Developing and validating in-house, standardised, English reading and listening placement/streaming and achievement tests aligned to the Common European Framework of Reference Levels A2–B1 at a Japanese university

By

Jack Victor Bower

B.A. Communication, University of Canberra

M.A. TESOL, University of Canberra

Department of Linguistics

Submitted on May 10, 2018

Research carried out at Hiroshima Bunkyo Women’s University in Hiroshima, Japan

Table of Contents

CHAPTERS

SUMMARY

The research project presented in this thesis investigates one aspect of redesigning a compulsory English as a foreign language curriculum at a Japanese university to match intended target CEFR proficiency levels. Specifically, this thesis focuses on the creation and validation of institutional standardised tests of English reading and listening proficiency intended to be aligned to levels A2 and B1 of the CEFR.

Kane's (2013a, 2013b) argument-based approach to test validation, which firstly elaborates an Interpretation/Use Argument (IUA), and secondly uses a validity argument to evaluate evidence gathered to support the IUA, is used as a framework for the validation of the tests in question, known as the Bunkyo English Tests (BETs). Inferences in the IUA are also drawn from Chapelle, Enright, and Jamieson (2008), and the warrant for the final inference is drawn from Bachman and Palmer (2010).

To substantiate the validity argument, data were collected from a variety of sources, including test results and test specifications, course outlines, surveys administered to teachers and students, interviews with teachers and university senior administrators, student course grades and assessment results, and student results from two other standardised tests, the Oxford Online Placement Test® and the Test of English for International Communication (TOEIC®).

Results indicate that the BETs seem to have functioned sufficiently as course placement tests, but not as class streaming tests to divide classes within courses. Further analysis shows that the tests did not function effectively as achievement tests aligned to CEFR levels A2 and B1. This study demonstrates that Kane's approach to test validation is viable for small-scale, in-house testing programs in the development phase, as it facilitated the selection of validity evidence, the analysis of which exposed areas of weakness in the tests and the test specifications, indicating clear avenues for improvement. Results also point to the need for further validation research on in-house tests which aim for CEFR alignment.

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.

The research presented in this thesis was approved by the Macquarie University Ethics Review Committee, reference number 5201401091 on December 17, 2014.

(Signed) Jock Bauer Date: May 10, 2018

ACKNOWLEDGEMENTS

Firstly, I would like to thank my wife, Yoko, for walking with me through the hard times.

Secondly, I wish to express enormous gratitude to my Primary Supervisor, Professor Mehdi Riazi, and to my Adjunct Supervisor, Associate Professor Rie Koizumi, for their guidance, feedback and reassurance.

I would also like to thank the three examiners of this thesis, Antony John Kunnan, James Dean Brown and Lynette May, whose feedback made the arguments herein more coherent and transparent.

My further thanks go to the dedicated and passionate teachers and learning advisors at the BECC, who worked so hard to create the new General English curriculum and the BETs, and who also gave their time to participate in this research.

Finally, thank you to the Hiroshima Bunkyo Women's University students and senior managers who participated in this study, and to the BECC staff who made administration of the BETs possible.

Research intended for inclusion in this thesis was supported by a TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment.

CHAPTER ONE

Introduction to the Study

1.0 Introduction

This chapter outlines the central elements of this study. Firstly, the main focus of the investigation is summarized. Secondly, an argument for the importance of the study is presented, and the study's research questions are listed. Thirdly, a detailed explanation of the context of the study and the history of the language tests which are the subject of this research is presented. This is followed by a description of the position of the researcher and my approach to research reflexivity in this study. The chapter ends with an outline of the chapters of the thesis.

1.1 The research problem

From 2014–2016 a general English as a foreign language curriculum at a small private Japanese women's university was redesigned with the aim of bringing it into alignment with levels A2 and B1 of the CEFR (see sections 1.4.4.3 and 1.4.4.4 for the reasons behind this CEFR-aligned renewal). A crucial component of this alignment project was designing and validating standardised language tests with the dual main purposes of placing students into two courses targeted at separate CEFR levels, and of measuring student English language proficiency against the curriculum's target CEFR levels. To achieve these purposes, a suite of three standardised tests of English targeted at three skills, and labelled as the Bunkyo English Reading Tests (BERTs), the Bunkyo English Listening Tests (BELTs), and the Bunkyo English Speaking Tests (BESTs) were designed. Of these three tests the BERTs and the BELTs are the focus of the research presented in this thesis. For the purpose of conciseness, when the BETs are referred to in this study it is in reference to the BERTs and the BELTs, not to the BESTs.

The research problem addressed and investigated by this thesis is the extent to which the BETs achieved their intended functions within the GE curriculum, that is the validity of the BETs. The process of evaluating the evidence-based interpretations or inferences made from test scores is known as validation. This study utilizes an argument-based approach to test validation to clarify what evidence needs to be gathered to assess if the BETs are valid and achieving their stated purposes in terms of test score interpretations. The argument-based approach used also guided the gathering of such evidence, and a critical evaluation the evidence to see if the BETs stated goals were achieved.

A significant challenge of this study is the validation of multiple test score interpretations within a single study, as different test score interpretations require different types of validity evidence. The BETs in the frame of this study were intended to be used both as streaming/placement tests, which require separating students into different groups, a norm-referenced decision, and as achievement tests, which require assessing if students meet a minimum defined standard, an absolute decision (Brown 2005, Brown and Hudson, 2002). In addition, the BETs examinded in the frame of this study were intended to have positive impact on English language education in the BECC. Validating these separate intended BET uses thus requires gathering and analysing evidence within both the norm-referenced and criterion-referenced paradigms, and also adopting a mixed-methods research approach.

1.2 The significance of this study

Although it has been the subject of some criticism (see Jin, Wu, Alderson & Song, 2017 for a useful overview of criticisms of the CEFR), the Common European Framework of Reference for languages (CEFR) commands a large and growing influence in the field of foreign language education (Figueras, 2012). As part of its increasing global uptake the CEFR is also gaining popularity in Asia, notably in China (Alderson, 2017), Vietnam (Huy & Hamid, 2015) and Japan. Evidence of the impact of the CEFR in Japan is the creation of a localized version of the CEFR called the CEFR-J (Negishi, Takada, & Tono, 2012; Tono & Negishi, 2012). Research has also begun to appear on the development and implementation of local, small-scale, in-house testing programs in Asia, which aim to either place students into courses with CEFR-based achievement goals, and/or to assess achievement of those goals (Y. Lee, 2011; Liu & Jia, 2017).

In spite of the CEFR's widespread use across language and test situations there remains an insufficiency of research in English into the creation and validation of localized tests based on the CEFR, such as localized tests in Japan. This thesis contributes significantly towards addressing this gap in the research literature. In addition, this thesis provides major benefits for the General English program at the Bunkyo English Communication Center at Hiroshima Bunkyo Women's University, by identifying rebuttal evidence found in BET validity argument, which needs to be addressed through revisions to the BETs, the BET specifications and some updates to the GE curriculum. These revisions will strengthen the evidence-based approach for important inferences which are made from BET scores.

This research provides evidence of the efficacy of the BETs for their placement and streaming functions, and also of the inefficacy of the BETs for their achievement function during the first two years of test development. Suggestions are made based on this evidence in section 7.3 for improvements to the BETs, which may lead to BETs better fulfilling their placement and streaming functions, and eventually fulfilling their achievement function. These suggestions may be beneficial for several stakeholder groups in the GE curriculum including teachers, students, and administrators. More specifically, these suggestions are likely to lead to improvements in the structure of individual test tasks, more accurate test specifications, better representation of the curriculum in BET tasks, and improved reliability and validity of the BETs. Another significant aspect of this study is the recommendations made for clearer communication about the BETs to various curriculum stakeholders (sections 7.3.7, 7.3.9 and 7.3.10).

In addition, the lessons learned from the first two years of BET development can serve as a very useful reference for other institutions in Japan and elsewhere, which may choose to adopt the CEFR as a framework for their foreign language curricula. This study exemplifies the challenges involved in aligning small-scale, institutional reading and listening assessments to the CEFR. Whilst several large scale commercial tests now claim alignment to the CEFR, for example the TOEIC and TOEFL® (Tannenbaum & Wylie, 2008), there are few examples of smaller institutions attempting to design in-house tests and claiming CEFR alignment of those tests. This study provides a useful example, along with some guidelines and lessons learned for other small institutions who may attempt similar projects. Given the increasing prevalence of the CEFR globally it is probable that many other institutions are attempting or will attempt similar projects.

Lastly, this study makes a valuable contribution to research in the field of language test validation in three areas. It provides a useful example of an effort to validate a dual purpose, in-house placement/streaming and in-house achievement test, and a rare example of how an argument-based approach to test validation can be applied to the validation of an in-house language test. Such research is still limited, particularly for language tests. Mattern and Packman (2009, p. 1) state that "research on the topic [of placement tests] is largely limited and has produced mixed results, which makes it difficult to make definitive statements regarding the utility of placement tests and subsequent course placement." In addition, this study provides a useful reference for validation of an in-house achievement test, as this remains an under-researched area in the field of language test validation. This thesis utilizes the most recent version of Kane's (2013a, 2013b) argument-based test validation framework, incorporating further inferences suggested and used by Chapelle, Enright and Jamieson (2008), and drawing on some warrants suggested in Bachman and Palmer’s Assessment Use Argument (AUA) (2010). Kane’s Assessment Use Argument consists of two main arguments: an Interpretation/Use Argument (IUA) and a validity argument. An IUA clearly details the inferences necessary for sound interpretations and uses of scores from a test, as well as the assumptions underlying each inference. The IUA also describes the kinds of evidence which are needed to back up each of the interpretations and assumptions. The second major argument in Kane’s framework, the validity argument, critically analyses evidence gathered to support the IUA.

This study represents a rare example in the literature of a full and detailed IUA and supporting validity argument for in-house reading and listening tests, which covers all of Chapelle et al.’s (2008) inferences. It also provides both comprehensive backing and rebuttal evidence for each inference, which remains uncommon in argument-based validation studies. The full and detailed examination of evidence in the BET validity argument presented here, allows readers to understand the BET validity argument in more depth than for previous comparable, argument-based validation studies, and it also provides a clear roadmap for insiders to revise the tests comprehensively, based on weakness found in the validity argument.

1.3 Research questions

1. *To what extent did the BETs within the frame of this study fulfil their functions as course placement and streaming tests in terms of inferences in the validity framework?*
2. *To what extent did the BETs within the frame of this study fulfil their function as achievement tests in terms of inferences in the validity framework?*
3. *Based on any weaknesses found in backing for inferences in the BET validity argument across two academic years from 2015–17, what recommendations can be*

made for changes to the BETs and their accompanying documents, procedures and policies?

1.4 Context of the study

1.4.1 Hiroshima Bunkyo Women’s University

Hiroshima Bunkyo Women’s University (HBWU) is a small, private women’s university in Japan. The university was chartered as a junior women's college in 1962, and became a four-year university in 1966. Like many universities in Japan HBWU faces challenges presented by a declining student population, a symptom of an overall demographic decline in Japan (Matsutani, 2012). This means increased competition amongst universities for a shrinking pool of students. It also means an overall lower standard of academic proficiency of university entrants to mid to low-level universities such as HBWU, as institutions are forced to lower their entry requirements in order to maintain student numbers. In this context, in 2007 HBWU was looking for ways to boost student enrolments after several years of decline by enhancing its English language programme.

1.4.2 The Bunkyo English Communication Center

In 2008, in spite of the declining tertiary student population in Japan, Kanda University of International Studies (KUIS) in Chiba, Japan had been able to increase its enrolments for several years and to be consistently oversubscribed.

In an effort to gain a competitive advantage in the shrinking tertiary education sector HBWU engaged the services of the Kanda External Language Consultancy Center (ELCC) to create an English language learning centre, to provide staff for the centre, and to create a general English language curriculum for all of HBWU’s first and second year students.

The Bunkyo English Communication Center (BECC) was founded at (HBWU) in 2008 with a mission of providing a communicative English language curriculum for all first and second year students at the university, along with a Self-Access Learning Center (SALC) to support student autonomous English study. The creators of the BECC were a team of educators from Kanda University of International Studies in Chiba working for the Kanda ELCC.

The original BECC academic staff consisted of four English teachers and one learning advisor. Seven years later when this study began in 2015 the BECC staff had grown to include ten English teachers and three learning advisors. The BECC is responsible for overseeing and renewing a General English curriculum for non-English-language majors from four of the university's departments of Early Childhood Education, Welfare, Nutrition and Psychology. The BECC also manages the curriculum for a new Global Communication Department of English language majors, established in 2010, and continues to improve the learning materials and learner support available in the SALC.

1.4.3 The General English Curriculum

The GE Curriculum is two years long and is broken into two courses: Freshman English (FE) for first years and Sophomore English (SE) for second years.

The original conception of the BECC General English Curriculum was a communicative curriculum, incorporating the principles of task-based learning. The curriculum emphasised theme-based project learning aimed at keeping students' interest by combining themes of interest with language learning materials that emphasised integrated learning skills (Thompson & Foale, 2008).

1.4.4 Evolution of the BECC General English Curriculum

From its inception in 2008 the evolution of the BECC GE curriculum can usefully be divided into four phases. The first phase was the creation of the curriculum in 2007 and 2008. The second phase was three, yearly rounds of renewal of lessons and assessment criteria based on teacher feedback from 2009–2011. The third phase was a renewal attempt based on the Common European Framework of Reference–Japan or CEFR-J from 2012–13, and the fourth phase which encompasses the period of this study, was a curriculum renewal undertaken from

2014–16, which aimed to remake the GE curriculum by creating two streams of courses at CEFR A1–A2, and A2–B1 levels based on two of the foundational documents for the CEFR: *Waystage 1990* (van Ek & Trim, 1998b) and *Threshold 1990* (van Ek & Trim, 1998a).

1.4.4.1 Phase 1 – Initial setup

Initially the GE curriculum and the accompanying lesson materials for first-year students had to be created within a window of a few months in late 2007. Subsequently, materials and curriculum for the second-year course were created in 2008 while the first-year curriculum was being taught.

