunity to perform a language task, next, the learner is exposed to model input of the same task, and finally the learner is asked to repeat their initial performance of the task. The process here, it is argued, is that in the initial output stage, the learner becomes aware of deficiencies in their interlanguage (i.e. they notice an IL gap). Next, when exposed to model input, the learner has the opportunity to mine that input for language to fill the gaps they noticed previously, and they also have the opportunity to notice how the language they used to complete the task differs from the way language was used in the model input to complete the same task (i.e. they notice an IL – TL gap). Finally, when asked to repeat their initial performance, the learner has the opportunity to incorporate language from the model to fill the gaps they noticed, thus demonstrating uptake.

Researchers have examined specifically whether L2 learners mine language presented to them and use it in the performance of a task (i.e. they have examined uptake in task performance). Boston (2008), for instance, set out to examine firstly, whether L2 learners mine written input, audio input, or both for language that they can use when subsequently producing output, and secondly, whether learners mine language features deliberately imbedded by the teacher into pre-task materials, and if so, does it make a difference if the teacher explicitly draws learners' attention to these features or not. Two classes of low-level (false-beginner) students at a Japanese vocational college were recorded performing the same task over a period of three weeks. A different task was used each week. One class was presented with two to four audio recordings prior to performing their task (the RPT group), while the other class was presented with the same recordings after performing their task (the RAT group). Students worked in pairs to complete the tasks. For both classes, pre-task and task materials along with task instructions were presented in written form to each pair of learners. Boston (2008) found that whether learners' attention was explicitly drawn to specific language features of the pre-task and task materials or not, learners still mined the material for language to use in their performance of the task. However, this was only the case for written materials. Boston (2008) found no evidence that the learners mined the audio input for language. One reason for this may have been due to the low-level of L2 English proficiency of the learners, and a reflection of their prior English language education in Japan where TL input is more likely to be presented in written than audio form.

In a follow-up study, Boston (2009) attempted to use pre-task activities to orient L2 learners' attention to a target structure (passive voice) with the aim of having learners use that structure during task production. Three separate classes of 72 high-beginner level learners at a Japanese university took part in the study. Twenty-four randomly selected students were recorded in each class, and these 24 students were divided into eight groups of three learners.

Each group in each class performed the same task. The task included a pre-task stage and a main-task stage.

To begin the pre-task stage for Class 1, learners were individually presented with a series of ten cartoon pictures depicting the 1911 theft of the Mona Lisa (a story they were familiar with having read about it in a previous class). The pictures were presented individually and out of order, and each learner was asked to reconstruct the story according to the order of events they believed occurred. Pictures were drawn in such a way that they could be described using either the active or passive voice. Ten minutes was given for this stage, after which the pictures were removed. The learners were then asked to sit in their groups of three. Fifteen statements about the series of events from the Mona Lisa story were projected onto a screen for the class to see. All statements were written in the passive voice. Learners in each group took turns to read the statements aloud and were asked to choose whether each statement was true of false according to the events that had been depicted previously. Results of this pre-task stage revealed that all learners in each group employed the passive voice when talking about statements that were not true according the depicted story (i.e. after reading an incorrect statement from the screen, the learner corrected the information in the statement using the passive voice).

Lastly, the statements that had been projected on the screen were removed, and each group was given the task of reconstructing the story from memory. Groups were audio recorded as they attempted to reconstruct the story.

The procedure for Class 2 was identical as that for Class 1 above. However, the 15 statements projected on the screen were all presented in the active voice. Class 3 was not presented with any true/false statements. Instead, as they worked individually to reorder the series of ten pictures, they were allowed to write picture descriptions as they went.

Furthermore, they were given 15 minutes for this stage (compared to 10 minutes for the other classes).

Boston (2009) found that the pre-task stage for Class 1 that was designed to prime learners to subsequently use the passive voice was unsuccessful in doing so, with only 15 out of a total of 101 utterances containing the passive voice structure. Boston (2009) notes that five of the eight groups in Class 1 recorded no instances of using the passive voice at all, and among the few instances it was used in the other three groups, it was used by one individual in each group only. Class 2 produced no utterances using the passive voice, while Class 3, who were not primed to use either active or passive voice, were almost as successful as Class 1 with 11 instances of passive voice use out of a total of 123 utterances.

When examining why the attempt to prime learners was unsuccessful, Boston (2009) postulates that the low-level of L2 English proficiency of the participants may have meant they were simply not at a stage developmentally where they were could readily use passive voice construction. Upon further reflection, Boston (2009) believes that the task used may have been flawed in that in order to complete the task, the passive voice was not more necessary nor more useful that the active voice (a structure that is arguably developed earlier and is more familiar to low-level learners of English).

3.3 CHAPTER SUMMARY

This chapter began with an overview of noticing in SLA and a look at different types of noticing. Studies that have investigated noticing in L2 writing were reviewed. While the majority of these studies have found support for the claim that the act of producing output triggers L2 learners to become aware of limitations in their IL (Cumming 1990; Swain & Lapkin, 1995; Uggen, 2012;), others have only found partial support for this claim (Izumi et al., 1999; Izumi & Bigelow, 2000). This chapter also highlighted that there is a paucity of

research on how learner-generated noticing impacts on speech performance (e.g. Lynch 2001, 2007; Sheppard & Ellis, 2018). However, these studies, each utilising enhanced TR, have found noticing to have a positive impact on speech performance in repeat performances of a task, thus providing support for Schmidt and Frota's (1986) concept of *noticing the gap* (in the case of Lynch, 2001, 2007), and Swain's (1995) noticing function of her Comprehensible Output Hypothesis (in the case of Sheppard & Ellis, 2018). The final section of this chapter took a closer look at the relationship between noticing and uptake.

4 Tasks and TBLT

4.1 INTRODUCTION

In the previous chapter we saw how researchers have investigated noticing in SLA. The main focus of most of these studies was to examine either one or both of the claims made by Swain's (1995) noticing function of the Comprehensible Output Hypothesis. These claims are, firstly, that the act of producing L2 output prompts learners to become aware of limitations in their L2, and secondly, that this increased awareness of limitations can orient learners to pay closer attention to language subsequently presented in relevant input. Some of the studies reviewed in the previous chapter also investigated Schmidt and Frota's (1986) claim that when given the chance to compare one's output to relevant input from a more proficient speaker (or writer), L2 learners have the opportunity to notice gaps between the language they used and the language used by the more proficient speaker, thus becoming aware of limitations (e.g. errors) in their IL.

As shown in Chapter 3, the claims above have received support from a majority of studies to date. However, while support for noticing has been found, few studies have examined the impact noticing has on L2 speech performance. The primary motivation for the research presented in this thesis is to investigate how noticing affects L2 speech performance with a view to informing task-based language teaching and learning. Therefore, having looked at the role of output in SLA and noticing in the two previous chapters respectively, this chapter provides an overview of the L2 teaching and learning context within which this research is situated, that is, a task-based language teaching/learning context.

This chapter begins, in Section 4.2, with a brief background of TBLT before exploring what is meant by the term 'task' in Section 4.3. Next, Section 4.4 examines how speech performance, in terms of complexity, accuracy, and fluency, is defined and measured in TBLT. This is followed, in Section 4.5, by an overview of literature that has investigated ways in which task design and implementation variables can be manipulated to influence L2 speech performance. Section 4.6 looks more closely at task repetition, its theoretical and psycholinguistic underpinnings, and the type of task repetition relevant to this study, namely, enhanced task repetition (Lynch, 2018). Presented lastly is an overall chapter summary.

4.2 TBLT

TBLT grew out the shift to communicative language teaching (CLT) in the 1970s and 80s with Prabhu's (1987) Bangalore/Madras Communicational Teaching Project (CTP) often credited as the first documented example of a TBLT syllabus being used (Ellis, 2009). One of the central tenants of CTP was that "form is best learnt when the learner's attention is on meaning" (Prabhu, 1982, p. 2). Since that time, TBLT has developed as an approach to CLT (Ellis, 2009) and remains "a thriving area of investigation in the field of second language acquisition" (Ahmadian, 2016, p. 377). Since Prabhu's claim that form is best learnt when the learner is attending to meaning, there has been debate in TBL literature about whether or not there should be an explicit focus on teaching of form; with those favouring a 'strong form' of TBLT rejecting any explicit focus on grammar, and advocates of a 'weak form' of TBLT believing that explicit attention to form is needed. Current opinion, based on empirical evidence, holds that while there should be a primary focus on meaning in TBLT, there is a need for some attention to form in order for an L2 learner to achieve a native-like level of proficiency (Ellis, 2016; Long, 2014).

4.3 WHAT IS A TASK?

Newton (2016) notes that "task research is a dynamic and expanding field" (p. 275), and because of this dynamism and continual growth, an important first step before embarking on a discussion on tasks is to define what exactly a *task* is. A range of definitions of 'task' has been proposed in TBL literature (Bygate, Skehan & Swain, 2001; Crookes, 1986; Ellis, 2003; 2009; Long, 1985; 2014; Nunan, 1989; 2006; Prabhu, 1987; Skehan, 1996; 1998; van den Branden, 2016; Willis & Willis, 2001), and while there is no single agreed upon definition, there are similarities across definitions. Ellis (2009, p. 223), for instance, states that although various definitions exist, commonalities suggest that the following 4 criteria must be satisfied if an L2 activity is to be called a 'task':

1. The primary focus should be on 'meaning' (by which is meant that learners should be mainly concerned with processing the semantic and pragmatic meaning of utterances).
2. There should be some kind of 'gap' (i.e. a need to convey information, to express an opinion or to infer meaning).
3. Learners should largely have to rely on their own resources (linguistic and non-linguistic) in order to complete the activity.
4. There is a clearly defined outcome other than the use of language (i.e. the language serves as the means for achieving the outcome, not as an end in its own right).

One criterion not included in the list above by Ellis (2009), although is included in Ellis, (2003) is that a task should have a real-world application (Long, 2014; Skehan, 1996), for example, to tell a story, solve a problem or give directions (Ahmadian, 2016). As the aforementioned criteria are largely common across definitions in TBL literature to date, a 'task'

as defined in the research presented in this thesis is one which satisfies the five criteria below, that is:

1. There is a primary focus on meaning.
2. There is an information gap which needs to be overcome.
3. Learners are required to rely on their existing stock of linguistic and non-linguistic resources.
4. There is a clear outcome other than use of language.
5. There is a clear real-world application.

4.4 SPEECH PERFORMANCE IN TBLT: MEANING VS FORM AND THE EMERGENCE OF CAF

The meaning vs form distinction that serves as the basis of CAF (complexity, accuracy and fluency) can be traced back to Brumfit (1979) who is often cited as being the first to distinguish fluency from accuracy in the L2 classroom (Hunter, 2011). Writing about L2 classroom activities, Brumfit (1984) states that 'accuracy' activities are those which focus on linguistic *form* and controlled production of grammatically correct L2 linguistic structures, while 'fluency' activities on the other hand, are *meaning*-based and foster spontaneous L2 speech production. In the mid 1990s Skehan (1996) proposed a proficiency model which took the meaning-vs-form distinction and further broke down form to include complexity in addition to accuracy, resulting in what has become known in the literature as CAF.

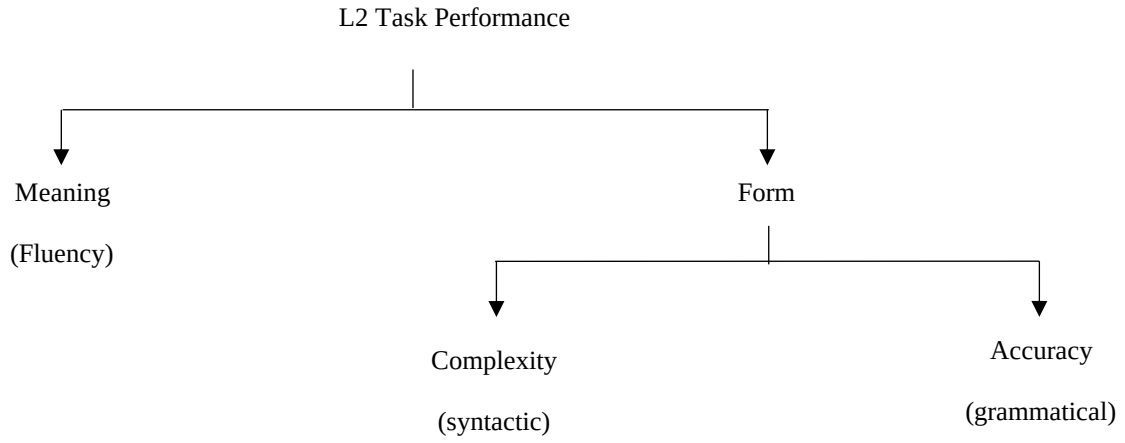


Figure 4.1 - *Skehan's (1996) three dimensions of L2 speech performance.*

According to Housen, Kuiken and Vedder (2012) "many L2 practitioners and SLA researchers now hold that L2 proficiency is not a unitary construct but, rather, that it is multicomponential in nature, and that its principal components can be fruitfully captured by the notions of complexity, accuracy, and fluency, or CAF for short" (p. 1). Although there exist a range of approaches to account for performance on language learning tasks (Skehan, 2009), complexity, accuracy, and fluency have provided the standard measurement in task-based L2 research for nearly two decades (Lambert & Kormos, 2014).

Before looking at how CAF can be measured, it is first important to define each construct, and although this has proved problematic in the literature, working definitions have been around since the 1990's and are still in use today (Housen et al., 2012). According to these working definitions, complexity is considered to be 'the extent to which the language produced in performing a task is elaborate and varied' (Ellis, 2003), accuracy concerns 'the extent to which the language produced in performing a task conforms with target language norms' (Ellis, 2003), and fluency refers to the ability to process the L2 with 'native-like rapidity' (Lennon, 1990).

While working definitions provide some guidance as to the nature of the CAF constructs, for several decades L2 researchers have been concerned with finding singularly accepted definitions plus valid methods of measurement. This has proved challenging as the next section will discuss.

4.4.1 What is fluency?

While none of the three dimensions of CAF are without controversy when it comes to finding a singularly accepted definition, even a brief review of the literature soon reveals that *fluency* is by far the most contentious of the three. Indeed, Thornbury (2000) describes *fluency* as "an elusive concept" (p. 139), and defining and measuring fluency consistently across different tasks and conditions has proved to be problematic (Wright & Tavakoli, 2016).

To some, the lack of a definition of fluency may seem surprising at first considering it is a commonly used term both inside and outside the L2 classroom. Outside of pedagogic circles people tend to use *fluency* to refer to someone's overall command of a language. This sense can be seen when someone asks, "Are you fluent?" after discovering their interlocutor speaks another language. Inside of pedagogic circles *fluency* is often used to refer to one isolatable component of language production. An important first step before a discussion on *fluency* can begin is to first identify which sense of the term *fluency* we are talking about.

One of the first to suggest that *fluency* can refer to different aspects of language in different contexts was Lennon (1990). In his study he suggests there are two senses of spoken fluency: one broad and one narrow. In the broad sense he states that 'fluency' is a synonym for oral proficiency. In this sense, "'fluent' represents the highest point on a scale that measures spoken command of a foreign language" (p. 398). Lennon describes the "ubiquitous" application form question on foreign language ability. Such a form, whether it be for employment or study, often

asks the applicant to place themselves on a scale from "fair" through "good" to "fluent" – which is the highest point on the scale.

In the narrow sense, Lennon (1990) argues that "fluency" in EFL refers to one isolatable component of oral proficiency. He states, "this sense is found particularly in procedures for grading oral examinations, where a learner can be fluent but grammatically inaccurate, or fluent but lack a wide and varied vocabulary" (p. 390). Alternatively, a learner may "speak correctly but not very fluently" (Lennon, 1990, p. 390). Looking at the following quote taken from band nine (the highest score) for "fluency and coherence" from the IELTS speaking test, one can get an idea of this narrow sense Lennon is referring to: "Speaks fluently with only rare repetition or self-correction; any hesitation is content related rather than to find words or grammar" (IELTS Speaking band descriptors, 2016). Other standardized tests are similar in that most descriptions of fluency on marking schemes for oral assessment refer to the absence of features such as hesitations and pauses (Chambers, 1997).

Kormos and Denes (2004) also suggest two forms of fluency: one which considers it as spoken language competence (i.e. the ability to communicate effectively) and another which regards it as a temporal phenomenon (i.e. relating to speaking rate and fluidity in delivery). Similarly, Derwing, Thomson and Munro (2006) posit that fluency is primarily characterised as either an indication of the degree of overall proficiency, or as a composite of temporal phenomena (e.g. speech rate, number and length of pauses). Tavakoli, Campbell and McCormack (2016) suggest that in a narrow sense fluency is "predominantly associated with the ability to communicate one's intended meaning effortlessly, smoothly, and with no or little disruption" (p. 448). For the purposes of the research proposed here, *fluency* will be used in a narrow sense and is considered to be a temporal phenomenon.

4.4.2 Defining fluency

Despite ambiguity in the interpretation, a number of researchers have sought to investigate fluency in L2 learners speech with the aim of finding a definition and valid methods for measurement. Fillmore (1979) was one of the earlier researchers to offer a definition of *fluency*. He defined it as having key aspects including the ability to fill time with talk, and the ability to talk at length with few pauses and hesitations while knowing what to say in a wide range of contexts. Others, such as Schmidt (1992), define fluent speech as being "automatic, not requiring much attention or effort" (p. 358). Gatbonton and Segalowitz (2005) define fluency as "the smooth and rapid production of utterances, without undue hesitations and pauses" (p. 326). Fluency has also been defined as "speech without (unnatural) hesitations" (de Jong, 2016, p. 113).

One approach of L2 researchers towards finding a definition of fluency has been to ask trained and/or untrained judges acting as listeners to provide a fluency rating to L2 speech samples. The researcher then quantitatively analyses those speech samples in an attempt to identify what factors influenced the judges in their subjective assessments (Lennon, 1990; Riggenbach, 1991).

Lennon (1990) used this method in a ground-breaking, albeit small-scale study. Lennon obtained speech samples from four German students who were on a six-month stay in England to improve their English language skills. One speech sample was taken from the beginning of their stay and one from the end. The speech samples were played to nine native-speaker teachers of EFL who gave fluency ratings to each of the recordings. The majority of raters were in agreement as to which was each participant's earlier recording and which was their later recording, and that the later rendering was more fluent than the earlier one.

Lennon then applied a battery of 12 quantifiable performance measures to each speech sample which were considered to be representative of fluency. Comparisons were made between the first and second recordings. Speech rate, filled pauses per T-unit (a main clause along with any subordinate clauses), and percentage of T-units followed by a pause were found to be significant indicators of a fluent speech performance. Lennon concluded that two key areas appeared to play a role in fluency: 1) speech-pause relationships, and 2) frequency of occurrence of dysfluency makers like filled-pauses and repetitions.

Similarly, Kormos and Denes (2004) asked three native-speaking and three non-native-speaking teachers to judge speech samples from L2 learners as to their perceived levels of fluency. It was found that for all six teacher-judges, the best predictors of judges' fluency scores were speech rate, mean length of utterance, phonation time ratio (calculated as the percentage of time spent talking as a proportion of the total time taken to produce the speech sample, de Jong & Perfetti, 2011) and number of stressed words uttered per minute.

Recent studies have attempted to break down fluency into subcomponents for further analysis. Segalowitz (2016) suggests fluency comprises cognitive fluency, which relates to the "speed, efficiency and fluidity of the cognitive processes thought to underlie the implementation of the speech act" (p. 79), and utterance fluency, which measures the "oral fluency of that speech act" (p. 79). Skehan, (2003) offers three subcomponents of fluency: 1) breakdown fluency, which includes measures such as number and length of filled/unfilled pauses, and total amount of silent time, 2) speed fluency, which typically involves measures of speech rate, articulation rate (i.e. total number of syllables produced divided by the time taken to produce them excluding pause time), and mean length of run, and 3) repair fluency, which includes repairs, false starts, and repetitions. More recently Skehan, Foster and Shum, (2016) suggest fluency can be represented by discourse fluency and clause fluency. Discourse fluency is associated with pauses that occur at clause boundaries and is assumed to be associated with macro-planning such as control of

content. Skehan et al (2016) claim that clause fluency, on the other hand, is associated with mid-clause pausing and is evidence of micro-planning such as grammatical encoding. Finally, although there is still no consensus on which measures are the best indicators of L2 fluency (Tavakoli et al., 2016), most studies in TBLT research agree that there should at least be some measure of speech rate (e.g. word or syllables per minute), and some measure of pauses (including, number and length, and whether filled or unfilled).

As shown above, while there has been a range of definitions presented in the literature to date, commonalities across definitions converge on the notion that fluent speech is produced at speed without undue pause or hesitation and is achieved without much conscious effort on the part of the speaker, and this will therefore be the definition of fluency adopted for the research presented in this thesis.

4.4.3 Defining accuracy

Although *accuracy* can be applied to a number of aspects of language production (e.g. pronunciation), here and throughout this thesis, *accuracy* refers to the grammatical accuracy of L2 speech. Defining accuracy (also known as *correctness*) in L2 oral performance has been less contentious than defining fluency. While L2 researchers often word their definitions of accuracy differently, such definitions are essentially the same in meaning. For example, Housen et al. (2012) define accuracy as the extent to which an L2 learner's speech deviates from a norm (usually the native speaker). Pallotti (2009) defines accuracy as the degree of conformity to certain norms, while Skehan (1996) states that accuracy refers to "how well the target language is produced in relation to the rule system of the target language" (p. 23).

Accuracy is not without criticism, however. For example, disagreement can arise when considering whether non-standard usage deemed to be acceptable in certain contexts or communities should be included in the 'norms' against which an L2 learner's speech is being

measured. This has led some to argue that accuracy should include not only grammatical accuracy but also 'appropriateness' and 'acceptability' (Housen et al. 2012). In addition, Foster and Wigglesworth (2016) argue for a weighted-clause ratio when measuring accuracy. Their justification for such a ratio is that not all grammatical inaccuracies should be treated as equal. They propose that errors should be classified into one of three levels, with level 1 errors being relatively minor and as such they do not interfere with meaning, while at the other end of the scale, level 3 errors are more serious and impact greatly on the meaning in a way that "makes the intended meaning far from obvious" (Foster & Wigglesworth, 2016, p. 106).

4.4.4 Defining complexity

There are a range of measures available to researchers examining syntactic complexity in L2 speech. Such measures include looking at the length of speech, amount of subordination, amount of coordination, variety and sophistication of grammatical forms, and frequency of use of certain grammatical forms (see Norris & Ortega, 2009 for a review). In the majority of TBLT studies examining speech performance, complexity as a dependent variable has most commonly been measured by amount of subordination.

One common measure of subordination is average length of error-free T-unit (minimal terminable unit), where a T-unit is an independent clause accompanied by any associated dependent clauses (Larsen-Freeman, 2009). The T-unit was borrowed from L1 research. However, more recently another measure, the AS-unit, was developed by Foster, Tonkyn, and Wigglesworth (2000) specifically for L2 research and this has been the most frequently adopted measure of complexity in L2 task-based research of late.

An AS-unit (analysis of speech), is defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated

with either" (p. 365). A sub-clausal unit consists of a segment of speech that can be elaborated into a full clause by recovering omitted information (Ellis & Barkhuizen, 2005).

4.4.5 What about lexis?

As mentioned in section 4.4, the three dimensions of CAF came together when Skehan (1996) added complexity to the already existing pairing of accuracy and fluency. More recently, a case has been made for the inclusion of lexis (Skehan, 2009) in order to gain a more complete picture of speech performance. Thus, while CAF is still used by many researchers, an increasing number are now including measurement of lexis, meaning CAF becomes CALF. Lexis often includes two separate measurements: 1) lexical complexity, which is typically measured using a type-token ratio analysis, and 2) lexical sophistication, which is typically measured by finding the number of words uttered that fall within the 500 most commonly used words in English, the number of words that fall within the 1,000 most commonly used words, and also the number of words uttered that fall within the 2,000 most commonly used words in English. This can be measured and calculated relatively easily by running a transcript of spoken language through an online lexical profiler such as Lextutor.com (Cobb, 2018). However, for the purposes of this study, lexis is not used as part of a measurement for speech performance, and a justification for this is included in section 6.8.

4.4.6 Measuring CAF

The validity of CAF as measures of speech performance have found support in a number of studies including Skehan and Foster (1997) where a factor analysis showed the three dimensions of CAF to be distinct. Furthermore, when referring to a large number of studies beginning in the mid-nineties that used CAF as dependent variables, Housen and Kuiken (2009) state that "From this diverse body of research, CAF emerge as distinct components of L2 performance and L2 proficiency which can be separately measured . . . and which may be

differentially developed by different types of learners under different learning conditions" (p. 462).

To date, the number of studies employing CAF as measures of speech performance continues to grow (e.g. Ahmadian & Tavakoli, 2010; Boers, 2014; Ellis, 2009; Kormos & Trebits, 2012; Robinson, 2001; Skehan & Foster, 1997; Skehan, 2009; Tavakoli, 2016; Tavakoli & Foster, 2008; Thai & Boers, 2015; Vercellotti, 2015; Yuan & Ellis, 2003).

As Norris and Ortega (2009) note, the predominant purpose of using CAF as measures of an L2 is to provide an insight into L2 developmental processes by enabling the documentation of what aspects of a learner's interlanguage change as acquisition unfolds. Within a TBLT context, measuring CAF allows researchers to examine how and why interlanguage develops in certain learners and under certain conditions (e.g. in certain tasks and with certain teachers).

4.4.6.1 Measuring complexity

Commonly used measures of complexity include the ratio of clauses to AS-units (Ahmadian, 2011; 2012; 2013; Ahmadian & Tavakoli, 2010; 2014; Awwad, Tavakoli & Wright, 2017; Bei, 2013), mean length of AS-unit (Ahmadian, 2011; 2012; Awwad et al., 2017), total number of different grammatical verb forms used (Ahmadian, 2013), ratio of subordinate clauses per AS-unit (Bui, Ahmadian & Hunter, 2019), number of words per AS-unit (Bui et al., 2019; de Jong & Vercellotti, 2016), number of subordinate clauses per T-unit (Mehnert, 1998), variety of past-tense verb forms (Foster & Skehan, 1996), and number of words per T-unit (Bygate, 2001).

4.4.6.2 Measuring accuracy

Compared to complexity and fluency, the range of dependent variables used to measure accuracy is much narrower. By far the most common measure is to look at the presence (or absence) of errors with one point of contention being the variety of English with which an L2 sample is to be compared in order to identify errors. Common measures used to date include percentage of error-free clauses (Ahmadian 2011, 2012, 2013; Ahmadian & Tavakoli, 2010,

2014; Awwad et al., 2017; Bei, 2013; Bui et al., 2019; de Jong & Vercellotti, 2015; Mehnert 1998; Skehan & Foster, 1997), number of errors per AS-unit (Gilabert, 2007), and number of errors per 100 words (Awwad et al., 2017; Bui et al., 2019).

4.4.6.3 Measuring fluency

As discussed earlier in this chapter, agreement on what constitutes fluency has been elusive in L2 literature. However, in recent times, there has been general agreement on measures of fluency in TBL research to the point it is now largely accepted that in order to gain a finer-grained analysis of fluency, as Skehan (2003) notes, there is a need for separate measures of speed fluency, (e.g. speech rate), repair fluency (e.g. reformulations, false starts and repetitions) and breakdown fluency (e.g. filled and unfilled pauses).

Speed fluency is most commonly measured in terms of number of syllables or words per 100 words or per minute of pruned and/or unpruned speech. Pruned speech refers to transcripts that have dysfluencies removed (e.g. false starts and repetitions) leaving only 'meaningful' words/syllables. Repair fluency looks at frequency of false starts, repetitions and reformulations, while breakdown fluency measures pauses in terms of their number, nature (e.g. either filled or unfilled), and locations (e.g. at clause boundaries or mid-clause).

4.4.7 Criticisms of CAF

Despite the wide-spread adoption of CAF measures to analyse L2 speech performance in TBLT research, the CAF model is not without criticism. A number of studies have raised both methodological and theoretical issues with the use of CAF (e.g. Lambert & Kormos, 2014). However, perhaps most notable is Pallotti (2009) who claims that there are a number of issues related to defining and operationalising CAF constructs. One issue she terms the "necessary variation fallacy" (Pallotti, 2009, p. 590). This, it is argued, is when researchers tend to believe the 'best' measures of language are those that show variation over time and across different tasks

among subjects. However, Palotti argues that of equal validity are measures that do not show any difference among groups of subjects. She states that:

> Research should be concerned with variations and differences, but also with constraints and similarities. If after an experimental treatment two groups of subjects do not show any difference, then this is not a nonresult, but a result just as interesting as their being different. Likewise, if a measure does not change over time, then this does not make it a poor measure, but perhaps a measure pointing to a trait that does not actually vary. (p. 591).

A second issue Pallotti (2009) raises is related to "the search for the *significant result*" (p. 591) among a range of measures. For instance, if two groups differ significantly on just one measure out of many, this may be an indication that the two groups are equal except for one characteristic, or in the case that *p* is close to 0.05, then that significance may be due to chance. Palotti (2009), therefore, cautions against reading too much into results from studies where variations in CAF are reported in only one or two measure among many.

One further criticism from Pallotti (2009) is that very few studies employing CAF as a measure of speech performance have also looked at how successful communication of the spoken message actually was. She provides the following examples of differing levels of communicative adequacy :

> If in an information gap task a learner were to utter unhesitatingly, 'colourless green ideas sleep furiously on the justification where phonemes like to plead vessels for diminishing our temperature', her production would score extremely high on CAF, in spite of being completely irrelevant, and probably counterproductive, for task success. In contrast, an

> utterance such as 'No put green thing near bottle. Put under table' is neither complex nor accurate, and may not be fluent either, but can turn out to be perfectly functional for achieving the speaker's (and the task's) intended communicative goal. (p. 596).

Palotti (2009) argues that communicative adequacy should be seen as a separate dimension of CAF, and should be used as a way of interpreting CAF measures. A counter to this argument is that first language speakers are not always communicatively accurate, therefore, an L2 speaker should not be judged on this either (P. Foster, personal communication, May 23, 2016). It is also argued that communicative adequacy can be achieved through gesture along (Foster & Wigglesworth, 2016); however, such communication would be very limited. To take communication further, language needs to be produced. Furthermore, Foster and Wigglesworth (2016) argue that in attempting to produce the TL, an L2 learner "arrives at a position of using grammatical morphemes which add nothing at all to meaning, such as the English third person, present tense verb ending. This shows language proficiency beyond adequacy, so it is necessary for an accuracy measure to be able to take account of what might be fully adequate but still inaccurate" (p. 106).

4.5 INFLUENCING L2 LEARNERS' SPEECH PERFORMANCE THROUGH MANIPULATION OF TASK DESIGN

One line of research in TBLT has aimed to quantitatively examine how the manipulation of task design and implementation variables can influence the speech performance of L2 learners as measured by CAF, and it is to this line of investigation that the research proposed here aims to contribute.

Manipulation of task design can occur in any one of three stages: pre-task, during-task, and post-task. In one of the earlier studies looking at manipulation at the pre- and during-task stage,

Yuan and Ellis (2003) investigated the effects of pre-task planning and online planning on task performance. In this study, forty-two undergraduate university students were asked to narrate a story based on a series of cartoon-frame picture-prompts. Participants were randomly assigned to one of three conditions: 1) a no-planning group, 2) a pre-task planning group, or 3) an online planning group. A time limit of six minutes was imposed on those in the no-planning and pre-task planning groups. Participants in the no-planning group were asked to perform the task immediately after seeing the picture prompts, and participants in the pre-task planning group were given ten minutes to plan before performing the task. Those in the online planning group were given no time to plan, however, they were given an unlimited amount of time to complete the task. Results showed that pre-task planning led participants to improved grammatical complexity, fluency and lexical diversity, while online planning positively affected participants' accuracy and grammatical complexity.

Tavakoli and Foster (2008) investigated the effects of level of complexity of picture prompt on L2 learners' oral output. They found that more complex picture prompts (i.e. pictures that contain events going on in both the background and foreground) resulted in learners attempting more complex language and producing longer mean lengths of utterance when narrating the story represented in those picture prompts. This is in contrast to less complex picture prompts (i.e. pictures that contain foreground events only) which elicited less complex language by comparison. The authors also found that picture prompts with a 'loose narrative structure' (a sequence of text-less cartoon picture frames that could be ordered in a number of ways to tell a story) produced more accurate language than 'tight narrative structure' prompts (picture prompts that could only be ordered in one way to tell the story).

The two studies above provide examples of ways in which task design can be manipulated. The kind of manipulation pertinent to the research in this thesis, however, is task repetition (TR).

4.6 TR AND DIFFERENT KINDS OF TR

Task repetition in a TBLT context asks an L2 learner to perform a task, and then repeat the same or a slightly different task one or more times after a given interval. Length of time before the repeat performance(s) can be broadly categorised as being either immediate or delayed. Researchers have examined different kinds of TR, however, as Bygate (2018) notes, regardless of type, they "all share an underlying theoretical and practical motivation for their use, centring on the ability of task repetition to promote the gradual integration of the various dimensions of relevant language knowledge into more fluent language use" (p. 24).

Although Bygate (2018) lists eight different types of task repetition (including 2 sub-types), and Manchon (2014) outlines five different types, they can all be categorised as fitting into one of the following three types as proposed by Patansorn (2010):

1. Procedural repetition – tasks that differ in content when repeated but maintain the same procedure.
2. Content repetition – tasks that keep the same content when repeated but differ in the procedure required to accomplish the communicative goal.
3. Task repetition – tasks that include the same content and procedure when repeated.

Research has shown that L2 learners' speech performance (in terms of CAF) can be impacted differently according to task type. Kim and Tracy-Ventura (2013), for instance, found that procedural repetition fostered development of syntactic complexity, whereas task repetition did not. Lynch and Maclean (2000) found procedural repetition facilitative of both accuracy and fluency, while results from Bygate's (2001) study found task repetition impacted positively on the fluency and complexity of participants' speech, but not on accuracy.

In addition to task-type, there are a number of other variables that have been manipulated in TR studies to date. For instance, number of repetitions has ranged from one (Ahmadian, 2013; Fukuta, 2016; Lynch, 2011; Mayo, Agirre & Azkarai, 2017) to 10 (Ahmadian, 2011), while duration between repetitions has ranged from no delay (i.e. immediate repetition) (Arevart & Nation, 1989; Bei, 2013; Boers, 2014; Fukuta, 2016; Lynch, 2001; Lynch & Maclean, 2000, Nation, 1991, Thai & Boers, 2015) to the longest duration reported in current TR literature of 3 months (Mayo et al., 2017). Finally, while some studies have used a dialogic task (Ahmadian, 2011; Lynch 2001; Lynch & Maclean, 2000; Mayo et al., 2017), others have used a monologic task (Ahmadian, 2013; Ahmadian & Tavakoli, 2010; Bei, 2013; Fukuta, 2016).

Before looking more closely at the empirical investigation of TR, it if first important to gain an understanding of the theoretical underpinnings of using TR as a pedagogic technique for developing L2 learners' IL.

4.7 THEORETICAL AND PSYCHOLINGUISTIC UNDERPINNINGS OF TR

In TR, a learner's first performance is regarded as preparation for subsequent performances (Ellis, 2005). As a number of authors have noted, on the surface this may seem reminiscent of drilling in behaviourist approaches to SLA (Ahmadian, 2012; Bygate, 2018), and it conjures up images of classrooms full of L2 learners repeating target phrases verbatim in unison until memorised. However, this is not the case with TR in TBLT as learners are placed in meaningful situations where they are free to use their stock of linguistic resources to complete a given task. Furthermore, because an L2 task has a real-world application and a desired outcome (e.g. to book a table at a restaurant or to tell a story) there is a clear target and reason for using language. When a task is repeated, motivation and value for the learner remain so long as the repetition has a clear purpose other than simply repeating the same language again. For example, a change of

interlocutor or a modification of the original task will keep the learner focused on achieving the outcome.

The theoretical underpinnings of TR are based on the assumption that humans have limited attentional capacity. As Skehan (2009) notes, this is a largely accepted assumption borrowed from contemporary cognitive psychology. Because attentional capacity is limited, when performing a speaking task there is competition between meaning and form for attentional resources, and a learner must prioritise which areas of performance will receive attention. In the case of a speaking task, learners tend to prioritise meaning as this is what is needed in order to complete the task. Because meaning is the priority, however, less attention is given to the grammatical form and complexity of speech. The central idea behind TR in TBL research is that during a repeat performance(s) of a speaking task, attentional capacity is freed up as the speaker is already familiar with the content of their talk and the procedural requirements of the task. This freeing up of attentional resources means more attention is available and can therefore be directed to the formal features of speech resulting in improved performance. As van den Branden (2007) states, "while the learners tend to focus on meaning construction during their first performance, they can free processing space during the second performance, allowing them to focus more on the forms they are using" (p. 170). It is claimed that, through repetition and improved performance, automatization of language is developed, thus driving interlanguage development forward.