1.4.4.2 Phase 2 – Piecemeal improvements

Problems were identified with the level of difficulty of much of the materials being pitched too high for the students, and a lack of appropriate scaffolding leading up to lesson final tasks. From 2009–2011 the GE curriculum went through three one-year cycles of curriculum renewal. For each of these cycles, the major focus was on improving the content of individual lessons to make them more engaging and to include more scaffolding to make the lessons suitable for lower proficiency students. Can do statements were also added as lesson goals for the second-year curriculum. In addition, there was an effort made to broaden the types of assessment included in students' final grades, and to standardize some assessments for the curriculum.

1.4.4.3 Phase 3 – Adoption of the CEFR-J

In 2012, a decision was made by BECC management to adopt the CEFR-J as the framework for further GE curriculum renewal. The CEFR-J is an ongoing project in Japan to create a localized version of the CEFR for the Japanese educational context. In 2012 the CEFR-

J primarily consisted of a self-assessment grid of localized CEFR can do statements in both English and Japanese and localized CEFR vocabulary lists. A defining feature of the CEFR-J is that it has split the lower CEFR proficiency levels into further sub-levels in order to give Japanese students a sense of achievement, and thus enhanced motivation, when moving up through these easier-to-achieve sub-levels. Specifically, the CEFR-J includes a Pre-A1 level. It also divides the A1 level into three sub-levels (A1.1, A1.2 and A1.3), the A2 level into two sub-levels (A2.1 and A2.2), and the B1 and B2 levels each into two sublevels (B1.1, B1.2 and B2.1 and B2.2). (See Negishi, Takada, & Tono, 2012, and Tono & Negishi, 2012 for succinct and useful overviews of the CEFR-J at this stage of its development.)

BECC management felt that the CEFR-J would provide increased structure, and transparency for the GE curriculum, which lacked a clear organizing framework. The CEFR-J also seemed to align with the existing underpinning pedagogical principles of the GE curriculum which were based on communicative language teaching, task-based learning and fostering learner autonomy (for a full description and evaluation of Phases 3 and 4 of the GE curriculum renewal see Bower et al., 2017).

Phase 3 of curriculum renewal was seen to be successful in terms of increased teacher knowledge of the CEFR-J, creation of clearer lesson and curriculum goals, and the introduction of a more structured approach to fostering learner autonomy through the use of self-assessment checklists and supporting SALC activities. However, after two years of this project there was a feeling among many teaching staff that little progress had been made in aligning lesson content to the CEFR-J.

During phase 3 of renewal, methodological problems became apparent with the attempt to align the GE curriculum to the CEFR-J self-assessment grid, which consisted of levelled can do statements across five language skills in both English and Japanese, with each can do statement being a single sentence of up to three clauses. Judging how these single-sentence can do statements should be actualized in the lesson materials, and the extent to which the content of existing lesson handouts matched the statements in the CEFR-J self-assessment grid proved to be quite problematic. (The CEFR-J self-assessment grid can be downloaded from this web site: http://www.cefr-j.org)

1.4.4.4 Phase 4 – CEFR-based renewal

In early May 2014, the Oxford Online Placement Test (OOPT) (Pollit, 2009; Purpura, 2010), a standardised commercial computer adaptive test of listening and spoken interaction designed to place learners at their English CEFR proficiency level, was administered to a representative sample of 71 entering GE students. The representative sample was formed by administering the OOPT to one class from each of three ability streams (high, middle and low), which had been made based on the 2014 BET results. The results of the 2014 test are not admissible for this study as informed consent forms were not signed by students. Therefore, the OOPT was run again at the beginning of 2015 and provided results very similar to the 2014 administration. The results of the 2015 OOPT are shown in the Figure 1.1 below.

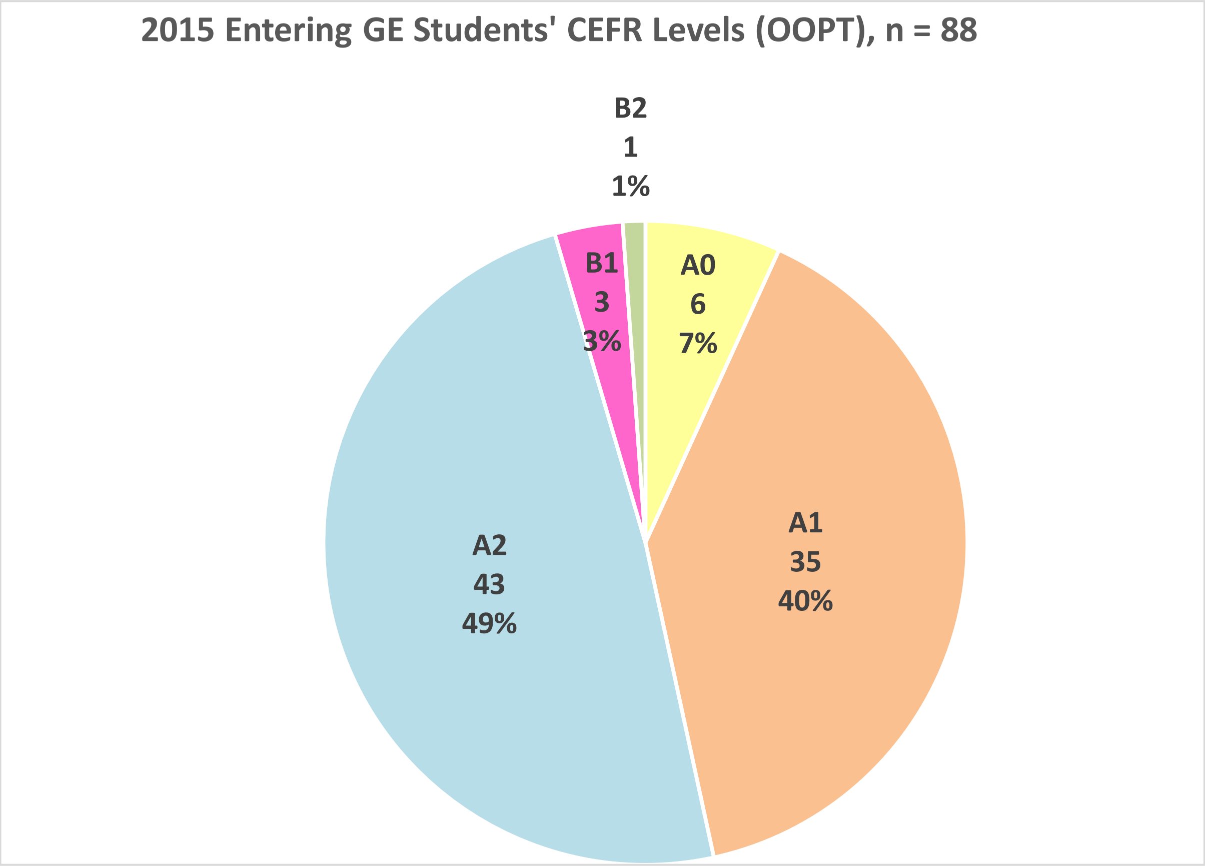


Figure 1.1. 2015 Entering GE Students' CEFR levels according to the OOPT

From Figure 1.1 it can be seen that according to the OOPT the vast majority of GE students enter at either the A1 or the A2 level. 47% of entering students are either at the pre-A1 or A1 levels, and 49% are at the A2 level. Just 4% of entering students are at the B1 and B2 levels. Based on this OOPT data, and also on the difficulties encountered in attempting to align the GE curriculum to the CEFR-J in phase 3, a further phase of curriculum renewal was initiated in 2014. In this phase course materials for two course streams were created for GE classes, a lower stream aimed at raising students from CEFR levels pre-A1 or A1 to A2 over two years of study, and a higher stream aimed at raising students from CEFR level A2 to level B1 over two years of study.

Table 1.1 summarizes the evolution of the BECC GE curriculum from 2008-2016.

Table 1.1. *Phases of BECC GE Curriculum Renewal*

2008	2009-2011	2012-2013	2014-2016
• First-year curriculum designed in late 2017, implemented in 2008 • Second-year curriculum designed in 2008 and implemented in 2009	• Yearly renewal cycles with improvements made to individual lessons • Lesson assessments improved • Can do statements added to second-year lessons	• Attempt to map the curriculum to the CEFR-J • Can do statements added as learning targets and self-assessment tools to all lessons • Course goals revised and clarified.	• Two levels of courses created (A1-A2 & A2-B1) • All lessons, assessments, and course materials redesigned with the aim of CEFR alignment

1.4.4.5 The CEFR

The CEFR is a product of the Council of Europe, an organization which promotes cooperation between European countries. The CEFR was preceded by an earlier Council of Europe document *The Threshold Level for Modern Language Learning in Schools* (van Ek, 1976) document, which aimed to give learners the minimum language competency required for living in a foreign language environment, such as being able to complete everyday transactions, and being able to build personal and work relationships (Trim, 2012). The Threshold was successful beyond the expectations of the group that organized its creation, and due to popular demand subsequently two more documents with specifications along the same lines as Threshold were created. Waystage detailed a level below Threshold, and Vantage detailed a level above. Waystage became the basis for the CEFR A2 level, and Threshold the basis for the CEFR B1 level.

The CEFR (Council of Europe, 2001a) was published commercially in French and English in 2001, and it has since had a huge impact on second language education not only in Europe, but in many countries around the world. Trim (2012) explains that the CEFR consists of three major parts.

1. scales of proficiency
2. a taxonomy of language use and competences
3. methodological options for learning, teaching and assessment

The CEFR has become best known for the first of these parts, its scales of proficiency or "can do statements", which describe second language learner proficiency divided into six levels of increasing proficiency from A1–C2, across many language subskills (see Little, 2006 for an accessible overview and description of the CEFR). The complete CEFR scales of proficiency, including the scales for the A2 and B1 levels, which were the target of the two new levels of GE courses and the BETs in this study, can be downloaded from a URL provided in the references section (Council of Europe, 2001b).

The GE curriculum designers drew on CEFR related resources such as *Waystage 1990* (van Ek & Trim, 1998b) for the A1–A2 course design, *Threshold 1990* (van Ek & Trim, 1998a) for the A2–B1 course design, the Core Inventory (North, Ortega, & Sheehan, 2010), and textbooks which claim CEFR alignment, in order to systematically redesign the curriculum based on target CEFR levels A2 and B1. The making of the new GE curriculum began with a planning phase in the first semester of 2014 and continued with one semester of GE materials being made each semester from second semester 2014, until the new curriculum was completed by mid-2016.

1.4.5 History of institutional English proficiency tests in the GE Curriculum

Table 1.2. *History of Institutional English Reading and Listening Proficiency Tests for the BECC's GE Curriculum*

Year	Curriculum Changes	Standardised Reading and Listening Tests (Administered in January and April)
2008	First-year GE curriculum based on the Kanda University of International Studies curriculum taught to first year students only Second-year GE curriculum developed with themes focusing on students' major subjects	Kanda English Proficiency Test (KEPT)
2009	New GE second-year curriculum taught to second year students	KEPT
2010	Individual lessons revised based on feedback	KEPT
2011	Individual lessons revised based on feedback	Kanda Assessment of Communicative English Test (KACE)
2012	Individual lessons revised based on feedback	Bunkyo English Achievement Test (BEAT)
2013	FE and SE revised based on feedback and to bring them closer to the CEFR-J	Bunkyo English Achievement Test (BEAT)
2014	Plan for recreating the GE curriculum based on the Threshold and Waystage created New first semester FE materials completed	Bunkyo English Test (BET)
2015	New FE second semester materials, and new SE first semester materials created	Revised BET based on KET and PET style questions
2016	New SE second semester materials to created. Feedback and revision process on the new GE curriculum began	Revised BET with lengthened reading and listening sections
2017	Further revisions based on teacher feedback	Revised BET with slightly lengthened listening section

From 2008–2010, a video-based English listening and reading proficiency test from KUIS known as the Kanda English Proficiency Test (KEPT) was used to assess the English proficiency of GE students. KEPT results were also used for class streaming. In 2011, another

KUIS test known as the Kanda Assessment of Communicative English Test (KACE) was used. However, because both tests were perceived to be too difficult for GE students, and were not able to measure achievement of the GE curriculum, new tests called the Bunkyo English Achievement Tests (BEATs) were trialled in 2012 and 2013. The name of these tests was changed to the Bunkyo English Tests (BETs) for their 2014 January and April administrations, and the specifications for the tests were revised with the intent of aligning them to levels A2–B1 of the CEFR. The structure of the BETs was revised in 2014 for their 2015 administrations and again slightly revised in 2016 for their 2017 administrations (for more detailed information about the history of institutional tests of reading and listening at the BECC see below and Bower, Rutson Griffiths, & Sugg, 2014).

1.5 The Bunkyo English Tests

1.5.1 Rationale for choosing BET task types

BET reading section testlet types were chosen by examining testlets from the reading and writing section of Cambridge’s Key English Test (KET), which is a criterion-referenced certification test of English at the A2 level, and Preliminary English Test (PET), which is a criterion-referenced certification test of English at the B1 level. Tasks which seemed to relate to a reading construct rather than to a writing construct, and which were able to be administered in a multiple choice or matching format amenable to machine grading with a bubble sheet reader were selected. Test specifications for these testlet types were then written by analysing official KET sample tests and by using information available in the KET Handbook for Teachers (University of Cambridge ESOL Examinations, 2012a), and the PET Handbook for Teachers (University of Cambridge ESOL Examinations, 2012b).

BET listening section testlet types were similarly selected by adapting testlet types from the listening section of the KET and PET, which were able to be scored using a bubble sheet reader.

1.5.2 Writing the BETs

The BETs are written and revised by a group of BECC teachers, all of whom have several years of teaching experience, and nearly all of whom have at least masters in TESOL or related field. The first BEATs administered in 2012 were made by a single teacher. Just two teachers, including the researcher, worked on the 2013 BEATs. In 2013, this team grew to three teachers to make the 2014 BETs. From the beginning of 2015 the team grew to four teachers and was officially named the General English Assessment Committee (GEAC), which is pronounced 'geek' with the pun intended. In 2016 the team grew to seven members. In 2015 and 2016 the GEAC was also responsible for making other assessments for the GE curriculum, including vocabulary quizzes, grammar quizzes writing and presentation rubrics, listening and reading assessments, and speaking tests known as the Bunkyo English Speaking Tests (BESTs).

Original material was written for the 2015 and 2016 BET testlets by GEAC members, based on the BET specifications and on GE curriculum content. Responsibility for writing testlets was divided between team members, and then a round of feedback was given on the items by all team members, and revisions made accordingly. Another round of feedback on the testlets and their items was given once all of the testlets had been assembled into test papers.

For the 2017 BET testlets, draft test papers were uploaded to web-based application called Moxtra (A mobile first, embeddable collaboration platform, n.d.). Each new or revised testlet then received feedback from at least three GEAC members, revisions were made based on the feedback, and a final round of feedback was given, and modifications made as needed.

Unfortunately, due to limited resources and a tight timeline it was not possible to pilot versions of the BET before using them with the actual student population, which means that changes to the BET structure were made quite slowly on a year by year basis in which statistical analysis and teacher feedback on one year's test administrations led to changes in the structure and content of following year's tests.

1.5.3 BET administrations

The 2015–17 BETs were administered using a DVD which was projected on a screen easily visible to all test takers. The DVD automated the timing and delivery of the test. All aural DVD test instructions and item instructions in the question booklets were given in Japanese, and the timing of the reading section was displayed using a countdown timer. All of the listening section testlet instructions and listening passages were also played from the DVD.

Students received a bubble sheet, a question booklet and a sheet of blank paper to make notes. After the test, the question booklets, and completed bubble sheets were collected. All scratch paper was also collected to maintain test security, so that test items could be used again for future test administrations.

There are two BET administrations each year. BETs 2 and 3 are administered at the end of each academic year in January. BET2 is used as an achievement test for students finishing the first-year curriculum, known as the Freshman English (FE) curriculum, and was worth 15% of students' second semester FE grades in the 2015/16 and 2016/17 academic years. BET2 results were also used for streaming students into their second-year Sophomore English (SE) classes. BET3 is intended to be an achievement test of the whole GE curriculum and it was worth 15% of students' second semester SE grades in the 2015/16 and 2016/17 academic years. The BET3 is also ultimately intended to form the basis of issuing the reading and listening components of proficiency certificates at CEFR levels A2 and B1 for students who have completed the GE curriculum.

BET1 was administered to new students at the end of March, before they started their GE classes in April. There are just a few days to run the bubble sheets through the bubble sheet reader, score the tests and stream FE classes before the first FE class of semester 1.

1.5.4 BET 2015 format

The 2015 BETs consisted of four sections, a reading section which had four tasks (or testlets) and 27 items, a short vocabulary section with a single task and five items, a grammar section which was a single task consisting of ten items, and finally a listening section which had five tasks and 26 items. Test takers were given 30 minutes for the reading section, 10 minutes for the vocabulary and grammar sections, and approximately 30 minutes for the listening section.

1.5.5 BET 2016 format

Changes were made to the test specifications for the 2016 iterations in attempt to increase the reliability of the tests by lengthening both the BERT and the BELT. The grammar and vocabulary sections of the BERT were changed to be subsections of the BERT, and a ten item testlet with a true/false response type and a longer reading passage was added to the BERT. In addition, a six item, true/false testlet was added to the BELT, and the target level of the first seven item section of the BELT was increased from A2 to B1 level.

1.5.6 BET 2017 format

Just one small change was made to the 2016 BET format for the 2017 BETs. Listening Part 1 was increased in length from seven items to ten items. The structure of the 2015 and 2016 BETs is summarized in the attached Appendix A.

1.5.7 BET versions

There are three versions of the BET. The three BETs are designed from the same test specifications so are of the same length, and contain the same sections and task types. This is to allow the BETs to be equated using horizontal equating (Skaggs & Lissitz, 1986). The thematic content of the tasks for each BET varies somewhat according to how far students have moved through the GE curriculum, but the BETs should be of the same approximate difficulty because the specifications of the tasks are written to make testlets of the same difficulty in terms of cognitive processing. For example, the text length for the focus texts in testlets of the same type for reading and listening and reading comprehension should be the same across all three BETs. The complexity of grammar, the vocabulary range, and the cognitive demands of items for equivalent testlets should also be similar across all three BETs although the topics and themes will change. The purpose and content of each of the three BETs is briefly outlined in Table 1.3.

Table 1.3. *Purposes and Contents of the BETs*

	BET1	**BET2**	**BET3**
Purposes	• To place entering students into the two FE courses, and stream them into classes within the courses • To give teachers and students diagnostic information about students' reading and listening ability	• To place second-year students into the two SE courses, and stream them into classes within the courses • To give teachers and students diagnostic information about students' reading and listening ability • To assess student achievement of the FE curriculum goals • To measure gains in students' reading and listening ability	• To assess student achievement of the overall GE curriculum goals • To form the basis for placing students into CEFR levels for reading and listening for achievement certificates • To measure gains in students' reading and listening ability
Content	50% based on the FE curriculum 50% based on the SE curriculum	All content based on the FE curriculum	50% based on the FE curriculum 50% based on the SE curriculum
Administration Time	Before the start of semester 1 in late March/early April	At the end of students' first year of study in mid-January	At the end of students' second year of study in mid-January

1.5.8 Presentation of BET grades

For the 2015 BETs students and teachers were provided with raw scores for the reading, vocabulary, grammar and listening sections. For the 2016 BETs students and teachers were given raw total BET scores and raw scores for the listening and speaking sections. Students were also provided with average BET total, listening and speaking scores for their class and stream in 2015 and 2016. For the 2017 BETs students were again given their raw scores for the whole test, and for the listening and reading sections. They were also given their relative position or ranking in their class and in their course stream.

1.5.9 Purposes of the BETs

The BETs are intended to serve four somewhat overlapping purposes.

1. To place students into appropriate English language courses and to stream them into classes within courses based on English language ability

2. To objectively measure improvements in students’ reading and listening ability after one year and two years of study in GE courses in terms of the CEFR designated levels

3. To form part of the basis for issuing proficiency certificates at the CEFR A2 and B1 levels upon completion of the GE program, including *can do* statements to show student competencies

4. To have positive impact on all stakeholder groups. Including positive impacts on student learning, teaching practice, and management attitudes to the BECC and GE assessment.

1.6 The scope of this study

This study focuses primarily on functions 1 and 2 of the BETs listed in the previous section, which are the BETs placement/streaming and achievement functions. There is also a minor focus on function 4, the impact function of the BETs. The intended function 3 of the BETs is not addressed, as achievement certificates were not issued to students within the frame of this study. The placement/streaming function of the BETs is the focus of research question

1 of this study, and the achievement function of the BETs is the focus of research question 2. Research question 3 focuses on the development phase of test validation, and details insights for further test development based on weakness found in the BET validity argument.

The timeframe of this study covers the first two academic years of BET development and implementation, which coincided with the design and implementation of the new GE curriculum from 2014 to 2016. Thus, five sets of reading and listening tests (referred to as the BETs in this study) are examined. The BETs which are examined in this study are summarized in the Table 1.4.

Table 1.4. *The BETs Covered by this Study*

Test	Administration Date	GE Student Cohort
BET1 2015	April 3, 2015	Students entering in 2015, or the 2015 cohort
BET1 2016	April 4, 2016	Students entering in 2016, or the 2016 cohort
BET2 2016	January 18, 2016	The 2015 cohort at the end of their first year of study in the GE curriculum
BET2 2017	January 19, 2017	The 2016 cohort at the end of their first year of study in the GE curriculum
BET3 2017	January 20, 2017	The 2015 cohort at the end of their second year of study in the GE curriculum

It is important to note that most frameworks of language test development and validation assume that the domain of a test can be well-defined before test development begins (Bachman & Palmer 2010; Kane 2006). However, in this case, out of necessity, the BETs were developed and revised at the same time as a new GE curriculum was being created and revised from 2015–16. Thus, there was two-way feedback in which the creation of the BET specifications influenced the design of reading and listening tasks in the GE lessons, and the curriculum also strongly influenced BET testlet creation, as lesson topics, tasks and vocabulary from the GE curriculum were used as a basis for making BET testlets.

1.7 The position of the researcher and efforts toward reflexivity

In qualitative and mixed-methods research it is well accepted that the position of the researcher within the context of research may cause bias in the research methodology chosen as well as in the interpretation of results. Kane (2012) also emphasises that there is a tendency toward "confirmationist bias" (p. 4) in validation studies such as the current one, which are in the development stage of validation. Efforts to acknowledge a researcher's position and point of view, and the possible influences of these on research methodology and conclusions, as well as to try to minimize this source of bias are known as reflexivity (see A. Lee, 2011 for a brief overview of reflexivity).

In the timeframe of this study the researcher worked firstly as the BECC Assistant Director for the 2014/15 academic year, and the first semester of the 2015/16 academic year, and then as the BECC Director for the second semester of the 2015/16 academic year and for the 2016/17 academic year. Over this time period the researcher was also the head of the General English Assessment Committee and was in charge of BET development. This position could lead to two possible sources of bias in the current study. These two possible sources of bias are described in the following paragraphs along with efforts made to mitigate them. The approach to reflexivity in this study is what A. Lee (2011, p 3) refers to as "methodological reflexivity" in which there is "A focus on the methods … deployed in research as well as an acknowledgment of the role of the researcher" (p. 3).

The first possible source of bias can be called *confirmationist bias* and may result from the researcher's role in leading the BET development process and a natural tendency to find evidence that supports the effectiveness of the BETs in their proposed uses. In order to minimize this type of bias, I have done my best to review the evidence in the BET validity argument in this study as objectively as possible. In addition, attempts were made to include the opinions of as many other stakeholders in the BETs as possible, such as student, teachers and managers.

The second source of possible bias in the results of this study is that teachers who responded to surveys and in the focus groups knew that the researcher was a BECC director, and this may have had an unconscious effect of causing them to answer survey items and to respond to focus group questions related to the BET in a positive way. Efforts to minimize this possible source of bias were made by making the teacher surveys anonymous, and by recruiting an outsider academic to lead the teacher focus groups.

1.8 Organization of the thesis

The remainder of this thesis is organized into six chapters. Chapter 2 firstly provides a broad review of the literature on language test validation across the 20th century and the beginning of the 21st century, and then explains the argument-based approach chosen for this study. Chapter 3 presents a review of exemplary argument-based language test validation studies, which have been conducted both on large-scale commercial tests and on small-scale, in-house language tests. Chapter 4 presents the details of the BET Interpretation/Use Argument, and Chapter 5 describes the methods used to seek backing in the BET validity argument. Chapter 6 presents, and analyses results found from backings sought for the BET validity argument.

In the final Chapter 7, the results of the BET validity argument for each inference in the BET IUA are summarized and conclusions are drawn. Each of the research questions of this study is then answered, followed by a description of the study's limitations. Reflections are presented on using Kane's argument-based approach in the context of this study, and contributions of this study to the literature on language test validation are outlined. Finally, suggestions for further research are made.

CHAPTER TWO

Literature Review: An Overview of Validity and Validation in Language Testing

2.0 Introduction

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) Standards (2014) define *validity* as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). According to Xi (2008), *validation* is the procedure of gathering and analysing evidence to support proposed interpretations and uses of test scores. Thus, validation describes the processes involved in evaluating test validity. Different conceptualizations of validity lead to different types of validation procedures and, in turn, to gathering different types of evidence for the conceived validity. Accordingly, validation frameworks give guidance on how to select, combine and appraise validity evidence.

This chapter first briefly outlines major periods in the history of test validity theory in the 20th and early 21st centuries. An overview of progress in validity theory in the wider fields of social science and education from the 1920s to the 1990s is presented, followed by a summary of how these advances were applied in the field of language testing. Secondly, the most recent advances in test validity theory into argument-based approaches are explained, followed by an overview of how argument-based approaches have been broadly adopted and adapted in the field of language assessment. Finally, the argument-based validation framework constructed for this study is explained.

2.1 Major periods in the history of test validation

2.1.1 Pre-Messick: componential perspectives on validity

From around 1920–1950 criterion validity along with content validity were the dominant validity models in educational testing. Criterion validity (or criterion-related validity) involves comparing the results of a designed test to some criteria or standards. In criterion validity, a newly designed test is considered valid to the extent to which it provides comparable results to those of a valid criterion.

Criterion validity is divided into two types: ‘predictive validity’ and ‘concurrent validity’. If the criterion is obtained sometime after the test, when the test taker performs in the target language domain, it is called predictive validity, and if the test score and the criterion are obtained around the same time it is called concurrent validity. An example of predictive validity in language testing is comparing students’ language proficiency test scores to their later academic performance. An example of concurrent validity in language testing is administering an in-house reading test and comparing the results with those of a commercial reading test (which has strong validity evidence). From a criterion validity perspective, validity is defined as the congruence between test takers’ performance on a newly designed test and on a valid criterion; the corresponding validation procedure is usually calculating correlations between the test takers’ performance on the test and its criterion, with high correlations considered as being strong evidence for validity.

Kane (2006) notes two major advantages of criterion-related validity. The first advantage is that in many cases the criterion is clearly related to a test’s uses and interpretations. For example, if a test claims to predict later academic performance then it is obviously useful to check this claim empirically by correlating test scores with later academically achieved grades. The second advantage, as Kane notes, is that an assessment of criterion validity is fairly objective, as long as the test is well designed, and the criterion is carefully defined and chosen.

The major limitation of criterion validity is that in some cases it is very difficult to find a suitable criterion (Cronbach & Meehl, 1955; Kane, 2006). For example, for abstract constructs such as ‘assertiveness’ or ‘attachment’ there is no real-world, easily-measured criterion available. There is also the problem of validating the criterion, which may lead to a circularity, or to an infinite chain, as further criteria are sought to validate each criterion.

Content validity was considered the major alternative to criterion validity from 1920–1950. Content validity allows for test validation without reference to a criterion. For content validity, a test is validated by how well it samples from its target domain. The target domain is the content, knowledge, or abilities which the test aims to assess. An example of content validity would be how well the questions on an achievement test represent the contents of a curriculum, or how well the tasks on an achievement test assess the course learning objectives.

For a test to have content validity it should elicit a “representative sample” of tasks from the target domain (Guion, 1977). Anastasi (1998) states that for a test to have content validity “the behaviour domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions” (p. 132).