TR studies to date frequently cite Levelt's (1989) model of speech production when outlining the psycholinguistic underpinnings. As mentioned in Chapter 1, Levelt's model (see Figure 4.2) includes three stages in the speech production process: 1) the conceptualiser, where the speaker comes up with the content of their speech, 2) the formulator, where the speaker encodes the content, and 3) the articulator, where the speaker verbalises their message. It is theorised that through TR, attention that was given to the conceptualiser during an initial

performance can be redistributed to the formulator and articulator during a repeat performance as the content of the speaker's talk is already familiar and is thus readily retrievable from short-term memory. This increased attention to the encoding and articulation of one's speech is thought to provide the speaker with the resources needed to improve in areas of speech performance including accuracy and complexity.

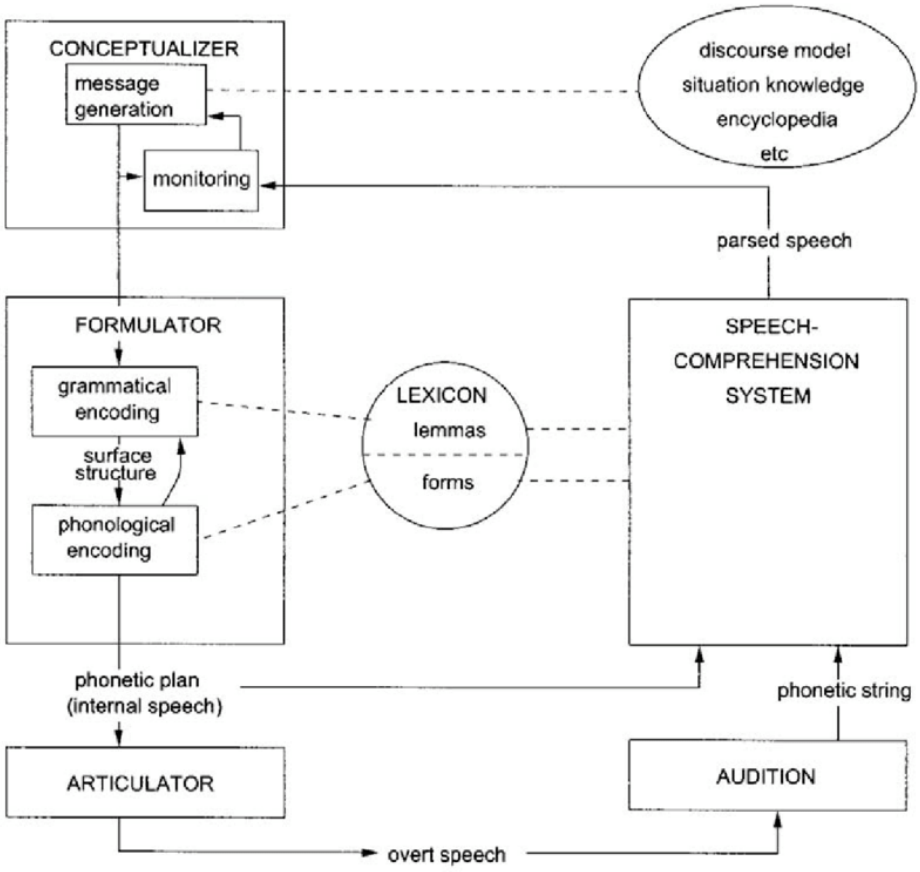


Figure 4.2 – *Levelt's (1989) model of speech production*

In one influential study investigating how different task design and implementation variables affect speech performance, Skehan and Foster (1997) asked L2 learners of English to perform three different speaking tasks, each one on a different occasion spaced one week apart. The tasks used were: 1) a 'decision-making task' where learner-learner dyads were required to

agree on advice to give to the authors of 'Agony Aunt' letters, 2) a 'narrative task' where students narrated a story to their partner based on 10 frames from a comic strip containing no text, and 3) a 'personal task' where learners had to describe to their partner what had most pleasantly or unpleasantly surprised them during their study-abroad experience in Britain.

Forty participants were randomly assigned to one of four groups. Groups A and B comprised the 'no-planning' groups, with participants in Group B performing the tasks in reverse order from Group A to counterbalance any task-order effect. Participants in Groups C and D were provided with 10 minutes of planning time. Again, task order was reversed between these two groups. In addition, in each group participants were randomly assigned to be 'post-taskers' or 'non-post-taskers'. The 'post-taskers' were told before they began the task that they would have to perform it a second time in front of the whole class. Thus, the 'post-taskers' performed their initial task with the knowledge that they would have to do a public performance later. The 'non-post-taskers' performed the task knowing that they would not be doing a public performance later.

All performances were audio recorded and recordings were later transcribed and coded. Transcripts were analysed for level of complexity (measured as total number of clauses divided by total number of c-units), accuracy (measured by percentage of error-free clauses), and fluency (measured by number of pauses, where a pause was considered a break in speech of one second or longer).

Results showed that participants who were allowed to plan produced significantly more accurate speech than those who were given no planning time, and 'planners' paused significantly less frequently. 'Planners' also spoke with greater complexity, although only in the personal and decision-making task, not the narrative task. With regard to the 'post-taskers', it was found that this condition had no effect on levels of fluency and complexity, except in the narrative task where 'post-taskers' paused more often than 'non-post-taskers'. Accuracy was improved in

'post-taskers', however, only for the decision-making task. After examining speech performance results, the authors concluded that learners are unable to equally prioritise all three dimensions of CAF while completing a speaking task, and that performing more highly in one aspect of CAF comes at the expense of the others. Results indicated that competition between accuracy and complexity was particularly apparent.

In another study examining the effects of type of planning and task repetition on the oral performance of EFL learners, Ahmadian and Tavakoli (2011) assigned 60 participants to one of four groups. Four different experimental conditions provided the groups with either 'pressured' online planning (i.e. restricted time to complete the task) or 'careful' online planning (i.e. unrestricted time to complete the task) along with either the dispensation to repeat the task or not repeat the task. After an initial performance, those in the task-repetition groups were required to repeat the same task one week later. The task given to participants was to watch a fifteen-minute silent movie and then narrate the story of that movie to a partner (who had not seen it) under the conditions of their group. Participants in the 'pressured' online planning groups were given a maximum of six minutes to narrate their story to their partner, while those in the 'careful' online planning groups were given no time restriction.

Participants' performances were audio recorded, transcribed and analysed for complexity, accuracy, and fluency. Complexity was measured by the ratio of clauses to AS-units, and syntactic variety measured by number of different grammatical verb forms used. Accuracy was measured by percentage of error-free clauses and percentage of correctly used verb forms (in terms of tense, aspect, modality, and subject-verb agreement), and fluency by total number of syllables uttered per minute of speech and total number of meaningful syllables (i.e. total number of syllables minus repeated, reformed or replaced words).

'Careful' online planning was found to improve both accuracy and complexity in participants' speech performance compared to the performances of those in the 'pressured' online

planning groups. Pressured online planners produced more fluent speech than careful online planners, while those in the repetition groups were significantly more fluent than those in the no-repetition groups. Finally, the authors reported that those in the careful online planning group with task repetition showed more accurate, complex and fluent speech than would otherwise have been shown under each condition alone.

4.7.1 Effects of TR on accuracy in L2 speech performance

While the majority of TR studies examining effects on CAF have found facilitative effects for fluency, the effects on accuracy and complexity are less clear. Furthermore, there have been suggestions that TR could be detrimental to an L2 learner's interlanguage in terms of accuracy as learners have been found to carry over errors from one performance to the next, leading to the possibility of those errors becoming fossilised (Boers, 2014).

In one TR study, for example, Nation (1989) investigated the effects of the 4/3/2 activity on learners' speech performance. In the 4/3/2 activity, a speaker is asked to deliver the same talk on a given topic three times, each time to a different speaker, and each delivery in a reduced amount of time (four minutes then three minutes, and then two minutes). Several studies have found this technique (originally devised by Maurice, 1983) to result in improved fluency from one delivery to the next (Arevart & Nation, 1991; Boers, 2014; Thai & Boers, 2015; Nation, 1989). Despite claims that the technique also facilitates accuracy, Boers (2014) notes that some errors made in the initial performance by participants in Nation's (1989) study were corrected in iterations, however, others were preserved, while new errors emerged.

Other studies, too, have found no benefit for TR on accuracy (Ahmadian, 2011; Sample & Michel, 2014). Performance results for complexity are similarly mixed. There were no clear improvements made by participants in Nation's (1989) study in terms of the amount of subordination in their speech from one delivery to the next. More recently, de Jong (2012) also

found no benefit for TR on the complexity of participants' speech. These studies highlight the need for a better understanding of the conditions under which TR is, and is not, facilitative of complexity and accuracy.

Findings to date suggest that while TR fosters development of fluency, perhaps what is needed in order to foster complexity and accuracy is *enhanced repetition* (Lynch, 2018), that is, the opportunity to engage in some kind of reflection on the initial performance before delivering the repeat performance.

4.8 ENHANCED TASK REPETITION

Lynch (2018) outlines a number of different ways in which an intervention could be applied before/between repeat performances of a task, such as the provision of feedback, examining a transcript of one's initial performance, and exposure to model input. He uses the term 'enhanced repetition' to refer to a repeat performance of a task after engaging in some kind of cognitive reflection on an initial performance. The studies using an output-input-output cycle reviewed in Chapter 3 fall under the definition of enhanced task repetition.

A key to enhanced TR, as Lynch (2018) explains, is that it provides L2 learners with an opportunity to redo their initial performance of a task in a way that does not "seem to the learners to be simple repetition of what they did before" (p.196). Enhanced TR mitigates the chance of a learner finding the repeat performance of a task boring as it adds value to the iteration. While Sheppard and Ellis (2018) state that there have been very few empirical investigations into the effects of intervention between repeat performances of a task, the number of studies reviewed in Chapter 3 using an output-input-output approach to examine noticing in written modality suggest that this may not be the case. Perhaps it is more accurate to say that there have been very few studies to date that have empirically investigated the impact of an intervention between repeat performances of a *speaking* task on L2 *speaking* performance.

Lynch (2018) provides examples of activities based on the idea of enhanced TR run at an English Language Education Institute of a university. In the one activity, called 'free talk', ESL students were given the opportunity to discuss a topic of their choosing with a group. Following the initial discussion, students reflected on their speech performance (from memory) and were provided with feedback (by Lynch who worked in the institute) before repeating their discussion with a new group. Following the repeat group discussion, students were asked to report on linguistic problems they had during the initial performance that were overcome in the repeat performance

'Poster carousel' (Lynch & Maclean, 1994) is another enhanced TR activity and was originally designed for students studying in a medical English course. To begin this activity, students were assigned a partner and were asked to prepare a poster based on a research article they had been given. The posters were displayed around a large room. Partner A (the 'host') from each pair stayed with the poster and waited, while Partner B students (the 'visitors') were assigned to a particular poster (other than their own). 'Visitors' spent about three minutes at their assigned poster asking questions to the 'host', who was instructed to respond to questions only, and not to initiate discussion. Once the three minutes had expired, the 'visitors' rotated clockwise in order to visit a new 'host' (until they had visited all six poster hosts, excluding their own). Once Partner B students had visited all posters, they switched roles with their partner, and they then hosted visitors to their poster. Lynch and Maclean (2000, 2001) conducted studies investigating the efficacy of the poster carousal activity as a technique for improving speech performance. They found that while participants in their study reported that they felt their speech had not improved as they spoke with successive 'visitors' to their poster, analysis of transcripts of participants' speech showed otherwise, with participants improving in various aspects of performance including syntactic accuracy, lexical accuracy, and pronunciation.

An important point to note with Lynch's poster carousal activity, is that there is no external intervention before or during the task (Lynch, 2018). Instead, learners are pushed towards more accurate L2 speech performance as a result of repeated interaction and internal reflection on their performance. As mentioned above, the studies reviewed in Chapter 3 also fall within the definition of enhanced TR.

4.9 CHAPTER SUMMARY

In this chapter, a brief background of TBLT was provided along with an outline of what is meant by the term 'task'. The following sections looked at the three elements of speech usually used to measure performance in TBL studies, namely, complexity, accuracy, and fluency. This was followed by a look at some criticisms of CAF. An overview of literature that has examined how manipulating the design and implementation variables of various tasks can influence L2 learners' speech was presented before a focus on the design feature most relevant to the present study – task repetition. Task repetition was further narrowed down in a discussion on enhanced task repetition, which included the observation that few studies have looked at the impacts on speech performance of enhanced task repetition.

Having identified gaps in current literature, in the following chapters this thesis will detail the experiment undertaken in an attempt to fill these gaps.

5 Methodology

5.1 INTRODUCTION

In this chapter, the methodological approach of the current study is presented, beginning with an outline of the research design, the context in which the study took place, the participants and the instruments and materials used. This is followed by an explanation of data collection procedures and the dependant variables.

5.2 DESIGN

This study used a factorial repeated-measures design. Data collection took place in three phases: a pre-test phase, a training phase and a post-test phase. Data recorded in the pre-test phase (Time 1) provided a baseline from which data collected subsequently could be compared. The training phase included three recording sessions (Times 2, 3 & 4) where a different intervention was introduced depending on group. Finally, the post-test phase, which was designed to measure the impacts of the training phase, included two recording sessions (Times 5 & 6). Recording sessions were conducted weekly over a total of seven consecutive weeks, with the exception of week 6 when no data was collected. Three intact classes were randomly assigned to one of three conditions:

Control group (C) (n = 12)

Unguided noticing group (UN) (n = 12)

Guided noticing group (GN) (n = 12)

The pre- and post-tests (Times 1, 5 and 6) were conducted in identical fashion for all participants regardless of group. The training phase of the study (Times 2, 3 & 4) included noticing training sessions for the two experimental groups, where participants were provided with either an unguided noticing prompt (UN group) or a guided noticing prompt (GN group). Participants in the control group (C group) did not receive noticing training. Instead, their training sessions consisted of pronunciation practice (unrelated to the narrative task used in the session) utilising Issues In English 2 software (2005) already installed on student computers (see section 5.6.2 for further detail). Table 5.1 outlines the schedule of activities over the seven weeks, and Figure 5.1 shows an overview of the study design with steps involved at each time (see Section 5.6 for a detailed description of each step).

Table 5.1 - *Seven-week duration of study*

Week Group	1	2	3	4	5	6
C	Pre-test	Pron training	Pron training	Pron training	Post-test	Delayed post-test
UN	Pre-test	UN training	UN training	UN training	Post-test	Delayed post-test
GN	Pre-test	GN training	GN training	GN training	Post-test	Delayed post-test

Note: C = control, UN = unguided noticing, GN = guided noticing, pron = pronunciation

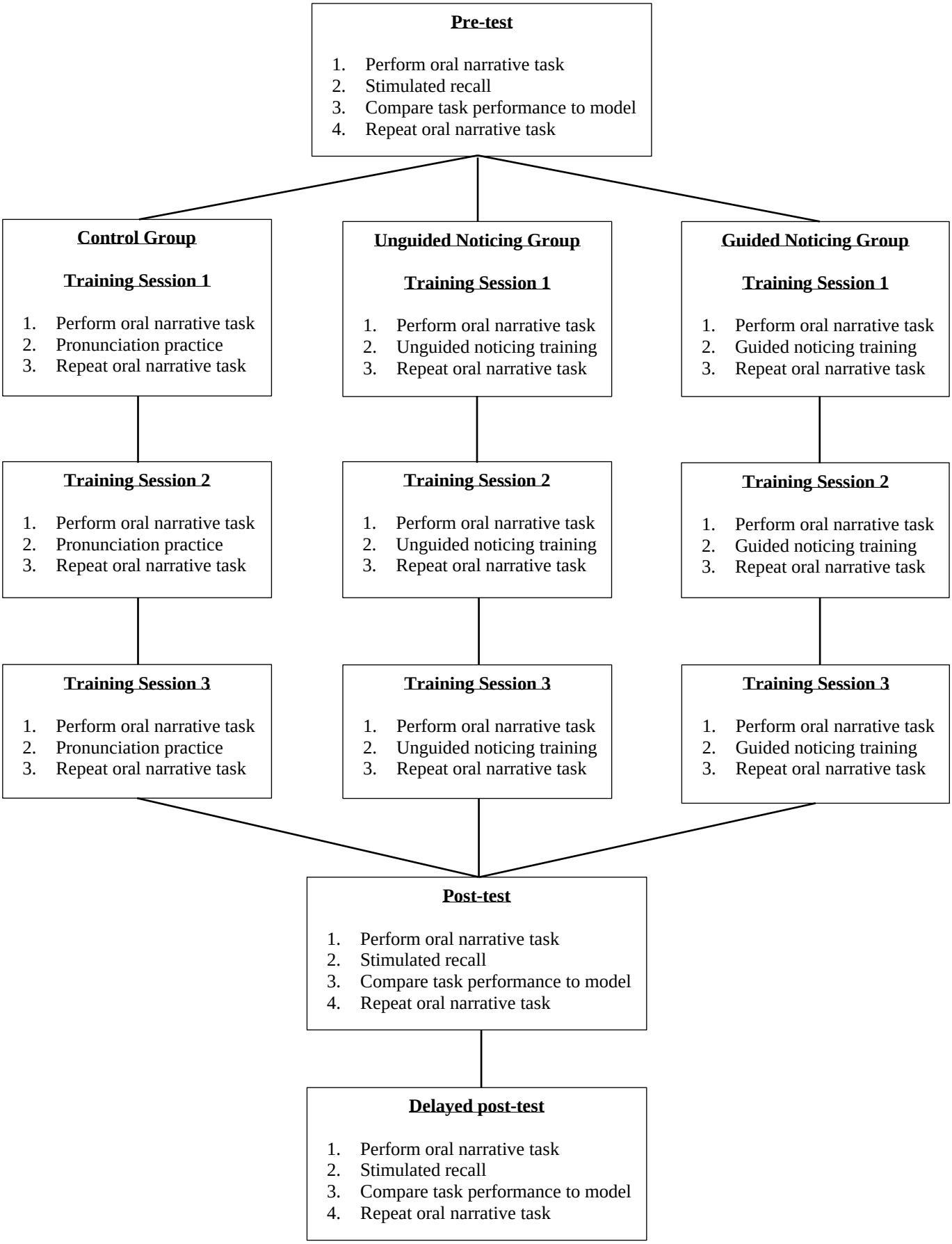


Figure 5.1 - *Overview of study design*

The pre-, post- and delayed post-test sessions were conducted one-on-one with each participant by the researcher. Training sessions were conducted with each group during participants' regularly scheduled class time (see sections 5.3 and 5.4 for more information on context and participants). Owing to the time needed to run one-on-one testing sessions, it was not possible to conduct the study with all three groups at the same time within the seven-week time frame. As a result, the C group and the UN group participated at the same time, while those in the GN group came from a separate intake of students and participated around three months later.

5.3 CONTEXT

This study took place at the English language centre of a university in Sydney, Australia. The language centre offers English Language Intensive Courses for Overseas Students (ELICOS) at four levels ranging from elementary to English for academic purposes (EAP). Each level comprises a ten-week course with 20 hours of classroom instruction per week and a maximum of 18 students per class. The vast majority of students enrolled at the language centre aim to matriculate into post-graduate programmes within the same university after successfully completing the EAP course. Successful completion of a course means achieving an overall passing grade comprising scores from in-class assessments (25%), mid-course exams (25%) and final exams (50%).

5.4 PARTICIPANTS

Participation in this study was completely voluntary. All participants were enrolled as students at the language centre and were at EAP level (from B2 to B2+ on the Common European Framework of Reference [CEFR]). Forty-three students initially agreed to participate, however, six were absent for one or more of the recording sessions and another had technical problems during one of the training sessions, which meant data recordings were lost. As a result, although

these seven students were allowed to continue their participation on subsequent occasions, their data were discounted from the study. Of the remaining 36 participants, 11 were male and 25 were female. This gender make-up reflected the student cohort of the centre at the time research took place. The age range of the participants was 22 to 36 years with an average age of 25.0 years. The first languages of participants included Chinese (16), Hindi (13), Nepali (3), Urdu (2), Bangla (1) and Thai (1). A summary of participant information can be found in Table 5.2.

Table 5.2 - *Summary of participants in the present study*

Characteristic		
Gender	Male	11
	Female	25
Age	Mean	25.0 years
	Range	22 - 36 years
L1 Background	Chinese	16
	Hindi	13
	Nepali	3
	Urdu	2
	Bangla	1
	Thai	1

Seventeen participants had successfully passed the upper-intermediate course at the language centre enabling them to enter the EAP program (i.e. they were existing students). The remaining 19 participants were newly arrived at the language centre and had been placed directly into EAP level based on scores from an in-house placement test and their recent IELTS scores. The length of time participants had spent in Australia varied. Existing students had arrived in Australia up to six months earlier, while newly arrived students had come to Australia as recently as two weeks prior to the commencement of the study. The relatively small pool of potential participants for this study meant that controlling for length of time in Australia was not possible.

At the language centre, students at any given level are assigned to classes by the course convenor with an attempt made, where possible, to include a mix of L1 backgrounds, gender, existing students and newly enrolled students in each class. Class A was randomly assigned to the Control condition, Class B to the Unguided Noticing (UN) condition, and Class C to the Guided Noticing (GN) condition. A breakdown of participants in each class is presented in Table 5.3.

Table 5.3 - *Summary of participants in each group*

Class	Participants		L1 background		Student status	
A	Male	3	Chinese	5	Existing student	6
	Female	9	Hindi	5	Newly arrived	6
			Nepali	1		
			Thai	1		
B	Male	5	Chinese	5	Existing student	5
	Female	7	Hindi	4	Newly arrived	7
			Urdu	2		
			Nepali	1		
C	Male	5	Chinese	6	Existing student	6
	Female	7	Hindi	4	Newly arrived	6
			Nepali	1		
			Bangla	1		

As mentioned above, using intact classes was necessitated by the decision to record participants during regularly scheduled class time in the three training sessions (Times 2, 3 & 4). This decision was made for three main reasons: Firstly, running intervention sessions with each group outside of scheduled class time would have been logistically difficult as some students have after-class commitments (e.g. a part-time job) and classrooms were often unavailable.

Secondly, although testing sessions at Times 1, 5 and 6 were conducted one-on-one with the researcher outside of scheduled class time, this was a time-consuming process and required

the booking of one of the university's consultation rooms for 3 to 4 hours a day. Consultation rooms are often in high-demand by university staff, therefore, running training sessions during class time minimised the interruption of the research on the day-to-day operation of the university by leaving consultation rooms free for others. Moreover, additional one-on-one sessions would have further impacted students' already busy schedules.

The final reason for conducting training sessions with intact classes was driven by the desire to see how task repetition, model input and noticing training could be used in an L2 classroom environment. This was vital if the results of this study are to inform L2 classroom teaching and learning practice. As a result of using intact classes, however, random assignation of participants to groups was not possible. Instead, as mentioned earlier, each of the three classes was randomly assigned to one of the three conditions.

Ethics approval was obtained for this research (see Appendix 1). All participants were given an information sheet outlining the study and signed a consent form to take part (see Appendix 2). Participants were made aware that they were taking part in research to investigate English language learners' speaking performance. They were told that their speech performance would be recorded, however, they were not told how their performances would be measured or analysed.

Because I am an English language teacher at the research site, it was stressed to participants that taking part in the study was completely voluntary, was not part of their EAP course, that neither the decision to participate or not to participate would have any impact on their assessments or final grades, and that they were free to withdraw from the study at any time without consequence. Any student in each of the three classes who chose not to participate still took part in the training sessions as the activities these sessions included (e.g. performing an oral narrative, pronunciation practice) were already part of the course syllabus. Their data, however, were not used for research purposes. In that way, within a class, any one student was unable to

identify another as being either a participant or a non-participant unless they chose to disclose the information.

5.5 INSTRUMENTS

In this section, a rationale for the task type employed in the present study is given along with a detailed description of the instruments used. Instruments consisted of six different picture sequences, six corresponding model-speaker recordings and three note-taking prompts (see Section 5.6.3 for information on recording equipment).

5.5.1 Rationale for type of task used in present study

Task type has been shown to influence speaking performance (Skehan & Foster, 1997). Each of the six separate tasks used in the research reported here were therefore designed to be of the same type in order to limit variability caused by task type when comparing participants' speaking performance over the course of the study.

Although there exist a range of possibilities for eliciting monologic speech from L2 learners, TBL researchers generally use one of three methods: 1) narration based on a picture sequence (de Jong & Vercellotti, 2016; Foster & Skehan, 2013; Fukuta, 2016; Gilabert, 2007; Tavakoli, 2009; Tavakoli & Foster, 2008; Yuan & Ellis, 2003), 2) narration based on a silent video (Ahmadian, 2012, 2013; Ahmadian & Tavakoli, 2014; Awwad, Tavakoli, & Wright, 2017; Ahmadian, Tavakoli, & Dastjerdi, 2015; Bygate, 1996; Skehan, Foster, & Shum, 2016), or 3) a question or statement to which participants are asked to respond (Arevart & Nation, 1991; Boers, 2014; De Jong & Perfetti, 2011; Nation, 1989; Tavakoli, 2016; Tavakoli, Campbell, & Mccormack, 2016; Thai & Boers, 2015).

Oral narratives based on a picture sequence (i.e. picture prompts) were used in the present study (see Appendices 3 – 8). One reason for this choice was that picture prompts are often used by TBL researchers. Thus, results from this study could be more reliably compared to those from

previous studies. In addition, as Yuan and Ellis (2003) point out, because oral narratives are monologic, they are not influenced by interactional variables. This was important as the focus in this study was on how noticing and exposure to model input influence speech performance. The additional influence of interaction would have made it difficult to ascribe improvements in performance to the intervention.

However, just as different task types can influence speaking performance, variations in characteristics of picture prompts also have an influence. For example, Tavakoli and Foster (2008) found that picture sequences with a tight narrative structure led to increased accuracy in participants' speech, and picture sequences with a complex storyline (in the form of background events in addition to foreground events in picture frames) led to more grammatical complexity. An attempt, therefore, was made to ensure the picture sequences used in this study were similar in terms of structure and complexity.

Originally each picture prompt contained eight frames. However, after piloting, it was found that some of the narrative tasks elicited quite short speech samples (some under 1 minute). As a result, either one, two or three frames were added to each prompt to lengthen the story, and thus, the amount of speech they elicited. In some frames, a small amount of text was added to provide contextual information including a label above the first frame on each picture sequence to indicate when the story took place (e.g. "last week"). Also, when piloting the prompts, it was found that the student-narrators had lengthy pauses at the beginning of each story as they tried to think of names for the characters. As a result, names were provided for the main characters in each prompt to help the narrators in this regard.

5.5.2 Picture prompts used in the present study.

After being given storylines written by the researcher, the six picture prompts for this study were drawn by a former colleague who had the requisite artistic talent. Each story was similar in that it had a clear beginning leading to a problem and then a final outcome. The picture

prompts used in this study can be found in Appendices 3 – 8 and a brief description of each is presented below.

Narrative task 1 (Ruined Dinner)

Narrative task 1 served as the pre-test and the basis for comparison of the experimental treatments. This prompt depicts the story of a couple inviting another couple over the phone to their house for dinner later in the week. The host couple then discuss what food they will prepare for the dinner before going to the supermarket to buy ingredients. On the day of the dinner, after preparing the food and putting it in the oven, unbeknown to the couple, the oven loses power when their pet dog steps on the electrical cord disconnecting it from the wall. At dinner time, just as their guests arrive, the hosts are shocked to discover their food is uncooked. In the end, they decide to order pizza for dinner instead.

Narrative task 2 (Hawaii Holiday)

Narrative task two was used in the first training session and took place in Week 2 of the study. The picture prompt for this task shows the story of a father surprising his wife and children with tickets for a family trip to Hawaii. After preparing for the trip and driving to the airport, the father discovers he has forgotten his passport. He leaves his family at the airport while he rushes home to collect his passport only to be stopped by the police for speeding. When he finally returns to the airport with his passport, he is too late to board the plane, and he watches through the window as the plane takes off with his family inside.

Narrative task 3 (Expensive Dinner)

Narrative task three took place in week three and served as the task for the second training session. This story shows a man meeting a woman and asking her on a dinner date later in the

week. The man then goes on a shopping spree with his credit card to buy new clothes, a new car and a new haircut in the hope of impressing the woman. On the day of the dinner date, the pair go to an expensive restaurant. After eating, the man attempts to pay for the dinner with his credit card, but it is declined. In the end, the woman reluctantly pays for the dinner with her credit card.

Narrative task 4 (No More Video Games)

Narrative four, used in the final training session in week four, shows the story of a student named Sara in her bedroom playing video games. Her mother knocks on Sara's bedroom door and asks her what she is doing. Sara lies, saying that she is studying for a test when in fact she is playing video games. Later at school, she scores poorly on a test. Her parents become angry when Sara reveals her low score. Her parents then confiscate her video games and tell her to study. After two weeks of studying every night, Sara scores well on the next test. Her parents are pleased and offer her video games back. However, Sara declines and says that she believes studying is more important.

Narrative task 5 (Wet Weekend)

Narrative task five was used as the post-test task in week five. This story shows a family trying to decide on a location for a weekend trip. The mother and the two children want to go to a beach, but the father disappoints his family by deciding that they will go on a camping trip in the forest instead. The father runs into problems at the camping site when the weather turns bad, he cannot put up the tent, a bear eats their food, and his family starts to complain. Frustrated, he tells his family that they can choose the location for the next trip. The following weekend the family can be seen relaxing on a beach.

Narrative task 6 (Wrong Day for a Meeting)

The final narrative task took place two weeks after the post-test (i.e. week 7 of the study) and served as the delayed post-test. This story shows Tim at home thinking about his plan for the next day. He thinks he has an important meeting at 9:30 am the following morning, so he plans the time he will wake up, eat breakfast, shower, get dressed and leave for the meeting. In the morning he is shocked when he realises he has overslept. He rushes to get ready but misses his bus. He then tries riding his bicycle to the meeting, but it has a flat tyre. He eventually takes a taxi, but when he arrives at the office, the lift is out of order, so he races up the stairs only to discover that he is a day early for his meeting.

5.5.3 Model Narratives

Model recordings were made for each of the six picture sequences. In an attempt to limit variability between model narrations, each was read by the same trained ESL-teaching colleague from a script. A further attempt was made to have general equivalency in speech rate, grammatical complexity and lexical complexity. To achieve this, each script (see Appendix 9) was analysed for degree of grammatical complexity by calculating number of clauses per AS-unit (Foster et al., 2000) and amount of lexical sophistication using Lextutor.com (Cobb, 2018). To achieve a similar speech rate in each narration, the total number of words in each script was determined and a target narration time was calculated in order to result in a speech rate of 100 words per minute. Actual results are displayed in Table 5.4.

Because the models were read from a script, they included no filled pauses (ums, ahs etc) and contained no grammatical errors. The lexical complexity measure shows the percentage of words in the model that fall within the 1,000 most commonly used words in English, the 2,000 most commonly used words, and the percentage of worlds that appear in the Academic Word List (Coxhead, 2000). Off-list words mostly included character names. Type-token ratios (TTR) were also determined. Although TTR can be influenced by text length (Kettunen, 2014), because

the text length of each model was quite similar (mean = 226.5 words, standard deviation = 23.2), TTR was deemed suitable in this instance.

Table 5.4 - *Characteristics of model narrations*

Measure / Narration	Duration (mins)	WPM	Grammatical complexity	Lexical complexity		Type-Token Ratio
1	2.48	108.72	1.41	1000 wds	76.01%	0.48
				2000 wds	7.01%%	
				AWL	1.11%	
				Off-list	15.87%	
2	2.28	95.47	1.48	1000 wds	77.68%	0.50
				2000 wds	6.70%	
				AWL	0.45%	
				Off-list	15.87%	
3	2.37	99.30	1.52	1000 wds	82.63%	0.42
				2000 wds	7.63%	
				AWL	2.54%	
				Off-list	7.20%	
4	2.03	102.30	1.54	1000 wds	87.62%	0.51
				2000 wds	3.81%	
				AWL	0.95%	
				Off-list	7.62%	
5	2.22	96.09	1.44	1000 wds	79.63%	0.52
				2000 wds	10.65%	
				AWL	0.93%	
				Off-list	8.80%	
6	2.20	97.73	1.51	1000 wds	85.25%	0.51
				2000 wds	8.76%	
				AWL	0.46%	
				Off-list	5.53%	

Note: WPM = words per minute, wds = words, AWL = Academic Word List, Off-list = words that do not fall in the 2000 most commonly used English words, nor in the Academic Word List.

5.5.4 Note-taking prompts

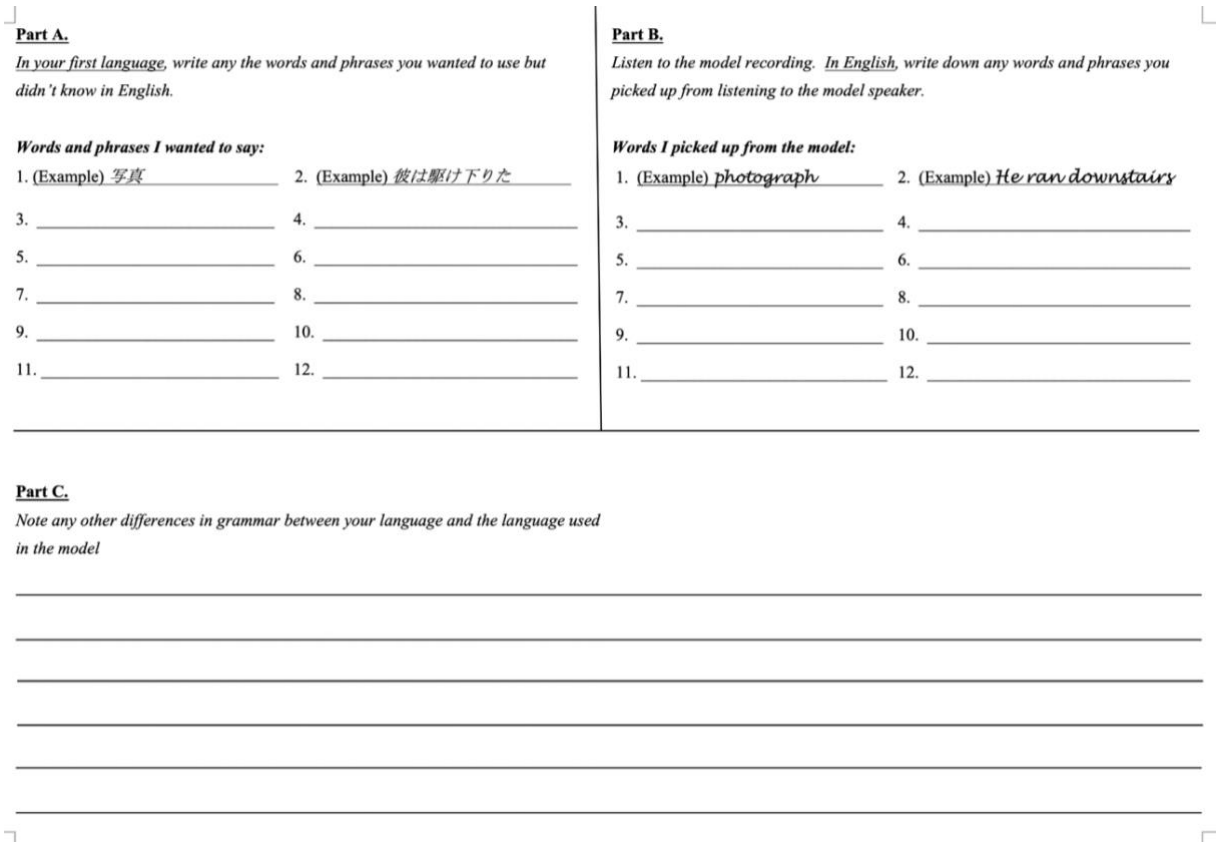
Three different note-taking prompts were used in the present study. Prompts 1 and 2 were used by the UN group and the GN group respectively in the three training sessions. Participants were required to use the prompts between performances of their oral narrative task, that is, after having delivered their first performance but before they were asked to deliver their repeat performance of the same task (see section 5.2.2 for details on training session procedure). Prompt 1, used by the UN group, (see appendix 10) consisted of lined paper with the following instruction at the top:

> *Listen to the recording of your story and the model speaker's story. While you listen, make notes of any language you think might be useful.*

Prompt 2, (see Figure 5.2 below and Appendix 11 for the full-size version) used by participants in the GN group during training sessions, included three parts: After delivering their first performance of the narrative task, Part A of the prompt asked participants to write in their L1 any words or phrases that they had wanted to use during delivery 1 of the narrative task but were not sure of in English. This was designed to encourage participants to notice IL gaps (i.e. to notice what it is they wanted to say in English but could not due to a lack of linguistic resources). Part B asked participants to listen to the model speaker's narration and write down any useful words or phrases they noticed. This part was designed to encourage participants to mine the model speaker narration for language to fill the gaps they identified in Part A of the prompt.

Finally, Part C of the prompt asked participants to note any differences in grammar in the language they used, and the language used by the model speaker. This final part of the prompt was designed to encourage participants to notice a difference between how they used language

to convey meaning, and how the model speaker used language to convey the same meaning. In other words, this part of the prompt encouraged participants to notice an interlanguage-target language gap (IL-TL gap).



Part A.

In your first language, write any the words and phrases you wanted to use but didn't know in English.

Words and phrases I wanted to say:

1. (Example) 写真 2. (Example) 彼は駆け下りた
3. 4.
5. 6.
7. 8.
9. 10.
11. 12.

Part B.