The validation procedure for content validity is generally based on expert judgment. For a test to have high content validity it is important that the target domain be clearly defined, that task specifications have been designed carefully to tap the skills or content of the target domain, and that the test is administered consistently and evaluated properly (Kane, 2006).

A major advantage of content validity is that it allows a test to be validated without having to refer to external criteria. For tests of a specific, clearly defined skill, such as writing ability, or piano playing, a good argument can be made that the samples obtained by the test are representative of the overall ability of the target domain. However, Messick (1989) pointed out that content validity, while providing useful evidence of a test’s representativeness of the target domain, does not address other important areas of validity such as inferences to be made from test scores and social consequences of a test. Thus “in a fundamental sense so-called content validity does not qualify as validity at all” (p. 17).

In addition, caution needs to be maintained with the use of expert judgment, as some studies have shown expert judgment to be inconsistent, and unreliable. For example, Alderson and Lukmani (1989) and Alderson (1990) conducted studies which elicited content area expert judgments of reading skill level, and specific reading skill targeted for three tests of English as a foreign language: an in-house reading test used at Bombay university, the Test of English for Educational Purposes (TEEP), and the English Language Testing Service (ELTS) a predecessor to the IELTS® test. The results showed disagreement by content experts on:

1. classifying reading test items by level of reading skill (higher order, middle order or lower order)
2. classifying reading test items by specific reading skill from a list
3. defining which reading skills items were testing

In addition, in the above Alderson studies, items defined by content experts as testing higher order skills did not correspond with higher item difficulty or discrimination. In a further study reported by Alderson (1993), in which test makers were asked to predict the difficulty of test items for a new National Certificate of English (NCE) test in Sri Lanka, found that “the judgments of experienced testers about item difficulty are simply too variable and inaccurate to be trustworthy” (p. 51).

Finally, a more recent study by Alderson and Kremmel (2013) replicated parts of a study by Shiotsu (2010) in which expert judges distinguished between items testing syntax and lexis. Alderson found that several of the items judged as testing syntax for a test in Shiotsu’s study should have been excluded from the test based on the expert judgment in the replication studies. These results cast doubt on the results of Shiotsu’s study, and also further highlight the need to include other methods in addition to expert judgment as sources of validity evidence.

2.1.2 Construct validity

In the early 1950s a third theory of validity and approach to test validation arose, known as construct validity. Cronbach and Meehl (1955) argued that for tests for which there is no clear criterion to predict and no available domain of content to sample from, and for which the construct to be measured may be abstract and only theoretically defined, construct validity is necessary. For construct validity, a theory of the construct must be clearly defined and the test results evaluated in light of how well they fit the theory. If test results do not match those features outlined by the theory, the measurement instrument may need to be revised. In addition, the theory may need to be re-evaluated if test results consistently contradict the theoretical positions. Cronbach and Meehl (1955) proposed a nomonological network, consisting of a theoretical framework of interrelated constructs, which lay above an empirical framework. The theoretical framework should be linked to the real world through observations, and the results of observations are used to confirm or modify the theoretical framework.

Cronbach and Meehl's theory of construct validity expanded validity theory to include a framework for validating the measurement of abstract psychological constructs. In addition, construct validity moved the focus of validation from a focus on validating tests to a focus on validating interpretations of test scores. As Cronbach and Meehl stated, it is not the test which is validated but "a principle for making inferences" (Cronbach & Meehl, 1955, p. 297).

Cronbach and Meehl (1955) suggested a number of methods which can be used in construct validation, including: (i) demonstrating differences between groups expected to measure high and low on the construct, (ii) correlation matrices with other measures of the same construct (iii) factor analysis to explore and confirm dimensions of the construct (iv) measures of internal consistency (v) studies of change or stability in the construct over occasions, for example, if the construct is stable over retests, and if the construct changes as predicted by experimental interventions, and finally (vi) observations of subjects' actual performance on the test, which may lead to the discovery of unexpected sources of test score variance, such as poor reading ability lowering math test scores.

Cronbach and Meehl’s seminal work created a context for the evolution of a unitary concept of validity as suggested by Messick, which is explained in the following section.

2.1.3 Post-Messick: A unitary conceptualization of validity (Late 1980s–1990s)

By the early 1990s the construct theory of validity had become widely accepted as a general theory of validity which subsumed the other two types of validity, namely, criterion and content related. Messick (1989) argued that the theory of validity had moved to an understanding that content validity, criterion-related validity (predictive and concurrent validity), and construct validity, were in fact all facets of a larger conceptualization of construct validity and not distinct from it. Thus, criterion and content validity were reduced to types of evidence to be gathered to support a larger argument for construct validity (Carr, 2011). Messick’s 1989 paper was massively influential, and it is widely regarded as a watershed in test validity theory.

Rather than slicing up the concept of validity into the three traditional categories of content, criterion and construct, Messick put forward the following conceptualization of validity denoting a unitary concept with four facets of *evidential basis*, *consequential basis, test interpretation and test use,* as illustrated in Table 2.1.

Table 2.1. *Messick's Unitary Model of Validity*

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity +Relevance/Utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

In addition, Messick (1989, 1996) described six components of construct validity. These components are firstly *content knowledge*, secondly *substantive* (skills and processes), *structural* (nature of the construct), and *generalizability across time and setting* (from Loevinger, 1957), and lastly *external* (comparison with other relevant criteria) and *consequential* (test impact*)*. Each is summarized briefly in the following paragraphs.

Messick emphasised that test *content* remained as a critical aspect of overall construct validity, including considerations of content relevance and representativeness. However, he also stressed that test content is insufficient by itself as a basis for validation. In addition, Messick argued that judgments of test content inevitably involve considerations of construct theories such as the processes test takers use to answer questions, making test content considerations essentially inseparable from construct considerations.

The substantive component according to Messick involves specifying the skills covered in the domain in terms of the construct theory, and then creating tasks, items and rubrics based on the theory-grounded domain specifications. After item analysis those items that do not function statistically as they should according to the construct theory should then be removed or revised. Evidence for this component should show that the processes engaged in by test takers and correlations are in line with those predicted by theory. This approach is in contrast to a purely content-based approach in which items are only included in a test based on content expert judgments, and to purely empirical approaches which only base the selection of items from a pool on statistical considerations.

The structural component of construct validity involves providing evidence that scores attached to a test make sense in terms of the distinct behaviours that the test claims to measure. Theoretical predictions that a construct is unidimensional or made up of multiple components or facets should be supported by data analyses of the test, and by correlations between test sections, and between test items and the overall test.

The generalizability component investigates how well test scores and interpretations based on test scores generalize across different test taker populations, in different settings, and in different task domains and contexts. This includes examining if relationships between the test and criterion remain stable across time and in different settings. Part of Messick's generalizability component overlaps with his content component, in that for generalizability, as for content, the test should include a wide representative sample of tasks from the target domain to allow for test scores to be generalized to the wider target domain. In addition, Messick (1996) distinguished between *generalizability as reliability*, which refers to traditional concepts of reliability in that "the consistency of performance across the tasks, occasions, and raters of a particular assessment" (p. 250) should be consistent, and *generalizability as transfer* which refers to "consistency of performance across tasks that are representative of the broader construct domain" (p. 250). Generalizability as transfer refers to the aspects of performance that the test task helps learners to learn, and also the aspects of performance that the test is predictive of. Evidence for generalizability could come from correlations with other tests that claim to measure performance of the same domain, or from correlations with later measures of performance on the actual domain tasks. Thus, Messick's generalizability as transfer component incorporated previous conceptions of criterion-related and predictive validity.

The external component of construct validity assesses the extent to which relationships between test scores and other external criteria matches that predicted by the construct theory. For example, a test of numeracy should correlate highly with other tests of numeracy, but not with a test of reaction time.

The final and most important aspect of Messick's (1989) framework was an emphasis on the consequence component of test utilization. Messick argued that both meaning and values should be considered when interpreting test results to make decisions. By this he meant that a validation framework should take into account both test score interpretations and the consequences of decisions made based on those interpretations. Social consequences of testing may include for example negative effects of a test on a particular social group. Another example of test consequences is washback (Messick, 1996) in which educational practices are changed in a positive or negative way as a result of the introduction of a test or changes to test content.

Messick also made the important point in his 1989 chapter that validation is a continuing process for the life of a test given the changes in theorisation of concepts as well as testing circumstances. There is no end point at which a test can be said to be valid. Rather validation continues as long as the test is in use.

2.1.4 Critics of Messick’s unitary conceptualization of validity

Messick’s unitary theory of validity is predominantly an interpretive theory that focuses on test score interpretations and consequences, rather than on the tests per se. Some validity theorists are critical of the dominant interpretive perspective that test validity should focus on the interpretations and uses of test scores, and also that validity should include considerations about the consequences of test use. These critics follow a realist perspective and prefer to focus on the test itself, viewing validity as the extent to which a test measures what it purports to measure. Borsboom, Mellenbergh and Heerden (2004), for example, argue for a simpler approach to validity, which focuses more narrowly on defining the theory of attribute, and establishing that variations in the attribute cause systematic variations in the measurement of the attribute from the test. They argue that validity should be considered a property of tests and not of the inferences made from test scores. To them the theory of a construct to be measured by the test should be the central concern, and making tests based on a theoretical idea of the construct to be measured should be the important part of validity, rather than having the primary focus on validating the test after construction. They accept that concepts such as reliability, predictive power and consequences are important to consider when analysing tests, but they insist that these should not be part of validity. To them (following the construct validity theory of Cronbach and Meehl, 1955) validity should only focus on defining the construct in terms of the processes involved and assessing if the test effectively measures variation in the test takers' performance in correspondence to the defined construct. As such, they argue that only direct causal evidence should be used to support the validity of a test and that the continued acceptance of correlational evidence for some components in the unitary concept of validity, such as correlation evidence for generalisability, and correlations with external criteria is flawed.

Another critic of the unitary concept of validity, Newton (2012), made an argument that several aspects of the definition of validity in the 1999 *Standards for Educational and Psychological Testing*, which draws on Messick's framework, needed to be clearer, and more consistent. In his 2012 article Newton argued against some aspects of Messick's validity framework. Firstly, Newton, in line with Borsboom et al. (2004), argued that validity should not include ethical and social evaluations of consequences of the use of test scores. For Newton, especially in the case of producers of commercial tests, these aspects of test evaluation should be left up to other stakeholders other than the test makers, such as test users, and should not be considered a facet of validity. Including these aspects would place too much burden on test makers.

Secondly, Newton argued that designers and users of tests should have separate validation responsibilities. Designers of commercial tests should only bear responsibility for validating their explicitly stated uses of a test. If tests are later employed for other purposes by test users, then the test users would bear responsibility for validating those uses.

In summary, the previous sections, and the current section described the evolution of the broad field of test validity theory in the 20th century. Initially, from a componential perspective, validity was generally conceived to consist of two types, criterion validity which consisted of correlating test results with a criterion, and content validity which assed how well test content represented and matched its target domain. A third type of validity known as construct validity emerged in the second half of the 20th century thanks to the work of Cronbach and Meehl (1955) which sought to validate abstract constructs through defining and verifying nomonological networks. Finally, at the end of the 20th century the field of test validation reached a wide consensus (although a few critics still remain) that validity is a unitary concept, and that what had previously been conceived as separate kinds of validity were actually facets of construct validity.

2.2 Validity in the field of language testing 1950s–1990s

Both Chapelle (1999) and Xi (2008) provide thorough overviews of how conceptions and approaches to language test validity evolved from the 1950s through to the 1990s. The consensus on validation theory for language tests in the 1950s and 60s is exemplified by Lado's (1961) classic text *Language testing*. The focus of language test validation at that time was on the test as a measurement instrument, and whether or not a test measured what it claimed to measure. Validation work mainly focused on validating multiple choice tests through criterion and content approaches, and reliability was seen as separate to validity.

In the 1970s more focus on communicative language tests rather than multiple-choice tests emerged. This trend was dubbed "communicative language testing" by Morrow (1979). Those in favour of communicative language testing argued that multiple-choice tests could not measure real communicative language ability. Communicative ability could only be measured by having test takers engage in communicative tasks similar to real-life situations, and by scoring their performance on these tasks. Such concerns had been raised in previous decades, but they came to the forefront of language testing in the 1970s and 1980s (Fulcher, 2000; Morrow, 2012). Communicative language testing advocated testing aspects of language use such as context and authenticity. Communicative language testing also advocated more qualitative assessment methods using human raters. However, validity approaches in the 1970s were still restricted to content and criterion types (Clark, 1975).

By the late 1970s and early 1980s the construct validation approach began to influence language testing research. New methods emerged to investigate the meaning of test scores, such as looking at test taking strategies (A. Cohen, 1984), comparing different test methods, (Shohamy, 1984), and analyzing test bias through item analysis (Z. Chen & Henning, 1985). The first book of papers on construct validation studies of language tests also appeared at this time (Palmer, Groot, & Trosper, 1981).

After 1990, thanks to Bachman's (1990) book *Fundamental considerations in language testing*, Messick's work on a unitary concept of validity also became the dominant validity model in language testing in the 1990s. Bachman's book introduced Messick's ideas to the language testing field, and was tremendously influential. One indication of the book's influence is that it was the most cited book in the respected journal *Language Testing* in the decade following its publication (McNamara, 2003). In the book's chapter on validity Bachman advocated the unitary concept of validity, along with the idea that the subject of validation should be the inferences made from test scores, not the test itself. Test consequences were also included in Bachman's conceptualization of validity.

2.2.1 Bachman & Palmer's test usefulness framework

Bachman followed up the important ideas expressed in his 1990 book with a second book co-authored with Adrian Palmer entitled *Language Testing in Practice* (1996). In the book, Bachman and Palmer's attempt to make Messick's work more interpretable and easy to implement for language testers by introducing the idea of 'test usefulness'. This is because although Messick's (1989) explanation of validity is theoretically exquisite, in practice it was quite challenging to comprehend and implement (McNamara, 2003). In their book, Bachman and Palmer describe test usefulness as comprised of six components: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Reliability corresponds to the traditional definition of this term in testing as the consistency of scores. Construct validity involves clearly defining the construct and assessing the extent to which test scores measure the construct. Authenticity reflects the extent to which test tasks are representative of tasks in the target language use domain, which was the major concern of the communicative language testing movement of the 1970s and 80s mentioned above. Interactiveness examines the individual test taker characteristics involved in completing test tasks. Impact refers to the effects of introducing a test on education systems and society. Finally, practicality involves doing a cost-benefit analysis of test development.

Bachman and Palmer argued that the relative importance of each of the six components in their framework will vary depending on the particular language test and the language testing situation. They further stated that none of the components should be neglected, but the emphasis placed on each component, and the type and amount of evidence to be gathered will vary depending on the test and context.

Language Testing in Practice was an important contribution to the field of language test validity, because it further introduced concepts from Messick's unitary conception of validity to a wider audience of language teachers, and it also presented test validity concepts in an easily comprehensible form. A couple of problems with this book however, were firstly a lack of connecting the components of their test usefulness framework to existing frameworks' terminology in the test validity field. Bachman and Palmer's new terminology for established concepts could be confusing for readers. Secondly, the book gave only a very cursory treatment of the types of evidence that can be gathered to support inferences from test scores, how to select appropriate types of evidence to gather, and how to prioritize sources of evidence. The types of feedback listed in the book focused only on test takers and not on other stakeholders, for example, for collecting washback evidence.

2.2.2 Further work on the consequences of test use in the language testing field

Advancing on Messick’s important contribution in recognizing the crucial nature of test consequences for test validity, in the 1990s a wider movement arose which argued that the consequences of testing should also be considered an important part of validity. This is aspect of validity is often called consequential validity. For example, Shephard (1993) argued that validation should not just assess whether a test measures what it claims to measure, but also whether a test *does* what it claims to *do*.

Other authors coming at validity from the perspective of “critical language testing” emphasised the importance of social consequences of tests. Shohamy (2001), for example, argued that tests can often be used as a means for exercising power in society. She claimed that tests are frequently used to maintain the exclusivity and power of elite groups. As a way of compromising test hegemony, Shohamy recommended using multiple methods of assessment rather than a big final standardised and centralized test. The use of multiple and varied assessment tasks should, Shohamy argued, decentralize, and democratize the testing process in favour of different stakeholders. She also emphasised, in agreement with Davies (1997), that test developers should bear some responsibility for how their tests are used, not just for making the test. Shohamy states that “Studies of the use of tests as part of test validation on an ongoing basis are essential for the integrity of the profession” (p. 390).

B. K. Lynch (2001) put forward a validity framework rooted in post-modern political theory. Lynch’s framework is intended for validating alternative assessments such as portfolios, and it places a strong emphasis on ethical considerations including the power relations involved in an assessment context. Lynch’s framework consists of five categories: fairness, ontological authenticity, cross-referential authenticity, impact/consequential validity, and evolved power relations. While Lynch’s framework does expand the boundaries of validity and created useful areas for debate, it suffers from a lack of specific examples of validation studies within this framework, making it difficult to interpret how the framework would be utilized in practice. In addition, the framework was criticized by Bachman (2005) as being simply a list of validity categories, which does not show practitioners a clear pathway for validating test interpretations to test uses.

Kunnan (2008) put forward a Test Fairness Framework (TFF), which he further elaborated ten years later in his book *Evaluating Language Assessments* (Kunnan, 2018). In the latest iteration of his framework Kunnan presents two main principles, each with accompanying sub-principles. The first main principle is the *principle of fairness*, which entails treating each test taker fairly and equally. The sub-principles of the fairness principle are:

1. An assessment *ought* to provide adequate opportunity to acquire the knowledge, abilities or skills to be assessed for all test takers.
2. An assessment *ought* to be consistent and meaningful in terms of its test score interpretation for all test takers.
3. An assessment *ought* to be free of bias against all test takers, in particular by avoiding the assessment of construct-irrelevant matters.
4. An assessment *ough*t to use appropriate access, administration and standard-setting procedures so that decision-making is equitable for all test takers. (Kunnan, 2018, p. 80, emphasis in original)

The second main principle of Kunnan's framework is the *principle of justice* which mandates that the test should have beneficial consequences for society, and promote justice and positive values . The sub-principals for the justice principle are:

1. An assessment institution *ought* to foster beneficial consequences to the test-taking community.

2. An assessment institution ought to promote positive values and advance justice through public reasoning of their assessment (Kunnan, 2018, pp. 80-81, emphasis in original).

Shohamy, Lynch and Kunnan's frameworks all helped to raise awareness of the importance of examining the consequences of language tests for individuals and minorities and to strengthen the case for including test consequences as part of language test validation.

Although Messick's framework and the following work on test fairness and consequences, and Bachman and Palmers concept of test usefulness were major steps forward for validation theory, there remained a lack of a comprehensive framework to give practical guidance on the best types of evidence to gather to support a specific test score interpretation and use, and guidance on when to stop a validation investigation. In order to address these deficiencies, argument-based frameworks arose as the next major advance in validation theory. These argument-based frameworks are described in the next section.

2.3 Argument-based validation frameworks 1990s–early 2000s

Argument-based validation frameworks appeared in the wider field of educational measurement in the late 20th century (see, e.g., Cronbach, 1988; Kane, 1992), and early 21st century (see Kane, 2001, 2002, 2004, 2006; Mislevy, Steinberg, & Almond, 2002, 2003). Argument-based test validation frameworks have also become prominent in the field of language testing, especially with the contributions from Bachman (2005), Chapelle, Enright, and Jamieson (2008) and Bachman and Palmer (2010).

Argument-based approaches to test validation do not add any new qualitative or quantitative methods to the test validation arsenal; however, they do provide a more logical, coherent, and organized framework for selecting and incorporating existing validation techniques within an argument, which is a big leap ahead compared to previous approaches (Chapelle, 2010). Proponents of an argument-based approach perceive it is as more practical and efficient than the preceding unitary construct validation approach (Messick, 1989), because it provides a framework for deciding what kind of validity evidence to gather, how much evidence to gather, and when to stop gathering and examining the evidence (Kane, 2013a, 2013b).

In argument-based approaches to validity, test developers or researchers must construct an argument for the selection or design of a test, for the interpretation and use of a test, and for the decisions to be made based on test results (Bachman & Palmer, 2010). The following sections summarize developments in argument-based approaches to validation both in the wider field of educational assessment and in the field of language assessment in the late 20th and early 21st centuries.

2.3.1 Toulmin's argument structure

Most argument-based approaches to test validation (Bachman, 2005; Bachman & Palmer, 2010; Kane, 2006, 2013a, 2013b; Mislevy, Steinberg, & Almond, 2002, 2003) are based on Toulmin's (1958) informal argument structure. Toulmin's informal argument structure describes how practical arguments in non-philosophical, and non-mathematical fields like the law, art criticism, and history are used to build a case for a particular conclusion. Toulmin divides the parts of an argument into a *claim* or conclusion, and *data* or grounds. The claim is the assertion of an argument, and the data are the facts on which the argument is based. Claims are hedged with the use of "qualifiers" such as 'most likely' 'almost certainly' etc. Claims and data are linked by "warrants", which are general rules, principles, or accepted procedures which authorize the inference from the data to the claim. Warrants are also supported by "backing", any evidence which can be collected in favour of or against the warrants. The type of backing depends on the field of the argument, and may include scientific theories, laws, precedents, statistics or other forms of evidence. In addition, conditions or exceptional circumstances under which the inference from data to claim may be rejected are known as *rebuttals*, or negation of warrants. Figure 2.1 shows a generic illustration of Toulmin's argument structure.

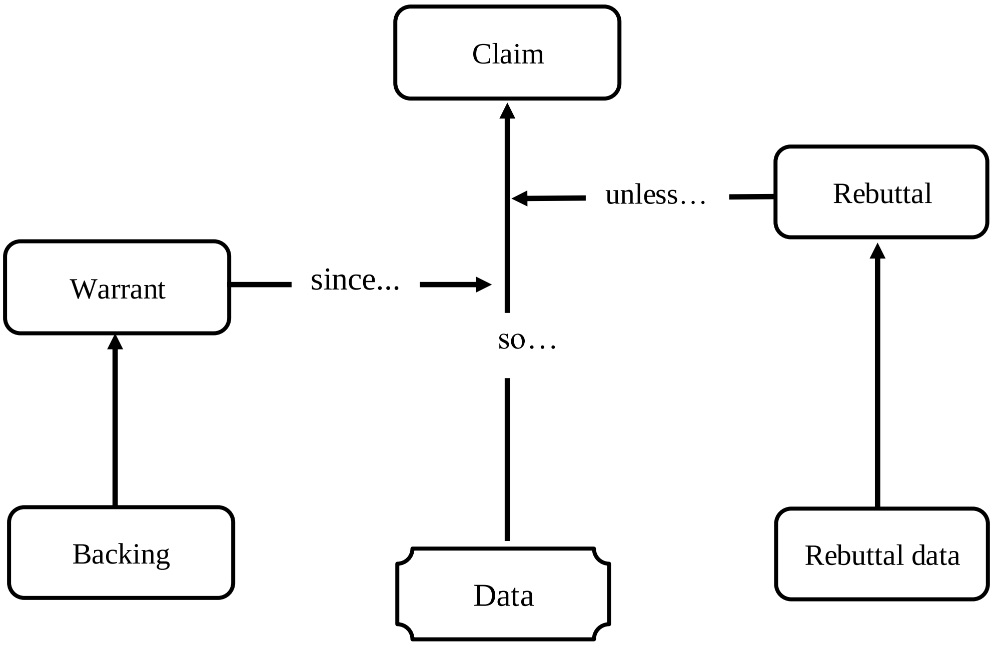


Figure 2.1 A generic illustration of Toulmin's informal argument structure (Toulmin, 1958)

To illustrate the categories in Toulmin's argument structure more clearly, an example of a specific argument is shown in Figure 2.2. In this case, the claim can be, for example that a woman named Elizabeth is eligible to vote in Australia. The data to back this claim is Elizabeth's Australian passport, which shows her age and citizenship. The warrant states that Australian citizens over the age of 18 are eligible to vote, and the backing for this warrant comes from Australian laws and statutes. A rebuttal to this claim is that Elizabeth has given up her Australian citizenship. Available rebuttal data is that Elizabeth has a Japanese passport, Japanese law stating that Japanese citizens must give up citizenship of another after the age of 22, and statements from Elizabeth's friends and family that she has relinquished her Australian citizenship. In this case the rebuttal data severely weakens the claim that Elizabeth is eligible to vote in Australia.

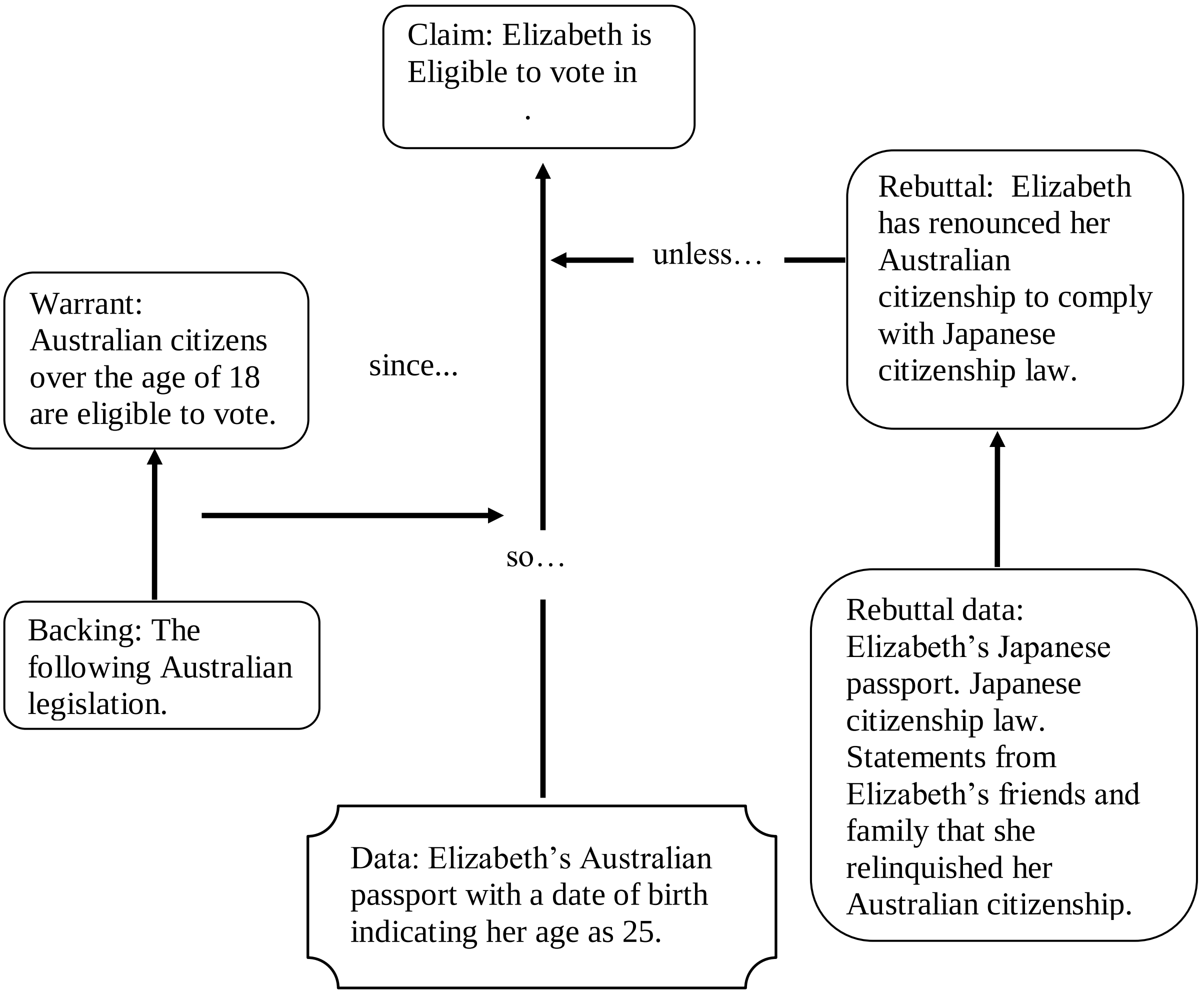

Claim: Elizabeth is Eligible to vote in .
Rebuttal: Elizabeth has renounced her Australian citizenship to comply with Japanese citizenship law.
unless…
Warrant: Australian citizens over the age of 18 are eligible to vote.
since...
so…
Backing: The following Australian legislation.
Rebuttal data: Elizabeth's Japanese passport. Japanese citizenship law. Statements from Elizabeth's friends and family that she relinquished her Australian citizenship.
Data: Elizabeth's Australian passport with a date of birth indicating her age as 25.