Listen to the model recording. In English, write down any words and phrases you picked up from listening to the model speaker.

Words I picked up from the model:

1. (Example) photograph 2. (Example) He ran downstairs
3. 4.
5. 6.
7. 8.
9. 10.
11. 12.

Part C.

Note any other differences in grammar between your language and the language used in the model

Figure 5.2 - *Note-taking prompt used by GN group in training sessions*

Finally, Prompt C was used by all participants, not in training sessions, but during the comparison stage of the testing sessions (see section 5.5.1 for further details on testing session procedure). This prompt consisted of lined paper with the following instruction at the top:

Listen to the two recordings. Make notes of any useful language

While prompts A and C remained unchanged from trialling and piloting through to the main study, the final design of prompt B was decided upon after trialling several versions. Earlier

versions required participants to write down the verbs they used, and the verbs used by the model speaker, but this proved to be a difficult and time-consuming task. Trialling also revealed other early versions of Prompt B to be too complicated and/or too long. The final version, and the version used in this study, was deemed suitable after successful trialling in a small pilot study.

5.5.5 Participants' familiarity with oral narratives and recording procedures.

The procedures used in both the testing and training sessions in this study involved a number of steps. Furthermore, in the training sessions, participants would be required to navigate software on their computer and record their oral narratives themselves. Because some of the steps involved would be unfamiliar, and therefore potentially confusing for participants, care was taken to ensure that the demands of performing an oral narrative combined with the procedural steps involved in the intervention sessions (e.g. using computer software, recording speech, saving files) did not negatively impact on their speaking performance. As noted in Chapter 2, TBL research to date has established that the cognitive demands of a task and changes to task conditions can impact upon allocation of attentional resources, which in turn can influence speaking performance. Because participants' speaking performance was to be measured over time in this study, it was important to gain a baseline in the pre-test that represented participants' true level of ability at that time. Also, because the same procedure would be used in all testing sessions (Times 1, 5 & 6), it was possible that participants would become familiar with task demands and procedural conditions as the study progressed. Becoming familiar with demands and conditions would lighten cognitive demands on participants, freeing up processing capacity which could be given to speech production. It was, therefore, important that participants were familiar with task demands and procedural conditions right from Time 1.

To minimise the impact that a lack of familiarity with the task requirements or procedures might have, two practice sessions were conducted with participants the week before data collection commenced. The purpose of the first practice session was to familiarise participants

with the task requirements (i.e. narrating a story based on a picture sequence), while the purpose of the second practice session was to familiarise participants with the procedural steps involved in the upcoming training sessions, namely, accessing the necessary software on the student computer, recording themselves performing a narration and saving their recording. The note-taking prompts were not used in the two practice sessions. Details of each practice session are presented below.

5.5.6 Practice Session 1

Practice Session 1 was held one day after participant recruitment had been finalised. Participants attended a 30-minute class, the purpose of which was to practice delivering an oral narrative based on a picture sequence. To begin the class, a picture sequence (see Appendix 12) taken from Sheppard (2006) was displayed on the projector screen at the front of the room for all participants to see. When asked if they had any prior experience telling a story based on a picture sequence like the one displayed, all participants reported that they had done so in previous English language classes. To set the story in time, participants' attention was drawn to the top of the first picture frame where a label stating "Last week" was printed. Participants were then instructed to look at the pictures and to discuss with a partner the story they thought the sequence told. After this, a recording of a model speaker telling the story was played for all participants to hear, and the script of the model narration was also displayed on the projector screen beside the picture prompt. The model speaker was an English language teacher at the university and the same model speaker used in the recordings for the main study.

Next, participants were told they were going to practice performing a narration based on a different picture sequence. To do this, each participant was assigned a partner. One student in each pair was given the role of 'speaker'. They were told they would narrate a story while their partner (the 'listener') would only listen. Once students' understanding of the requirements had been confirmed, the 'speakers' were handed a picture sequence (see Appendix 13) taken from

Heaton (1975). They were then given one minute to look at the pictures in order to gain a general understanding of the story depicted before beginning their narration. No corrective feedback was given to any speaker. Once completed, 'speaker' and 'listener' roles were reversed and a different picture sequence (see Appendix 14) also taken from Heaton (1975) was handed to the new 'speaker' and the procedure was repeated as before.

5.5.7 Practice Session 2

The second practice session was held in the computer room at the language centre one day after the first practice session. The purpose of this session was to allow participants to become familiar with the procedural steps that would be involved in the training sessions of the study. This session began with a demonstration, led by the researcher, of how to voice record using the university-owned headsets provided and Audacity recording software (Audacity Team, 2017) already installed on each computer. In addition, participants were shown how to save recordings in the appropriate folder. After the demonstration, participants were given five minutes to practice recording and saving their speech using Audacity. Existing students were already familiar with this process having used the software in classes at lower levels, so they were seated next to newly arrived students to provide assistance if needed. Two English teachers from the university helped monitor during these five minutes and assisted participants in using the software when necessary.

Next, students were told that they would have the opportunity to record themselves performing a narration based on a picture prompt similar to the one used the previous day in the first practice session. A new picture prompt, again taken from Heaton, (1975) was displayed on the projector screen at the front of the room, and participants' attention was again drawn to the label above the first picture frame which read "last week" and thus set the story in time. The same picture sequence was then handed out to each participant printed on A4 paper (see Appendix 15). Participants were given one minute to look at the pictures to gain a general

understanding of the plot and were then asked to record themselves telling the story before saving it in the designated folder on their computer desktop. At the end of this session, once it had been confirmed that all participants had successfully recorded and saved their oral narrative, all recordings were discarded.

After completion of the two practice sessions, all participants had heard and read a model for one narrative task and had the opportunity to perform two different narrations of their own. All participants had also successfully recorded and saved an oral narrative performance using computers in the university computer lab in the same manner that would be required in their upcoming training sessions. As a result, it was felt that participants had become familiar with the requirements of the task and task conditions that they would encounter during the data collection period (from Time 1 to Time 6).

5.6 DATA COLLECTION

Data were collected from three testing sessions (Times 1, 5 & 6) and three training sessions (Times 2, 3 & 4). Data collection procedures are explained in detail below.

5.6.1 Testing sessions

All participants took part in testing sessions at Times 1, 5 and 6 (weeks 1, 5 and 7 of the study). The purpose of the pre-test (Time 1) was to provide baseline data from which data subsequently collected in the post-test (Time 5) could be compared. The delayed post-test (Time 6) was designed to examine whether any gains made in noticing and/or speech performance from the pre- to the post-test were maintained. Testing sessions lasted around 20-25 minutes each. They were conducted one-on-one by the researcher with each participant from each group in a small office outside participants' regularly scheduled class time. During testing weeks (i.e. weeks 1, 5 & 7), six 30-minute timeslots were made available each day outside of class time

(12pm – 1pm & 2pm – 4pm, Monday – Friday). Participants were asked to sign-up for one of the 30 thirty-minute timeslots each testing week in order to complete their test.

All testing sessions were conducted in identical fashion with the exception of the picture prompt used, which was unique for each testing week. Testing sessions began with the researcher explaining the procedure to the participant. For the purposes of consistency, this explanation was read from a script (see Appendix 16) and when completed, the participant was given an opportunity to ask any clarifying questions. A short summary of the testing procedure was also presented to the participant on paper to outline each step (see Figure 5.3).

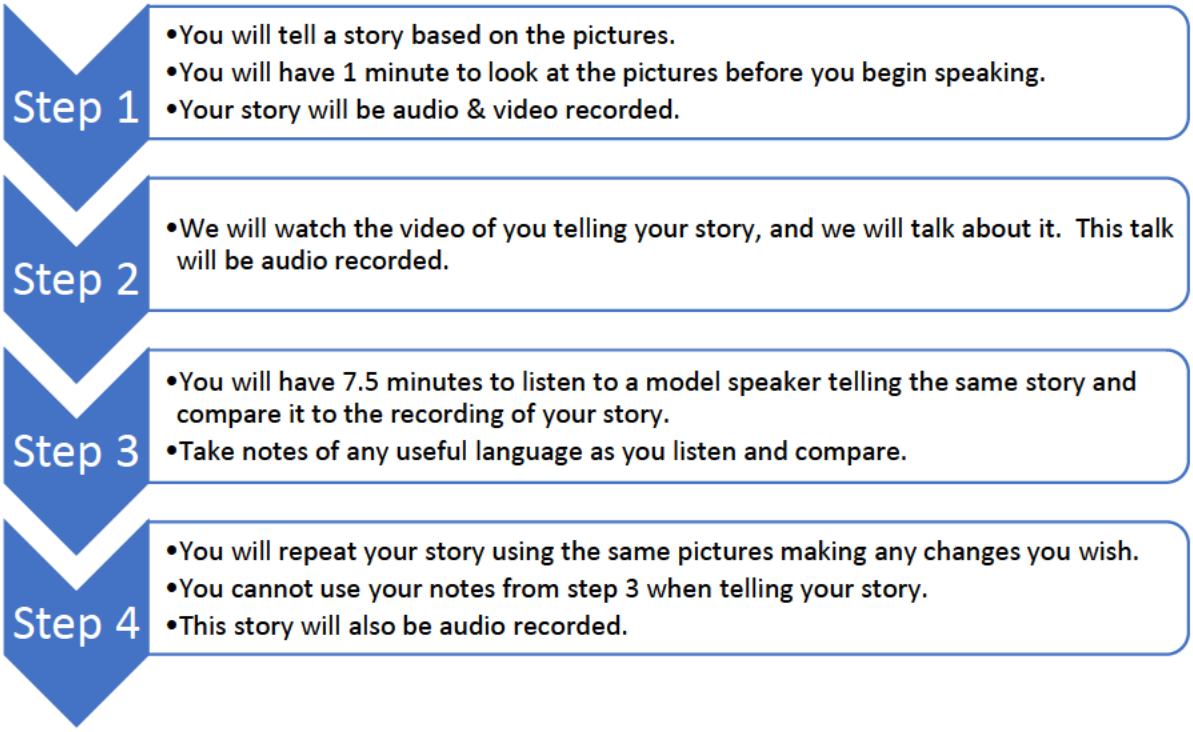


Figure 5.3 - *Summary of steps in tests*

After explaining the procedure and checking understanding, the participant was then given the picture prompt for the testing session printed on A4 paper. The order of picture prompts used over the course of the study was decided at random before the study began. Participants were given one minute to preview the story during which time note-taking was not permitted. After

the minute had expired, the participant was asked if they felt they understood the story depicted. On the rare occasions that participants were unsure, they were allowed to ask clarifying questions. In Testing Session 1, for example, a participant asked, "Are they a couple?", referring to the two main characters, and another participant asked, "Is Jack the father?" in another testing session. Answers to contextual clarification questions such as these were provided. On a small number of other occasions, participants pointed to something in a picture frame and asked, "What's this in English?", however, no assistance with regard to grammar or lexis was provided. Instead, participants were told, "just do your best to tell the story".

Next, the participant was asked to tell the story shown in the picture sequence (delivery 1). Each participant's first delivery was audio recorded using a small digital voice recorder and video recorded (including audio) on an Apple Macbook Air laptop computer. No assistance was given to the participant while delivering their narration. Once the participant had finished, the recording was stopped on both devices.

Step 2 consisted of a stimulated recall (SR) session (see section 5.9.1 for more on SR methodology). The aim of the SR session was to "elicit data about the thought processes involved in carrying out the task" (Gass & Mackey, 2001, p.1). The SR began when the following script (adapted from Sheppard, 2006) was read to the participant:

> *Next, let's watch the video recording of your story. I know what you said in your story, however, I'm interested to know what you were thinking while you were speaking. So let's watch the video, and if you remember what you were thinking at any time, you can press the spacebar to pause the video and tell me. Also, I might pause the video to ask you some questions as well.*

Once the participant's understanding of the SR procedure had been confirmed, the video was started, and the participant was asked to push the spacebar in order to demonstrate that they knew how to pause the video. The video was then restarted. Participants were free to pause and comment at any stage. In addition, the researcher paused the recording when the participant appeared to be encountering a problem in telling their story. Problems were signalled by a lengthy or unnatural pause (either filled or unfilled), or by some other form of dysfluency such as a false start, reformulation or repetition. Below is an example of an interaction between the researcher (R) and a participant (P) during a delayed post-test SR session initiated by a trigger (an apparent linguistic problem encountered by the participant).

Trigger: *The uh, he uh, he he take his clothes quickly, uh, but uh there is some problem uh, so he's not dress well, uh not good dress.*

R: *Here you paused a little bit, do you remember what you were thinking at that time?*

P: *Yeah, problem here due to uh the problem, I wasn't sure how uh, I want to say he dressed up a little bit dirty or something, but dirty is not the correct word, uh, maybe 'untidy'.*

Following the SR, the testing session moved to the comparison stage (i.e. step 3). To begin this stage, the participant was told that they would have 7.5 minutes to listen to the audio recording of their story and compare it to a recording of a model version of the same story. The decision to allocate 7.5 minutes was made after trialling and piloting, during which time it was found that allowing more than 7.5 minutes resulted in some learners trying to dictate word-for-word from the model recording, whereas allowing fewer than 7.5 minutes did not provide some learners with enough time to complete their comparison.

For this comparison stage, participants were asked to take notes of any useful language on the paper provided, and they were also reminded at this point that after comparing the two recordings, they would be asked to repeat their narration a second time (delivery 2). This reminder provided encouragement for participants to mine the model narration for language to fill the gaps they had identified in the SR session and to notice differences in how they used language to convey meaning and how the model speaker used language to convey the same meaning. Had participants not been reminded that they would be required to perform the narration again, they may not have felt a need to seek language from the model input to fill the gaps they had noticed.

Before the comparison stage began, the two recordings were opened in separate windows on a Dell desktop computer in the testing room. The participant was told that they could play, pause, stop, rewind, fast-forward and change between recordings as often as they liked during the 7.5 minutes.

The participant was then asked to demonstrate that they knew how to control the playback of the two recordings. Once this had been confirmed, note-taking paper was provided (see Appendix 17) along with a headset for listening. A small digital countdown timer was placed on the desk in front of the participant to display the time remaining in the comparison stage. In order to be less intrusive, the researcher moved to another desk in the office while the participant was listening to recordings.

At the conclusion of the comparison stage, the participant was given one minute to review their notes before they were taken away by the researcher. The test then moved to the final stage where instructions for the participant's second narrative performance were read from the following script:

Now I'd like you to tell your story again. If you want to change anything from your first story, you can. I'll audio record this story, but it won't be video recorded. When you're ready, please begin

After the participant had finished their second narration, the audio recording was stopped, and the testing session ended.

5.6.2 Training sessions overview

Three training sessions were conducted with each group separately and were designed to introduce the different interventions depending on group. Training sessions were held weekly on the same day for each group in weeks 2, 3 and 4. Training sessions lasted around 20 minutes and included three stages regardless of group: 1) participant's first delivery of an oral narrative task, 2) a training stage, and 3) participant's second delivery of the same oral narrative task. All groups used narrative tasks 2, 3 and 4 (see Appendices 4, 5 and 6) in training sessions 1, 2 and 3 respectively. The conditions of the training stage, which took place between deliveries in each training session, differed depending on group. For the GN group, the training stage included use of a guided noticing prompt (see Appendix 11) designed to encourage participants to notice gaps in their interlanguage, and to direct their attention to the linguistic form of their oral output and of the model narrative input. Training for the UN group included use of an unguided noticing prompt (see Appendix 10) which was designed *not* to direct participants' attention to any particular aspect of their oral output nor any particular aspect of the language used in the model narration input. In other words, they were free to attend to whichever aspects of their output and of the model input that they liked. The control group received no noticing training; instead they took part in pronunciation practice in their training stage using Issues in English 2 software (2005) already installed on student computers. The pronunciation practice was unrelated to the oral narrative tasks and required participants to select a word from a list which was then presented

in a sentence. Participants could then listen to a recording of the sentence being read by a model speaker and record themselves reading the sentence for comparison.

Each training session was run by the researcher during regularly scheduled class time in the computer lab at the research site. Another ESL staff member was also present to assist with any issues (e.g. technical problems). The computer lab consisted of a room with 25 computers for students plus a central computer for the teacher. The latter included NetSupport (2017) management software which enabled the teacher to take over operation of student computers remotely. This meant the teacher was able to 'lock' student computers (which meant they could not be operated by students) as well as deliver and collect files electronically.

Training sessions began when participants were assigned to a computer in the room. Care was taken to leave as much space as possible between participants in order to reduce possible distractions and limit background noise in recordings. As mentioned above, all participants, regardless of group, completed three stages in the training sessions: 1) delivery 1 of the oral narrative task, 2) a training stage, and 3) delivery two of the same oral narrative task. Stage 2 (the training stage) differed according to group. Details of the training sessions for the experimental groups (GN & UN) and the C group are outlined in more detail in the following section.

5.6.2.1 Training session procedure for UN and GN groups.

GN and UN groups took part in training sessions separately, however, the procedure was identical with the exception of the note-taking prompt used during the training stage. Sessions began when students were each seated at a computer and computers were 'locked' remotely via the teacher's computer. The researcher then explained the session procedure to the class with the aid of a table summarising the steps involved which was presented on the projector screen at the front of the room (see figure 5.4).

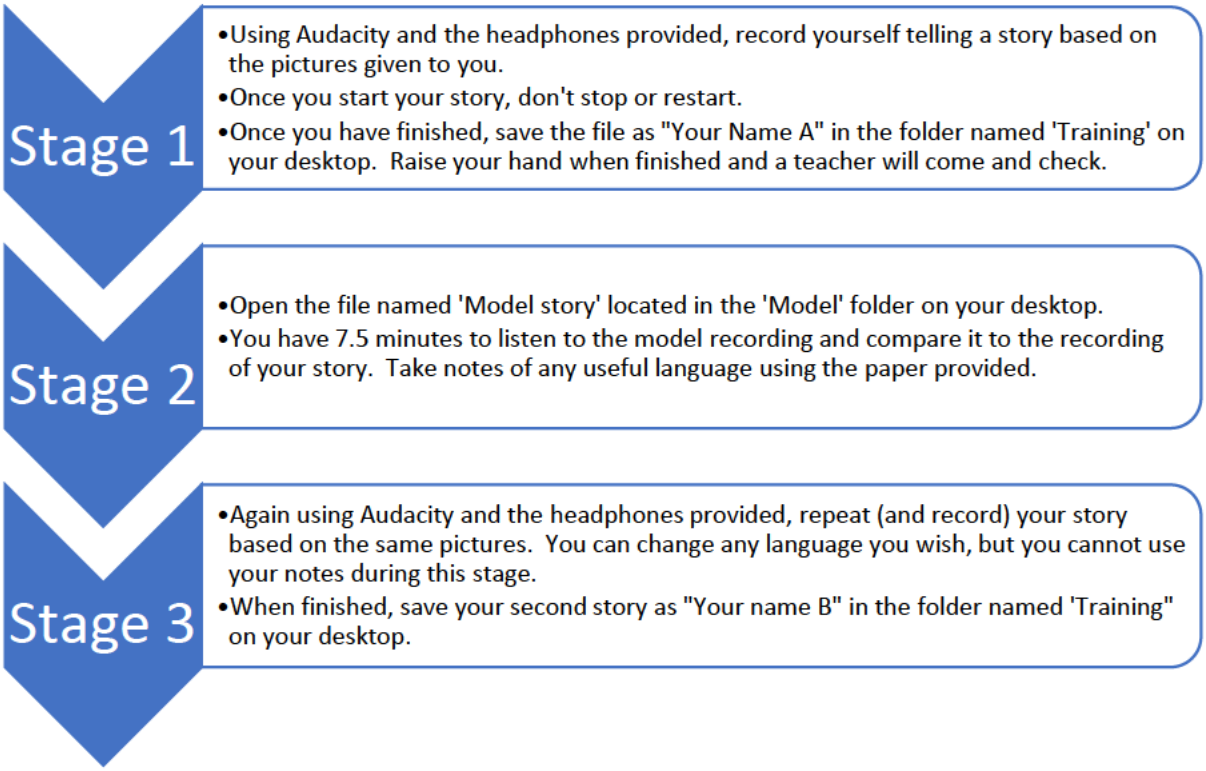


Figure 5.4 - *Summary of training procedure for participants in experimental groups*

Once the procedure had been explained, participants were given the opportunity to ask questions so that any uncertainties could be clarified. Next, a printout of the picture prompt for the session was handed out to participants. They were given one minute to preview the picture sequence without taking any notes. An opportunity was given for participants to ask any clarifying questions about the picture sequence; however, as with testing session procedure, while answers to questions seeking contextual clarification were given, any questions related to lexis or grammar were not provided. Instead, participants were instructed to "just do your best to tell the story". The researcher then 'unlocked' the student computers, and participants were asked to record their narration and save it in the designated folder when finished.

After finishing their narration and saving it to the specified folder, participants raised their hand and either the researcher or the assisting teacher checked to ensure the file had been saved

correctly. Upon confirmation, each participant's computer was 'locked' again until all had finished and saved their recordings.

Next, Step 2 was conducted in identical fashion for participants in the UN and GN groups except that the UN group used note-taking prompt A (the unguided noticing prompt – see Appendix 10), while the GN group used note-taking prompt B (the guided noticing prompt – see Appendix 11). Before starting this step, participants in each group were reminded they would listen to the recording of their story and a recording of a model speaker telling a story based on the same picture sequence. They were advised that they would have 7.5 minutes to listen to both recordings, and that they could stop, play, rewind, fast-forward and switch between recordings as often as they liked. They were also asked to take notes on the note-taking paper given to them (i.e. prompt A for the UN group, prompt B for the GN group). On the projector screen, the researcher modelled where participants could access both recordings. Note-taking prompts were then handed to each student, and their computers were then 'unlocked'. The researcher and the assisting teacher checked to ensure that all participants had successfully opened both the audio file of the model narrator recording and the audio file of their own recording. Once confirmed, the 7.5-minute comparison stage began and a countdown timer on the projector screen displayed the time remaining.

After the 7.5-minute comparison time had expired, participants were given one minute to review their notes before they were collected by the teacher. The final step (delivery 2) then began with the researcher reminding the participants they would now retell their story and record it in the same fashion as in Step 1. The training session ended once the researcher and teacher had confirmed that each participant had successfully recorded and saved the file of their second oral narrative performance in the correct location. These files were collected electronically via the teacher's computer, saved to a password-protected folder on Google Drive before being deleted from all computers in the room.

5.6.2.2 Training session procedure for C group.

The training session procedure for the C group also included three steps. Steps 1 and 3 were identical to those of the UN and GN groups and were explained in the same way. The 3 steps for the C group were displayed for participants on the projector screen (see figure 5.5).

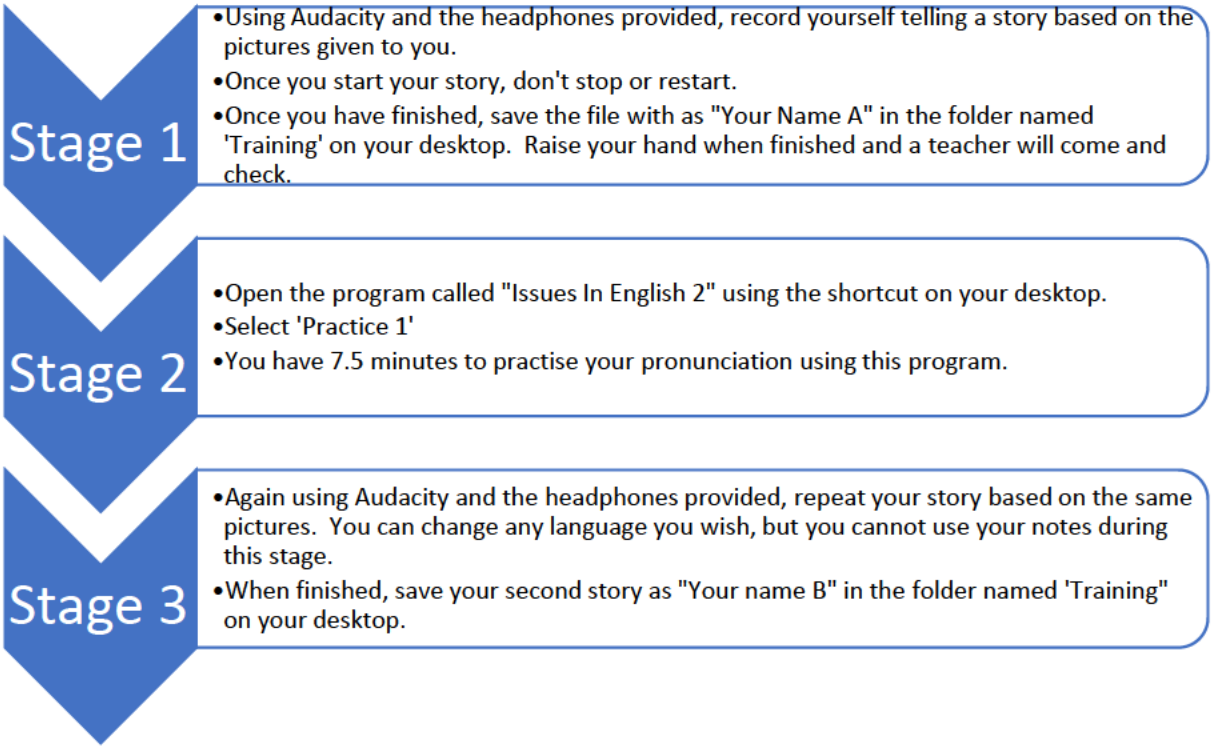


Figure 5.5 - *Summary of training procedure for participants in the control group*

After completing Step 1 in the same manner as the UN and GN groups, participants in the C group were reminded that they now had pronunciation practice on their computers (Step 2). The researcher modelled on the projector screen using the teacher's computer where the software could be found. Although participants were already familiar with this software having used it in previous classes, the focus of the activity was different from that of previous sessions in that a different activity within the program was used. These activities (different on each training session occasion) focused on pronunciation of a pre-selected phoneme unrelated to the oral

narrative task. Following Step 2, Step 3 (repeat performance of the oral narrative task) was conducted in the same manner as Step 3 for the UN and GN groups.

5.6.3 Recording equipment

In the testing sessions, audio recording was done using a small digital voice recorder placed on the table in front of the participant. Video recording was done on an Apple Macbook Air laptop computer using Quicktime software already installed. In training sessions, participants recorded themselves using Logitech headsets and Audacity recording software preinstalled on each computer in the university computer lab.

5.7 RESEARCH QUESTIONS, HYPOTHESES, AND DEPENDENT VARIABLES

This study was guided by three broad research questions:

1. *Can learners be trained to notice gaps in their output?*
2. *How does noticing training influence subsequent incorporation of input?*
3. *What are the impacts (both immediate and long-term) of noticing training and exposure to model input on learners' speech performance?*

Each of these broad research questions then led to more specific research questions and hypotheses for each as the following table shows.

Table 5.5 - *Research questions and hypotheses guiding the present research*

Broad Research Question	**Specific Research Question**	**Hypothesis**
1. Can learners be trained to notice gaps in their output?	1a. Does the number of form- and/or meaning-related IL gaps that learners notice change following noticing training (i.e. from pre-test to post-tests)?	1a. Following training, the number of form-related IL gaps noticed will increase for the GN group but not for the UN or C groups.
		1b. Following training, the number of meaning-related IL gaps noticed will decrease for the GN group, increase for the UN group, and remain unchanged for the C group.
	1b. Does the number of IL-TL gaps that learners notice change following training intervention?	1c. Following training (i.e. in post-tests), the number of IL-TL gaps noticed will increase for the GN group to a larger extent than for the UN group, who in turn will outperform the C group.
2. How does noticing training influence subsequent incorporation of input?	2. In testing sessions, what percent of solvable IL gaps noticed in delivery 1 are filled in delivery 2 after exposure to model input, and how does this change following intervention?	2. Following training (i.e. in post-tests), after noticing IL gaps in their first delivery, the GN group will incorporate language from subsequently presented model input to fill a higher percentage of solvable gaps during their second delivery compared to the UN group. The C group will not change from pre- to post-tests.
3. What are the immediate impacts of noticing training and exposure to model input on learners' speech performance?	3. What changes in CAF occur in learners' speech performance from delivery 1 to delivery 2 following intervention?	3a. Following training (i.e. in post-tests), compared to the UN and C groups, the GN group will speak with greater accuracy in delivery 2 than in delivery 1 of their narrative task.
		3b. Following training, compared to the UN and C groups, the GN group will speak with greater complexity in delivery 2 than in delivery 1 of their narrative task.
		3c. Following training, compared to the GN group, the C and UN groups will speak with greater fluency in delivery 2 than in delivery 1 of their narrative task.
4. What are the long-term impacts of noticing training and exposure to model input on learners' speech performance?	4. What changes in CAF occur in learners' speech performance from delivery 1 of the pre-test to delivery 1 in each post-test?	4a. Compared to the C and UN groups, the GN group will speak with improved accuracy and complexity from delivery 1 of the pre-test to delivery 1 of the post- and delayed-post tests.
		4b. Compared to the GN and UN groups, the C group will speak with improved fluency from delivery 1 of the pre-test to delivery 1 of the post- and delayed-post tests.

In order to answer the specific research questions, a number of dependent variables were used to measure two broad areas: 1) participants' speech performance, and 2) noticing.

Dependent variable measures of speech performance targeted three dimensions of oral output: grammatical complexity, grammatical accuracy, and fluency (CAF) (see Chapter 4 for a discussion on CAF including definitions and methods of measurement). Noticing was further divided into 'noticing an IL gap' (i.e. problems noticed during the speech production process when a learner realises they lack the linguistic resources to express what it is they want to say) and 'noticing an IL-TL gap' (i.e. problems a learner notices when later comparing their speech to that of a model speaker) (refer to Chapter 2 for a discussion on the different types of noticing). A breakdown of the three components of CAF and the two types of noticing is displayed in figure 5.6.

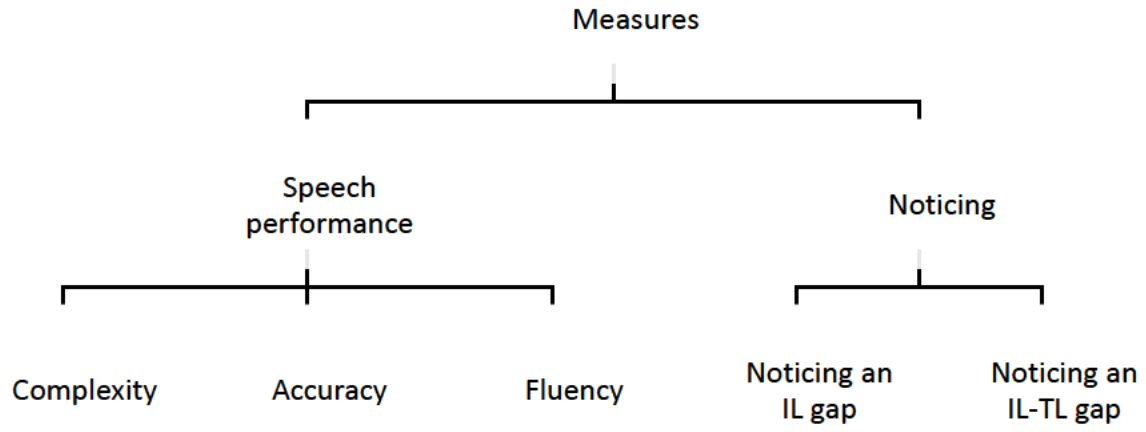


Figure 5.6 - *Dependent variables used in the present study*

5.7.1 Measuring speech performance - CAF

As explained in Chapter 4, the three components of CAF have provided the standard dependent variable measures of speech performance in TBL research for several decades and "all

three must be considered if any general claims about learners' L2 performance and proficiency are to be made" (Housen et al., 2012, p. 3). An issue presents itself, however, when deciding what measures of each of the CAF elements to use as "the sheer number of CAF measures currently available is somewhat daunting" (Housen et al., 2012. p. 8). Although a discussion on CAF, including definitions and methods of measurement was presented in Chapter 4, the three constructs are briefly reviewed again below together with a more detailed explanation of how each was measured using examples from the present study where necessary.

5.7.2 Accuracy

Of the three dimensions of CAF, accuracy has been labelled the simplest (Pallotti, 2009), and the most transparent and consistent (Housen & Kuiken, 2009). When measuring accuracy, researchers generally look at the presence (or absence) of errors in relation to target language norms (Housen & Kuiken, 2009; Skehan, 1996; Wolf-Quintero, Inagaki & Kim, 1998).

In the present study, following other TBL researchers (See Chapter 4) accuracy was calculated using three separate measures: 1) number of errors per 100 words, 2) number of self-repairs per 100 words, and 3) percentage of errors self-repaired. All errors in syntax, morphology and lexical choice were counted, but errors in pronunciation and intonation were not considered as they are not included in the model of speech performance measurement used in this study. Each measure of accuracy is briefly outlined below.

5.7.2.1 Number of errors per 100 words.

In order to determine number of errors per 100 words, errors were marked on participants' transcripts, totalled and then divided by the total number of words delivered including dysfluencies such as false starts, repetitions and reformulations (i.e. unpruned speech) (Lennon, 1990). The result of this calculation was then multiplied by 100 to arrive at total number of errors per 100 words.

5.7.2.2 Number of self-repairs per 100 words.

Although some studies include self-repairs as a measure of (dys)fluency, for the purposes of this study, and following de Jong and Vercellotti (2016), Gilabert (2007) and Michel, Kuiken and Vedder (2007), self-repairs were considered a measure of accuracy as they indicate an awareness of form (Ellis & Barkhuizen, 2005). A self-repair was defined as "changing and reformulating a phrase or linguistic unit previously uttered for the purpose of correction", (Tavakoli, Nakatsuhara & Hunter, 2017, p.54). It should be noted that only 'correct' self-repairs were counted in this measure. Self-correction attempts that included an error(s), as can be seen in the following example, were not counted (but were counted as an error):

...but then he forget, uh no uh, then he was forget his passport

Because the attempted self-correction in the above example was erroneous, it was not counted as a self-correction. It was, however, counted as a reformulation (see below for a definition). The example above, therefore, includes two errors (forget, was forget) and one reformulation.

5.7.2.3 Percentage of errors self-repaired.

This measure was defined as the number of self-repairs as a percentage of the total number of errors committed (Wigglesworth, 1997). In order to calculate this, each participant's total number of self-repairs was divided by their total number of errors and then multiplied by 100.

5.7.3 Complexity

For nearly two decades, number of clauses per AS-unit (Foster et al., 2000) has been the most commonly adopted measure of grammatical complexity in TBL research. It was therefore used in the present study to allow for results to be compared to existing TBL studies. An AS-unit is defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal

unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365). AS-units were determined by following the guidelines laid out in Foster et al. (2000) which include details on how to handle dysfluencies such as false starts, repetitions and self-corrections.

5.7.4 Fluency

As discussed in Chapter 4, of the three dimensions of CAF, fluency is arguably the most controversial. Although it is a multi-level, multidimensional construct (Lahmann, Steinkrauss & Schmid, 2017), the level the present study is concerned with is utterance fluency (Segalowitz, 2010) which relates to temporal measures of speech (Huensch & Tracy-Ventura, 2017).

While a single agreed-upon definition of fluency remains elusive in L2 literature, the definition adopted here is "the ability to produce the L2 with nativelike rapidity, pausing, hesitation, or reformulation (Housen, Kuiken, & Vedder, 2012, p. 2). In order to gain a more detailed picture, Skehan (2003) believes fluency needs to be separated into three sub-dimensions; speed fluency (e.g. speech rate), breakdown fluency (e.g. pauses and hesitations) and repair fluency (e.g. self-repairs and reformulations). Five measures covering these three aspects were used in this study (see below), and more detail about each is given in the following section:

1. number of words uttered per minute
2. number of silent pauses per 100 words
3. mean length of silent pause (in seconds)
4. number of filled pauses per 100 words
5. number of reformulations per 100 words

5.7.4.1 Speed Fluency

As the name suggests, speed fluency is the measure of the speed of one's speech (Revesz, Ekiert & Torgersen, 2014). Speech rate is generally acknowledged as an important global

measure of fluency. Number of words uttered per minute (WPM) was used as a measure of speed fluency in this study and calculated by taking total number of unpruned words uttered (i.e. all words including dysfluencies such as repetitions and reformulations) divided by total time taken to utter those words (in minutes). Although some studies use pruned speech when analysing speech rate (i.e. speech with dysfluencies such as false starts and repetitions removed), this is more common in studies with participants at a lower-level of proficiency because such learners often employ strategies such as repetition of words and phrases in an attempt to come across as being more fluent (R. Gilabert, personal communication, July 17, 2018). Because the participants in this study were upper-intermediate – advanced level, using unpruned words per minute was deemed appropriate.

5.7.4.2 Breakdown Fluency

The following three measures of breakdown fluency were used in the present study:

Number of filled pauses per 100 words.

This was calculated by taking the total number of non-lexical interjections (e.g. hmm, um ah etc) during a participant's narrative delivery divided by the total number of words in that delivery, and then multiplying by 100.

Number of silent pauses per 100 words.

This measure was determined by calculating the total number of pauses longer than 200 milliseconds that were not interrupted by any sound, divided by the total number of words in the participant's oral narrative, and then multiplying by 100. The amount of time used to determine the cut-off for a silent pause has varied considerably in L2 research, with a cut-off of 100 milliseconds at the low end, and 1,000 milliseconds at the high end (de Jong & Bosker, 2013). The decision to take 200 milliseconds as the cut-off for this study was made following de Jong and Perfetti (2011) who cite research by Lennon (1990) which found that pauses longer than 200 milliseconds sound dysfluent.

Mean length of silent pause.

This was determined by calculating total silent pause time (i.e. adding together all silent pauses longer than 200 milliseconds) during a participant's delivery, and then dividing by the total number of silent pauses in that delivery.

5.7.4.3 Repair Fluency

The following two measures of repair fluency were used:

Number of reformulations per 100 words.

A reformulation was defined as "modifying/reformulating a linguistic unit that has been uttered" (Tavakoli et al., 2017, p. 54). The number of reformulations per 100 words was calculated by taking the total number of reformulations during a participant's delivery of an oral narrative task divided by the total number of words in that delivery, and then multiplying by 100.

Number of repetitions per 100 words.

A repetition was defined as "exact repetition of a word or phrase previous uttered" (Tavakoli et al., 2017, p. 54). Number of repetitions per 100 words was determined by taking the total number of repetitions during a participant's delivery divided by the total number of words uttered in that delivery, and then multiplying by 100.

A summary of the dependent variable measures of CAF along with definitions and calculations used is shown in table 5.6.

Table 5.6 - *Speech performance measures, definitions and calculations used in the present study*

Dimension	Measure	Definition	Calculation
Complexity	Amount of subordination	Total number of clauses per AS-unit where an AS-unit was determined following guidelines set out in Foster et al (2000).	Total number of clauses divided by total number of AS-units
Accuracy	Frequency of errors	An error included any error in syntax, morphology and lexical choice (errors in pronunciation and intonation were not considered).	(Total number of errors / total number of words) * 100
	Frequency of self-corrections	"The act of changing and reformulating a phrase or linguistic unit previously uttered for the purpose of correction" (Tavakoli et al., 2017, p.54).	(Total number of self-corrections / total number of words) * 100
	Percentage of errors self-corrected	Total number of self-corrections (as defined above) expressed as a percentage of total errors.	(total number of self-corrections / total number of errors) * 100
Speed Fluency	Number of words per minute (WPM)	Total number of words (including repetitions, reformulations, repairs) divided by the total time taken to utter them (in mins).	Total number of words / total time in minutes
Breakdown Fluency	Frequency of filled pauses	"A pause filled with non-lexical interjections such as hmm, uh etc" (Tavakoli et al., 2017, p.54).	(Total number of filled pauses / total number of words) * 100
	Frequency of silent pauses	A pause of 200 milliseconds or longer that is not interrupted by any sound.	(Total number of silent pauses / total number of words) * 100
	Mean length of silent pause	The average length (in seconds) of all silent pauses of 200 milliseconds or longer.	Total silent pause time / total number of silent pauses
Repair Fluency	Frequency of reformulations	"Modifying/reformulating a linguistic unit that has been uttered" (Tavakoli et al., 2017, p. 54).	(total number of reformulations / total number of errors) * 100
	Frequency of repetitions	"Exact repetition of a word or phrase previous uttered" (Tavakoli et al., 2017, p. 54).	(total number of repetitions / total number of errors) * 100

5.8 MEASURING NOTICING

In the present study, measurement of noticing took place during the three testing sessions at Times 1, 5 and 6. As outlined in Section 5.6.1, each testing session included the following four steps:

Step 1: *Delivery 1.* The participant delivers their first performance of an oral narrative task based on a given picture sequence.

Step 2: *Stimulated recall session.* The participant and researcher watch a video recording of the participant's first delivery of the narrative task. The recording is stopped by the researcher when the participant appears to be encountering a problem in performing the oral narrative task (in the recording). The participant is asked if they recall what they were thinking at the time the apparent problem was encountered. Following the participant's explanation (if any), playing of the recording is then continued. The video is stopped, and problems are recalled by the participant in this way until the end of the recording. Participants are also free to stop the video and comment at any time.

Step 3: *Comparison stage.* The participant listens to a recording of their first delivery and compares it to a model recording of the same task. The participant takes notes of any useful language as they listen. The participant can stop, pause, rewind, fast forward and switch between the two recordings as often as they like within the 7.5-minute time limit for this stage.

Step 4: *Delivery 2.* The participant performs the same oral narrative task a second time. The notes taken during the previous stage are not permitted to be used in this final stage.

Noticing an IL gap was measured during the stimulated recall stage (step 2) of each testing session while noticing an IL-TL gap was measured during the comparison stage (step 3). Before going into more detail about how these forms of noticing were measured, the following sections provides a rationale for SR as a methodology for measuring noticing.

5.8.1 Stimulated recall (SR) methodology in L2 research

A discussion of SR methodology was provided in Chapter 2 (Section 2.5.1), including an outline of two areas of potential concern when using this method: firstly, veridicality (i.e. the accuracy of the thought process recollected by the participant), and secondly, variability in how linguistic problems encountered by participants are identified by the researcher, and how questions eliciting participants thought processes are asked.

In an attempt to limit these problems associated with how SR sessions are run, this study followed the protocols set out in Gass & Mackey (2017). All SR sessions were conducted immediately after the original task for all participants in order to limit as much as possible issues related to veridicality. Also, all instructions regarding SR procedure were read to participants from a script (see section 5.6.1 for an example script), and the only question used to prompt recall after identifying an apparent linguistic problem was, "Do you remember what you were thinking at that time?" (see section 5.6.1 for a sample SR exchange between researcher and participant).

A general discussion of noticing including definitions and measures was presented in Chapters 2 and 3. In this section, a brief review of the types, definitions and measures is

provided using specific examples from this study. The section begins with the noticing of IL gaps, before moving on to the noticing of IL-TL gaps

5.8.2 Measuring instances of noticing an IL gap

Noticing an IL gap – this type of noticing, as Izumi (2013) points out, occurs *during* the speech production process when the speaker realises they lack the linguistic resources to express what they want to say. Although the gap is noticed by the participant during the speech production process, it is reported immediately after the speaking task during the SR session.

In this study, noticing an IL gap was measured by identifying any instance during the stimulated recall session that a participant reported a linguistic problem they encountered while performing delivery 1 of their oral narrative. The reporting of problems could either come directly from the participant (i.e. participant initiated) or as a response to a question from the researcher (i.e. researcher initiated).

In order to avoid arbitrarily 'fishing' for comments from the participant during SR sessions, and to achieve consistency across all SR sessions, participants were only prompted for comment following the guidelines set out by Fukuta (2013). These state that the researcher should prompt the participant for a response when some form of dysfluency occurs in the participant's speech (e.g. a self-repair, an undue pause etc) as it is an indication that the speaker has encountered a problem in producing speech.

All stimulated recall sessions were transcribed and analysed by the researcher. Measurement of noticing an IL gap came from examination of transcripts of participants' stimulated recall comments. IL gaps were identified and classified as either a) noticing a content-related problem, or b) noticing a grammar-related problem (non-linguistic problems noticed were not included). Rules for classification into one of the categories were determined following Hanaoka (2007) who modelled his classifications from Williams (2001). Each category is described below.

a) Content-related problems – these included two subcategories, 'lexis' and 'other content'. Lexis provided the vast majority of comments from the 'content-related' category and included any comment with reference to specific vocabulary, about not knowing the correct word to use, not knowing the word in English, not knowing how to describe something, or not knowing how to say something in English. Below are two example comments from participants in this category.

1. *"I I have many words to say, but it's difficult to explain in English, because, you know, it's second language, that's why, so here I know what to say for this machine in my language, but I don't have vocabulary in English." [referring to 'oven']*

2. *"Yes, I have a problem that I don't know how to uh the vocab for this, so I was trying to think of the vocab for this." [referring to 'doorbell']*

'Other content'-related instances included comments about L2 problems related to meaning (other than lexis). An example is provided below.

3. *"Here I want to say how he feels now because now he know he is wrong, uh he know he made wrong decision to go camping, and his family had a right decision to go to the beach, so now uh the man know his decision is wrong."* (referring to the character's realisation that he had made the wrong decision to go camping, and that he should have followed the wishes of his family at the outset by going to a beach instead.)

Classification into this category was difficult at times. In example 3 above, it is apparent that the participant lacks the vocabulary to express their intended meaning. However, as in a similar example in Hanaoka (2007), the participant stopped short of mentioning any specific lexical items (or lack of).

b) Finally, grammar-related problems included any explicit comment about grammar including verb tense, word form and word order problems, examples include:

4. *"Yeah, I pause here because I think I made a mistake with the verb. But I already made the mistake, and I thought to correct it or not, but I just continue even though I think I made a mistake. Maybe I should use past verb but I said present verb, but I just continue."*

5. *"This one I don't know about the correct grammar, because this is the next weekend, but this story is still in the past, so I don't know if I can use past or future to describe this next trip in this picture, you know what I mean? Like, it's still past, but how can I say the future from the past?"*

c) Non-linguistic problems (i.e. problems not related to participants' knowledge of L2 English) were not counted (e.g. any occasion where something depicted was unclear to a participant). Examples of non-linguistic problems included:

6. *"This one is not clear for me. I don't know if this food is fish or vegetable or something he is cutting. It's not clear for me"*

7. *"Actually here, I don't know if they are all friends or if they are a family, so I paused while I think about their relationship."*

5.8.3 Solvable IL gaps

One of the goals of this study was to identify whether, in their second delivery of the oral narrative task, learners use language from model input to fill the IL gaps they had noticed in their initial delivery. In order to do that, the IL gaps learners noticed in their first delivery had to be analysed to determine whether there was actually language available in the model input to fill those gaps. If there was language available to fill an IL gap noticed, this was deemed to be a 'solvable IL gap'. The following example, taken from a SR session of the pre-test, illustrates a solvable IL gap when a participant did not know the word 'oven'.

Trigger: *Next, she uh put uh the uh chicken in uh in uh in the uh machine, in the machine.*

R: *Here you paused a little bit, do you remember what you were thinking at that time?*

P: *Yeah, I don't know the name for this in English. First I think microwave but then I think that's not right, so I just say 'machine', I know it's not right, uh, but because I don't know the real word*

The participant clearly realised she did not know the word 'oven' in English, so she called it 'the machine'. In the subsequently presented model input, the language needed to fill gap was available as the narrator said, "Debbie put the chicken into the oven". Therefore, the IL gap noticed by the participant could be filled with language from the model input, therefore, the gap was 'solvable'.

5.8.4 Measuring instances of noticing an IL-TL gap.

Noticing an IL-TL gap – this type of noticing occurs *after* speech production when a speaker has the opportunity to compare the way in which they used language to convey meaning with the way in which a more proficient speaker uses language to convey the same meaning (Izumi, 2013).

Noticing an IL-TL gap was defined as any note (word or phrase) made by a participant during the comparison stage (Step 3) of the testing sessions at Times 1, 5 and 6. During this stage, participants were asked to listen to an audio recording of the first delivery of their oral narrative and compare it to a model recording of the same narrative while taking notes of any useful language. Below is an example of notes taken by a participant during the comparison stage. They include ten instances of noticing an IL-TL gap. Each instance, however, could not be categorised as either being meaning- or grammar-related using the same methodology used to identify IL gaps because it was not possible to determine the nature of the gap noticed under the conditions of the present study. Although a second SR session could have been conducted to determine the nature of the IL-TL gaps noticed, but this would have added even more time to the testing session. Furthermore, it would have delayed the time before the participant was asked to deliver the repeat performance of the narrative task, and this would have detracted from the benefits of immediate TR as a pedagogical technique.

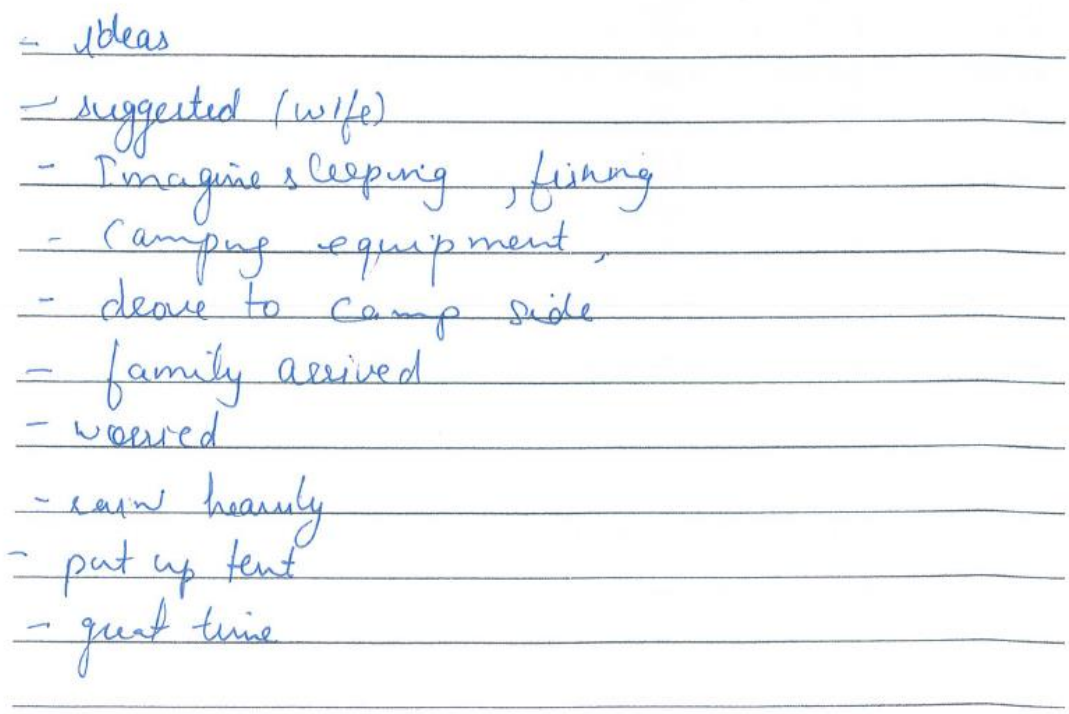


Figure 5.7 - *Sample of notes from comparison stage*

5.9 INTERRATER RELIABILITY

Interrater reliability was calculated for all measures of speech performance and noticing, and results are shown in Table 5.7. A randomly selected sample of 10%, along with written instructions for how to measure each dependent variable, was given to two trained-ESL teachers who also had post-graduate research experience in applied linguistics. For measures of speech performance, the randomly selected sample of 10% of transcripts came from both testing and training sessions, whereas the randomly selected sample for noticing measures came only from testing sessions (as noticing was not measured in training sessions). Measurement of noticing of IL gaps (both form -and meaning-ralated) came from analysing SR transcripts, while measurement of noticing of IL-TL gaps came from counting the number of notes learners made on the note-taking paper provided in testing sessions.

Table 5.7 - *Interrater reliability: Mean percentage of rater agreement*

	Dependent variable	N	Mean % of agreement
Accuracy	Number of errors per 100 words	43	94.29
	Number of self-repairs per 100 words	43	93.62
	Percentage of errors self-repaired	43	93.90
Complexity	Number of clauses per AS-unit	43	91.27
Fluency	Number of words per minute	43	99.12
	Number of silent pauses per 100 words	43	100
	Mean length of silent pause	43	98.73
	Number of filled pauses per 100 words	43	100
	Number of reformulations per 100 words	43	98.08
Noticing	Number of form-related IL gaps noticed	43	96.26
	Number of meaning-related IL gaps noticed	43	96.68
	Number of IL-TL gaps noticed	43	100

5.10 CHAPTER SUMMARY

In this chapter, the methodology followed for the experimental research presented in this thesis was provided. Both testing and training procedures were outlined in detail along with a description of how data were collected. Dependent variables for each of the broad areas under investigation (noticing, and speech performance) were detailed along with how participants' transcripts were coded and analysed. Rationalisation for the procedures used was given in light

of previous studies and current trends in related literature. The results of the analyses explained in this chapter are described in detail in the following chapter.

6 Results

6.1 INTRODUCTION

Presented in this chapter are the results of quantitative and qualitative analyses conducted on the data from the study. Results are divided into three main sections. Firstly, in Section 6.2, results that relate to noticing and its impacts on participants' incorporation of language from model input are presented. Secondly, in Section 6.3, speech performance results from testing sessions are outlined, followed lastly, in Section 6.4, by speech performance results from training sessions.

Noticing data from the testing sessions were analysed in a series of mixed two-way analyses of variance (ANOVAs) using SPSS Version 24, with time as a within subjects factor and group as the between subjects factor. Speech performance data were analysed in a series of mixed three-way ANOVAs, with time and delivery as within subjects factors and group as the between subjects factor. All ANOVAs were carried out using Bonferroni adjusted alpha levels of 0.002 to control the familywise error rate.

6.2 NOTICING

6.2.1 Screening of noticing data

To begin the screening of noticing data, an examination of stem-and-leaf-plots showed no outliers (defined as being more than two standard deviations away from the mean) for either of the two measures of noticing (noticing of grammar and noticing of content) in any of the testing sessions (i.e. Times 1, 5 & 6). Furthermore, the assumption of homogeneity of variance was met as assessed by Levene's test ($p > .05$). Box's test of equality of covariance revealed

there was homogeneity of covariances ($p > .05$), and the assumption of sphericity was met according to results of Mauchly's test of sphericity.

6.2.2 Two types of noticing used

As discussed in Chapter 4, two types of noticing were measured in this study: Noticing an interlanguage gap (IL gap) and noticing an interlanguage–target language gap (IL–TL gap). The distinction between these two types is explained in Table 6.1. Each type of noticing came from a different stage of the 4-stage testing sessions. In Stage 1, participants gave their first delivery of the narrative task. Next, in Stage 2, a stimulated recall was conducted by the researcher with the participant in order to identify IL gaps. Stage 3 (the comparison stage) required learners to listen to a recording of the first delivery and compare it to a recording of a model speaker's performance of the same narrative task while noting differences in language used (i.e. identify IL-TL gaps). The final stage, Stage 4, required the participant to deliver their repeat performance of the same oral narrative task.

Table 6.1 - *Explanation of the two types of noticing used in the present study*

Type of noticing	Explanation
Noticing an IL gap	This type of noticing, as Izumi (2013) points out, occurs *during* the speech production process when the speaker realises they lack the linguistic resources to express what they want to say. In the present study, instances of participants noticing an IL gap were measured through examination of stimulated recall comments from testing sessions.
Noticing an IL-TL gap	This type of noticing occurs *after* speech production when a speaker has the opportunity to compare the way in which they used language to convey meaning with the way in which a more proficient speaker used language to convey the same meaning (Izumi, 2013). Instances of noticing an IL-TL gap in the present

study were measured by analysis of participants' note-taking during the comparison stage in testing sessions.

6.2.3 Noticing an IL gap

The underlying question guiding this study was whether L2 learners could be trained to notice. Hence, research question 1a asked: *Does the number of form- and/or content-related IL gaps that learners notice change following noticing training (i.e. from pre-test to post-tests)?*

In order to determine the features of their own linguistic output that learners notice as problematic when performing an oral narrative task, a qualitative assessment of stimulated recall transcripts from pre-, post- and delayed post-tests was conducted. Based on this assessment, as explained in Chapter 5, comments were categorised as either noticing of grammar (NoG) or noticing of content (NoC). No language-related comments fell outside these two categories, and any non-language related comments were discounted. Group means are displayed in Table 6.2 (see Appendix 18 for ANOVA results).

Table 6.2 - *Means (and SDs) for instances of noticing of grammar (NoG) and noticing of content (NoC)*

Variable	Group	Pre-test Mean (SD)	Post-test Mean (SD)	Delayed Post Mean (SD)
NoG	Control	0.33 (0.65)	0.17 (0.39)	0.17 (0.39)
	Unguided Noticing	0.25 (0.45)	0.17 (0.39)	0.42 (0.51)
	Guided Noticing	0.33 (0.49)	2.67 (0.65)	3.00 (0.85)
NoC	Control	4.00 (0.74)	2.67 (0.78)	2.33 (1.07)
	Unguided Noticing	4.17 (1.70)	3.42 (0.79)	2.75 (1.54)
	Guided Noticing	4.08 (1.31)	3.42 (1.93)	3.17 (1.40)

ANOVA results revealed a statistically significant interaction between time and group for number of instances of noticing grammar (NoG; $F[4, 66] = 41.739$, $p < .001$, $\eta_{p2} = .717$). As can be seen from Table 6.2, all groups performed similarly in the pre-test; however, the GN group made larger gains than the C and UN groups in the post and delayed post-tests, that is, the GN group noticed significantly more grammar-related gaps in their output after having gone through noticing training. In contrast, the C and UN groups showed little change from pre- to post-tests.

When examining instances of noticing content (NoC), a simple main effect of time was found, $F(2, 66) = 13.319$, $p < .001$, $\eta_{p2} = .289$. Looking at the means, it can be seen that all groups showed a reduction in NoC from the pre-test to the post-test, and again from the post-test to the delayed post-test.

6.2.4 Noticing an IL – TL gap

Research question 1b: *Does the number of IL-TL gaps that learners notice change following training intervention?*

A qualitative assessment of participants' notes during the comparison stage of testing sessions (i.e. Times 1, 5 & 6) was undertaken to count instances of noticing an IL-TL gap. Inspection of the means shown in Table 6.3 reveals a small improvement in noticing for the UN and GN groups from pre-test to post-test, accompanied by a small decrease for the C group participants. However, ANOVA results revealed no statistically significant interaction between time and group, and no simple main effect of time or group.

Table 6.3 - *Average instances of noticing an IL – TL gap for each group at each assessment*

Variable	Group	Pre-test Mean (SD)	Post-test Mean (SD)	Delayed Post Mean (SD)
Noticing IL-TL gap	Control	8.25 (4.97)	6.17 (5.04)	6.58 (5.35)
	UN	8.18 (4.83)	9.27 (5.05)	9.64 (4.93)
	GN	8.42 (3.45)	10.08 (6.93)	9.67 (6.71)

Note: UN (unguided noticing), GN (guided noticing), SD (standard deviation).

6.2.5 Incorporation of model input into repeat performance

Research question 2: What percent of solvable IL gaps noticed in delivery 1 are filled in delivery 2 after exposure to model input, and how does this change following intervention?

After noticing IL-gaps in their first delivery output, this question investigated: first, whether participants filled the solvable gaps in their repeat performance using language presented in the model input; and second, whether the percentage of solvable gaps filled changed following intervention. In order to address these questions, an ANOVA was conducted for the dependent variable of percentage of solvable IL-gaps filled, with time as a within subjects factor and group as the between subjects factor. Results (see Appendix 19) revealed no significant interactions (means are shown in Table 6.4).

Table 6.4 - *Average percentage of solvable IL gaps filled for each group at each assessment*

Variable	Group	Pre-test Mean (SD)	Post-test Mean (SD)	Delayed Post Mean (SD)
% of solvable IL gaps filled in repeat performance	Control	85.6 (13.2)	87.5 (19.0)	86.1 (18.2)
	Unguided Noticing	83.1 (13.3)	82.6 (15.7)	84.7 (16.2)
	Guided Noticing	84.6 (14.2)	73.1 (12.8)	72.7 (12.4)

6.3 SUMMARY OF NOTICING RESULTS

In sum, the results for noticing show that training significantly increased the number of grammar-related IL gaps noticed by the GN group but not by the C or UN groups. By contrast there was no significant effect of training on the number of content-related IL gaps noticed by any group. Furthermore, results suggest that the intervention had no significant impact on the number of IL-TL gaps noticed by any group, nor on the incorporation of language from model input to fill solvable IL gaps noticed. In other words, having gone through noticing training, when asked to deliver their first performance of the oral narrative task in post-tests, the GN group noticed significantly more grammar-related gaps in their output than the C and UN groups. However, noticing training did not have an impact on the number of content-related gaps noticed by any group. With regard to number of IL-TL gaps noticed, results suggest that noticing training had no substantial impact for any group. Finally, looking at the percentage of solvable IL-gaps noticed in first deliveries that were filled in the repeat performances using language from the model input, results show that noticing training had no significant impact.

6.4 SPEECH PERFORMANCE RESULTS FROM TESTING SESSIONS

In order to determine the impacts of noticing training on participants' speech performance, transcripts of recordings of each delivery of the oral narrative task in testing sessions (i.e. Times 1, 5 and 6) were examined and coded for measures of complexity, accuracy, and fluency (CAF).

Analysis of learners' speech performance was guided by the following research questions:

Research question 3: What changes in CAF occur in learners' speech performance from delivery 1 to delivery 2 following training intervention?

Research question 4: What changes in CAF occur in learners' speech performance over the duration of the study?

Research question 3 related to the immediate effects of the intervention on learners' speech performance; that is, changes in performance from delivery 1 to delivery 2 within a particular session. Research question 4 related to the longer-term effects; in other words, changes in speech performance over the course of the seven-week study.

After a description of data screening, speech performance results are presented beginning with accuracy, followed by fluency, and then complexity.

6.4.1 Screening of speech performance data from testing sessions

Examination of stem-and-leaf-plots for each measure of speech performance from testing sessions (Times 1, 5 and 6) showed one outlier (defined as being more than 2 standard deviations away from the mean) from the C group for speech rate (number of words uttered per minute), and one outlier from the GN group for number of filled pauses per 100 words. The data were initially analysed twice, once with outliers and once without. Because the pattern of significant results did not change, a decision was made to include both outliers in the final analyses. No other outliers were found in any other measure of speech performance. Results from Box's test revealed that the assumption of equality of covariance was not met for all dependent variables. Nevertheless, as this study involves groups equal in size, and because ANOVA is robust to departures from the assumption of equality of covariance (Tabachnick & Fidell, 1989), a decision was made to proceed with analyses. The assumption of homogeneity of variance was met as assessed by Levene's test ($p > .05$), and according to results of Mauchly's test of sphericity, the assumption of sphericity was met on all but two occasions.

For the two instances that violated the assumption of sphericity, a Greenhouse-Geisser correction was applied and is reported in the appropriate section below.

Mixed three-way ANOVAs were conducted separately for each measure of complexity, accuracy, and fluency with group as the between-subjects variable, and time and delivery as the two within-subjects factors (see Appendices 20 - 22 for ANOVA results). A series of one-way ANOVAs using Bonferroni adjusted alpha levels of 0.002 was used to tease apart the nature of any significant interactions involving the group variable by determining under which particular conditions significant group differences occurred. Two planned comparisons were conducted for each ANOVA; one comparing the C group to the UN group, and a second comparing the GN group to the C and UN groups combined.

6.5 ACCURACY

Three measures were used to assess accuracy in participants' speech performance: number of errors per 100 words, number of self-repairs per 100 words, and percentage of errors self-repaired. This section reports the results from analyses of these measures, which were collected during testing sessions at Times 1, 5 and 6 of the study (i.e. pre-test, post-test and delayed post-test). Means are displayed in Table 6.5 and ANOVA results are in Appendix 20.

Table 6.5 - *Mean group scores (and SDs) for each measure of accuracy in delivery 1 (D1) and delivery 2 (D2) at times 1, 5 and 6 (Pre-, Post-, and Delayed Post-test)*

Variable	Group	Pre-test Mean (SD)		Post-test Mean (SD)		Delayed post-test (SD)	
		D1	D2	D1	D2	D1	D2
Number of errors/100 words	Control	11.61 (3.03)	10.79 (2.62)	12.58 (3.63)	11.57 (2.08)	13.18 (3.06)	11.91 (2.69)
	Unguided Noticing	11.30 (2.31)	10.64 (2.67)	12.70 (4.09)	10.97 (2.47)	13.17 (3.05)	11.39 (2.95)
	Guided Noticing	11.42 (1.46)	10.56 (1.09)	10.87 (1.26)	6.16 (1.32)	12.11 (2.14)	7.60 (1.87)
Number of self-repairs/100 words	Control	0.53 (0.59)	0.55 (0.96)	0.36 (0.57)	0.25 (0.37)	0.20 (0.21)	0.16 (0.20)
	Unguided Noticing	0.47 (0.59)	0.64 (0.46)	0.71 (0.64)	0.49 (0.47)	0.71 (0.89)	0.67 (0.36)
	Guided Noticing	0.48 (0.51)	0.47 (0.64)	0.70 (0.87)	1.21 (0.96)	0.94 (0.86)	1.36 (0.60)
Percentage of errors self-corrected	Control	4.87 (5.71)	4.98 (8.78)	3.63 (6.71)	2.31 (3.14)	1.52 (1.60)	1.29 (1.66)
	Unguided Noticing	3.84 (4.21)	5.35 (3.05)	5.72 (5.28)	6.64 (6.38)	5.07 (5.49)	8.44 (6.70)
	Guided Noticing	4.31 (4.46)	5.07 (5.42)	7.58 (8.60)	13.78 (10.81)	7.84 (6.91)	14.97 (8.07)

6.5.1 Number of errors per 100 words

Table 6.5 displays mean group data for number of errors per 100 words. ANOVA results (see Appendix 20) revealed a significant two-way interaction between group and delivery (F[2, 33] = 10.703, p =.001, η_{p2} = .393). Means in Table 6.5 show no obvious group difference on first deliveries with combined means of 12.46, 12.39 and 11.47 for the C group, UN group and GN group respectively. However, the GN group performed with significantly fewer errors per 100 words on second deliveries (combined mean = 8.11) than the C and UN groups combined (mean = 11.21; t[33] = 4.577, p < .001).

6.5.2 Number of self-repairs per 100 words

No statistically significant results were found for number of self-repairs per 100 words.

6.5.3 Percentage of errors self-repaired

ANOVA revealed a significant interaction between group and time for the dependent variable of percentage of errors self-repaired, F(4, 66) = 8.373, p < .001, η_{p2} = .585. Further testing revealed no statistically significant difference between groups at the pre-test, F(2, 33) = .024, p =.967, although there was a significant difference at the post-test, F(2, 33) = 15.319, p < .001, and the delayed post-test, F(2, 33) = 37.694, p < .001. Contrasts between the GN group and the C and UN groups combined reached statistical significance in the post-test, t(33) = 4.858, p < .001, and in the delayed post-test, t(33) = 7.705, p < .001. Contrasts between the C and UN groups reached statistical significance in the post-test, t(33) = 2.215, p = .034, and in the delayed post-test, t(33) = 4.887, p < .001.

6.5.4 Summary of accuracy results from testing sessions

To sum up the accuracy results from testing sessions, in all three measures, the GN group significantly outperformed the C and UN groups following intervention in terms of number of

errors per 100 words and percentage of errors self-repaired. In other words, while there was no difference in any accuracy measure at the outset (i.e. in the pre-test), after noticing training, compared to the UN and C groups, the GN group spoke with fewer errors per 100 words in their second delivery of the post-tests (both tests combined), they also self-repaired a higher percentage of their errors (both deliveries combined).

6.6 FLUENCY

Following Tavakoli and Skehan (2005), measurement of fluency was broken down into speed fluency, breakdown fluency, and repair fluency. Measures used in each of these three sub-categories are shown in Table 6.6. Results from each sub-category are addressed in turn below, and ANOVA results are displayed in Appendix 21. Alpha levels were set at .01 to compensate for use of multiple comparisons.

Table 6.6 - *Dependent variable measures for each fluency sub-category*

Sub-category	Measure
Speed fluency (speech rate)	Number of words per minute
Breakdown fluency	Number of filled pauses per 100 words
	Number of silent pauses per 100 words
	Mean length of silent pause
Repair fluency	Number of repetitions per 100 words
	Number of reformulations per 100 words

6.6.1 Speed fluency

6.6.1.1 Number of words per minute

No statistically significant results were found for number of words uttered per minute (see Table 6.7 for means).

Table 6.7 - *Mean group scores for speech rate (words per minute) for delivery 1 (D1) and delivery 2 (D2) at Times 1, 5 and 6 (Pre-, Post- and Delayed Post-test)*

Variable	Group	Pre-test Mean (SD)		Post-test Mean (SD)		Delayed post-test Mean (SD)	
		D1	D2	D1	D2	D1	D2
Speech rate	Control	114.26 (22.53)	107.28 (18.36)	116.33 (22.92)	120.41 (21.58)	121.51 (23.62)	119.70 (20.90)
(WPM)	Unguided Noticing	117.48 (19.23)	114.31 (17.16)	117.50 (16.34)	116.89 (18.41)	120.83 (19.87)	117.27 (16.58)
	Guided Noticing	110.67 (19.21)	104.18 (19.10)	107.27 (27.45)	110.55 (22.82)	115.28 (28.35)	113.90 (24.86)

6.6.2 Breakdown fluency

Table 6.8 presents mean group scores for the 3 measures of breakdown fluency: number of filled pauses per 100 words, number of silent pauses per 100 words, and mean length of silent pause. Analyses of each measure are outlined in turn.

6.6.2.1 Filled pauses

A simple main effect of time (F[2,66] = 22.931, $p < .001$, $\eta_{p2} = .410$) was found for the breakdown fluency measure of filled pauses per 100 words. This result reflected a reduction in the number of pauses per 100 words both from pre-test (M=12.16, SD=6.95) to post-test (M=9.50, SD=5.53; t=(35) = 4.280, $p < .001$), and from post-test to delayed post-test (M=8.25, SD=5.89; t=(35) = 2.834, $p = .008$).

A simple main effect of delivery was also found for number of filled pauses per 100 words (F[2,66] = 9.733, $p = .004$, $\eta_{p2} = .228$), with participants averaging 10.57 filled pauses per 100 words in delivery 1, and 9.36 filled pauses in delivery 2.

6.6.2.2 Silent pauses

No statistically significant results were found for number of silent pauses per 100 words.

6.6.2.3 Mean length of silent pause

No statistically significant results were found for mean length of silent pause.

6.6.2.4 Summary of breakdown fluency results

In sum, results from analyses of breakdown fluency show that participants spoke with fewer filled pauses per 100 words from one testing session to the next (both deliveries combined) regardless of group. Also, participants spoke with fewer filled pauses per 100 words in their second delivery of the narrative task compared to their first delivery, in each testing session, regardless of group.

Table 6.8 - *Mean group scores for three measures of breakdown fluency for delivery 1 (D1) and delivery 2 (D2) at Times 1, 5 and 6 (Pre-, Post-, and Delayed Post-test)*

Variable	Group	Pre-test Mean (SD)		Post-test Mean (SD)		Delayed post-test Mean (SD)	
		D1	D2	D1	D2	D1	D2
Number of filled pauses/100 words	Control	11.96 (5.42)	11.68 (5.97)	10.58 (5.01)	9.08 (6.10)	9.42 (6.46)	6.78 (6.02)
	Unguided Noticing	13.74 (9.04)	12.86 (9.33)	11.16 (4.59)	9.78 (6.76)	10.31 (7.16)	8.12 (5.66)
	Guided Noticing	11.50 (6.09)	11.19 (6.03)	8.77 (6.03)	7.63 (4.82)	7.74 (4.83)	7.12 (5.39)
Number of silent pauses/100 words	Control	22.11 (9.96)	23.78 (8.73)	21.86 (9.03)	18.82 (6.52)	18.81 (9.12)	20.70 (5.74)
	Unguided Noticing	17.96 (5.39)	18.89 (5.42)	18.77 (6.40)	18.61 (6.09)	15.51 (5.30)	17.35 (5.02)
	Guided Noticing	23.09 (9.14)	25.14 (8.89)	23.57 (9.62)	23.40 (8.29)	23.49 (10.03)	22.07 (9.05)
Mean length of silent pause (seconds)	Control	0.55 (0.23)	0.55 (0.17)	0.54 (0.17)	0.55 (0.18)	0.54 (0.16)	0.52 (0.13)
	Unguided Noticing	0.53 (0.08)	0.54 (0.11)	0.57 (0.11)	0.55 (0.10)	0.53 (0.10)	0.52 (0.10)
	Guided Noticing	0.66 (0.11)	0.76 (0.21)	0.68 (0.16)	0.64 (0.13)	0.68 (0.35)	0.62 (0.12)

6.6.3 Repair fluency

Table 6.9 shows mean group scores for the two measures of repair fluency: number of repetitions per 100 words and number of reformulations per 100 words. Analyses of each are explained below.

6.6.3.1 Repetitions

There was a simple main effect of delivery for number of repetitions per 100 words (F[1,33] = 8.862, p = .005, η_{p2} = .212), with means showing that all groups combined produced somewhat fewer repetitions per 100 words in their second delivery (M = 2.39) than their first (M = 2.77).

6.6.3.2 Reformulations

No statistically significant results were found for number of reformulations per 100 words (see Table 6.9 for mean data).

6.6.3.3 Summary of repair fluency results

Results for repair fluency show that, on average, across all testing sessions, participants spoke with fewer repetitions in their second delivery than in their first delivery, regardless of group. Number of reformulations per 100 words in participants' speech did not differ significantly by group, time or delivery.

6.7 COMPLEXITY

6.7.1 Number of clauses per AS-unit

Table 6.10 displays group means for complexity as measured by total number of clauses per AS-unit. ANOVA results (see Appendix 22) for the number of clauses per AS-unit with a Greenhouse Gasser correction showed a significant two-way interaction between time and delivery (F[4,66] = 4.601, p = .013, η_{p2} = .122). An examination of the means in Table 6.10 shows that at pre-test, the average number of clauses per AS-unit decreased from delivery 1

(*M* = 1.51, *SD* = 0.15) to delivery 2 (*M*=1.47, *SD*= 0.10), whereas at post-test, there was no change from delivery 1 (*M* = 1.55, *SD* = 0.11) to delivery 2 (*M* = 1.55, *SD* = 0.12), and at delayed post-test the number of clauses per AS-unit increased somewhat from delivery 1 (*M* = 1.53, *SD* = 0.12) to delivery 2 (*M* = 1.62, *SD* = 0.17). However, all of these differences were small.