Figure 2.2. An example of Toulmin's informal argument structure

Toulmin's informal argument structure has been adopted by several theorists in the field of test validation in educational assessment, to become the basic building block of argument-based approaches to validity. For test validation, multiple inferences join together to form a "web of inferences" (Mislevy et al., 2003, p. 3), or "chain of inferences" (Kane, 1992, p. 530). This web or chain of inferences forms the structure of a test validity argument, with each of the multiple inferences in the larger validity argument being built using Toulmin's informal argument structure.

2.3.2 Kane’s multiple inferences

Kane (1992) argued that in practice for test validation multiple inferences link together to form a coherent conclusion. Later Kane, Crooks and Cohen (1999) elaborated this idea to show how an interpretive argument could be constructed for a performance assessment. In Kane’s description, three inferences are represented as bridges to be crossed on the way from an example of test-taker performance, to a judgment about test taker ability in the target domain. The target domain is the areas of performances about which one wishes to make inferences based on test scores. Each of the bridges is built on assumptions, which should be explicitly stated, and which require support from evidence.

The first bridge was first called ‘observation’ (Kane, 1992), and later ‘scoring’ (Kane, Crooks and Cohen, 1999, Kane, 2008) and ‘evaluation’ (Kane, 2001, 2002, 2004). This is an inference from an observation of test-taker performance to an observed score. This inference rests on two assumptions. The first assumption is that the criteria used to score the performance are suitable and have been properly applied, which means that, for example, (a) the answer keys for a multiple choice test have one clear answer for each item, (b) rubrics for constructed responses have clear and appropriate criteria, and (c) are applied consistently by the raters. The second assumption is that the test was conducted under suitable conditions to assess test-taker ability or skill. For example, that there were no problems with equipment, that test takers were motivated, and that no test-takers received any extra help. The backing for the first assumption can come from, for example, a sound answer key for a reading test, or a rubric for constructed response answers in speaking and writing, and intra- and inter-reliability indices. Instructions for test administration in the test specifications and reports of suitability of test administration locations can provide backing for the second assumption too.

The second inferential bridge in Kane’s framework is called “generalization”. It is usually not possible to include all possible tasks from the target domain on a test, so a narrower range of possible tasks from the target domain is selected for testing. This narrower range of tasks is called the universe of generalization. Kane, Crooks and Cohen (1999, p. 8) "refer to the subdomain for which it is plausible to consider the observed performances to be a random or representative sample as the universe of generalization and … refer to an individual's expected score over the universe of generalization as the individual's universe score for the assessment procedure."

The generalization inference allows interpretation to extend from interpretations about scores on a single test or the 'observed score' to interpretations about scores on other possible tests or tasks in the same universe of generalizations, or the 'universe score'. The universe score refers to the score from a universe of possible observations that would be obtained from other versions of the test focusing on the same narrow area of the target domain. One assumption supporting the generalization inferential bridge is that tasks on the test are representative of tasks in the 'universe of generalization'. Evidence for this assumption could come from an analysis of test content compared to the defined universe of generalization. Another assumption underlying the generalization inference is that the test-taker would score the same on alternate, parallel versions of the test or the test tasks. Evidence to support the generalization inference can come from reliability studies which show the consistency of scores across raters, tasks and occasions. Further backing for this inference can come from evidence showing that the test specifications are clear and specific, and that the test tasks in alternate forms of the test match the test specifications. In addition, statistical analyses providing evidence for equating of alternate test forms can provide backing for this inference, as can generalizability studies.

The third and final inferential bridge is an 'extrapolation' from the universe score to the target score. Kane, Crooks and Cohen define the target score as, a test taker's "expected score over all possible performances in the target domain" (1999, p 7). In other words, for the extrapolation inference, test scores from the narrower sample of tasks from the universe of generalization, are taken to indicate test-taker performance in the wider target domain. Evidence to support this inference can come from criterion related studies in which other measures of performance in the target domain are correlated to scores on the test to be validated. Support can also come from a coherent argument that the tasks in the narrower universe of generalization cover a large area of the target domain, as the more of the target domain that is covered by the universe generalization, the stronger the support for the extrapolation inference will be.

A diagram of Kane, Crooks and Cohen's three inferential bridges is provided in Figure 2.3.

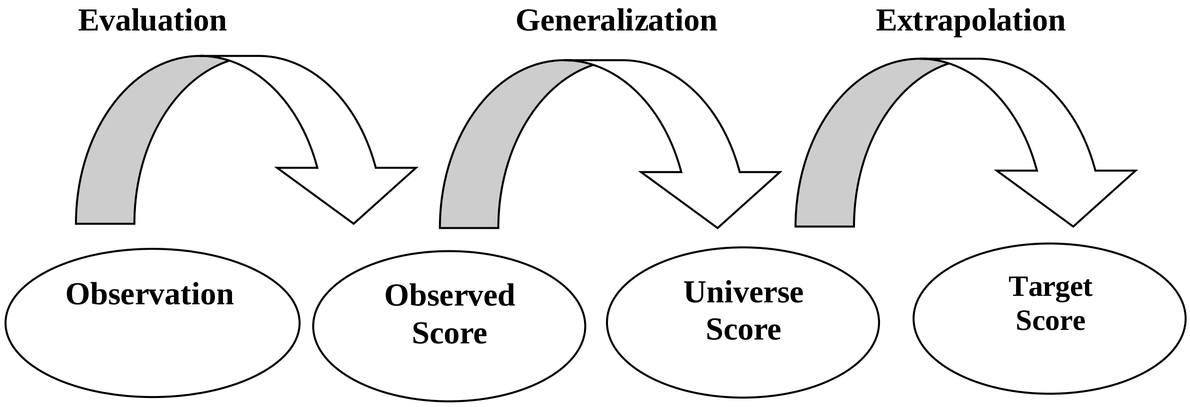


Figure 2.3. Kane, Crooks, and Cohen's (1999) inferential bridges in an argument-based framework

2.3.3 Advances in Kane's approach—interpretive and validity arguments and a further inferential bridge

In his later work, Kane (2001, 2002) clarified that his argument-based approach consists of two main arguments. The first is an interpretive argument and the second is a validity argument. The interpretive argument details the proposed interpretation and use of test scores, and the validity argument assesses the credibility of the assumptions and backing for the interpretive argument.

Subsequently, Kane (2004, 2006) adopted Toulmin’s informal argument structure as means to provide logical and clear support for each of the inferences in the interpretive argument. Each inference is supported by a warrant, which is a rule leading from data to a claim. In turn, each inference is supported by assumptions. Assumptions are the postulates upon which the warrant and ultimately the inference rests, and Kane makes it clear that these postulates should be explicitly stated along with their proposed supporting evidence. This supporting evidence for assumptions is known as backing in Toulmin’s informal argument structure and Kane’s argument-based validation framework. A useful analogy could be to consider each assumption as a building block forming the foundation of each inference in the interpretive argument, each assumption is in turn supported by building blocks which are backing as shown in Figure 2.4.

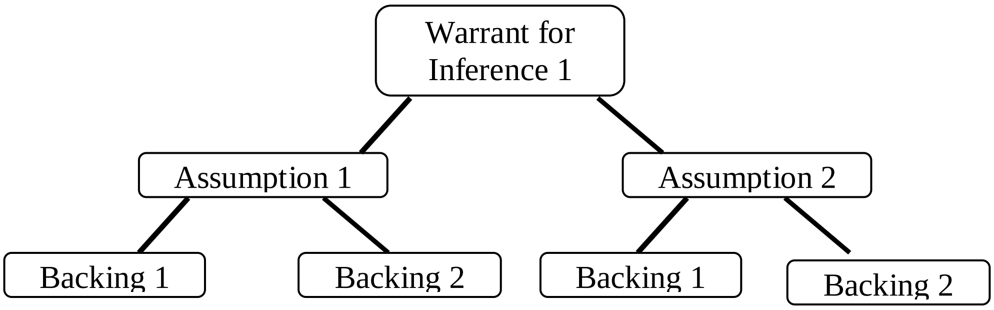


Figure 2.4. Argument structure for inferences in Kane's interpretive argument

Kane (2001, 2002, 2004, 2006) further extended his framework to include the consequences of test use for common decisions which are made on the basis of test scores including acceptance into courses of study, and awarding of certifications. Kane’s work has been very influential in the field of educational assessment, and has also become influential in the field of language assessment. In the following sections three major efforts by language testing scholars to adapt and apply Kane's framework are described and discussed.

2.4 Developments of argument-based approaches in language testing

The following sections outline the major contributions, Bachman (2005), and Bahman and Palmer (2010), and Chapelle et al. (2008) to bringing argument-based approaches to test validity to the field of language testing.

2.4.1 Bachman 2005—more fully integrating test uses and consequences into argument-based validation

In an important article in the journal *Language Assessment Quarterly* in 2005 Bachman argued that previous argument-based approaches to test validation, although providing a useful framework for examining assertions made based on inferences from test scores, neglected to include matters of test use and consequences. Bachman also observed that preceding work on test validation in the language testing field (Bachman & Palmer, 1996; Kunnan, 2003) mainly presented lists of questions or qualities to be addressed, without providing a means for assimilating these into a clear set of procedures for language testers to follow.

Bachman (2005) thus proposed that an overall 'Assessment Use Argument (AUA)' should be created to validate the uses of language tests. Bachman's AUA as described in his 2005 article consists of two parts an 'assessment utilization argument', and a 'validity argument', both of which are based on Toulmin's (1958) argument structure. The validity argument in Bachman's AUA links test scores to interpretations, while the assessment utilization argument is similar to Kane's (2002, p. 32) "decision based interpretation" which aims to make clear links between the score-based interpretations of the validity argument and test use decisions made based on these interpretations. Bachman's AUA, however placed more emphasis on the decision inference than in Kane's work, and it also provided a new and useful breakdown of types of warrants which are needed to support an assessment use argument. Figure 2.5 shows Bachman's structure of an AUA.

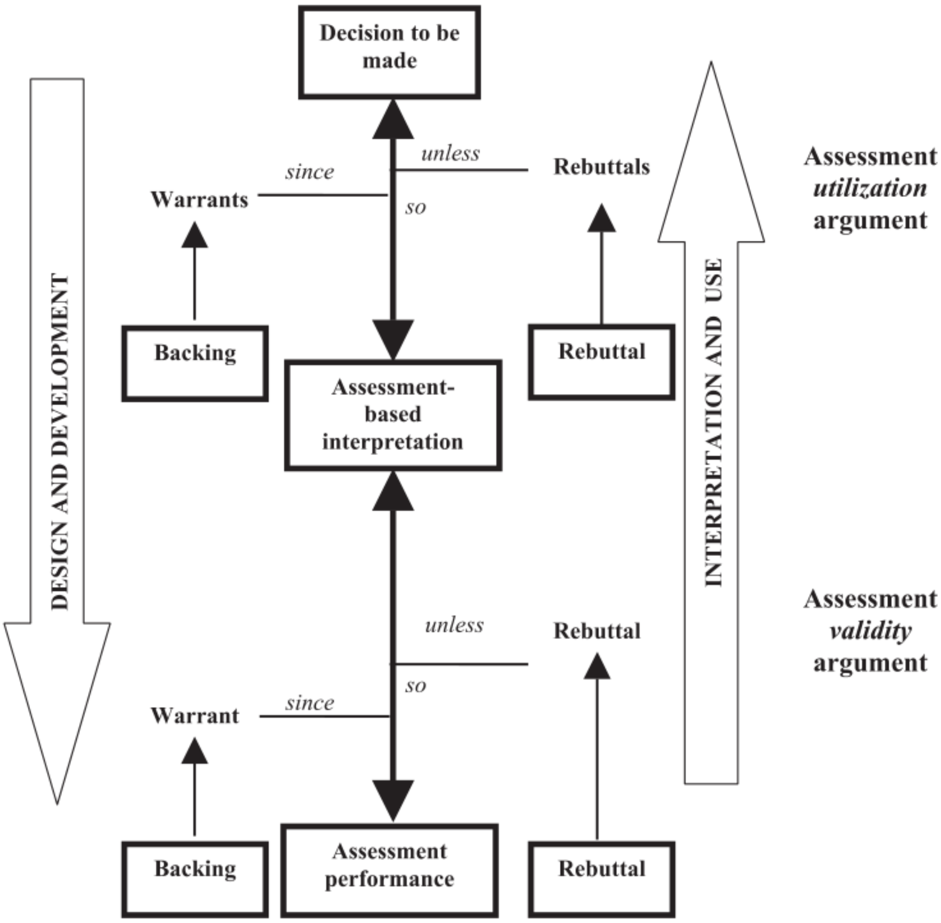


Figure 2.5. The structure of an Assessment Use Argument (AUA). Reprinted from "Building and supporting a case for test use", by L. F. Bachman, 2005, *Language Assessment Quarterly: An International Journal*, *2*(1), p. 25. Copyright 2005, Lawrence Erlbaum Associates, Inc. Reprinted with permission.

The four types of warrants which Bachman argues are needed to support an assessment use argument are 'relevance', 'utility', 'intended consequences' and 'sufficiency'. Relevance deals with how closely the test tasks represent tasks in the Target Language Use (TLU) domain. Utility focuses on how much the score-based interpretation increases the likelihood of making suitable decisions. Intended consequences relate to the beneficial outcomes the test is intended to bring about. Finally, sufficiency deals with the test providing sufficient information for the decisions to be made. Each of these warrants requires backing which can consist of various forms of evidence depending on the particular warrant.