Table 6.9 - *Mean group scores for two measures of repair fluency for delivery 1 (D1) and delivery 2 (D2) at Times 1, 5 and 6 (Pre-, Post-, and Delayed Post-test)*

Variable	Group	Pre-test Mean (SD)		Post-test Mean (SD)		Delayed post-test Mean (SD)	
		D1	D2	D1	D2	D1	D2
Number of reformulations/100 words	Control	1.14 (1.12)	1.51 (1.11)	1.73 (0.99)	1.23 (0.90)	1.05 (0.84)	0.88 (0.48)
	Unguided Noticing	1.57 (0.94)	1.38 (0.71)	1.44 (0.89)	1.40 (0.83)	1.23 (0.69)	1.12 ().74)
	Guided Noticing	1.53 (1.14)	1.24 (0.75)	1.69 (1.12)	1.04 (0.66)	1.74 (0.86)	1.32 (1.12)
Number of repetitions/100 words	Control	2.68 (1.74)	2.59 (2.00)	2.37 (1.98)	1.93 (1.69)	2.02 (1.79)	2.24 (1.75)
	Unguided Noticing	2.75 (2.32)	1.95 (1.74)	2.24 (2.06)	1.94 (1.75)	2.95 (1.76)	1.80 (1.35)
	Guided Noticing	3.52 (1.87)	3.35 (2.39)	3.87 (2.59)	3.23 (2.55)	2.55 (1.74)	2.48 (2.29)

Table 6.10 - *Mean group scores for complexity delivery 1 (D1) and delivery 2 (D2) at Times 1, 5 and 6 (Pre-, Post-, and Delayed Post-test)*

Variable	Group	Pre-test Mean (SD)		Post-test Mean (SD)		Delayed post-test Mean (SD)	
		D1	D2	D1	D2	D1	D2
Total number of clauses per AS-unit	Control	1.52 (0.17)	1.50 (0.10)	1.58 (0.11)	1.52 (0.12)	1.51 (0.09)	1.62 (0.12)
	Unguided Noticing	1.50 (0.09)	1.46 (0.08)	1.58 (0.11)	1.59 (0.09)	1.53 (0.08)	1.60 (0.13)
	Guided Noticing	1.51 (0.17)	1.46 (0.12)	1.50 (0.12)	1.55 (0.14)	1.55 (0.18)	1.63 (0.24)

6.8 A JUSTIFICATION FOR THE EXCLUSION OF LEXIS IN MEASURING SPEECH PERFORMANCE IN THE PRESENT STUDY

As mentioned in section 4.4.5, lexis is sometimes included by researchers as part of a measurement of speech performance. However, for the purposes of this study, lexis was not measured, and there are two main reasons for this. Firstly, in the pilot study it was found that both lexical complexity, as measured by a type-token ratio analysis, and lexical sophistication measured using Cobb's (2018) online lexical profiler lextutor.com, showed no significant change from each participant's initial performance to their repeat performance. Secondly, in the main study, it was found that many participants were already speaking with a higher degree of lexical complexity and lexical sophistication than the language used in the model recordings. Therefore, it was unlikely they were going to notice more complex or sophisticated language than they were already using.

6.9 SUMMARY OF SPEECH PERFORMANCE RESULTS FROM TESTING SESSIONS

To summarise, the results for speech performance in testing sessions revealed a significant interaction involving group for two of the three measures of accuracy. There was an immediate, positive intervention effect (from delivery 1 to delivery 2) for the GN group for number of errors per 100 words. There was also a positive interaction effect between time and group for percentage of errors self-corrected. In other words, after intervention, from delivery 1 to delivery 2 in both post-tests combined, the GN group spoke with fewer errors per 100 words and self-corrected a significantly higher percentage of their errors compared to the C and UN groups. By contrast, no significant interaction effects involving group were found for any measure of fluency or complexity, therefore suggesting no significant impact of the intervention on these aspects of speech performance.

6.10 SPEECH PERFORMANCE RESULTS FROM TRAINING SESSIONS

Results of participants' speech performance from the three training sessions (Times 2, 3 and 4) of the study are presented below. These results indicate the extent to which the different intervention conditions impacted upon the complexity, accuracy, and fluency of participants' speech during the three training sessions. Results for accuracy are presented first, followed by fluency and then complexity. One-way ANOVAs were used to tease apart the nature of any significant interactions involving the group variable. For measures of accuracy and complexity, two planned contrasts were conducted, one comparing the C group to the UN group, and another comparing the GN group with the C and UN groups combined. However, unlike the planned comparisons in testing sessions, in order to reflect the hypothesis that the C group would outperform the other groups in fluency during training sessions, two planned contrasts for all measures of fluency involved comparing the GN group to the UN group, and comparing the C group with the GN and UN groups combined.

6.10.1 Screening of training session data

Examination of stem-and-leaf plots for measures of CAF in training session data revealed two outliers. One from the C group for number of words uttered per minute and another from the GN group for number of repetitions per 100 words. Results revealed no significant difference when running analyses both with and without outliers. Therefore, a decision was made to keep the outliers in the final data analyses. The assumption of homogeneity of variance was met as assessed by Levene's test ($p > .05$). As with the data screening for speech results from testing sessions reported earlier in this chapter, results of Box's test for training session data revealed violations of the assumption of equality of covariance. However, for the same reasons stated in Section 6.4.1 a decision was made to proceed with analyses. Finally, the assumption of sphericity was violated on one occasion according to results of Mauchly's test

of sphericity, in which case a Greenhouse-Geisser correction was applied as is reported in the appropriate section.

6.10.2 Accuracy

As with accuracy in testing sessions, accuracy in training sessions was measured by number of errors per 100 words, number of self-corrections per 100 words and percentage of errors self-corrected. Although there was a simple main effect of time for number of errors per 100 words, the means showed no consistent change from pre-test (*M*=13.43) to post-test (*M*=11.86) to delayed post-test (*M*=12.99). Furthermore, since the main effect of time was not significant in any other analysis, and there were no significant interactions involving time, group means for the three training sessions were combined and are presented in Table 6.11. They show that the GN group made clear and considerable improvements on each measure from delivery 1 to delivery 2 for the three training sessions combined. ANOVAs revealed significant two-way interactions between group and delivery for each of the three measures of accuracy (see Appendix 23 for ANOVA results). Analysis of each measure is addressed in turn in the following sections.

Table 6.11 - *Mean accuracy scores according to Group (GN, UN, C) and delivery (1 vs. 2) for all training sessions combined (times 2, 3 and 4)*

Variable	Group	D1	D2	% change
Number of errors/100 words	Control	13.79	13.86	+0.51
	Unguided Noticing	14.20	11.61	-18.24
	Guided Noticing	14.06	9.04	-35.70
Number of self-repairs/100 words	Control	0.77	0.49	-36.36
	Unguided Noticing	0.73	0.58	-20.55
	Guided Noticing	0.72	1.29	+79.17
% of errors self-repaired	Control	5.79	3.06	-47.15
	Unguided Noticing	4.67	5.15	+10.28
	Guided Noticing	5.24	14.52	+117.10

6.10.2.1 Number of errors per 100 words

ANOVA results revealed a significant two-way interaction between group and delivery for number of errors per 100 words, $F(2,33) = 36.120$, $p < .001$, $\eta_{p2} = .686$. Means in Table 6.11 show that, on average across all three training sessions, the C group spoke with little change in number of errors per 100 words from one delivery to the next. On the other hand, the UN group spoke with fewer errors in their second deliveries combined, and the GN group improved the most from delivery 1 to delivery 2 with a clear and considerable reduction in number of errors per 100 words.

While it can be seen that all groups performed similarly in terms of average number of errors per 100 words across all first deliveries, when examining delivery 2 means, the GN group significantly outperformed the C and UN groups combined, $t(33) = 3.737$, $p = .001$, whereas the C and UN groups did not differ significantly from one another ($t[33] = 1.963$, $p = .058$).

6.10.2.2 Number of self-repairs per 100 words

A significant two-way interaction was also found between group and delivery for the number of self-repairs per 100 words (SR/100 words), $F(2,33) = 13.628$, $p < .001$, $\eta_{p2} = .452$. Means in Table 6.11 show no obvious group difference on first deliveries, with scores ranging from 0.73 to 0.87. However, the GN group self-repaired markedly more on second deliveries (0.92) than did the C and UN groups combined (combined mean = 0.54; $t[33] = 5.888$, $p < .001$).

6.10.2.3 Percentage of errors self-repaired

Group and delivery also showed a significant interaction for the measure of percentage of errors self-repaired, $F(2,33) = 34.002$, $p < .001$, $\eta_{p2} = .673$. Means again show no obvious difference between groups in first deliveries (means range from 4.67 – 5.79); however, when looking at second deliveries the GN group self-repaired a significantly greater number of errors than the C and UN groups combined (M=4.12; $t[33] = 10.027$, $p < .001$).

6.10.2.4 Summary of accuracy results from training sessions

In sum, compared to the C and UN groups, across all training sessions the GN group spoke with fewer errors per 100 words from delivery 1 to delivery 2. They also self-corrected more errors per 100 words and self-corrected a higher percentage of errors in their second delivery compared to the C and UN groups. No significant interaction involving time was found, suggesting that there were no significant changes in accuracy from one training session to the next.

6.10.3 Fluency

Mirroring measures of fluency in testing sessions, fluency in training sessions was broken down into three sub-categories. First, speed fluency was measured by speech rate in terms of number of words uttered per minute. Second, breakdown fluency was measured by number of filled pauses per 100 words, number of silent pauses per 100 words, and mean length of silent

pause; and third, repair fluency was measured by number of repetitions per 100 words and number of reformulations per 100 words.

An examination of ANOVA results (see Appendix 24) revealed significant two-way interactions between group and delivery for each of the six dependent variables across the three sub-categories of fluency (i.e., speed fluency, breakdown fluency, and repair fluency). No significant three-way interactions were found. Inspection of the group means in Table 6.12, shows that the three groups performed similarly on all measures in delivery 1. With the exception of number of reformulations per 100 words and mean length of silent pause, there was a trend for the C group to make clear and considerable gains in fluency from delivery 1 to delivery 2, the UN group to make smaller yet still considerable gains, and the GN group to make the smallest gains, and, for some measures, to show a reduction in fluency.

Table 6.12 - *Mean fluency scores according to group (GN, UN, C) and delivery (1 vs 2) for all training sessions combined (times 2, 3 and 4)*

Variable	Group	D1	D2	% change
WPM	Control	110.59	131.78	+19.16
	Unguided Noticing	113.66	124.08	+9.17
	Guided Noticing	112.50	103.58	-7.93
FP/100 words	Control	7.03	3.26	-53.63
	Unguided Noticing	6.78	4.69	-30.83
	Guided Noticing	6.48	7.77	+19.91
SP/100 words	Control	21.22	17.86	-15.83
	Unguided Noticing	20.68	18.69	-9.62
	Guided Noticing	21.41	24.34	+13.69
MLSP	Control	0.62	0.56	-9.68
	Unguided Noticing	0.61	0.53	-13.11
	Guided Noticing	0.63	0.66	+4.76
Reform/100 words	Control	2.20	0.73	-66.82
	Unguided Noticing	2.05	1.60	-21.95
	Guided Noticing	2.02	1.61	-20.30
Reps/100 words	Control	2.62	1.12	-57.25
	Unguided Noticing	2.40	1.73	-27.92
	Guided Noticing	2.52	2.82	+11.90

Note: WPM (words per minute), FP (filled pauses), SP (silent pauses), MLSP (mean length of silent pause), Reform (reformulations), Reps (repetitions).

Each two-way interaction is described in detail below, through the examination of relevant group means.

6.10.3.1 Speed fluency

Speech rate

ANOVA results (see Appendix 24) for the dependent variable of speech rate (words per minute) showed a highly significant interaction between group and delivery with a

Greenhouse-Geisser correction applied, $F(2,33) = 75.268$, $p < .001$, $\eta_{p2} = .820$. Group means in Table 6.12 show similar delivery 1 speech rates across all groups, with means ranging from 110.59 to 113.66 words uttered per minute. Delivery 2 means, however, show a decrease in speech rate for the GN group, a slight increase for the UN group, and a considerable increase for the C group.

6.10.3.2 Breakdown fluency

Table 6.12 displays group means for the three measures of breakdown fluency. Analysis of each measure is presented in turn, and ANOVA results are provided in Appendix 24.

Filled pauses

When looking at the number of filled pauses per 100 words in training sessions, group and delivery again showed evidence of a significant interaction, $F(2,33) = 17.231$, $p < .001$, $\eta_{p2} = .511$. It can be seen from the means in Table 6.12 that the C group showed a considerable improvement (reduction) in average number of filled pauses per 100 words from delivery 1 to delivery 2. Slightly smaller improvements were evident for the UN group from delivery 1 to delivery 2, while the GN group showed a small *increase* in the average number of filled pauses per 100 words from delivery 1 to delivery 2.

Silent pauses

For the dependent variable of number of silent pauses per 100 words, a significant interaction was found between group and delivery, $F(2,33) = 18.448$, $p < .001$, $\eta_{p2} = .528$. Means revealed a trend similar to that of filled pauses, with the C group making an average of 21.22 silent pauses per 100 words across all first deliveries in the three training sessions, which was reduced to an average of 17.86 silent pauses per 100 words in all second deliveries. A small improvement was also apparent for the UN group from delivery 1 to delivery 2, whereas the GN group had an almost identical average number of silent pauses per 100 words in delivery 1 and delivery 2 (see Table 6.12).

Mean length of silent pause

No statistically significant results were found for mean length of silent pause.

Summary of breakdown fluency results from training sessions

To summarise breakdown fluency results from training sessions, compared to the UN and GN groups, from delivery 1 to delivery 2 across all training sessions, the C group spoke with fewer filled pauses per 100 words and fewer silent pauses per 100. Furthermore, the UN group clearly outperformed the GN group in filled pauses per 100 words and number of silent pauses per 100 words. No significant interaction involving time was found, meaning that there were no significant changes in any measure of breakdown fluency from one training session to the next.

6.10.3.3 Repair fluency

As the means in Table 6.12 show, the C group made clear gains in number of reformulations per 100 words, and number of repetitions per 100 words. Statistical analyses of both measures are discussed in turn. See Appendix 24 for ANOVA results.

Reformulations

ANOVA results revealed a significant interaction between group and delivery for number of reformulations per 100 words, $F(2,33) = 10.337$, $p < .001$, $\eta_{p2} = .386$. Inspection of the means revealed that the three participant groups did not differ from one another in the number of reformulations per 100 words across all first deliveries in the training sessions; however, second delivery means show that the C group made significantly fewer reformulations (M=0.73) than the UN and GN groups combined (M =1.61; $t(33) = 3.753$, $p = .001$).

Repetitions

A significant interaction was found between group and delivery for number of repetitions per 100 words ($F[2,33] = 12.405$, $p < .001$, $\eta_{p2} = .429$). Once again, all groups performed similarly on their first delivery. However, the C and UN groups improved in second deliveries,

making fewer repetitions; whereas the GN group spoke less fluently in their second deliveries, making more repetitions per 100 words on average. No statistically significant difference was found between groups when contrasting first delivery means. However, for second delivery means, comparison of the C group with the UN and GN groups combined revealed a statistically significant difference favouring the C group, $t(33) = 2.988$, $p = .005$. On the other hand, although the UN group improved by recording fewer repetitions per 100 words in their second delivery compared to the GN group who showed an increase, comparison between these two groups did not reach statistical significance ($t[22] = 2.205$, $p = .038$).

Summary of results for repair fluency from training sessions

While there was no significant interaction effect involving time, across the three training sessions combined, the C group spoke with fewer reformulations per 100 words, and fewer repetitions per 100 words from delivery 1 to delivery 2 compared to the UN and GN groups. Also, from delivery 1 to delivery 2 across all training sessions, while the UN and GN groups made similar improvements in number of reformulations per 100 words, when looking at number of repetitions per 100 words, the UN group outperformed the GN group (with fewer repetitions), however, statistical comparisons did not reach significance.

6.10.4 Complexity

No statistically significant results were found for number of clauses per AS-unit (see Appendix 25 for ANOVA results). Means are displayed in Table 6.13.

Table 6.13 - *Mean complexity scores according to group (GN, UN, C) and delivery (1 vs 2) for training sessions (times 2, 3 and 4)*

Variable	Group	Time 2 Mean (SD)		Time 3 Mean (SD)		Time 4 Mean (SD)	
		D1	D2	D1	D2	D1	D2
Total number of clauses per AS-unit	Control	1.54 (0.17)	1.50 (0.10)	1.45 (0.10)	1.54 (0.14)	1.52 (0.12)	1.50 (0.11)
	Unguided Noticing	1.49 (0.09)	1.47 (0.10)	1.50 (0.13)	1.52 (0.12)	1.58 (0.12)	1.52 (0.11)
	Guided Noticing	1.50 (0.17)	1.48 (0.12)	1.48 (0.14)	1.47 (0.11)	1.56 (0.11)	1.49 (0.09)

6.11 SUMMARY OF SPEECH PERFORMANCE RESULTS FROM TRAINING SESSIONS

To sum up, the results for speech performance in training sessions revealed an immediate, positive effect (from delivery 1 to delivery 2 for all training sessions combined) for the C group on all measures of fluency compared to the UN and GN groups with the exception of number of silent pauses per 100 words. The UN group, who took part in unguided noticing training during their intervention sessions, also made gains from delivery 1 to delivery 2 in all measures of fluency except number of silent pauses per 100 words, but to a lesser extent than the C group. Finally, although the GN group, whose intervention sessions involved guided noticing training, made small gains from delivery 1 to delivery 2 in number of reformulations per 100 words, averaged across the three training sessions, they performed worse in all other measures of fluency (i.e. they became less fluent in their second delivery compared to their first).

Finally, in terms of accuracy, in the three training sessions combined, the GN group improved significantly in all three measures from delivery 1 to delivery 2 compared to the C and UN groups.

6.12 OVERALL SUMMARY OF KEY FINDINGS

In this section the key findings from the study are summarised, starting with noticing. As a result of noticing training, the GN group noticed significantly more form-related IL gaps than the UN and C groups, and this increase in noticing was reflected in immediate improvements in accuracy of speech performance (i.e. improvements from delivery 1 to delivery 2 within a session). Compared to the C and UN groups, the GN group spoke with significantly fewer errors per 100 words in their second than in their first deliveries. Also, noticing training had a significant longer-term impact on speech performance with the GN

group self-correcting a higher percentage of total errors compared to the C and UN groups when deliveries 1 and 2 combined in post-tests are compared.

Results from training sessions show that the C group outperformed the UN and GN groups in terms of fluency, while the GN group outperformed the C and UN groups in all measures of accuracy. These results are interpreted and discussed in the following chapter.

6.13 CHAPTER SUMMARY

In this chapter, quantitative results from the study were presented. The first section (Section 6.3) outlined results related to noticing training and the incorporation of language from model input into participants' repeat performance of their oral narrative task. Section 6.4 presented results related to the impacts of noticing training on speech performance by examining changes in the accuracy, fluency, and complexity of participants' speech from the pre-test to the post-tests (Times 1, 5 and 6). Lastly, Section 6.5 dealt with results related to changes in participants' speech performance during the 3 intervention sessions (Times 2, 3 and 4). In each section, repeated measures ANOVAs were used firstly to identify any significant interactions, before one-way ANOVA results or t-tests were reported to tease apart the nature of significant interactions. The following chapter discusses the results reported here.

7 Discussion.

7.1 INTRODUCTION

This study began with the broad, overarching aim of examining whether L2 learners could be trained to notice, and what impact (if any) such training might have on speech performance. In this chapter, the results are interpreted, and the findings discussed. The chapter is presented in five main sections beginning in Section 7.2 with a discussion of the findings related to noticing training. This is followed in Section 7.3 by a discussion of the immediate impact of noticing training on speech performance, and then in Section 7.4, the longer-term effects of noticing training on speech performance. In Section 7.5, findings for speech performance in training sessions are discussed. Each section begins with a table summarising the research question(s), the original hypothesis(es) and the findings related to that particular section. The chapter ends, in Section 7.6, with an overall summary of the main findings.

7.2 DISCUSSION OF FINDINGS RELATED TO NOTICING

In this section, findings are discussed in terms of whether or not support was found for each of the hypotheses related to noticing training. Chapter 5 presented a table summarising research questions and hypotheses, and this has been expanded below to also include a summary of findings related to noticing (Table 7.1).

After the summary table is a discussion of the nature and number of IL gaps L2 learners notice when performing a narrative speaking task. This is followed by interpretation and discussion of the effects of noticing training on IL gaps noticed. Finally, the findings related

to noticing training and incorporation of model input in learners' repeat performance of their narrative task are discussed.

Table 7.1 - *Overview of research questions, hypotheses and findings related to noticing training*

Broad Research Question	Specific Research Question	Hypothesis	Finding
5. Can learners be trained to notice gaps in their output?	1a. Does the number of form- and/or meaning-related IL gaps that learners notice change following noticing training (i.e. from pre-test to post-tests)?	1a. Following training, the number of form-related IL gaps noticed will increase for the GN group but not for the UN or C groups.	1a. Support for this hypothesis was found. The number of form-related IL gaps noticed increased significantly for the GN group from pre-test to both post-tests, while the UN group decreased pre-test to post-test but made a slight improvement in the delayed post-test. The C group decreased from pre- to both post-tests.
		1b. Following training, the number of meaning-related IL gaps noticed will decrease for the GN group, increase for the UN group, and remain unchanged for the C group.	1b. No support was found for this hypothesis. Although the GN group decreased as expected, the control and UN groups also decreased from pre-test to post-tests.
	1b. Does the number of IL-TL gaps that learners notice change following training intervention?	1c. Following training (i.e. in post-tests), the number of IL-TL gaps noticed will increase for the GN group to a larger extent than for the UN group, who in turn will outperform the C group.	1c. No support for this hypothesis was found. Although the GN group improved from pre- to post-tests, as did the UN group to a lesser extent, results did not reach statistical significance. The C group decreased from pre- to post-tests.
6. How does noticing training influence subsequent incorporation of input?	2. In testing sessions, what percent of solvable IL gaps noticed in delivery 1 are filled in delivery 2 after exposure to model input, and how does this change following intervention?	2. Following training (i.e. in post-tests), after noticing IL gaps in their first delivery, the GN group will incorporate language from subsequently presented model input to fill a higher percentage of solvable gaps during their second delivery compared to the UN group. The C group will not change from pre- to post-tests.	2. No support for this hypothesis was found. However, this may be due to the number and nature of IL-gaps noticed by the GN group compared to the C and UN groups (see section 6.3.5 for a discussion).

7.2.1 IL Gaps Noticed

As stated earlier, the noticing of IL gaps occurs during the speech production process when a learner becomes aware that they lack the linguistic resources to express what it is they want to say. While these gaps were noticed internally by learners when delivering their first performance of the narrative task, they were identified for research purposes during the stimulated recall session of each test which followed immediately after each participants' first delivery. At this point, participants had only delivered their first performance of the narration and were reviewing the recording with the researcher (i.e. they hadn't yet had the opportunity to compare their performance to that of a model speaker or deliver their repeat performance). IL gaps noticed, therefore, came only from each participant's first delivery of their oral narrative. This is in contrast to the noticing of IL-TL gaps which comes *after* the stimulated recall session when participants have the opportunity to compare a recording of their first delivery with a recording of a model speaker's performance. In this section, only findings related to IL gaps are discussed (findings for IL-TL gaps are discussed in section 7.3.6).

7.2.2 What is the nature and number of IL gaps noticed without intervention?

Before looking at whether L2 learners can be trained to notice IL gaps in their oral output, an important first step is to establish the number and nature of the IL gaps they notice without any instruction or direction. Noticing studies to date have consistently found that L2 learners notice lexis-related problems when writing. However, as far as the number and nature of IL gaps noticed by L2 learners performing a narrative speaking task is concerned, research has not yet established whether IL gaps noticed are generally form-related, meaning-related, or a combination of the two. Furthermore, as discussed in Chapter 3, the vast majority of studies on the relationship between output, input and noticing have taken a teacher/research-led

approach to noticing (e.g. Izumi et al., 1999; Izumi & Bigelow, 2000; Uggen, 2012), rather than learner-led noticing as was the focus in this study.

Given L2 learners' tendency to direct attention towards meaning when producing speech in L2 classroom communicative tasks, it was hypothesised that learners would notice more meaning- than form-related IL gaps. This prediction was based on Skehan's Trade-off Hypothesis (1996, 1998, 2009). According to the Trade-off Hypothesis (also known at The Limited Attention Capacity Hypothesis - refer to Chapter 4 for a review of literature), because human attentional capacity is limited, when L2 learners are performing a speaking task increased attention given to one aspect of speech performance (e.g. meaning) comes with a corresponding drop in another aspect (e.g. form). Current literature suggests that the tendency is for L2 learners to prioritise meaning as this is what will help them to complete the task (Ahmadian, 2012; Skehan 1998; 2007). It is, therefore, reasonable to assume that without any intervention, this focus on meaning during the speech production process would lead a learner to notice mostly meaning-related IL gaps in their output.

As shown in the previous chapter, the results supported this assumption. Expressed as a percentage, meaning-related instances of noticing an IL gap in the pre-test represented 92.4% of total instances for the C group, 94.3% for the UN group, and 92.5% for the GN group. A closer examination of meaning-related IL gaps revealed that the overwhelming majority of these were lexical gaps. These results are in line with those of previous studies investigating written output (e.g. Hanaoka, 2007; Mackey, Gass, & McDonough, 2000). However, they add to existing research by providing clarity about what is it L2 learners notice as being problematic in their *oral* output in a TBLT context.

Since the results from the pre-test showed that the L2 learners in this study primarily noticed meaning-related IL gaps when performing an oral narrative task, the next step was to establish what kind of impact noticing training would have on the number and nature of IL

gaps noticed. In the following section a discussion of the findings related to noticing training is presented.

7.2.3 Effects of noticing training on IL gaps noticed

The relative percentages of form- and meaning-related IL gaps noticed in the post- and delayed post-test show a clear increase post-intervention in form-related instances of IL noticing for the GN group, but not for the UN or C groups. Thus, support for hypothesis 1a was found. As reported in Section 6.2.3, for the C group, 7.62% of IL gaps noticed in the pre-test were form-related. In the post-test, this dropped to 5.66%, before increasing slightly to 7.48% in the delayed post-test. For the UN group, of all their IL gaps noticed, 5.66%, 4.74% and 13.25% were form-related in the pre-, post- and delayed post-test respectively. The GN group, on the other hand, showed a dramatic increase in percentage of form-related IL gaps noticed from 7.48% in the pre-test, to 43.84% in the post-test, and 48.62% in the delayed post-test.

As discussed in Chapter 3, other studies have found that intervention can increase L2 learners' noticing of form (e.g. Hanaoka, 2007). Such studies, however, have looked at noticing during the input and/or second output stage of an output-input-output sequence. The results from this study differ in that after intervention, learners in the GN group noticed significantly more form-related IL gaps in their language production during the *first* stage of an output-input-output sequence.

It can, therefore, be said with some confidence that the guided noticing training given to the GN group during the three intervention sessions in this study was successful in training them to allocate more attention to form when producing output and, as a result, notice more form-related IL gaps in their output. The unguided noticing training given to the UN group in this study, on the other hand, had no impact on noticing of form-related IL gaps. This is

unsurprising given that it was established in the pre-test that all learners focused primarily on meaning in terms of the features of their output that they noticed as being problematic. Unlike the guided noticing training given to the GN group that was designed to shift their attention towards form, the unguided noticing training received by the UN group in the intervention sessions was not designed to shift participants' attentional focus. Instead, it was designed to allow participants to attend to whichever aspects of their output and the model input that they liked. It is therefore unsurprising that the focus on meaning-related problems in their output in the pre-test carried over to post-tests.

7.2.4 Effects of noticing training on number of meaning-related IL gaps noticed

As stated in hypothesis 1b, it was thought that the C group would not change the degree to which they noticed meaning-related IL gaps from pre- to post-tests because they received no noticing training. However, it was hypothesised that the UN group would notice more meaning-related IL gaps as they had the opportunity to practice 'unguided' noticing during training sessions, and this might reinforce their focus on meaning-related aspects of their output and of the model input. For the GN group, it was thought that their noticing of meaning-related IL gaps would decrease as a result of a shift in attentional focus from meaning to form brought about by the guided noticing training.

No support for this hypothesis was found. As expected, there was a reduction in the number of meaning-related instances of IL gaps noticed from pre-test to post-tests for the GN group. Again, this was likely due to the shift in attention from meaning to form brought about by their guided noticing training. Unexpectedly, however, there was also a decrease, albeit a small one, for both the C and the UN groups in meaning-related IL gaps noticed from one testing session to the next.

An explanation as to why the C and UN groups did not notice more meaning-related IL gaps in their post-tests may be because participants in these groups did not see value in the type of TR used in this study. There is general agreement in the literature that a common issue with TR (both in empirical studies and in use of TR as a pedagogical technique in L2 classrooms) is boredom and fatigue among learners in subsequent iterations (Bygate, 2001, also see Ahmadian, Mansouri, & Ghominejad, 2017 for an opposing view). As Lambert, Kormos, & Minn (2017) point out, it is necessary for learners to see value in repeating tasks in order to avoid boredom and fatigue. Value can be designed into the task by, for example, changing the audience for the repeat performance which can motivate the speaker to repeat their output while remaining focused on meaning (Arevart & Nation, 1989; Nation, 1990, Thai & Boers, 2015).

In this study, there was no change of audience for participants, so they may not have seen the value in repeating the same task, and therefore lacked motivation to notice. Although the GN group (who did improve in number of meaning-related IL gaps noticed in post-tests) also had no change in audience for their repeat performance, they may have seen the value in repeating the task as a result of their improved grammatical accuracy from one delivery to the next.

More specifically, those in the GN group were making clear gains in grammatical accuracy from one performance to the next while those in the C and UN groups were not (see Section 7.3.1 for a discussion on the impacts of noticing training on the accuracy of participants' speech performance). It is possible that the training helped the learners in the GN group to see the link between noticing form-related IL gaps in their first delivery, mining the model input for language, and improving their grammatical accuracy in their second delivery as a result. At the same time, because those in the UN and C groups were not making clear improvements in speech performance from delivery 1 to delivery 2, they may not have seen

the same value in the output-input-output sequence, and they did not have the same high-level of motivation to notice IL gaps as a consequence.

As an interim summary, an interesting point to note here is that, although the GN group noticed significantly more form-related IL gaps in post-tests compared to the other groups, this did not come at the cost of noticing meaning-related IL gaps. Results showed that meaning-related IL gaps dropped slightly from pre-test to post-tests for the GN group. However, not only did they still account for just over half of all their IL gaps noticed, they also noticed more meaning-related IL gaps than the other two groups (with the exception of the post-test where the GN and UN groups noticed an identical number). In other words, the noticing of form-related IL gaps by the GN group in post-tests came *in addition to* the noticing of meaning-related gaps rather than *instead of* noticing meaning-related IL gaps. This suggests that as a result of going through guided noticing training, those in the GN group balanced their allocation of attentional resources between both meaning and form during the act of producing speech in their first delivery of the narrative task to a much higher degree than they did prior to training. It also suggests that learners can not only be trained to notice certain kinds of IL-gaps (e.g. form-related), but they can also be trained to notice a larger number of IL gaps than they would without training.

7.2.5 Effects of noticing training on number of solvable IL-gaps filled

As mentioned in Chapter 3, one claim related to the relationship between output, input and noticing is that the initial production of output can lead a learner to notice what it is they cannot express accurately in their L2, and as a result lead them to pay greater attention to language in subsequently presented model input (Uggen, 2012). If then given the chance to repeat their initial performance incorporating language from the model input, learners are in a position to fill gaps in their IL and thus drive their L2 acquisition forward. No research to date, however, has examined whether learners can be trained to both notice and fill their IL gaps

over a period of time. Research question 2a therefore asked: In testing sessions, what percent of solvable IL gaps noticed in delivery 1 are filled in delivery 2 after exposure to model input, and how does this change following intervention?

To answer this question, a qualitative analysis of participants' comments from stimulated recall sessions was conducted in order to determine whether the IL gaps they noticed in their first delivery could be filled with language available in the model input, if so, these were deemed 'solvable IL gaps' (see Chapter 5, Section 5.8.3 for an example) . The next step was to determine whether the solvable gaps were actually filled by the speaker in the repeat performance of the task. This was done by way of a qualitative analysis of participants' delivery 2 transcripts. It was hypothesised that the GN group would fill a higher percentage of solvable IL-gaps in their second delivery following intervention. However, results showed that this was not the case, as such, no support for hypothesis 2a was found. Results showed that while the C and UN groups maintained their percentage of solvable IL gaps filled from pre-test to post-tests, the GN group surprisingly decreased in percentage of solvable IL gaps filled from pre-test to post-tests. An explanation for this may be linked to increasing demands on working memory.

As discussed in Chapter 3, previous studies have investigated a link between noticing, incorporation of input and working memory (e.g. Hanaoka & Izumi, 2012; Shephard, 2006). These authors found that larger working memory capacity resulted in more incorporation of language from model input when asked to repeat a speaking task. This suggests that after directing working memory capacity to the demands of linguistic production as well as to the procedural demands of a given task, those with a larger working memory capacity still have cognitive resources available to direct to the uptake of input presented in a model.

In the present study, the combination of linguistic and procedural demands (including noticing) placed on learners in the GN group may have overloaded working memory and

hampered their ability to incorporate language from the model input into their repeat performance, thus providing a possible explanation for why the GN group did not improve in percentage of solvable gaps filled as hypothesised. Section 6.2.3 reported results of noticing IL gaps and showed that on average in post-tests, the GN group noticed around twice the amount noticed by the UN and C groups (6.13 IL gaps, 3.38 IL gaps and 2.67 IL gaps respectively). The fact that the GN group had noticed a larger number of IL gaps meant that they therefore needed to recall a larger amount of language in the model input in order to fill those gaps when delivering their repeat performance. Therefore, in addition to the cognitive demands of producing language for the second delivery of their narrative, these learners had the added burden on working memory of trying to recall more language from the input than learners in the C and UN groups. In other words, the task of recalling language from model input, and the consequent burden on working memory, is lower for a learner who has only two or three IL gaps to fill as opposed to a learner who has six IL gaps to fill. It could be, therefore, that the more IL gaps that are noticed, the more attention is taken away from subsequent speech production and given to recollection and retrieval of language from model input that is stored in short-term memory.

Another consideration is the nature of solvable IL gaps noticed. In the pre-test, the overwhelming majority of solvable IL gaps noticed by all groups were meaning related, and of those, the vast majority were lexical. This trend was carried through to post-tests for the C and UN groups. However, those in the GN group transitioned from a majority of meaning-related IL gaps noticed in the pre-test, to a close-to-equal number of meaning and form-related IL gaps in post-tests. It is plausible that meaning-related IL gaps are comparatively easier to fill than form-related, especially if the majority of meaning related IL gaps noticed are lexical. Filling a lexis-related IL gap would require identifying the necessary word(s) in model input, and then holding that word(s) in short-term memory until it is needed during the repeat performance. A

form-related IL gap, on the other hand, may consist of a string of lexical items which would require more attention to accurately hold in short-term memory before being needed in the repeat performance. To illustrate this point, the example of a lexis-related IL gap noticed reported in Chapter 5 is repeated here.

Trigger: *Next, she uh put uh the uh chicken in uh in uh in the uh machine, in the machine.*

R: *Here you paused a little bit, do you remember what you were thinking at that time?*

P: *Yeah, I don't know the name for this in English. First I think microwave but then I think that's not right, so I just say 'machine', I know it's not right, uh, but because I don't know the real word.*

In order to fill the IL gap she noticed, the participant needed to identify the word 'oven' which was used in the model input, hold that word in working memory, and then retrieve it for use at the appropriate time during her repeat performance. Filling this IL gap would be comparatively easier than filling the IL gap below which was noticed in the post-test by a participant from the GN group.

Trigger: *Bob, uh Bob, Bob ask to Sally uh uh the uh the the cooking time.*

R: *Here you paused a little bit, do you remember what you were thinking at that time?*

P: *Here I wanted to say the question that Bob ask to Sally, you know, uh like, 'Bob asked Sally how long the cooking time', but I know this grammar is not*

correct, uh because we study this in class about changing the words' order, and verb position, so I try to think of the grammar, but uh I can't think, so I just say that.

In the above example, the student was making reference to the grammar of reporting questions. The narrator in the model input said, "Bob asked how long the chicken would take to cook". This was the language the participant needed to fill the IL gap she had noticed. However, identifying this sentence in model input and holding it in short-term memory ready for retrieval in the repeat performance is clearly more cognitively demanding than holding a single word (e.g. 'oven').

7.2.6 Effects of noticing training on number of IL-TL gaps noticed

The second type of noticing investigated in this study was the noticing of IL – TL gaps. These were the gaps learners noticed when they had the opportunity to compare what they said in their first delivery of the narrative task to a recording of a model speaker performing the same task. The number of IL-TL gaps noticed came from an analysis of participants' note-taking during the comparison stage of the testing sessions.

As discussed in Section 5.8.4, unlike IL gaps noticed, IL-TL gaps noticed could not be clearly categorised as related to either meaning or form. To illustrate this point, a learner noted the phrasal verb '*put up*' on their note paper when comparing their recording to the model narration. During tests, however, it was impossible to know if the learner made this note because they had noticed the verb form (i.e. a grammar-related note) or they had noticed the use of the appropriate word(s) to describe putting up a tent (i.e. a meaning-related note). As mentioned in Chapter 5, while it could have been possible to conduct a second stimulated recall session following the comparison stage to find out why the participant made each note, this would have added a significant amount of time to an already long testing session. Furthermore,

a second stimulated recall session would have increased the period of time between the participant's first delivery and their repeat performance. This would have increased the likelihood of memory decay and taken away from the value of using immediate task repetition as a pedagogic technique to improve speaking performance.

As outlined in Section 5.6.1, to begin the comparison stage, participants were given lined paper and were asked to listen to both recordings (their first performance and the model narration) and to make a note of any differences between the language they used in their narration and the language the model speaker used to express the same meaning. The example below (used in Section 5.8.4) shows a participant's note taking with 10 instances of noticing an IL-TL gap.

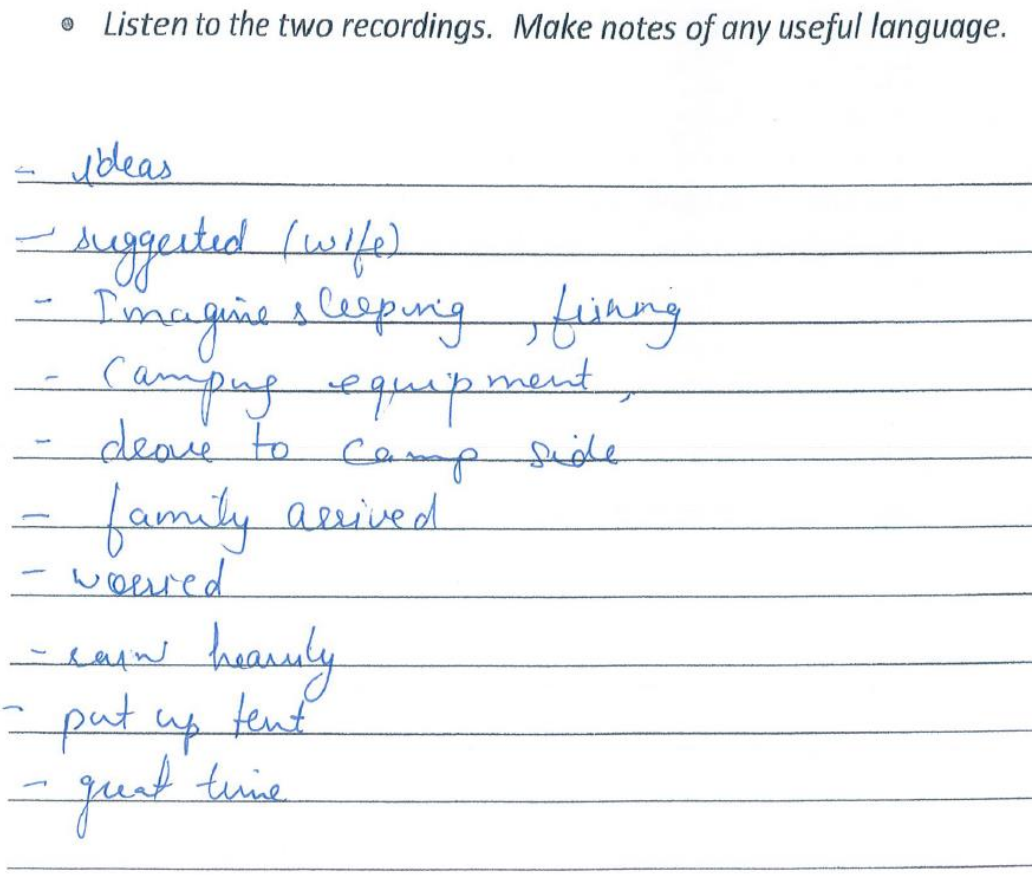


Figure 7.1 – *Example of IL-TL gaps noticed during the comparison stage of a testing session.*

It was hypothesised that following training (i.e. in post-tests) the GN group would notice more IL-TL gaps than the UN group. It was further hypothesised that the UN group would notice more IL-TL gaps than the C group, who would show no increase. As seen in Section 6.2.4, although the GN group increased from 8.42 instances in the pre-test to 10.08 instances in the post-test, this increase of 1.66 was only marginally better than the increase of 1.09 for the UN group who recorded 8.18 instances in the pre-test and 9.27 in the post-test. The control group, on the other hand, recorded 2.08 fewer instances of noticing an IL-TL gap from the pre-test to the post-test (8.25 to 6.17). In the delayed post-test, a small drop was seen from the post-test mean for the GN group, while slight increases were seen for the C and UN groups.

It is possible that the GN group did not improve as hypothesised because of the impact of the high number of total IL gaps noticed on the IL-TL gaps they noticed. As discussed in section 7.2.1, those in the GN group noticed significantly more form-related IL gaps than the other groups in post-tests, and as many or more meaning-related IL gaps and this resulted in a higher total number of IL gaps noticed for the GN group than the UN and C groups (6.08, 3.58 & 2.83 respectively). A similar pattern was seen in the delayed post-test where the GN group outperformed the UN and C groups in total instances of noticing (6.17, 3.17 & 2.50 respectively).

During the comparison stage of the post-tests, when participants were asked to listen to their first delivery recording and the model speaker's recording and note differences in language used (i.e. asked to notice IL-TL gaps), it is possible that those in the GN group used the time only to mine the model speaker's recording for the language they had noticed they needed to fill the IL gaps, rather than to compare the two recordings to find IL-TL gaps. In other words, they were preoccupied with finding language in the model to solve the IL gaps they had noticed earlier, and, as a result, spent less time trying to identify IL-TL gaps. This

suggests that what learners actually do when presented with a model is still unclear, and it may be that training learners to notice IL gaps, and training them to notice IL-TL gaps needs to be conducted separately in order for the former not to take away from the latter.

Support for this possibility was found in an analysis of delivery 1 transcripts and learners' notes from the comparison stage of post-tests. The analysis suggested that some of the notes recorded by GN group learners were indeed related to language they needed to fill their previously noticed IL gaps (rather than to notice differences in language between their recording and that of the model). The following extract, for example, shows a form-related IL gap noticed by a participant in the GN group during the stimulated recall session of the post-test.

> "Yeah, here I have a problem with the verb, with the tense, uh because uh uh in this picture he uh he is driving and the police stop him, but uh I know the story is in the past, but I don't if uh I don't know if I say the past verb or 'ing' for this one."

An analysis of this participant's note-taking from the comparison stage of the post-test shows he wrote, *"he drove too fast"* which is what was said in the model input, and this was the language needed to fill the IL gap. However, as mentioned above, there was not enough evidence in participants' note-taking during the comparison stage to conclusively say that the majority of the comparison time was used to mine model input for language to fill IL gaps, rather than to find new gaps (IL – TL gaps) that had not been noticed when producing output. Because both IL gaps and IL – TL gaps were measured in this study, it is likely that the former impacted on the latter. For instance, it may be that noticing a large number of IL gaps limits a learner's ability to then notice IL – TL gaps because attention is focused on the former. Future

studies may wish to look at how different types of noticing in isolation impact on speech performance. Furthermore, more investigation is needed to find out how different types of noticing may impact on each other (see Chapter 8 for suggestions for future research).

7.2.7 Summary of discussion of findings related to noticing training

To summarise, the overarching question related to noticing was whether learners can be trained to notice. The results from this study strongly suggest that learners can be trained to notice IL gaps. Furthermore, not only can learners be trained to notice more IL gaps, they can be trained to shift their attention in order to notice more form-related IL gaps. However, the methodology used in this study to train learners to notice IL – TL gaps did not result in more noticing.

When it comes to filling IL gaps using language from subsequently presented model input, results indicate that the increase in number of IL gaps noticed, and the change in nature from meaning- to form-related gaps noticed, puts extra burden on working memory resulting in a lower overall percentage of solvable IL gaps filled during the repeat performance.

Having discussed the findings related to training learners to notice, in the following section there is discussion of the impact of such training on their short-term speech performance.

7.3 IMMEDIATE IMPACTS OF NOTICING TRAINING ON LEARNERS' SPEECH PERFORMANCE.

This section discusses the findings related to the immediate effects of noticing training and exposure to model input on learners' speech performance in terms of complexity, accuracy and fluency (CAF). As outlined in Chapter 5, the immediate impacts were found by comparing learners' speech performance in delivery 1 of their oral narrative during testing sessions with their second delivery in the same session. Firstly, Table 7.2 gives an overview of the research

question, hypotheses and findings related to the immediate impacts of noticing training and exposure to model input on CAF. This is followed by a discussion of each finding in turn.

Table 7.2 - *Overview of the research question, hypotheses and findings related to the immediate impacts of noticing training and exposure to model input on learners' speech performance*

Broad Research Question	Specific Research Question	Hypothesis	Finding
3. What are the immediate impacts of noticing training and exposure to model input on learners' speech performance?	3a. What changes in CAF occur in learners' speech performance from delivery 1 to delivery 2 following intervention?	3a. Following training (i.e. in post-tests), compared to the UN and C groups, the GN group will speak with greater accuracy in delivery 2 than in delivery 1 of their narrative task.	3a. Support for this hypothesis was found. The GN group showed significant gains in number of errors per 100 words and percentage of errors self-repaired from delivery 1 to delivery 2 in post-tests compared to the UN and C groups.
		3b. Following training, compared to the UN and C groups, the GN group will speak with greater complexity in delivery 2 than in delivery 1 of their narrative task.	3b. No support was found for this hypothesis. No significant change was seen from delivery 1 to delivery 2 by any group in any test.
		3c. Following training, compared to the GN group, the C and UN groups will speak with greater fluency in delivery 2 than in delivery 1 of their narrative task.	3c. No support for this hypothesis was found as no significant improvements in fluency were found between groups in post-tests.

7.3.1 Immediate impacts of noticing training on CAF

7.3.1.1 Impacts on accuracy

As reported in Chapter 6, there was a significant improvement in two of the three measures of accuracy (number of errors per 100 words and percentage of errors self-repaired) from the first delivery to the second delivery for the GN group over the C and UN groups following treatment (i.e. in post-tests). Therefore, support for hypothesis 3a was found. Also, Chapter 4 included a review of studies showing that current research on the effects of TR on accuracy are unclear (Bui et al., 2019). Some prior TR studies have found no improvement in accuracy in repeat performances (e.g. Bygate, 2001; Boers, 2014; Thai & Boers, 2016), while others have found evidence that TR does lead to improvements in accuracy in iterations (e.g. Lynch & McLean, 2001; Hawkes, 2012; Ahmadian & Tavakoli, 2011). Results from this study support the latter, and, furthermore, add to existing knowledge by showing that participants can be trained to attend to form and incorporate form-related language from model input resulting in improved accuracy in their repeat oral performance in an O-I-O sequence.

7.3.1.2 Impacts on complexity.

Section 7.7 showed there was no significant difference in the grammatical complexity of learners output from delivery 1 to delivery 2 in post-tests between groups, so hypothesis 3b was not supported. The original hypothesis was based on Levelt's (1989) model of speech production (see Chapter 4) which holds that because the content of one's talk is already familiar, pressure is taken off the 'conceptualiser' during a repeat performance of a task allowing more attentional resources to be directed to the 'formulator' which is responsible for grammatical encoding. It was thought that this freeing up of attentional resources, combined with the language presented in the model input and guided noticing training would provide learners in the GN group with the capacity to speak with improved complexity in their second delivery.

There are two potential reasons why no support for hypothesis 3b was found. Firstly, the majority of immediate TR studies have examined effects on speech performance without an intervention between performances. As a result, learners have been able to begin the repeat performance of their given task with their first performance still fresh in mind (whether it be content from the initial delivery or task procedure). As mentioned previously, this familiarity with the task allows more attention to be given to grammatical encoding, including the complexity of the grammar being encoded, during the repeat performance. In order to do this, a learner must hold information from their initial performance, and, as Ahmadian (2013) notes, "the whole process of maintaining a piece of information and doing cognitive work upon it is executed by the working memory system" (p. 41). However, in this study there was an input stage, consisting of an SR session and the opportunity to compare recordings before the task was repeated. The cognitive demands of these activities may have impacted learners' ability to hold information from the initial delivery in working memory. In other words, when delivering the repeat performance, the task of recalling IL gaps and IL-TL gaps, plus recalling language from the model narration to fill gaps may have added load to working memory. This, in turn, may have resulted in the need for the speaker to pay more attention to the conceptualisation of their repeat talk than they otherwise would have had there been no intervention between deliveries.

Furthermore, in the present study, learners were not permitted to use any notes during the repeat performance of the task. They therefore relied on memory for a number of functions: 1) to recall the content from their initial delivery, 2) to recall the IL gaps they had noticed, 3) to recall the IL-TL gaps they had noticed, and 4) to recall language presented in the model input. Trying to recollect the aforementioned information would clearly place pressure on working memory, and this, combined with the cognitive demands of L2 speech production, means that the repeat delivery of the learners' oral narrative becomes a very cognitively

complex task. Any freeing up of attentional resources as a result of being familiar with the content of one's talk after the initial delivery may have been directed to recollection of gaps and model input language rather than to the 'formulator' which is responsible for the complexity of grammatical encoding.

A second reason that may help explain why no support for hypothesis 3b was found relates to the level of complexity of the language presented in the model input. As explained in Chapter 4, level of grammatical complexity was measured in total number of clauses per AS-unit. In the three model narrations used in the testing sessions of this study, the level of complexity was 1.41, 1.44, and 1.51 clauses per AS-unit respectively. Upon examination of learners' output, it was found that the level of complexity in the model narrations was in fact lower than participants' mean complexity scores for first deliveries across all groups, which were 1.51, 1.55 and 1.53 total clauses per AS-unit in the pre-, post- and delayed post-test respectively. In an O-I-O sequence there is an assumption that model input provides more fluent, accurate and complex language (on average) than learners use in the initial output stage, and this more fluent, accurate and complex language can be incorporated by learners in their repeat performance of a task. However, because the participants in this study were already producing more complex speech than what was provided in the model narrations, in their second delivery, on average, it was unlikely participants would have incorporated more complex language from the model input than they had already used themselves.

7.3.1.3 Impacts on fluency

Hypothesis 3c stated: Following training, compared to the GN group, the C and UN groups will speak with greater fluency in delivery 2 than in delivery 1 of their narrative task. This was based on the assumption that the GN group would have a greater focus on form during speech performance and, therefore, would speak with greater accuracy but less fluency owing to a trade-off effect (Skehan, 1996, 1998, 2009), than the C and UN groups who would remain

meaning focused. No support for this hypothesis was found as no significant results involving group emerged from statistical analyses. As a result, there was no evidence in this study that noticing training had any effect on learners' fluency from delivery 1 to delivery 2. However, an interesting point here is that, although the GN group did not improve significantly in any fluency measure, neither did they become less fluent. Considering the gains made by the GN group in accuracy, this is an important point, and is discussed further in the following section.

7.3.1.4 Impacts of noticing training on CAF in light of the Trade-off Hypothesis.

A number of TR studies have found a trade-off in that an increase in fluency results in a decrease in accuracy and/or complexity when comparing speech from an initial performance to a repeat performance (Ahmadian & Tavakoli, 2011; Lynch & Maclean, 2000; Michel et al., 2007; Sample & Michel, 2014; Yuan & Ellis, 2003). In this study, however, the GN group maintained levels of fluency and complexity while they made large improvements in grammatical accuracy from delivery 1 to delivery 2 in post-tests. This gain in accuracy with no corresponding drop in fluency and/or complexity does not provide support for the Trade-Off Hypothesis (Skehan 1996, 1998, 2009). Results from training sessions on the other hand, indicate that the inclusion of model input works to mitigate improvements in fluency that would otherwise be made by an L2 speaker in their repeat performance had they not been exposed to model input (see section 7.6 for a more detailed discussion).

7.3.2 Summary of immediate impacts of noticing training on speech performance.

To sum up the immediate impacts of noticing training on speech performance, after having gone through guided noticing training, learners in the GN group were able to identify more form-related IL gaps in their initial output and fill those gaps with language from model input resulting in significantly improved accuracy during their repeat output stage. Importantly, the GN group was able to speak with improved accuracy in their repeat

performance without a corresponding loss in fluency. The C group who received no noticing training, and the UN group who received unguided noticing training did not show any improvements from their first to their repeat performance despite exposure to the same model input as the GN group. No improvements were seen in complexity for any group.

According to the results of this study, it can therefore be concluded that learners are able to be trained to notice form-related gaps in their output, plus notice and incorporate language presented in model input resulting in more accurate speech in the short term.

7.4 LONGER-TERM EFFECTS OF NOTICING TRAINING ON LEARNERS' SPEECH PERFORMANCE

While the short-term impacts of noticing training on speech were explored through changes in performance from delivery 1 to delivery 2 in post-tests, the longer-term effects were determined by investigating changes in learners' speech over time, that is, by comparing participants' delivery 1 performances over the three testing sessions. Delivery 1 performances were assumed to represent participants' baseline level of proficiency at each time. Therefore, in order to find longer-term effects, improvement would need to be seen from delivery 1 of the pre-test to delivery 1 of the post- and delayed post-tests. However, no significant longer-term changes in speech performance were found (i.e. participants' delivery 1 performance in the pre-test and their delivery 1 performance in the post-tests did not differ significantly according to group). There are two possible explanations why no longer-term improvements in speech performance were seen.

Firstly, it may be that the seven-week duration of this study was not long enough for longer-term improvements to become apparent. Secondly, having only one iteration may not have been enough to allow for automatisation and thus, development of interlanguage. A central tenant of TR as a pedagogical technique is that it can promote automatisation of language (de Jong & Perfetti, 2011). However, the number of repetitions required to achieve

automatisation is unknown. As discussed in Chapter 4, the number of iterations used in TR research to date has varied from one to eleven. It can be assumed that in each iteration, an L2 learner is refining and optimising their language performance (Lambert et al., 2017). However, the process of refinement and optimisation may take longer when a learner has to take into account model input before a repeat performance as was the case in this study. The task of reorganising their talk to incorporate language from input may require more attentional resources to be allocated to the 'conceptualiser' than would otherwise be the case had the learner been required to repeat their performance without exposure to model input. After exposure to model input, a learner may still be finalising the conceptualisation and encoding processes, and it may be that a third or even fourth iteration is needed in order for the learner to become familiar with the content of their talk, and as a result, to free up attentional resources required in order to see the benefits of TR on speech performance.

7.5 CONSIDERATIONS REGARDING THE USE OF A BONFERRONI ADJUSTMENT IN THE PRESENT STUDY

A decision was made to use a Bonferroni adjustment to control familywise error rate when determining levels of significance in the present study. The resulting alpha level of 0.002 meant that results in this study were very conservative. However, it should be noted that there is a certain amount of disagreement in the literature, even among statisticians, about when and if a Bonferroni adjustment should be applied (e.g. Nakagawa, 2004; Perneger, 1998). Had the Bonferroni adjustment not been used in this study, several measures of speech performance from testing sessions would have reached the level of significance with an alpha set at 0.05 (see ANOVA results in appendices 18 – 22).

7.5.1 Summary of discussion of longer-term impacts of noticing training on speech performance.

No longer-term impacts of noticing training on speech performance were found in this study. This may, however, be due to the relatively short duration of the study (7 weeks) and/or because only one iteration was asked for. It is possible that in their repeat performance, having made adjustments after being exposed to model input, learners were still finalising the content of their talk. If so, according to Levelt's (1989) model of speech production and the Limited Attention Capacity Hypothesis (Skehan 1996, 1998, 2009), further iterations may be needed in order to provide the freeing up of attentional resources necessary to then be directed to aspects of speech performance (e.g. CAF).

7.6 SPEECH PERFORMANCE RESULTS FROM TRAINING SESSIONS

Chapter 5 outlined in detail the different conditions in training sessions under which the three groups in this study took part. In brief, all training sessions involved the same three steps, and Steps 1 and 3 were the same for all participants regardless of group. In Step 1, participants recorded themselves performing an oral narrative task based on the same picture sequence (a different picture sequence was used in each training session). In Step 3, all participants repeated their oral narrative task based on the same picture sequence. Between these performances, in Step 2, the C group spent 7.5 minutes doing pronunciation practice which was unrelated to the narrative task. For the UN group, step 2 involved 7.5 minutes of 'unguided' noticing training, that is, they were given a noticing prompt that asked them to compare the recording of their oral narrative to a recording of a model speaker performing the same narrative task. In this stage, those in the UN group were instructed to note any differences in language between the two recordings, whereas the GN group were provided with a 'guided' noticing prompt. This included three parts. Part A targeted the noticing of IL gaps by asking

learners to note down in their first language anything they wanted to but did not know how to say during their first performance (i.e. they were asked to identify grammar-related IL gaps). In part B, participants were asked to listen to the model recording to see if they could find language to fill the gaps they noticed in part A (i.e. they were asked to mine the model input for language to fill their previously noted IL gaps). Lastly, in part C they were asked to compare their recording to that of the model speaker and note any differences in *grammar* between the two recordings (i.e. they were asked to identify IL – TL gaps).

Having recalled the conditions of the training sessions, the following sections discuss the impacts of those conditions on the accuracy, fluency and complexity of their speech performance respectively.

7.6.1 Accuracy.

It was hypothesised that the GN group would outperform the UN and C groups on all measures of accuracy from the first delivery to the second delivery, and that the UN group would in turn outperform the C group. Results partially supported this hypothesis. With training conditions for the GN group aimed at directing participants' attention to form, it was unsurprising to see the GN group outperforming both the C and UN groups on all measures of accuracy (number of errors per 100 words, number of self-repairs per 100 words and percentage of errors self-repaired). Somewhat surprisingly, however, the C and UN groups did not differ from one another. This suggests that the opportunity to listen to a model speaker after performing an oral narrative (as was the case for the UN group) offers no advantage for accuracy unless learners' attention is directed towards the formal features of their output and of the input (as was the case for the GN group).

7.6.2 Fluency.

It was hypothesised that the C group would outperform the UN and GN groups in improvements in fluency from delivery 1 to delivery 2 because they did not listen to model input before their iteration, and therefore had less burden on working memory compared to the UN and GN groups when performing their repeat narrative. Results supported this hypothesis. The C group made significant improvements from their first to their second delivery in speech rate, number of filled pauses, number of silent pauses, number of reformulations and number of self-repairs. The UN group also improved in each measure of fluency, but to a lesser extent than the C group. As with gains made by the C group, improvements made by the UN group are consistent with the psycholinguistic underpinnings of TR as represented by Levelt's (1989) model of speech production (outlined in Chapter 4).

For the GN group, the shift of attention from meaning to form, as evidenced by marked improvements in measures of accuracy, came at the cost of fluency. With the exception of number of reformulations per 100 words, which can be an indication of attention to form, the GN group performed less fluently from delivery 1 to delivery 2 in all measures of fluency. Unlike fluency results from testing sessions, the fluency results here do provide support for Skehan's (1996, 1998, 2009) Trade-off Hypothesis and are consistent with Levelt's (1989) model of speech production.

7.6.3 Complexity.

It was hypothesised that the GN group would improve in grammatical complexity compared to the C and UN groups. However, no support for this hypothesis was found. A likely explanation is the same as was given for complexity in testing sessions, that is, on average, the participants in this study were already performing the speaking tasks with greater complexity than the model speaker. Furthermore, as Skehan (2014) points out, studies

investigating speech performance in terms of CAF have largely agreed that improvements can be made in one aspect of form but not both (i.e. improvement in accuracy *or* complexity, but not accuracy *and* complexity), and the results of this study are consistent with this.

7.7 SUMMARY OF SPEECH PERFORMANCE RESULTS IN TRAINING SESSIONS

Although each group began and ended each training session performing an oral narrative task, the intervention between performances differed. Results showed those conditions had an impact on participants' speech performance that was consistent with Levelt's (1989) model of speech production. Those in the C group were not required to process model input during their intervention, learners in the UN group processed model input maintaining the tendency that they showed in the pre-test to attend to meaning rather than form (i.e. semantic processing), and those in the GN group were required to process input with their attention directed towards form (i.e. syntactic processing). As a result of the differing intervention conditions, looking at the three training sessions combined, the C group made clear and considerable gains in all measures of fluency over the UN and GN groups with the exception of mean length of silent pause. The UN group also made gains in all measures of fluency, but to a lesser extent. The GN group, on the other hand, became more dysfluent from delivery 1 to delivery 2 in all measures of fluency with the exception of number of reformulations per 100 words. In terms of accuracy, the GN group clearly outperformed the C and UN groups on all measures, while the UN group also made small gains in number of errors per 100 words and percentage of errors self-corrected. The C group performed more poorly from delivery 1 to delivery 2 on all measures of accuracy. Finally, with regard to complexity, no significant changes were seen from delivery 1 to delivery 2 for any group.

7.8 CHAPTER SUMMARY

In this chapter, the findings as they relate to noticing, immediate impacts on speech performance and longer-term impacts on speech performance have been discussed. This study contributes to existing noticing research by showing that while learners have a natural tendency to notice meaning-related gaps in their output along with meaning-related language presented in model input, they can in fact be trained to notice form-related gaps, and this has a positive impact on accuracy in the short-term. Crucially, significant short-term gains made in accuracy for the GN group came while maintaining levels of fluency.

The lack of gains in fluency in this study (contrary to what is found in the majority of TR studies) was explained in light of the Limited Attention Capacity Hypothesis (Skehan, 1996, 1998, 2009) and by the increased cognitive demands of the task brought about by the input stage in tests. In training sessions, when cognitive demands were lower for the C and UN groups owing to their respective intervention conditions, gains in fluency were apparent. The same fluency gains were not seen by the GN group whose intervention conditions were more cognitively demanding due to the syntactic processing of form required. However, this did lead to significant gains in accuracy.

No longer-term impacts of noticing training on speech performance were evident in the results which may be because the seven-week duration of this study did not allow enough time for longer-term improvements to become apparent.

In the following chapter, implications for L2 theory and pedagogy and presented along with a discussion of the limitations of the research presented here and suggestions for future research.

8 Conclusion

8.1 INTRODUCTION

This study makes an important contribution to existing knowledge on noticing in SLA by showing that learners can be trained to notice, and this results in improved accuracy in speech performance. After a brief summary of key findings, this chapter presents a discussion of where the results lie in relation to previous research in the same area. Next, methodological, theoretical and pedagogical implications of the findings are explored. The chapter concludes with a discussion of the limitations of this study along with identification of possible avenues for future research.

This study brought together two key constructs from SLA literature, namely, task repetition, and noticing. The key motivation for this research was to investigate whether these two features could be used to help improve L2 speech performance in a TBLT context. Chapter 1 highlighted the need for pedagogic techniques in TBLT that are designed to direct L2 learners' attention towards form while maintaining a primary focus on meaning during speaking tasks. TR has been used as a tool to achieve this redirection/reallocation of attention. As outlined in Chapter 4, the theoretical and psycholinguistic underpinnings of TR include the notion that attentional resources are freed up after delivering an initial performance of a speaking task, and these resources can then be reallocated to the grammatical encoding of speech during the repeat performance, resulting in more accurate and/or complex L2 speech. However, as was also highlighted in Chapter 1, results from TBL research to date have produced unclear findings on the impacts of TR on accuracy and complexity, and what might, therefore, be needed is some kind of intervention between performances designed to direct learners' attention towards form.

Chapter 3 explained how influential *noticing* has been in SLA theory and research. However, it was also noted that despite the high degree of influence, *noticing* has received relatively little empirical investigation, especially in oral modality. The little research conducted to date has found that the act of producing L2 speech can be a trigger for learners to notice gaps in their interlanguage, and that subsequent exposure to relevant input can orient learners' attention to language in the input that can be used to fill those gaps. Findings from noticing research involving L2 output, whether written or spoken, have shown that learners notice mainly meaning-related aspects of their output and of model input. However, no studies have sought to investigate whether learners can be trained to overcome this tendency to notice meaning and notice form-related features of their output and of model input instead.

By combining noticing and TR, results from this study show that L2 learners can be trained to notice form-related gaps in their output, and then fill those gaps during a repeat performance of the same task after being exposed to relevant model input, and that this, in turn, results in improved accuracy in speech performance.

8.2 THEORETICAL IMPLICATIONS

Firstly, results from this study shed light on the nature of the IL gaps that learners notice when asked to perform an L2 speaking task. As stated in Chapter 1, prior to conducting the experiment reported in this thesis, previous research has established that the vast majority of the IL gaps L2 learners notice when writing are lexical, however, what it is that learners notice when producing oral output had yet to be established. Results reported here show that when left undirected, a vast proportion of the IL gaps learners notice in oral output are also lexical.

Results from the research reported in this thesis also support the claim that the act of producing L2 output can lead a learner to discover what it is they do not know in their L2. Furthermore, after identifying what it is they do not know (i.e. after identifying IL gaps),

learners then attend more closely to linguistic features of subsequently presented relevant input in order to find language to fill gaps noticed. However, this research adds to existing knowledge by demonstrating that L2 learners can be trained to not only notice *more* IL gaps when producing L2 output, but to also notice more *form-related* IL gaps. Furthermore, and of high importance, is the finding that as a result of noticing training and the subsequent shift in attentional focus, learners significantly improve the accuracy of their speech performance when given the chance to repeat a task, and this higher degree of accuracy comes while maintaining levels of fluency and complexity. This ability to improve accuracy in the short-term without a corresponding drop in fluency and/or complexity is of importance to TBLT as researchers have long been concerned with finding ways to direct L2 learners' attention towards form while maintaining a primary focus on meaning during L2 tasks in order to achieve a better balance of CAF.

Another theoretical implication from this study is that it provides support for the idea of different types of noticing. As outlined in Section 2.4, since Schmidt and Frota's (1986) original idea of *noticing*, and Swain's (1995) original *noticing function*, a number of different types of noticing have appeared in L2 literature (e.g. also in Section 2.4, Izumi [2013] outlines 4 different types). Because the two types of noticing investigated in this study reacted differentially to the treatment, results add support for the claim of different types of noticing.

8.3 PEDAGOGICAL IMPLICATIONS

As mentioned previously, one of the motivations for this study was to examine a technique for improving L2 learners' performance that could be applied in the L2 classroom. Results from this study show that the opportunity to produce initial output, followed by exposure to model input, and then the provision of repeating initial output has a role in improving the formal features of L2 speech in the classroom. However, findings also suggest

that formal features are only improved as a result of an intervention designed to shift learners' attention from meaning towards form. Without directing learners' focus, the natural tendency is to attend to meaning-related features (e.g. lexis) of both output and input. Furthermore, findings show that exposure to model input before repeating a speaking task mitigates positive impacts on fluency that would have occurred had the model input not been presented. In other words, the largely consistent finding from previous research that L2 fluency is enhanced when learners are given the opportunity to immediately repeat the same speaking task was not found in the testing sessions of this study (when learners were exposed to model input before their repeat performance). Fluency was significantly improved, however, in training sessions but only by participants in the C Group, who received no model input, and by participants in the UN group, who received model input, but did not have their attention shifted away from meaning.

Findings suggest that as far as fluency is concerned, exposure to input before a repeat performance interferes with working memory processes (e.g. the ability to hold information from an initial performance in working memory ready for retrieval in a repeat performance) as a result of the need to recall content from one's first delivery plus recall language from model input. As mentioned in Chapter 4, the theoretical underpinnings of TR include the notion that attentional resources that are devoted to the 'conceptualiser' during an initial performance are freed up during the repeat performance(s), and, as a result, can be reallocated to the 'formulator' and articulator' resulting in more accurate and complex speech. However, results from this study suggest that as the L2 learner makes changes to their talk following exposure to model input, these changes are such that attentional resources still needed to be allocated to the 'conceptualiser' as participants reorganise their new talk. It may be, therefore, that further repetitions are needed in order for the learner to refine their talk to the point where pressure is taken off the 'conceptualiser' and more attention can be directed to grammatical encoding.

Results from TR research have consistently found that learners are able to improve in fluency when given the opportunity to repeat a speaking task. However, improvements in accuracy and/or complexity as a result of TR alone have been hard to come by, thus the balance between CAF mentioned in the quote above has remained elusive. Results from this study, however, show that manipulating the design further by providing exposure to model input in addition to TR can achieve a much better balance between CAF than when providing TR alone. However, this better balance in CAF only comes if learners' attention is directed towards the formal features of their output and, as a result, to the formal features in model input.

There are a number of possible issues with regard to the pedagogic implications in light of the findings of this study. Firstly, a long-standing issue has been that L2 classroom do not often make use of the task design variable of TR (Rossiter, Derwing, Manintim & Thomson, 2010). A second issue is that, in my experience in L2 teaching and teacher training, it is rare that L2 teachers make use of model *spoken* input in the L2 classroom. However, if a teacher were to design a speaking task utilising TR and model input, results from this study show that this would not work to achieve a balance in CAF as is needed if TBLT is to be successful in promoting acquisition (Ellis & Shintani, 2013). Findings from this study show that one way to shift learners' attention towards form is to train them to notice the formal features of their output and of language presented in model input. Only then will TR and the provision of model input lead to a better balance of CAF.

8.4 LIMITATIONS OF THE PRESENT STUDY

As with all studies, there are limitations that should be considered. Firstly, with 36 participants divided into three groups, the sample size for the present study is considered small. However, employing a larger sample size was not possible for two main reasons. Firstly, the pool of potential participants was limited as they came from a relatively small group of students

all at the same level (EAP) within the same language centre. It was necessary to have participants from the same school in order to conduct the study as planned (e.g. to be able to collect data from participants during their regularly scheduled class time).

A second reason for using a relatively small sample size was because of the time that was anticipated to be needed to analyse data. Indeed, the intensive analysis of speech performance in this study was highly time consuming. With 36 participants each narrating a story twice on six occasions, there were 432 narrations to manually transcribe, code and analyse. On top of this, there were 108 stimulated recall sessions to analyse. While some analysis could be done with the help of automated technology, this option was not always available. For instance, when counting and measuring silent pauses, the 'textgrid silences' function in PRAAT (Boersma & Weenink, 2007) can be used. While this was suitable for recordings made in the quiet office environment during testing sessions of this study, for recordings from training sessions which were conducted with each class as a whole in the computer room of the university, due to background noise the automatic identification of silent pauses was not possible. Furthermore, even when the function was used, pause boundaries identified by PRAAT still had to be checked manually to ensure accuracy. The amount of time that was anticipated to be needed to transcribe, code and analyse meant that any more participants would have made data collection coding and analysis unrealistic.

Another limitation relates to the measuring of noticing. As mentioned in Chapter 2, quantifying what has been noticed has been problematic as a data elicitation method because it requires the measurement of learner-internal processes. In this study, dysfluent markers in participants' speech, such as undue pauses, hesitations, reformulations, repetitions and false starts were taken as a potential indication that the learner had encountered a linguistic problem. During stimulated recall sessions, while watching a playback of the learner's performance, participants were asked if they recalled what they were thinking at the time the problem

appeared to be encountered. Although the participants in this study were at a B2/B2+ (CEFR) level of English proficiency, for some it was difficult to articulate the problem encountered.

The level of difficulty in reporting their thought processes at the time the problem was encountered was likely dependent on two main factors. Firstly, the level of L2 English of the learner, and secondly, the nature of the problem encountered. Two examples from the present study can be used to illustrate this point. Firstly, one would assume it is easier for Learner A to say that they paused during the speech production process because they did not know a particular word, compared to Learner B who paused because they did not know how to express the idea that a character in the picture prompt felt a sense of frustration and regret after realising that his planned camping weekend was a disaster and that he should have agreed to his wife's original suggestion of going to the beach. Assuming equivalency in L2 English proficiency, it is clearly more difficult for Learner B to report the problem they encountered than it is for learner A. In fact, it took learner B several minutes and multiple attempts to report the idea above.

As a result of difficulty in reporting gaps noticed and thought processes, there is the resulting possibility that participants in this study noticed gaps in their output while narrating their story, but did not report them because either they did not have the confidence to do so accurately in their L2, or because the problem went unidentified by the researcher (i.e. there was no obvious dysfluency marker to indicate a problem). When conducting the stimulated recall sessions for this study, it soon became apparent that the reporting of gaps noticed came from researcher-initiated questions, rather than through information volunteered by the learner. Therefore, the only instances of noticing that were likely to be reported by learners came after the researcher stopped the recording to ask if the learner recalled what they were thinking at the time the speaker appeared to be encountering a problem.

In addition, there were a few instances when a learner stopped the recording to talk about a problem they encountered and what they were thinking at the time. This was despite the fact that there was no dysfluency marker to indicate a problem. Had the learner not stopped the recording, the noticed gap would have gone unreported.