Bachman (2005) argued that his AUA approach to language test validation was able to accommodate the expanded list of concerns raised about the uses and consequences of test scores in the years preceding the paper. These included test usefulness (Bachman & Palmer, 1996), ethics and validity (B. K. Lynch, 2001), fairness framework (Kunnan, 2003) and critical language testing (Shohamy, 2001). The ways in which Bachman argued that these previous frameworks which addressed the consequences of test use could be incorporated into an AUA are briefly summarized in the following three paragraphs.

Bachman argued that test usefulness as previously defined in Bachman and Palmer (1996) as consisting of six qualities of reliability, construct validity, authenticity, interactiveness, and impact (this list excludes practicality, which lies outside an AUA) could be included in an AUA by making warrants for each of the first five qualities part of an assessment validity argument, and a warrant or warrants for the impact quality as part of the assessment utilization argument. Bachman also argued that four of the aspects of B. K. Lynch's (2001) framework (ontological authenticity, cross-referential authenticity, impact/consequential validity and evolved power relations) could be incorporated into an AUA as warrants of the 'intended consequences' type. However, Bachman also claimed that the fairness category as defined in Lynch's framework was more an aspect of overall test design, development and use that as such it was not related to a particular test use, and should not form a part of an AUA.

Bachman further contended that Kunnan's five qualities of fairness, as outlined in his 2003 paper, of validity, absence of bias, access, administration, and social consequences could also be integrated into an AUA. The qualities of validity and administration in the 2003 version of Kunnan's test fairness framework can be covered by warrants supporting a validity argument in an AUA and the other three qualities of absence of bias, access and social consequences can be covered by warrants of the intended consequences type in an assessment utilization argument.

Finally, Bachman claimed that the key concerns of critical language testing (Shohamy, 2001) can also be addressed in an AUA. For example, the focus of critical language testing on the purposes for which tests are used can be addressed via warrants of the intended consequences type, and critical language testing's call for multiple types of assessment to be used for decisions about test takers rather than a single test, can be addressed through warrants of the sufficiency type in the AUA.

2.4.2 Chapelle, Enright & Jamieson 2008—a further two inferential bridges

A relatively recent, outstanding example of an application and adaptation of Kane's argument-based framework to language test validation which integrated Bachman's utilization argument is the validation argument for the new TOEFL test (Chapelle, Enright, & Jamieson, 2008). The authors of this study added an additional two inferences to those suggested by Kane (2001, 2002, 2004). These additional inferences are 'domain definition', and 'explanation'. In addition, Chapelle et al. included a final 'utilization' inference, which took its name from Bachman's AUA, and which links test scores to test uses.

Figure 2.6 on the next page shows the chain of inferences for the interpretive argument in Chapelle et al.'s (2008) TOEFL validation study.

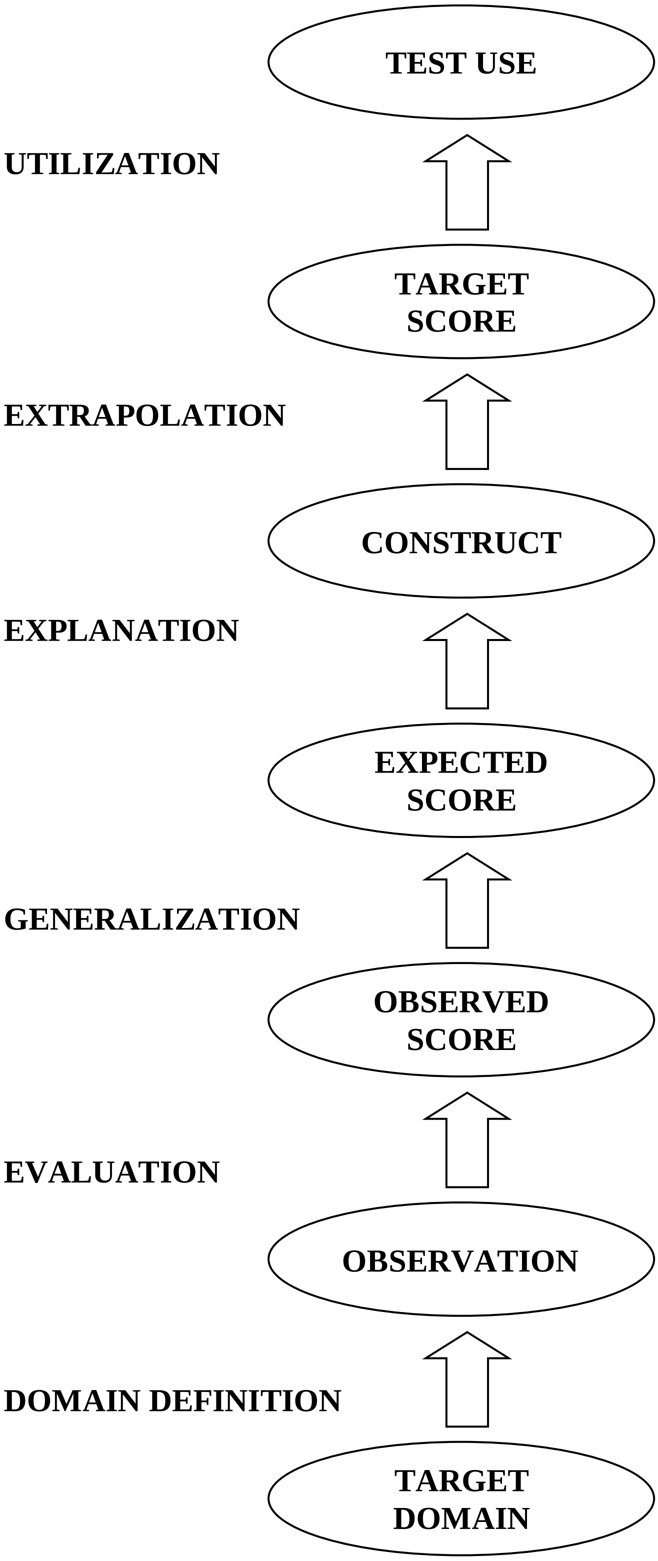


Figure 2.6. TOEFL validation interpretive argument. Adapted from *Building a Validity Argument for the Test of English as a Foreign Language* (p. 15), by C. A. Chapelle, M. K. Enright, & J. Jamieson, 2008, New York, NY: Routledge. Copyright 2008 by Routledge. Adapted with permission.

The domain definition inference links test tasks, which are defined in the domain definition, to real-world target domain tasks. The warrant supporting this inference states that the test tasks are representative of those language use situations for which the test seeks to measure language proficiency. Bachman and Palmer (2010) define a “Target Language Use (TLU) Domain” as “a specific setting outside of the test itself that requires the language user to use language use tasks” (p. 60). This corresponds to Kane’s (2013b) “target domain”, which “specifies the kinds of tasks and the ranges of contexts and conditions of observation associated with the observable attribute” (p. 22). For Chapelle et al.’s TOEFL validation study the domain was “language use in the English medium institutions of higher education” (Chapelle et al., 2008, p. 60). The warrant to supporting the domain definition inference is that the test tasks elicit “relevant knowledge, skills and abilities in situations representative of those in the target domain” (Chapelle et al., 2008, p. 14). Backing for this warrant can come from a thorough analysis of the target domain, and evidence that the test tasks are a representative sample of the target domain.

The explanation inference links observed test performance to a construct of language proficiency. For language testing the constructs will come from theories of second language acquisition and theories of second language competence. Theories of language proficiency provide definitions of what a test measures, and also the criteria that should distinguish learner ability at different levels of proficiency. Evidence to support the explanation inference can come from correlations with other measures of the same trait the test claims to measure (what used to be known as concurrent validity), factor analysis to confirm that the different theoretical language components the test claims to measure exist in the test data, analysis of task difficulty, and cognitive processing studies to confirm that the cognitive processes elicited by the test are actually those predicted by the theory (Weir, 2005a).

Finally, a decision-making or utilization inference links the target score to decisions made about test takers based on the test scores. This inference focuses on aspects of "consequential validity" as was suggested by Messick (1989) in his unitary approach to validity. The utilization inference assesses the extent to which the test results are effective for their intended uses, for example, whether the use of a test for making immigration decisions is appropriate or not. The importance of this inference is emphasised by both Bachman and Palmer (2010) and Kane (2013a, 2013b). Indeed, Chapelle et al. took the term 'utilization' for their interpretive argument from Bachman's (2005) work.

The utilization inference may also include aspects of test consequences, or consequential validity. As noted in section 2.3.2, consequential validity emphasizes the importance of test scoring not being biased against any particular social group (Xi, 2010), and that test scores should not have unfair negative consequences for some test takers (Bachman & Palmer, 2010). Consequential validity is also concerned with test washback. According to A. Green (2013), "Washback refers to the impact that a test has on the teaching and learning done in preparation for it" (p. 40). Cheng, Sun and Ma (2015) state that "When researchers use the term WASHBACK, they tend to ask questions about the influence of the introduction of a new test, or an existing test, on classroom teaching and learning" (p. 437). For example, the introduction of a new test may have positive or negative effects on classroom teaching and curriculum content. (Cheng, Sun, & Ma, 2015 give a useful overview of washback research in language assessment).

For the TOEFL validation study the warrant to support the utilization inference was that "estimates of the quality of performance in English-medium institutions of higher education from the TOEFL are useful for making decisions about admissions and appropriate curricula for the test takers" (Chapelle et al., 2008, p. 22). This inference is based on assumptions that test scores are easy to understand for all stakeholders, and that the tests have a positive influence on the way English is taught. Backing for the first assumption can come from demonstrating that materials are available to explain the test clearly to all stakeholders, while backing for the second assumptions underlying the utilization inference can come from washback studies.

2.4.3 Bachman and Palmer 2010—a comprehensive guide to building an AUA

In 2010, an updated version of Bachman and Palmer's previous influential work *Language Testing in Practice* (1996) was published, re-entitled *Language Assessment in Practice*. In this book, the authors present a thorough framework for creating or selecting language assessments, and for justifying their use. The major change from their 1996 book is a move to structuring test development around an updated version of Bachman's (2005) Assessment Use Argument (AUA). As in Bachman's 2005 paper, Toulmin's argument structure forms the backbone of the AUA, and the main focus of Bachman and Palmer's particular argument-based approach is on justifying assessment use, rather than on interpreting test scores. Similar to Kane's model Bachman and Palmer's model involves two iterative stages, which feedback into each other. The first stage is clearly describing the AUA, and the second stage is gathering evidence to support it, which Bachman and Palmer call *justification*. Bachman and Palmer's (2010) argument-based framework is illustrated in Figure 2.7.

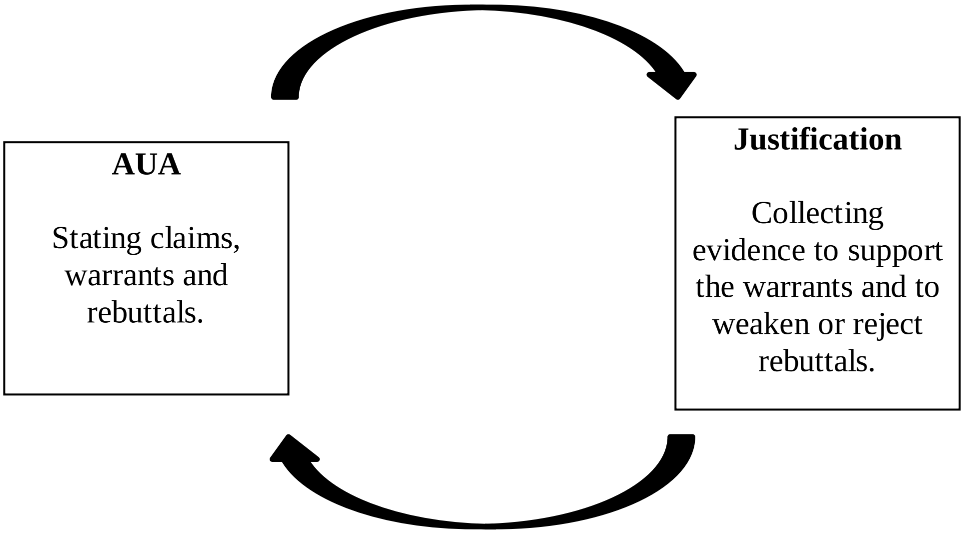

AUA
Stating claims, warrants and rebuttals.
Justification
Collecting evidence to support the warrants and to weaken or reject rebuttals.

Figure 2.7. The Structure of Bachman and Palmer's (2010) Assessment Use Argument

Bachman and Palmer detail four claims that any AUA should include. These are:

1. *That the consequences of the test are beneficial to stakeholders.*

   Stakeholders that can be affected by assessments include the test takers, teachers, educational programs, and institutions, and even society at large.

2. *That the decisions made based on test results take into account community values and are equitable for stakeholders affected.*

   Test scores should not show any systematic bias against a particular group. Test scores should not be contrary to local cultural norms.

3. *That interpretations about the ability to be assessed are:*

   a) *meaningful* with respect to a particular learning syllabus, an analysis of the abilities needed to perform a particular task in the TLU domain, a general theory of language ability, or any combination of these

   b) *impartial* to all groups of test takers

   c) *generalizable* to the TLU domain about which decisions are to be made

d) *relevant* to the decisions to be made

e) *sufficient* for the decisions to be made.

4. *That assessment records are consistent across different assessment tasks and aspects of the assessment procedure* (e.g. forms, occasions, raters).

Figure 2.8 from Bachman and Palmer illustrates the structure of their AUA.