Also related to stimulated recall sessions is the probability that if given the choice, participants would rather have conducted the stimulated recall session in their first language in order to allow more freedom in reporting the L2 problems they encountered when narrating their story. Several previous studies in the same area have been in an EFL context where all participants share the same L1, and stimulated recall sessions were conducted using participants' L1. In Sheppard's (2006) study, of the 81 participants who took part in stimulated recall sessions, all but two chose to do so in their L1 (Japanese), presumably because they felt more comfortable reporting the necessary information. With the range of L1s represented by the participants in this study, conducting stimulated recall sessions in each participant's L1 would have required the help of research assistants who spoke the necessary L1 and could be trained to conduct stimulated recall sessions, something that was unrealistic for this study.

8.5 AVENUES FOR FUTURE RESEARCH

There are a number of avenues that could be considered for future research in light of the findings from the results presented in this thesis. Firstly, future investigations into the kind of noticing examined in this study may wish to look at whether level of L2 proficiency plays a role in what gaps are noticed and whether they are filled after exposure to model input. For instance, after identifying IL gaps, whether lower-level learners have the necessary L2 listening skills to accurately identify language presented in model input to fill those gaps is something that needs to be answered before any real-world, L2 classroom applications can be considered for lower-level learners.

Another potential area for investigation in future research is to look at how noticing training impacts more specifically on 'accuracy' of speech performance. This study used 'accuracy' as a global measure, however, future studies may want to explore whether there are certain types of errors that L2 learners can be trained to notice and fill after exposure to input.

Also, in this study two types of noticing were examined (IL gaps and IL – TL gaps). Although each of these types of noticing occurs at a different time, it is possible the former impacted upon the latter in this study. As mentioned in Chapter 7, for participants in the GN group who had noticed a larger number of IL gaps when producing their initial output, it is possible they spent their time during the comparison stage mostly (or totally) trying to mine the model input for language to fill their previously noticed gaps, rather than to identify new gaps that they had not noticed when producing output (i.e. IL – TL gaps). As explained in Chapter 7, this may be one reason why those in the GN group noticed a higher number of IL gaps than the other groups but did not notice a significantly higher number of IL – TL gaps. Therefore, there is a need for further, separate examination of noticing training for these two different types of noticing.

Results from this study showed that while those in the GN group noticed a significantly higher number of IL gaps than those in the other groups, they actually filled a lower percentage of solvable gaps during their repeat performance than the other groups. They also filled a lower percentage of solvable gaps in post-tests compared to the percentage they filled in the pre-test. As discussed in Chapter 7, this is possibly because of the added cognitive burden resulting from the need to hold more information in working memory during their repeat performance. Therefore, future studies might explore the relationship between the number of IL gaps noticed when producing output, the percentage of solvable IL gaps filled after exposure to model input, and working memory capacity. Along the same lines, more investigation is needed into the nature of solvable IL gaps filled. In this study, as mentioned above, while the number of form-

related IL gaps noticed increased for the GN Group, their percentage of solvable gaps filled went down. It is plausible that form-related solvable gaps noticed are more demanding to fill than meaning-related gaps (refer to Section 7.2.5 for a discussion).

Finally, the extent to which the findings from the research presented here can be generalised beyond the context within with they were collected needs to be considered, as it does with all studies. Although participants in this study represented six different L1s, of the total of 36, 29 were from Chinese or Hindi L1 backgrounds (16 and 13 participants respectively). All participants had completed a university degree in their home country before arriving in Australia to continue their L2 English education, and the age range participants represented was relatively narrow. It remains to be seen whether the findings of this study would apply for L2 learners in other contexts and with target languages other than English.

Lastly, whether results of noticing training and its impacts on immediate gains in speech performance as reported in this study apply to other types of tasks, can be shown to impact speech performance over time, and can be shown to transfer to a new task are all questions that require further investigation.

REFERENCES

Adams, R. (2003). L2 output, reformulation and noticing: Implications for IL development. *Language Teaching Research*, *7*(3), 347-376.

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: does it transfer to a new task? *The Language Learning Journal*, *39*(3), 269-280.

Ahmadian, M. J. (2012). Task repetition in ELT. *ELT Journal*, *66*(3), 380-382.

Ahmadian, M. J. (2013). Working memory and task repetition in second language oral production. *Asian Journal of English Language Teaching*, *23*(1), 37-55.

Ahmadian, M. J. (2016). Task-based language teaching and learning. *The Language Learning Journal*. *44*(4), 377-380

Ahmadian M. J., García Mayo M (eds.) 2018. *Recent Perspectives on Task-based Language Learning and Teaching.* Trends in Applied Linguistics. De Gruyter Mouton

Ahmadian, M. J., & Tavakoli, M. (2010). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, *15*(1), 35-59.

Ahmadian, M. J., & Tavakoli, M. (2014). Investigating what second language learners do and monitor under careful online planning conditions. *Canadian Modern Language Review*, *70*(1).

Ahmadian, M. J., Mansouri, S. A., & Ghominejad, S. (2017). Language learners' and teachers' perceptions of task repetition. *ELT Journal*, *71*(4), 467-477.

Ahmadian, M. J., Tavakoli, M., Ketabi, S., & Kasaian, Z. (2010). On the benefits of careful within-task planning and task repetition in EFL classrooms. *English Language Teaching*, *3*(1), 155-158.

Arevart, S., & Nation, P. (1991). Fluency improvement in a second language. *RELC Journal*, *22*(1), 84-94.

Audacity Team (2017). Audacity®: Free Audio Editor and Recorder [computer application]. Version 2.1.3 retrieved June 8th 2017, from https://audacityteam.org/

Awwad, A., Tavakoli, P. and Wright, C. (2017). 'I think that's what he's doing': Effects of intentional reasoning on second language (L2) speech performance, *System*, *67*(1), 158-169.

Bei, X.G. (2013). Effects of immediate repetition in L2 speaking task: A focused study. *English Language Teaching*, (*6*), 11–19.

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal*, *45*(3), 221-235.

Boersma, P., & Weenink, D. (2007). Praat (Version 4.5.25) [Software].

Boston, J. S. (2008). Learner mining of pre-task and task input. ELT Journal, 62(1), 66-76.

Boston, J. S. (2009). Pre-task syntactic priming and focused task design. ELT Journal, 64(2), 165- 174.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. London: Routledge.

Brumfit, C. J. 1979. 'Notional syllabuses—a reassessment'. System 7(2), 111–6.

Brumfit, C. J. 1984. *Communicative Methodology in Language Teaching*. Cambridge University Press.

Bui, G., Ahmadian, M. J., & Hunter, A. M. (2019). Spacing effects on repeated L2 task performance. *System*, (*81)*, 1-13.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23-48).

Bygate, M. (2016). Sources, developments and directions of task-based language teaching. *The Language Learning Journal*, *44*(4), 381-400.

Bygate, M. (2018). *Learning Language through Task Repetition*. Amsterdam: John Benjamins

Bygate, M., Skehan, O., & Swain, M. (Eds.). (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow, England: Longman.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*(4), 535-544.

Cobb, T. (2018). *Range for texts v.3* [computer program]. Accessed 15 Feb 2016 at https://www.lextutor.ca/cgi-bin/range/texts/index.pl

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Crookes, G. (1986). Task classification: A cross-disciplinary review (Tech. Rep. No. 4). Honolulu: University of Hawaii at Manoa, Social Science Research Institute, Center for Second Language Classroom Research.

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing. *Written Communication, 7*(4), 482- 511.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132.

de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In The 6th workshop on disfluency in spontaneous speech (diss) (pp. 17-20).

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533-568.

de Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*(3), 387-404.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*(2), 183-193.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition, 20*, 349–386.

Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-261). Cambridge: Cambridge University Press.

Egi, T. (2008). Investigating stimulated recall as a cognitive measure: Reactivity and verbal reports in SLA research methodology. *Language Awareness, 17*, 212 – 217.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ellis, R. (ed.). 2005. *Planning and Task Performance in a Second Language*. John Benjamins.

Ellis, R. (2008). *The Study of Second Language Acquisition*. Oxford University Press.

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. (2015). Understanding second language acquisition 2nd Edition-Oxford applied linguistics. Oxford university press.

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, *20*(3), 405–428.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Ellis, R., & Mifka-Profozic, N. (2013). Recasts, uptake, and noticing. In J. Bergsleithner, S. Frota, & J. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor*

of Richard Schmidt (pp. 61-79). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. Routledge.

Fillmore, C. J. (1979). On fluency. In D. Kempler, and W. S. Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85-102). New York: Academic Press.

Foster, P., & Skehan, P. (1996). The Influence of Planning and Task Type on Second Language Performance. *Studies in Second Language Acquisition 18*, 299-323.

Foster, P., & Skehan, P. (2013). Anticipating a post-task activity: The effects on accuracy, complexity, and fluency of second language performance. *Canadian Modern Language Review*, *69*(3), 249-273.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.

Foster, P., & Wigglesworth, G. (2016). Capturing Accuracy in Second Language Performance: The Case for a Weighted Clause Ratio. *Annual Review of Applied Linguistics*, *36*, 98-116.

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research*, *20*(3), 321-340.

García Mayo, M. D. P., & Imaz Agirre, A. (2016). Task repetition and its impact on EFL children's negotiation of meaning strategies and pair dynamics: An exploratory study. *The Language Learning Journal*, *44*(4), 451-466.

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Erlbaum.

Gass, S. M., & Mackey, A. (Eds.). (2013). *The Routledge handbook of second language acquisition*. Routledge.

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). New York: Routledge.

Gass, S. (2015). Comprensible input and output in classroom interaction. In N. Markee (Ed.), *The handbook of classroom discourse and interaction* (pp. 182-197). Malden, MA: Wiley-Blackwell.

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review*, *61*(3), 325-353.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), 215-240.

Hama, M., & Leow, R. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*(3), 465-491.

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research*, *11*(4), 459-479.

Hanaoka, O., & Izumi, S. (2012). Noticing and uptake: Addressing pre-articulated covert problems in L2 writing. *Journal of Second Language Writing*, *21*(4), 332-347.

Hawkes, M. L. (2011). Using task repetition to direct learner attention and focus on form. *ELT Journal*, *66*(3), 327-336.

Heaton, J. (1975). *Beginning Composition through Pictures.* London: Longman.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics, 30*(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Huensch, A., & Tracy–Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal, 101*(2), 275-293.

Hunter, J. (2012). "Small Talk": Developing Fluency, Accuracy, and Complexity in Speaking. *ELT Journal, 66*(1), 30-41.

IELTS speaking band descriptors (public version). (2016). Retrieved April 4, 2017, from https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

Issues In English 2 (2005) ProTea Textware [Computer software]. Retrieved from https://www.proteatextware.com/shopexd.asp?id=39&bc=yes

Izumi, S. 2002. 'Output, input enhancement and the noticing hypothesis,' *Studies in Second Language Acquisition 24*, 541–77.

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis. *Applied Linguistics*, *24*(2), 168-196.

Izumi, S. (2013). Noticing and L2 development: Theoretical, empirical and pedagogical issues. In J. Bergsleithner, S. Frota, & J. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 25-38). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *Tesol Quarterly*, *34*(2), 239-278.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition*, *21*(3), 421-452.

Izumi, Y., & Izumi, S. (2004). Investigating the effects of oral output on the learning of relative clauses in English: Issues in the psycholinguistic requirements for effective output tasks. *Canadian Modern Language Review*, *60*(5), 587-609.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? Journal of Quantitative Linguistics, 21 (3), 223-245.

Kim, Y., & Tracy-Ventura, N. (2013). The role of task repetition in L2 performance development: What needs to be repeated during task-based interaction? *System*, *41*(3), 829-840.

King, J. (Ed.) (2015). *The dynamic interplay between context and the language learner*. Basingstoke: Palgrave Macmillan.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality and aptitude in narrative task performance. *Language Learning*, 62(2), 439–472.

Krashen, S. D. (1980). The input hypothesis: Issues and implications. In J.Alatis (Ed.), *Current issues in bilingual education* (pp. 144-158). Washington, D.C.: Georgetown University Press.

Krashen, S. D. (1985). *The Input Hypothesis: Issues and implications*. London: Longman.

Lahmann, C., Steinkrauss, R., & Schmid, M. S. (2017). Speed, breakdown, and repair: An investigation of fluency in long-term second-language speakers of English. *International Journal of Bilingualism*, *21*(2), 228-242.

Lambert, C., & Kormos, J. (2014). Complexity, Accuracy, and Fluency in Task-based L2 Research: Toward More Developmentally Based Measures of Second Language Acquisition. *Applied Linguistics*, *35*(5).

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167-196.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, *30*(4), 579-589.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language earning*, *40*(3), 387-417.

Lennon, P. (1990). The advanced learner at large in the L2 community: Developments in spoken performance. *IRAL-International Review of Applied Linguistics in Language Teaching*, *28*(4), 309-324.

Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning, 47*, 467– 506.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware vs. unaware learners. *Studies in Second Language Acquisition, 22*, 557-584.

Leow, R. P. (2018). Noticing hypothesis. In *The TESOL Encyclopedia of English Language Teaching*, eds J. I. Liontas, TESOL International Association, and M. DelliCarpini. Hoboken, NJ: Wiley-Blackwell.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Levelt, W. J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Loewen, S. (2004). Uptake in Incidental Focus on Form in Meaning-Focused ESL Lessons. *Language Learning, 54*(1), 153-188.

Long, M. H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77-99). Clevedon, England: Multilingual Matters.

Long, M. (2014). *Second language acquisition and task-based language teaching*. John Wiley & Sons.

Lynch, T. (2001). Seeing what they meant: Transcribing as a route to noticing. *ELT Journal*, *55*(2), 124-132.

Lynch, T. (2007). Learning from the transcripts of an oral communication task. *ELT Journal*, *61*(4), 311-320.

Lynch, T. (2018). Perform, reflect, recycle. Learning Language through Task Repetition. In M. Bygate (Ed), *Learning language through task repetition* (pp. 193-232). John Benjamins Publishing Company.

Lynch, T., & Maclean, J. (1994). Poster carousel. In K. Bailey & L. Savage (Eds.), *New ways of teaching speaking* (pp.108-109). Washington, DC.

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research*, *4*(3), 221-250.

Lynch, D., & Maclean, J. (2001). A case of exercising: effects of immediate task repetition on learners' performance. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing.* Harlow, Essex: Addison Wesley Longman.

Mackey, A. 2002. 'Beyond production: Learners' perceptions about interactional processes,' *International Journal of Educational Research 37*, 379–94.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, *27*(3), 405-430.

Mackey, A., & Gass, S. M. (2016). *Stimulated recall methodology in applied linguistics and L2 research*. Routledge.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*(4), 471-497.

Mackey, A., Philip, J., Egi, t., Fuji, A. y Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback, and L2 development. En P. Robinson, (Ed.), *Individual differences and instructed language learning*, (pp. 181- 209). Philadelphia: John Benjamins.

Maurice, K. (1983). The fluency workshop. *TESOL Newsletter, 17*(4), 29.

Mayo, M. D. P. G., Agirre, A. I., & Azkarai, A. (2017). Task Repetition Effects on CAF in EFL Child Task-Based Oral Interaction. *Recent Perspectives on Task-based Language Learning and Teaching*, *27*, 11.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83-108.

Mennim, P. (2007). Long-term effects of noticing on oral output. *Language Teaching Research*, *11*, 265 – 280.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(3), 241-259.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044-1045.

Nation, P. (1989). Improving speaking fluency. *System*, *17*(3), 377-384.

NetSupport (2017). NewSupport®: [computer application]. Retrieved June 15th 2017, from https://netsupportsoftware.com

Newton, J. (2016). Researching tasks. *Language Teaching Research, 20*(3) 275–278.

Nguyen, B. T. T., & Newton, J. (2019). Learner proficiency and EFL learning through task rehearsal and performance. *Language Teaching Research*. https://doi.org/10.1177/1362168818819021

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.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge University Press.

Nunan, D. (2006). Task-based language teaching in the Asia context: Defining 'task'. *Asian EFL Journal*, *8*(3),12-18.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590-601.

Patanasorn, C. (2010). *Effects of procedural content and task repetition on accuracy and fluency in an EFL context*. Northern Arizona University.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal*, *316*(7139), 1236-1238.

Philp, J. & N. Iwashita (2013). Talking, tuning in and noticing: Exploring the benefits of output in task-based peer interaction. *Language Awareness 22*(4), 353–370.

Prabhu, N. S. 1982. 'The Communicational Teaching Project, South India.' Madras: The British Council.

Prabhu, N. S. (1987). *Second language pedagogy* (Vol. 20). Oxford: Oxford University Press.

Rebuschat, P. & Williams, J.N. 2013. Implicit learning in second language acquisition. In *The Encyclopedia of Applied Linguistics*, C. A. Chapelle (ed.). Oxford: Wiley Blackwell, Vol. 20.

Révész, A., Ekiert, M., & Torgersen, E. N. (2014). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*(6), 828-848.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423-441.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition, 19*, 223-247.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Sample, E., & Michel, M. (2014). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. *TESL Canada Journal*, 23-23.

Sato, M. (2007). Social relationships in conversational interaction: A comparison between learner-learner and learner-NS dyads. *JALT Journal, 29,* 183 – 208.

Schmidt, R. (1983). Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. *Sociolinguistics and Language Acquisition*, *137*, 174.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied linguistics 11*(2): 129–158.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, *14*(4), 357-385.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. AILA Review, 11, 11-26. Retrieved from http://www.aila.info/download/publications/review/AILA11.pdf#page=11 308

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge, England: Cambridge University Press.

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker, Proceedings of CLaSIC 2010, Singapore, December 2-4 (pp. 721-737). Singapore: National University of Singapore, Centre for Language Studies.

Schmidt, R. & Frota, S. 1986. Developing basic conversational ability in a second language: a case study of an adult learner of Portuguese. In Richard R. Day (ed.) *Talking to learn: conversation in second language acquisition*. Rowley, MA: Newbury House, pp. 237–326.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79-95.

Truscott, J., & Sharwood Smith, M. (2011). Input, intake and consciousness: The Quest for a Theoretical Foundation. *Studies in Second Language Acquisition, 33*(4), 497-528.

Sheppard, C. 2006. 'The effects of instruction directed at the gaps second language learners noticed in their oral production'. Unpublished PhD thesis, University of Auckland, New Zealand.

Sheppard C., & Ellis, R. (2018) The effects of awareness-raising through stimulated recall on the repeated performance of the same task and on a new task of the same type. In M. Bygate (Ed), *Learning language through task repetition* (pp. 171-192). John Benjamins Publishing Company.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38-62.

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics*, *18*, 268-286.

Skehan, P. (2003). Task-based instruction. *Language teaching*, *36*(1), 1-14.

Skehan, P. (2007). Language instruction through tasks. In *International handbook of English language teaching* (pp. 289-301). Springer, Boston, MA.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics*, *30*(4), 510-532.

Skehan, P. (2011). Researching tasks: Performance, assessment and pedagogy. *Shanghai Foreign Language Education Press, Shanghai*, *18*, 184-189.

Skehan, P. (Ed.). (2014). *Processing perspectives on task performance* (Vol. 5). John Benjamins Publishing Company.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185-211.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 97-111.

Smith, B. (2012). Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology*, *16*, 53 – 81.

Song, M. J., Suh, B. R. (2008). The effects of output task types on noticing and learning of the English past counterfactual conditional. *System 36*, 295–312.

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* (pp. 235-252). Rowley, MA: Newbury House.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64-81). New York: Cambridge University Press.

Swain, M., 1995. Three functions of output in second language learning. In: Cook, G., Seidlhofer, B. (Eds.), *Principle and Practice in Applied Linguistics: Studies in Honor of H.G. Widdowson*. Oxford University Press, Oxford, pp. 125–144.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J.P. Lantolf (Ed.) *Sociocultural theory and second language learning* (pp. 97–114). 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). Routledge.

Swain, M., Lapkin, S., 1995. Problems in output and the cognitive processes they generate: a step towards second language learning. *Applied Linguistics 16* (3), 371–391.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133-150.

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *Tesol Quarterly*, *50*(2), 447-471.

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.

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2017). Scoring validity of the Aptis Speaking Test: Investigating fluency across tasks and levels of proficiency. ARAGs Research Reports Online.

Thai, C., & Boers, F. (2015). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *Tesol Quarterly*, *50*(2), 369-393.

Thornbury, S. (2000). Targeting accuracy, fluency and complexity. *ETProfessional*, *16*, 3-6.

Truscott, J. (1998). Noticing in second language acquisition: A critical review. *Second Language Research, 14*(2), 103-135.

Uggen, M. S. (2012). Reinvestigating the noticing function of output. *Language Learning, 62*(2), 506-540.

Van den Branden, K. (2007). Practice in perfect learning conditions? In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 161-179). Cambridge: Cambridge University Press.

Van den Branden, K. (2016). The role of the teacher in task-based language teaching. *Annual Review of Applied Linguistics, 36*, 164-181.

Vercellotti, M. L. (2015). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 1–23.

Watanabe, Y. (2008). Peer–peer interaction between L2 learners of different proficiency levels: Their interactions and reflections. *The Canadian Modern Language Review, 64*, 605–635.

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research, 11*, 121–142.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106.

Williams, J. (2001). Learner-generated attention to form. *Language Learning, 51*, 303-346.

Willis, D. and Willis, J. (2001). Task-based language learning. In R. Carter and D. Nunan (Eds.), *The Cambridge Guide to Teaching English to Speakers of Other Languages*. Cambridge: Cambridge University Press.

Wolf-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. Honolulu: U.

Wright, C. and Tavakoli, P. (2016) New directions and developments in defining, analyzing and measuring L2 speech fluency. *International Review of Applied Linguistics in Language Teaching, 54* (2). pp. 73-78.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied linguistics*, *24*(1), 1-27.

Appendices

Appendix 1: Ethics Approval

23 July 2019

MACQUARIE University

Professor Lynda Yates
Department of Linguistics
Faculty of Human Sciences
Macquarie University NSW 2109

Reference: 5201700463

Dear Professor Yates

FINAL APPROVAL

Re: "The effects of noticing training, model input and task repetition on L2 speech production" (5201700463)

Thank you very much for your response. Your response has addressed the issues raised by the Faculty of Human Sciences Human Research Ethics Sub-Committee and approval has been granted, effective 16th June 2017. This email constitutes ethical approval only.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research

The following personnel are authorised to conduct this research:

Professor Lynda Yates
Professor Linda Cupples
Mr Matt Campbell

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 16th June 2018
Progress Report 2 Due: 16th June 2019
Progress Report 3 Due: 16th June 2020
Progress Report 4 Due: 16th June 2021
Final Report Due: 16th June 2022

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

1

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/current_research_staff/human_research_ethics/resources

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Sub-Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Sub-Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/current_research_staff/human_research_ethics/managing_approved_research_projects

5. Please notify the Sub-Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites:

http://www.mq.edu.au/policy
http://www.research.mq.edu.au/current_research_staff/human_research_ethics/managing_approved_research_projects

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of approval to an external organisation as evidence that you have approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of ethics approval.

Yours sincerely,

Dr Naomi Sweller
Chair
Faculty of Human Sciences
Human Research Ethics Sub-Committee

2

HUMAN RESEARCH ETHICS COMMITTEE

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

Notification of Expedited Approval

To Chief Investigator or Project Supervisor:	**Professor Lynda Yates**
Cc Co-investigators / Research Students:	**Professor Linda Cupples** **Mr Matthew Campbell**
Re Protocol:	**The effects of noticing training, model input and task repetition on L2 speech production.**
Date:	**05-Sep-2017**
Reference No:	**H-2017-0227**
Date of Initial Approval:	**05-Sep-2017**

Thank you for your **Response to Conditional Approval (minor amendments)** submission to the Human Research Ethics Committee (HREC) seeking approval in relation to the above protocol.

Your submission was considered under **Expedited** review by the Ethics Administrator.

I am pleased to advise that the decision on your submission is **Approved** effective **05-Sep-2017**.

In approving this protocol, the Human Research Ethics Committee (HREC) is of the opinion that the project complies with the provisions contained in the National Statement on Ethical Conduct in Human Research, 2007, and the requirements within this University relating to human research.

Approval will remain valid subject to the submission, and satisfactory assessment, of annual progress reports. *If the approval of an External HREC has been "noted" the approval period is as determined by that HREC.*

The full Committee will be asked to ratify this decision at its next scheduled meeting. A formal *Certificate of Approval* will be available upon request. Your approval number is **H-2017-0227**.

If the research requires the use of an Information Statement, ensure this number is inserted at the relevant point in the Complaints paragraph prior to distribution to potential participants You may then proceed with the research.

Conditions of Approval

This approval has been granted subject to you complying with the requirements for *Monitoring of Progress*, *Reporting of Adverse Events*, and *Variations to the Approved Protocol* as detailed below.

PLEASE NOTE:
In the case where the HREC has "noted" the approval of an External HREC, progress reports and reports of adverse events are to be submitted to the External HREC only. In the case of Variations to the approved protocol, or a Renewal of approval, you will apply to the External HREC for approval in the first instance and then Register that approval with the University's HREC.

- ***Monitoring of Progress***

Other than above, the University is obliged to monitor the progress of research projects involving human participants to ensure that they are conducted according to the protocol as approved by the HREC. A progress report is required on an annual basis. Continuation of your HREC approval for this project is conditional upon receipt, and satisfactory assessment, of annual progress reports. You will be advised when a report is due.

- ***Reporting of Adverse Events***

1. It is the responsibility of the person **first named on this Approval Advice** to report adverse events.
2. Adverse events, however minor, must be recorded by the investigator as observed by the investigator or as volunteered by a participant in the research. Full details are to be documented, whether or not the investigator, or his/her deputies, consider the event to be related to the research substance or procedure.
3. Serious or unforeseen adverse events that occur during the research or within six (6) months of completion of the research, must be reported by the person first named on the Approval Advice to the (HREC) by way of the Adverse Event Report form (via RIMS at https://rims.newcastle.edu.au/login.asp) within 72 hours of the occurrence of the event or the investigator receiving advice of the event.
4. Serious adverse events are defined as:
    - Causing death, life threatening or serious disability.
    - Causing or prolonging hospitalisation.
    - Overdoses, cancers, congenital abnormalities, tissue damage, whether or not they are judged to be caused by the investigational agent or procedure.
    - Causing psycho-social and/or financial harm. This covers everything from perceived invasion of privacy, breach of confidentiality, or the diminution of social reputation, to the creation of psychological fears and trauma.
    - Any other event which might affect the continued ethical acceptability of the project.

5. Reports of adverse events must include:
    - Participant's study identification number;
    - date of birth;
    - date of entry into the study;
    - treatment arm (if applicable);
    - date of event;
    - details of event;
    - the investigator's opinion as to whether the event is related to the research procedures; and
    - action taken in response to the event.

6. Adverse events which do not fall within the definition of serious or unexpected, including those reported from other sites involved in the research, are to be reported in detail at the time of the annual progress report to the HREC.

- ***Variations to approved protocol***

If you wish to change, or deviate from, the approved protocol, you will need to submit an *Application for Variation to Approved Human Research* (via RIMS at https://rims.newcastle.edu.au/login.asp). Variations may include, but are not limited to, changes or additions to investigators, study design, study population, number of participants, methods of recruitment, or participant information/consent documentation. **Variations must be approved by the (HREC) before they are implemented** except when Registering an approval of a variation from an external HREC which has been designated the lead HREC, in which case you may proceed as soon as you receive an acknowledgement of your Registration.

Linkage of ethics approval to a new Grant

HREC approvals cannot be assigned to a new grant or award (ie those that were not identified on the application for ethics approval) without confirmation of the approval from the Human Research Ethics Officer on behalf of the HREC.

Best wishes for a successful project.

Associate Professor Helen Warren-Forward
Chair, Human Research Ethics Committee

For communications and enquiries:
Human Research Ethics Administration

Research & Innovation Services
Research Integrity Unit
The University of Newcastle
Callaghan NSW 2308
T +61 2 492 17894
Human-Ethics@newcastle.edu.au

RIMS website - https://RIMS.newcastle.edu.au/login.asp

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: +61 (0) *2 9850 9646*

Email: lynda.yates@mq.edu.au

Chief Investigator's / Supervisor's Name & Title: Professor Lynda Yates.

Information for Participants

Name of Project: The effects of noticing training, model input and task repetition on L2 speech production.

I wish to invite you to participate in my research on improving English learners' speaking performance. The details of the study follow, and I hope you will consider being involved. I am conducting this research as part of my Doctor of Philosophy degree at Macquarie University in Sydney. My supervisors are Professor Lynda Yates and Professor Linda Cupples who both work in the Linguistics Department of the Faculty of Human Sciences at Macquarie University in Sydney. Professor Yates can be contacted by email at lynda.yates@mq.edu.au or by phone at +61 (0)2 9850 9646. Professor Cupples can be contacted by e-mail at linda.cupples@mq.edu.au or by phone at +61 2 9850 8788. I can be contacted at [redacted].

Aim of the Study:
The aim of this research is to investigate speaking performance in second language learners of English.

Time Requirements:
Information will be gathered during your regularly scheduled classes at the University of Newcastle during your current course. You will also be asked to participate on three occasions outside of regular class time. Each time will take about 15 minutes.

Speaking Activities:
As in all English classes at the University of Newcastle, you will participate in various speaking activities. Speaking activities are designed to improve your English-speaking ability. In class, as part of my research, I hope to voice-record you while you perform some of these activities. In addition, I plan to video/audio record you while you perform speaking activities on three different occasions. Each occasion will take about 15 minutes and will happen outside of class time. Your recordings will be transcribed, and a transcript will be provided to you if you wish to see one. Copies of your recording(s) can also be provided upon request. Any information or personal

details gathered in the course of the study will remain confidential and will only be accessed by the research team. You will not be identified by name in any publication of the results. Your name will be replaced by a pseudonym; this will ensure that you are not identifiable.

Participation is completely voluntary. Your decision to participate will in no way impact on your grades for this course and will not give you any additional advantage nor disadvantage. If you decide to participate, you are free to withdraw your consent from the project and discontinue at any time without having to give a reason and without consequence. You will be participating in the in-class speaking activities as part of routine classes, and you will be recorded as a normal part of the activity. However, only if you agree to participate in the study will you take part in the video recording sessions outside of class. You will not be assessed in any of the speaking activities that are recorded. Only if you agree to participate in the study will the recordings be transcribed for research purposes. If you choose not to participate, you will still be recorded during in-class speaking activities for teaching purposes, however, those recordings will not be used for research purposes. This will mean you will not be able to be identified as either a participant, or a non-participant.

It is unlikely that this research will raise any personal or upsetting issues, but if it does you may wish to contact the University of Newcastle Student Counselling Centre at 02 4921 5801.

Any information or personal details gathered in the course of the study are confidential, except as required by law. The voice recordings will be kept in a password protected file on Google Drive and/or on a password protected computer at Macquarie University. The transcriptions and other data will be kept in the same manner for five (5) years following thesis submission and then destroyed. Only the investigators will have access to the data. It is anticipated that a summary of the findings from this research will be available at the end of 2018. A copy of the summary will be e-mailed to you upon request.

Research Process:
This research will be completed within your current 10-week English course. The results may be presented at conferences or written up in journals without any identifying information.

This project has been approved by the Human Research Ethics Committees of Macquarie University (Approval No. ……., Valid to ../../….) and University of Newcastle (Approval No. H-2017-0227 Valid to ../../….).

Thank you for considering this request.

Regards,

Matthew Campbell

Consent Form

I, _____ have read *(or, where appropriate, have had read to me)* and understand the information above and any questions I have asked have been answered to my satisfaction. I agree to participate in this research, knowing that I can withdraw from further participation in the research at any time without consequence. I have been given a copy of this form to keep.

I wish to receive a copy of my audio/video recordings. (Please tick) Yes ☐ No ☐

I wish to receive a copy of the summary of the findings. (Please tick) Yes ☐ No ☐

Participant's Name:
(Block letters)

Participant's Signature: _____Date:_____

Investigator's Name:
(Block letters)

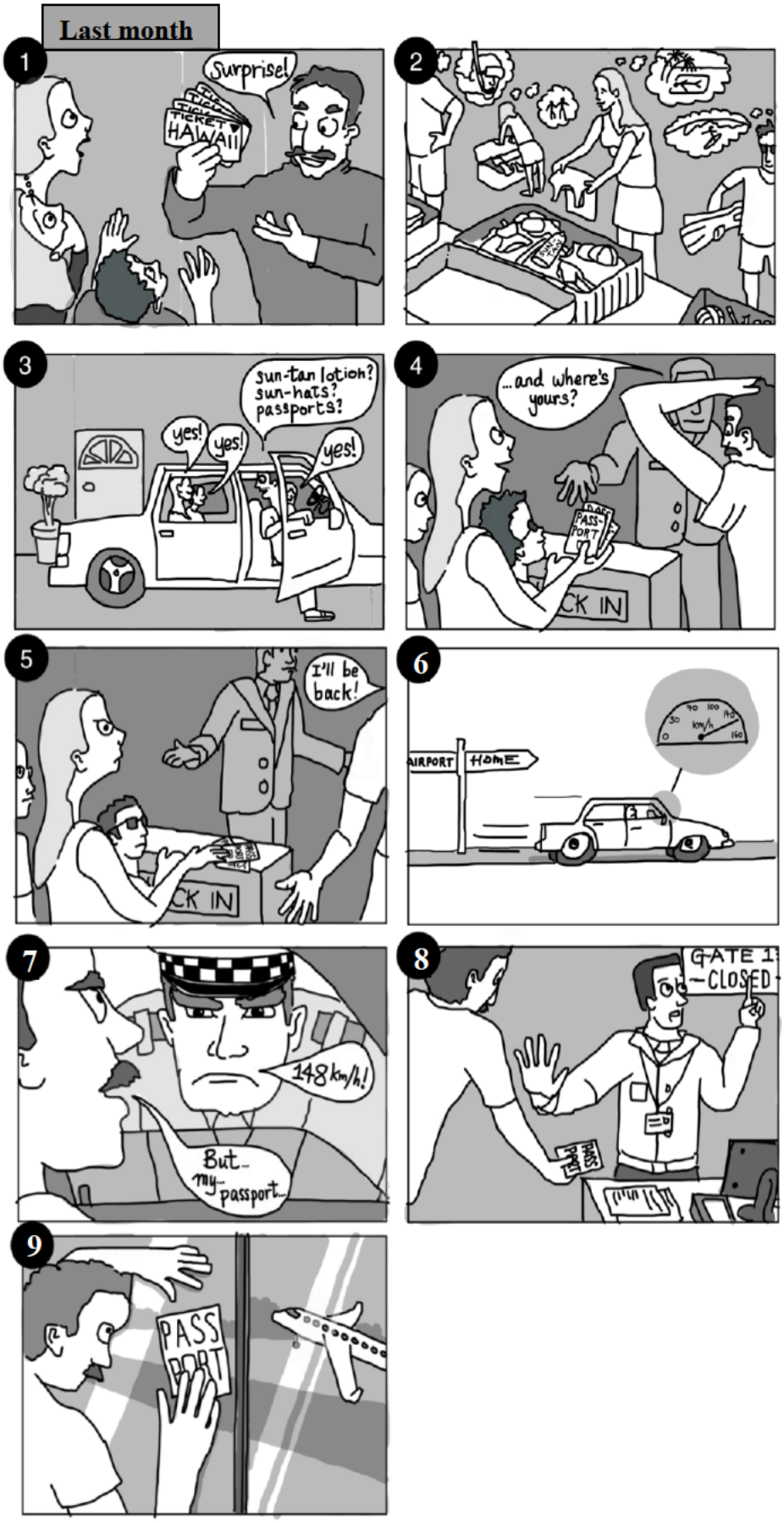
Investigator's Signature:_____ Date:_____

The ethical aspects of this study have been approved by the Macquarie University Human Research Ethics Committee (approval No. ..... and University of Newcastle Human Research Ethics Committee (approval No. H-2017-0227). If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Committee through the Director, Research Ethics & Integrity (telephone (02) 9850 7854; email ethics@mq.edu.au). Any complaint you make will be treated in confidence and investigated, and you will be informed of the outcome.

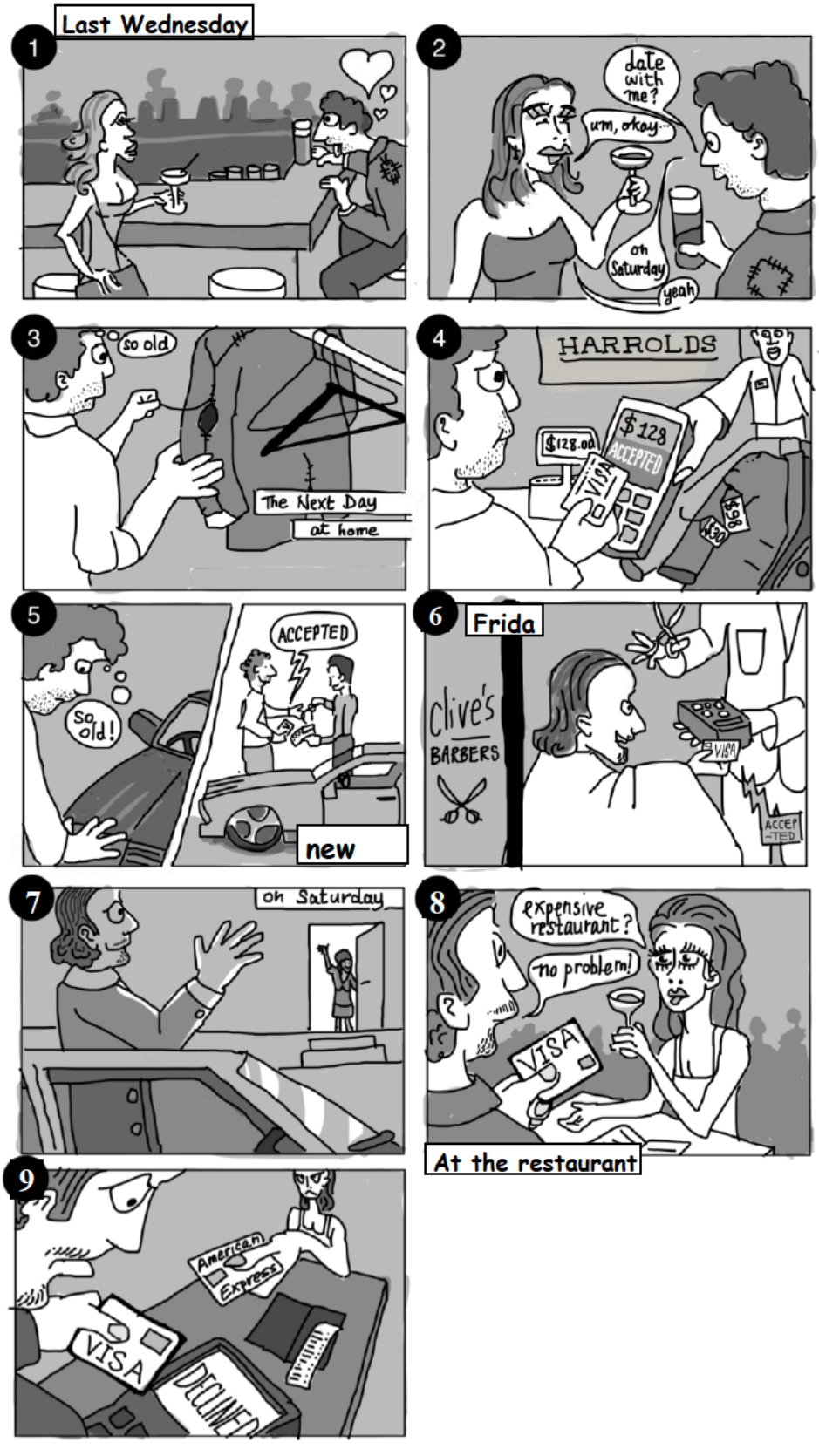
(INVESTIGATOR'S [OR PARTICIPANT'S] COPY)

Appendix 3: Time 1 picture prompt - 'Ruined Dinner'.

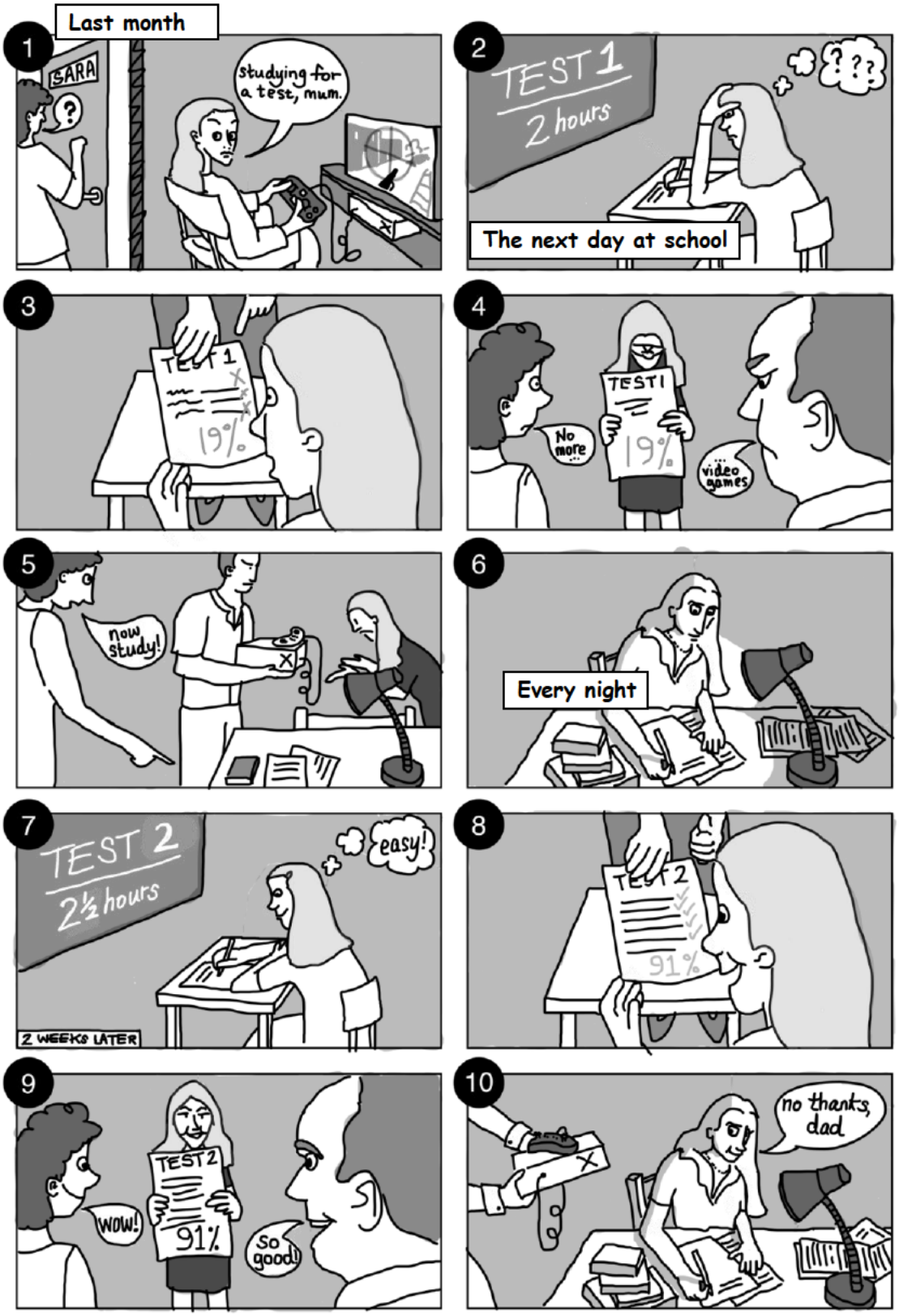
Appendix 4: Time 2 picture prompt - 'Hawaii Holiday'.



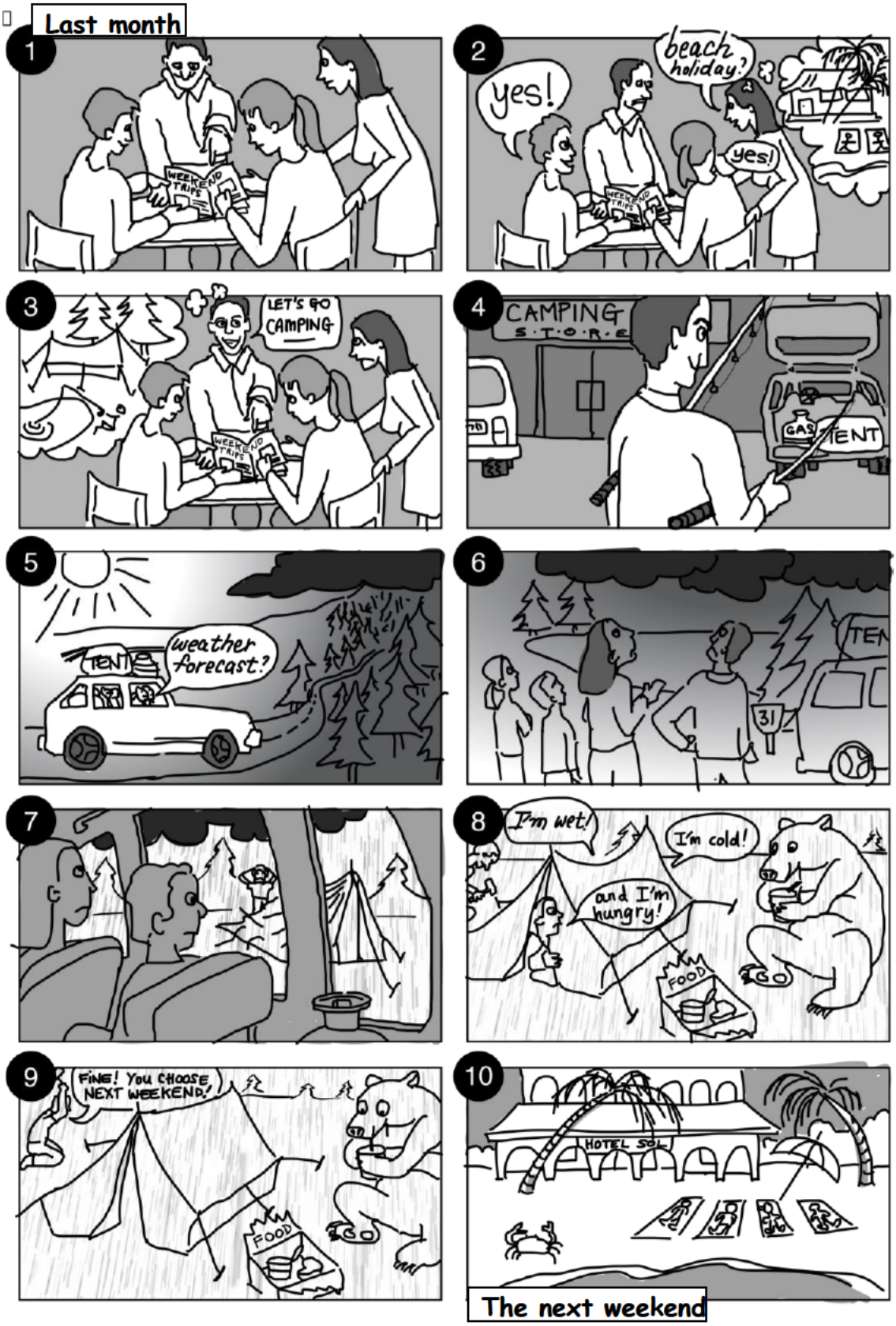
Appendix 5: Time 3 picture prompt - 'Expensive Dinner'.



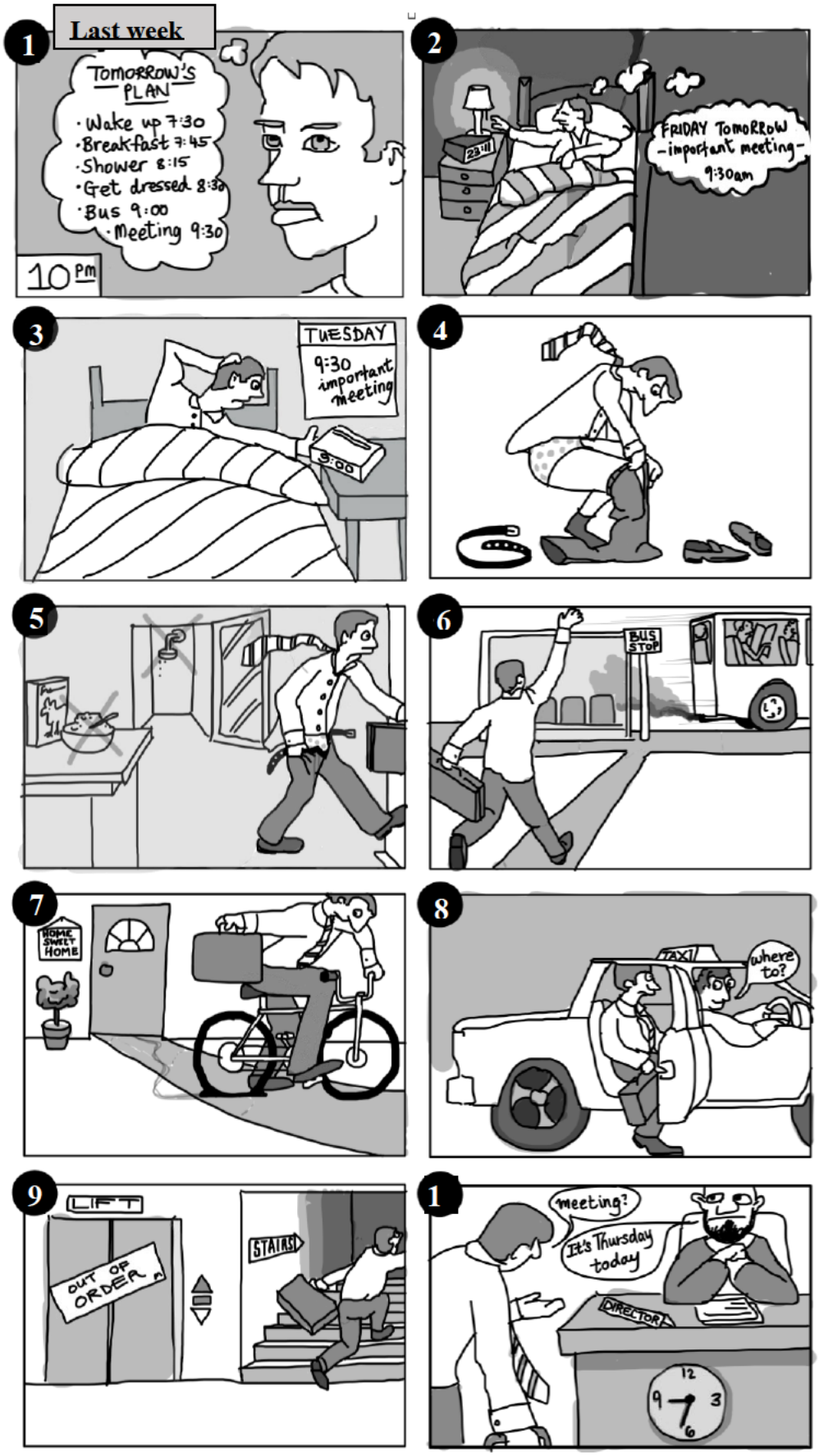
Appendix 6: Time 4 picture prompt - 'No More Video Games'.



Appendix 7: Time 5 picture prompt - 'Wet Weekend'.



Appendix 8: Time 6 picture prompt - 'Wrong Day For A Meeting'.



No more video games script

One month ago, Sara was in her bedroom playing video games. Her mother knocked on her door and asked, “What are you doing?” Sara replied, “I’m studying for a test”. The next day, she had a test at school. The test was for two hours, but Sara didn’t know any answers. Later, she received her test result, and she was shocked to see that she got a score of only nineteen percent. She went home and showed her parents her test score. Her parents were really angry, and they said, “No more video games”. Sara’s dad took her video game machine off her, and Sara’s mum told her to study. Sara stayed up late and studied hard every night for the next two weeks. Then she had a second test that was two and a half hours long. She was happy during the test because she thought it was easy. When she received her score, she was glad to see that she got ninety one percent. She went home and happily showed her parents her test score. They were both really pleased. So Sara’s dad offered to give her video game machine back, but Sara said, “No thanks, dad”. She decided studying was more useful than playing games.

Ruined dinner script

Last week Bob and Sally were relaxing at home with their dog, and Sally called her friend Debbie to invite her and her husband for dinner on Saturday at 6pm. Debbie said, “Yeah, great”. So, while Bob was playing with the dog, he asked what food they should cook for their friends, and Debbie thought roast chicken would be good. The next day Bob and Sally went to the supermarket to buy some food. On Saturday at about 2pm they began preparing the food in their kitchen. Bob cut some vegetables and Debbie put the chicken into the oven. Bob asked how long the chicken would take to cook. Debbie replied, “About three hours”. Later Bob was in the dining room setting the table, he kicked the dog’s ball into the kitchen and the dog chased after it. As the dog was chasing the ball, it stepped on the oven’s electrical cord and it disconnected from the wall, so the oven had no power. Just before 6pm, Debbie took the chicken out of the oven, she found that it wasn’t cooked at all, and Bob found that the electrical cord was not plugged in. Then, their friends, Debbie and Max arrived and rang the doorbell. Bob and Sally were worried, and Sally said, “We have no food, what are we going to do?”, and then Bob had a good idea. He called the pizza company and ordered four large pizzas. Twenty minutes later, the pizzas arrived and Bob paid the delivery man. Finally, they all sat down at the dinner table and ate the pizza. Everyone had a great evening.

Family trip to Hawaii script

Last month, Sam surprised his family with tickets for a family holiday in Hawaii. Sam's wife and children packed their suitcases for their trip and thought about what they would do in Hawaii. Sam's son thought about surfing, Sam's daughter thought about playing on the beach, Sam's wife thought about sunbathing, and Sam thought about snorkelling. After they finished packing, they got in their car and Sam asked his family if they had their sun-tan lotion, sun-hats and passports. Everyone replied, "Yes!". Then they arrived at the airport and went to the check-in. The man at the check-in counter asked for their passports. Everyone had their passport, but Sam suddenly realised that he forgot his. Sam's wife was angry, Sam said, "I'll be back" and he drove from the airport back to his house to get his passport. But he drove too fast and he was stopped by a police officer. The officer said, "You were driving at 148 kilometres per hour, and Sam said "But I had to get home quickly to get my passport because I'm flying to Hawaii with my family". Finally Sam arrived back at the airport with his passport, but the gate had closed, so Sam missed the flight. Sam watched through the window as the plane took off with his family inside.

Wrong day for a meeting script

Last week Tim was at home thinking about his plan for the next day. He planned to wake up at seven thirty, eat breakfast at seven forty-five, take a shower at eight fifteen, get dressed at eight thirty, catch the bus at nine o'clock, and go to a meeting at nine thirty. Later that night, Tim went to bed and thought about his important meeting the next day. In the morning Tim was shocked when he woke up. He overslept. It was two minutes past nine. He quickly got dressed, but he didn't have time to take a shower or eat breakfast. He ran to the bus stop, but he was too late, so he went back to his house and tried to ride his bicycle to work, but the bike had a flat tire, so Tim took a taxi. The driver asked "where would you like to go?" Tim replied, "To work, quickly please!" When Tim arrived at work, the elevator was out of order, so he had to run up the stairs. When he finally arrived at the office, it was just after nine thirty. He asked his boss "What about the meeting?" his boss said, "Today is Tuesday, the meeting is tomorrow". Tim had the days confused. He thought it was Wednesday!

Expensive date script

Last week James was drinking at a bar. He saw a beautiful woman and he said, “Would you like to go on a date with me?” the woman said, “um, sure”, so James said, “How about on Saturday?”, and the woman agreed. The next day, James was at home looking for clothes to wear for his date. But his clothes were really old. So he went shopping and bought a new jacket a shirt. The clothes cost one hundred and twenty eight dollars and he paid with his credit card which was accepted. Next, James looked at his car and thought it was really old, so he went and bought a new car. Again, he paid with his credit card which was accepted. After that, James went to the barber to get a haircut for his date. The haircut cost fifty dollars which James paid for with his credit card. And then, on Saturday, James looked good with his new haircut, he wore his new clothes and arrived at his dates house in his new car. James and his date were enjoying dinner at the restaurant, and his date asked, “Isn’t this a really expensive restaurant?” James said “No problem, I have my credit card”. James tried to pay for the dinner but his credit card was declined! James was really embarrassed and his date was angry because she had to pay for the dinner.

Soggy weekend trip script

Last Thursday, Jack showed his family a magazine with ideas for weekend trips. Jack's wife, Kim, suggested a beach holiday and the children said "Yes!", but Jack said, "No, let's go camping". Jack imagined sleeping in a tent in the forest and fishing. The next day, Jack went to the camping store and bought some camping equipment, including fishing rods, a tent and some gas for cooking. Early on Saturday morning Jack and his family drove to the campsite but there were dark clouds in the sky. Kim asked Jack if he checked the weather forecast. The family arrived at the campsite and they looked worried because there were even more dark clouds in the sky. Then, it started raining really heavily. Kim and the children stayed inside the car while Jack tried to put up the tent, but he was having trouble. Finally he put the tent up and his family was inside. The children said, "I'm cold" and "I'm wet", and Kim said "I'm hungry", but outside a bear was eating all their food. Jack said, "Fine, next weekend you can choose where you want to go for a trip". So the next weekend they went to the beach and stayed in a nice hotel and had a really great time.

Appendix 10: Unguided noticing prompt (prompt 1)

Narration comparison.

Listen to the recording of your story and a model speaker's story. While you listen, make notes of any language you think might be useful.

Appendix 11: Guided noticing prompt (prompt 2)

Part A.

In your first language, write any the words and phrases you wanted to use but didn't know in English.

Words and phrases I wanted to say:

1. (Example) 写真
2. (Example) 彼は駆け下りた
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____

Part B.

Listen to the model recording. In English, write down any words and phrases you picked up from listening to the model speaker.

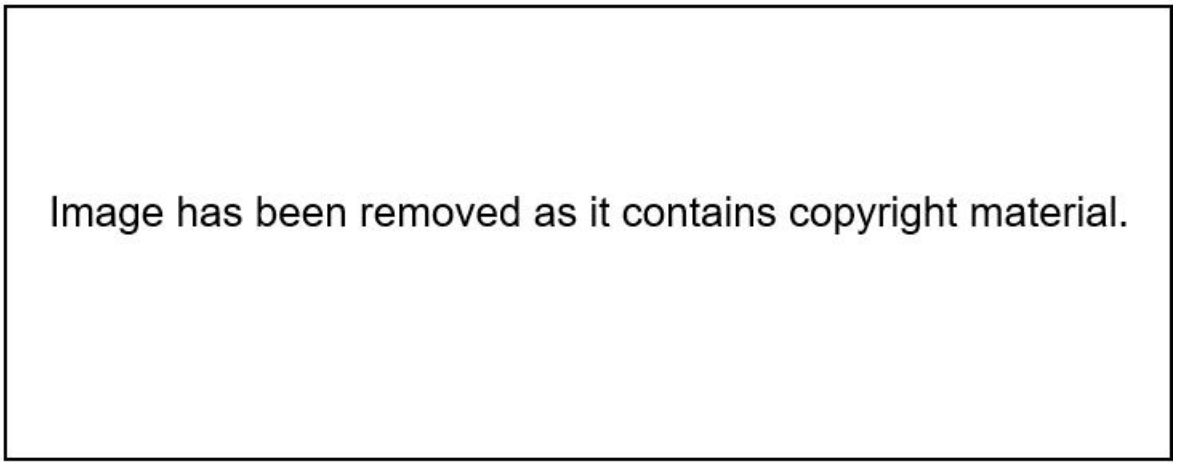
Words I picked up from the model:

1. (Example) photograph
2. (Example) He ran downstairs
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____

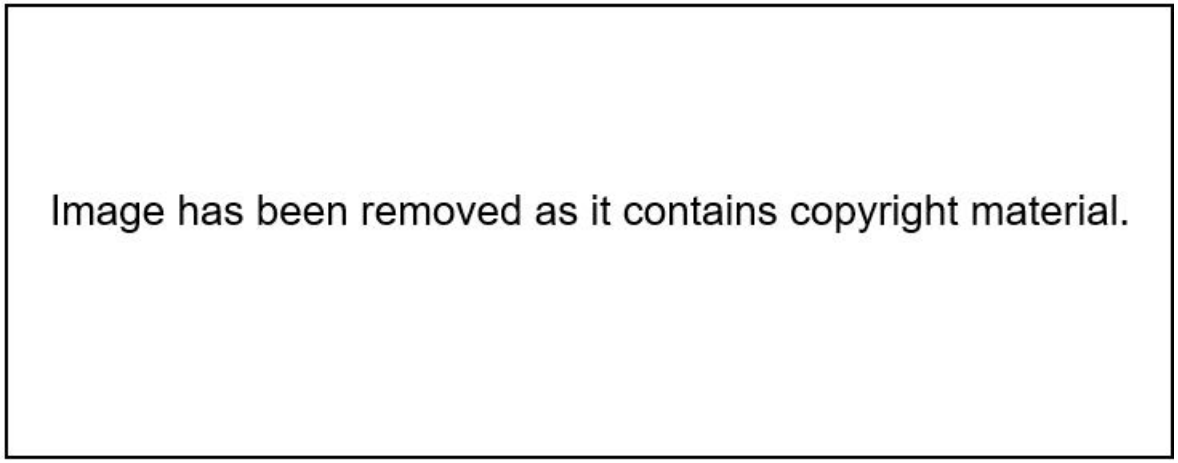
Part C.

Note other differences in grammar between your language and the language used in the model

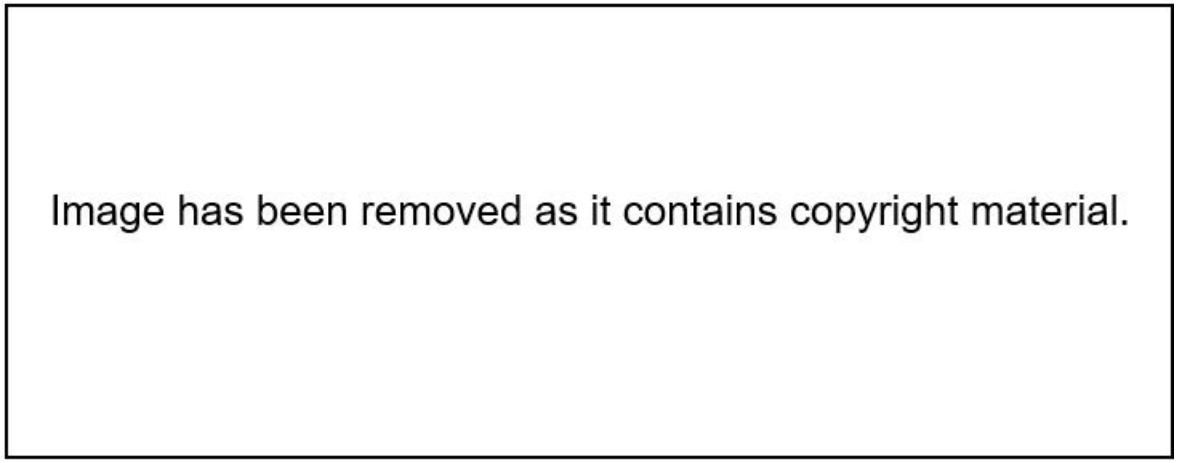
Appendix 12: Picture prompt from practice session 1 **(taken from Sheppard, 2006)**



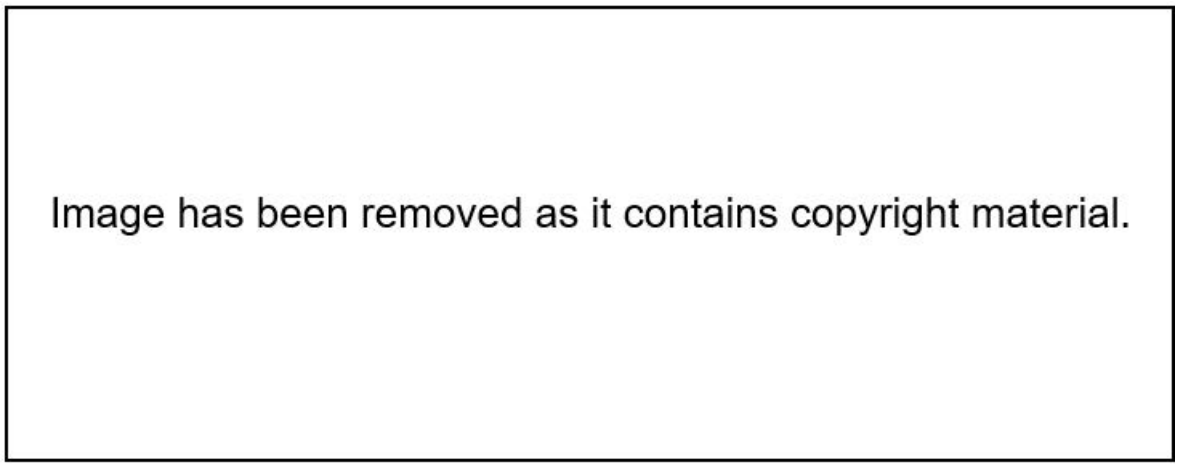
Appendix 13: Picture prompt A from practice session 1 (taken from Heaton, 1975)



Appendix 14: Picture prompt B from practice session 1 (taken from Heaton, 1975)



Appendix 15: Picture prompt from practice session 2 (taken from Heaton, 1975)



Appendix 16: Script read to participants for procedure of testing sessions

---

Hi _____, thanks for coming and helping me with my research. Today, I'd like to record you while you do a speaking task, OK?

There are four steps to go through, and it should take around 20 minutes in total. I'll explain each step now, and if you have any questions, please ask at any time.

In step 1, I'm going to ask you to tell a story based on a picture sequence like we practised last week. After I give you the pictures, you'll have 1 minute to look at them to gain an understanding of the story, and then I'll ask you to begin your story. I'll audio record your story using this digital recorder, and I'll video record you telling the story using this computer.

After you have finished your story, in step 2 we'll watch the video of you telling your story, and as we watch we'll talk about it. I'll also audio record our discussion of your story.

In step 3, I'll give you a chance to listen to the audio recording of your story again and compare it to an audio recording of a model speaker. While you listen, you can take notes of any useful language. You'll have 7.5 minutes for this step.

Lastly, in step 4, I'll ask you to repeat your story again using the same picture sequence. I'll audio record your story again, but I won't video record it. You won't be able to use your notes from the previous step while you tell your story again.

Do you have any questions?

Appendix 17: Note taking paper for the comparison stage of testing sessions

Narration comparison.

Listen to the recording of your story and a model speaker's story. While you listen make notes of any language you think might be useful.

Appendix 18: ANOVA results for measures of noticing IL gaps in testing sessions (times 1, 5 & 6)

Variable	Source	F	p	η_{p2}
Noticing of grammar (NoG)	Time	36.913	.000	.528
	Group	75.331	.000	.820
	Time*Group	41.739	.000	.717
Noticing of content (NoC)	Time	13.319	.000	.289
	Group	1.152	.328	.065
	Time*Group	.522	.720	.031

Appendix 19: ANOVA results for percentage of solvable IL gaps filled in testing sessions (times 1, 5 & 6)

Variable	Source	F	p	η_{p2}
Percentage of IL gaps filled	Time	0.915	.346	.027
	Group	4.545	.018	.216
	Time*Group	1.662	.205	.091

Appendix 20: ANOVA results for measures of accuracy in testing sessions (times 1, 5 and 6)

Variable	Source	F	p	η_{p2}
Errors per 100 words	Time	1.832	.168	.053
	Group	4.572	.018	.217
	Time*Group	4.198	.004	.203
	Delivery	76.319	.000	.698
	Delivery*Group	10.703	.000	.393
	Time*Delivery	10.413	.000	.240
	Time*Delivery*Group	3.844	.007	.189
Number of self-repairs per 100 words	Time	1.358	.264	.040
	Group	5.619	.008	.254
	Time*Group	3.409	.014	.171
	Delivery	.371	.547	.011
	Delivery*Group	.382	.685	.023
	Time*Delivery	.228	.797	.007
	Time*Delivery*Group	.077	.989	.005
Percentage of errors self-repaired	Time	3.838	.026	.104
	Group	23.276	.000	.585
	Time*Group	8.373	.000	.337
	Delivery	11.091	.002	.252
	Delivery*Group	5.903	.006	.263
	Time*Delivery	1.307	.278	.038
	Time*Delivery*Group	1.115	.357	.063

Appendix 21: ANOVA results for measures of fluency in testing sessions (times 1, 5 and 6)

Variable	Source	F	p	η_{p2}
Speech rate (WPM)	Time	7.825	.001	.192
	Group	.459	.636	.027
	Time*Group	1.103	.363	.063
	Delivery	2.710	.109	.076
	Delivery*Group	.071	.931	.004
	Time*Delivery	5.202	.008	.136
	Time*Delivery*Group	.629	.643	.037
Number of filled pauses per 100 words	Time	22.931	.000	.410
	Group	.372	.692	.022
	Time*Group	.198	.938	.012
	Delivery	9.733	.004	.228
	Delivery*Group	.456	.638	.027
	Time*Delivery	1.927	.154	.055
	Time*Delivery*Group	.465	.761	.027
Number of silent pauses per 100 words	Time	7.588	.001	.187
	Group	1.744	.191	.096
	Time*Group	1.393	.246	.078
	Delivery	.932	.341	.027
	Delivery*Group	.321	.727	.019
	Time*Delivery	5.713	.005	.148
	Time*Delivery*Group	3.109	.021	.159
Mean length of silent pause	Time	.788	.459	.023
	Group	4.580	.018	.217
	Time*Group	.518	.723	.030
	Delivery	.085	.773	.003
	Delivery*Group	.007	.993	.000
	Time*Delivery	1.926	.154	.055
	Time*Delivery*Group	1.195	.321	.068

Number of reformulations per 100 words	Time	1.291	.282	.038
	Group	.270	.765	.016
	Time*Group	1.276	.288	.072
	Delivery	5.561	.024	.144
	Delivery*Group	1.539	.230	.085
	Time*Delivery	1.013	.369	.030
	Time*Delivery*Group	.673	.613	.039
Number of repetitions per 100 words	Time	3.631	.032	.099
	Group	.959	.394	.055
	Time*Group	2.909	.028	.150
	Delivery	8.862	.005	.212
	Delivery*Group	2.216	.125	.118
	Time*Delivery	.116	.891	.004
	Time*Delivery*Group	1.564	.194	.087

Appendix 22: ANOVA results for the measure of complexity in testing sessions (times 1, 5 & 6)

Variable	Source	F	p	η_{p2}
Number of clauses per AS-unit	Time	5.870	.005	.151
	Group	.083	.921	.005
	Time*Group	.783	.540	.045
	Delivery	1.245	.272	.036
	Delivery*Group	.112	.894	.007
	Time*Delivery	4.601	.013	.122
	Time*Delivery*Group	.647	.631	.038

Appendix 23: ANOVA results for measures of accuracy in training sessions (times 2, 3 and 4)

Variable	Source	F	p	η_{p2}
Errors per 100 words	Time	6.020	.004	.154
	Group	1.793	.182	.098
	Time*Group	.545	.703	.032
	Delivery	105.927	.000	.726
	Delivery*Group	36.120	.000	.686
	Time*Delivery	1.229	.229	.036
	Time*Delivery*Group	.426	.789	.025
Number of self-repairs per 100 words	Time	2.119	.128	.060
	Group	5.666	.008	.256
	Time*Group	.746	.564	.043
	Delivery	.406	.529	.012
	Delivery*Group	13.628	.000	.452
	Time*Delivery	.311	.734	.009
	Time*Delivery*Group	1.006	.411	.057
Percentage of errors self-repaired	Time	.821	.444	.024
	Group	25.698	.000	.609
	Time*Group	.468	.759	.028
	Delivery	14.470	.001	.305
	Delivery*Group	34.002	.000	.673
	Time*Delivery	.402	.671	.012
	Time*Delivery*Group	.214	.930	.013

Appendix 24: ANOVA results for measures of fluency in training sessions (times 2, 3 and 4)

Variable	Source	F	p	η_{p2}
Speech rate (WPM)	Time	.131	.878	.004
	Group	1.738	.192	.095
	Time*Group	.949	.441	.054
	Delivery	55.479	.000	.627
	Delivery*Group	75.268	.000	.820
	Time*Delivery	.254	.776	.008
	Time*Delivery*Group	.939	.447	.054
Number of filled pauses per 100 words	Time	.489	.615	.015
	Group	1.056	.359	.060
	Time*Group	.472	.756	.028
	Delivery	18.058	.000	.354
	Delivery*Group	17.231	.000	.511
	Time*Delivery	2.823	.067	.079
	Time*Delivery*Group	.672	.614	.039
Number of silent pauses per 100 words	Time	1.355	.265	.039
	Group	.326	.326	.066
	Time*Group	.291	.883	.017
	Delivery	3.277	.079	.090
	Delivery*Group	18.448	.000	.528
	Time*Delivery	2.587	.083	.073
	Time*Delivery*Group	.500	.736	.029
Mean length of silent pause	Time	.306	.738	.009
	Group	1.753	.189	.096
	Time*Group	.125	.973	.008
	Delivery	7.489	.010	.185
	Delivery*Group	6.214	.005	.274
	Time*Delivery	.989	.377	.029
	Time*Delivery*Group	.152	.962	.009

Number of reformulations per 100 words	Time	.663	.519	.020
	Group	1.365	.269	.076
	Time*Group	.532	.712	.031
	Delivery	53.900	.000	.620
	Delivery*Group	10.377	.000	.386
	Time*Delivery	.538	.587	.016
	Time*Delivery*Group	.165	.956	.010
Number of repetitions per 100 words	Time	3.901	.025	.106
	Group	2.266	.120	.121
	Time*Group	.345	.847	.020
	Delivery	17.854	.000	.351
	Delivery*Group	12.405	.000	.429
	Time*Delivery	1.583	.213	.046
	Time*Delivery*Group	1.656	.171	.091

Appendix 25: ANOVA results for the measure of complexity in training sessions (times 2, 3 & 4)

Variable	Source	F	p	η_{p2}
Number of clauses per AS-unit	Time	1.802	.173	.052
	Group	.408	.668	.024
	Time*Group	.576	.681	.034
	Delivery	2.069	.160	.059
	Delivery*Group	.889	.421	.051
	Time*Delivery	2.281	.110	.065
	Time*Delivery*Group	.399	.809	.024