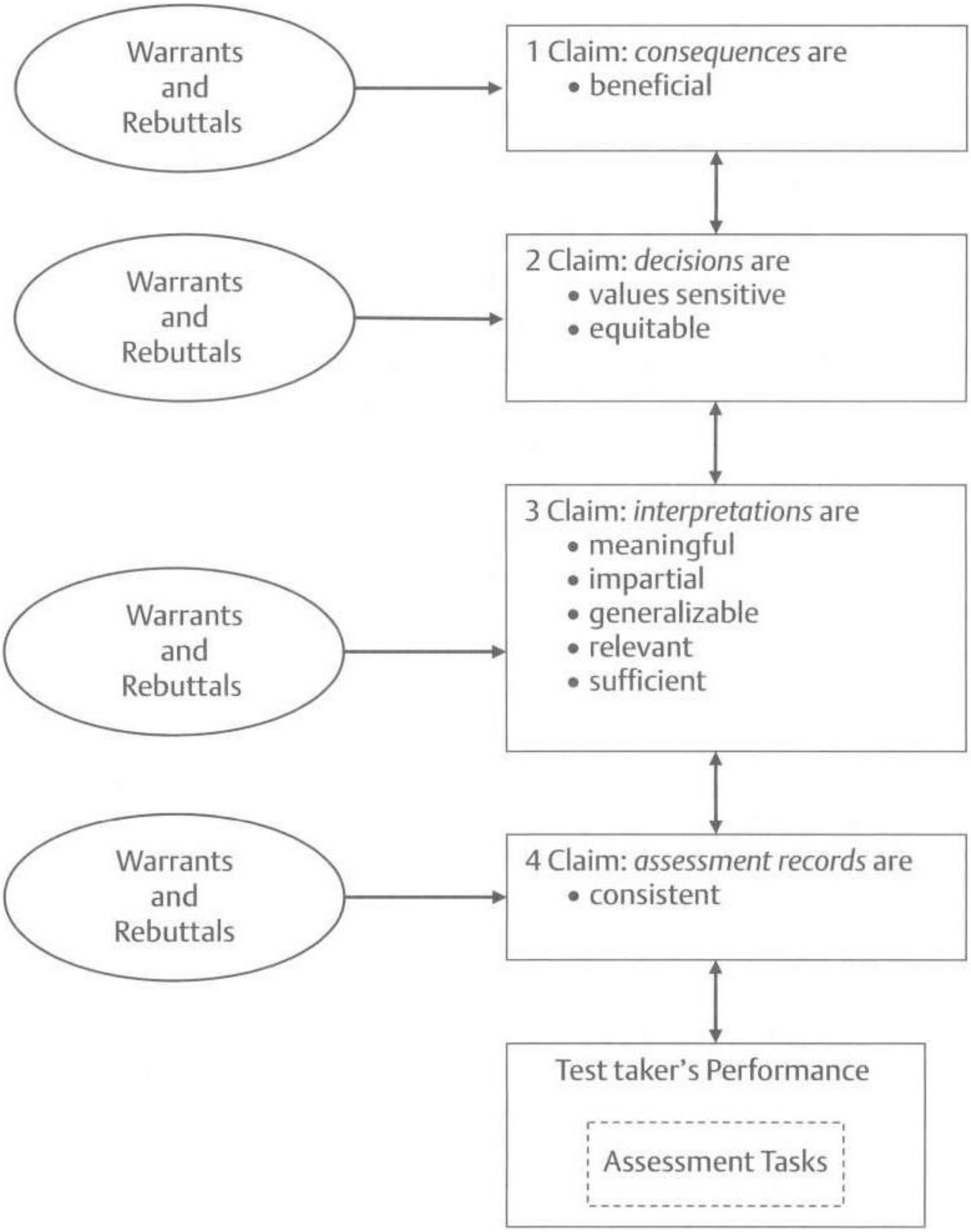


Figure 2.8 Claims and warrants in an Assessment Use Argument. Reprinted from *Language Assessment in Practice: Developing Language Assessments and Justifying their Use in the Real World* (p. 104), by L. F. Bachman, 2010, Oxford, England: Oxford University Press. Copyright 2010 by Oxford University Press. Reprinted with permission.

2.4.4 Kane's Interpretation/Use Argument (IUA)

In the most recent iteration of his argument–based approach to validity (2012, 2013a, 2013b), perhaps influenced by Bachman and Palmer's (2010) work, Kane places more emphasis on the importance of test use, while maintaining the centrality of the importance of test score interpretations. Kane had mentioned the importance of test use in his earlier work, however in his 2013 papers, to indicate a greater importance of test use in his theoretical framework, Kane changed the name of the first of the two major arguments in his model from 'interpretive argument' to an Interpretation/Use Argument (IUA). According to Kane (2013b) his preference is to give interpretations and uses of test scores "equal billing" (p. 2) in an IUA.

Kane divides test consequences into three categories:

1. Intended outcomes

These are decisions that are based on test scores, for example, certification, hiring or enrolment or placement in a course.

2. Adverse impact

This refers to negative consequences of test score uses for certain groups, for example minority groups.

3. Systemic effects

This refers to effects on an educational system, or what is commonly known as washback.

In his 2013 explanations Kane gives a good summary of the history of considering test consequences as part of validity and how the list of consequences to be included in a validity argument has grown over time, however, he failed to give clear examples of specific kinds of warrants for a decision inference or of the kinds of evidence that can be gathered to support warrants for the decision inference. Other authors such as Bachman and Palmer (2010) do a more thorough job of elaborating the kinds of warrants and support needed to support inferences about test use. Indeed, Kane's most recent conception of evaluating the uses of a test in a validity argument is far narrower than Bachman and Palmer, because Kane mainly focuses on test takers for evaluating negative consequences (Kane, 2013b). Bachman and Palmer (2010) on the other hand clearly argue that all stakeholders should be considered in an Assessment Use Argument, and they suggest a warrant that "The consequences of the decisions will be beneficial for *each group* [italics in original] of stakeholders" (p. 186). In short, while Kane's 2013 IUA goes further toward explaining the role of test uses in his argument-based approach than his previous versions, more elaboration and examples are needed to help practitioners to plan the decision inference part of an IUA, and to choose the types of evidence needed to support warrants for a decision inference.

2.4.4.1 Development and appraisal stages of test validation

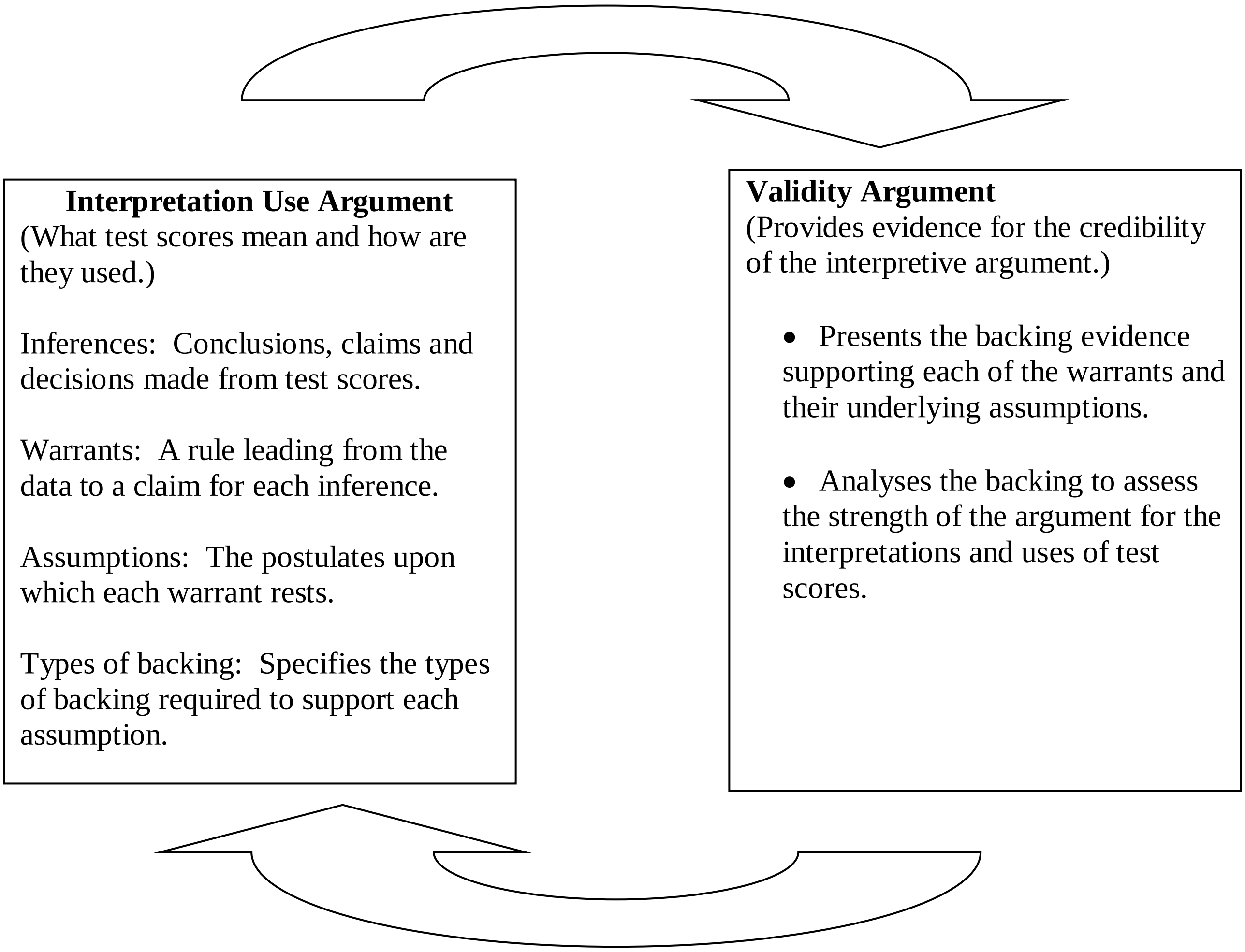
Kane (2013b) explains that although the distinction between the IUA and validity arguments is a useful way to conceptualize the test validation process, in practice the two are interwoven and are not strictly sequential. Kane divides validation into two stages. The first stage is the development stage in which usually happens at the beginning of developing or adopting a test as stated below.

> the goal is to develop (or adopt) a testing program and to develop an IUA that represents the proposed interpretation and use of the scores and is consistent with the characteristics of the test. If any assumptions are found to be untenable, the test, the IUA, or both can be modified to resolve the discrepancies. This iterative process of development and revision continues until the fit between the test and IUA is considered acceptable (Kane, 2013b, pp. 16–17).

After the development stage comes the appraisal stage. The appraisal stage occurs once development of the test and the IUA are complete. In the appraisal stage the IUA should be critically analysed and challenged preferably by a third party, who was not involved in the test development.

Kane’s development stage of test validation is particularly relevant to this thesis, which focuses on the development and simultaneous early validation of the BETs.

Figure 2.9 summarizes the development stage of Kane’s argument-based approach to test validity. The boxes summarize the IUA and the validity argument and the arrows represent a feedback loop in which problems revealed by analysis of the validity argument can lead to modifications to the test and the IUA, which in turn entail changes to the validity argument.



Interpretation Use Argument
(What test scores mean and how are they used.)

Inferences: Conclusions, claims and decisions made from test scores.

Warrants: A rule leading from the data to a claim for each inference.

Assumptions: The postulates upon which each warrant rests.

Types of backing: Specifies the types of backing required to support each assumption.

Validity Argument
(Provides evidence for the credibility of the interpretive argument.)

- Presents the backing evidence supporting each of the warrants and their underlying assumptions.
- Analyses the backing to assess the strength of the argument for the interpretations and uses of test scores.

Figure 2.9 The structure of the development stage of Kane’s (2013a, 2013b) *argument-based approach to test validation*

2.5 The uptake of argument-based approaches to validation in language testing

In recent years, thanks to the introduction of argument-based approaches to language testing through the work of Bachman (2005), Bachman and Palmer (2010), and Chapelle et al. (2008), argument-based approaches to validation have become more popular in language testing. This section briefly summarizes this trend by outlining Chapelle and Voss's (2014) historical review of validation approaches in language testing. Some more recent exemplary research in language testing employing an argument-based approach to validation for tests of second/foreign language reading and listening proficiency are summarized in Chapter 3.

Chapelle and Voss (2014) conducted a review of all validation research in the journals *Language Testing* and *Language Assessment Quarterly* from 1984–2011 to get an overview of how approaches to test validation had changed over this period. They searched these two key journals' databases through their webpages for terms related to test validation, and then examined each study to confirm that it was an empirical validation study related to interpretation and use. This resulted in a list of 123 studies from 1984–2011. Each paper was then classified into one of four approaches to test validation:

1) *One question three validities* – which indicates studies that used the approach to validation dominant before the 1980s consisting of criterion, content or construct validity as are outlined in sections 2.1.1 and 2.1.2 of this thesis.
2) *Evidence gathering* – refers to studies which used Messick's (1989) framework to gather evidence to support construct validity which is summarized in section 2.1.3 of this thesis.
3) *Test usefulness* – refers to studies utilizing Bachman and Palmer's (1996) validation framework from in their important book *Language Testing in Practice*, which is briefly outlined in section 2.2.1.

4) *Argument-based approach* – refers to studies using the approaches outlined in sections 2.3 and 2.4 of this thesis.

The results of Chapelle and Voss survey of studies in the two key language testing journals are reproduced in Table 2.2.

Table 2.2. *Types of Validation Studies Appearing in the Journals Language Testing and Language Assessment*

Time Period	*No. of articles*	*Not explicit %(n)*	*One question %(n)*	*Gathering evidence %(n)*	*Test usefulness %(n)*	*Argument-based %(n)*
1984–1990	20	70.00% (14)	30.00% (6)	0.00% (0)	0.00% (0)	0.00% (0)
1991–1995	18	61.11% (11)	11.11% (2)	27.78% (5)	0.00% (0)	0.00% (0)
1996–2000	18	88.89% (16)	0.00% (0)	11.11% (2)	0.00% (0)	0.00% (0)
2001–2005	22	77.27% (17)	0.00% (0)	13.64% (3)	9.09% (2)	0.00% (0)
2006–2011	45	66.67% (30)	0.00% (0)	13.33% (6)	8.89% (4)	11.11% (5)
Total	123	88	8	16	6	5

Note. Adapted from "Evaluation of Language Tests Through Validation Research" by C.A. Chapelle and E. Voss. in A. Kunnan (Ed.), *The Companion to Language Assessment, 3*(2) 2014, Hoboken, NJ: Wiley-Blackwell. Copyright 2014 by Wiley-Blackwell. Adapted with permission.

These results show that language test validation studies increased markedly in the period from 2006–2011, slightly more than doubling from the previous five-year period from 2001–2005. They also show that argument-based approaches to test validation made their first appearance in these journals at this time, although the overall number was still relatively small.

2.6 The Argument-based validation framework used for this study

This study utilizes a combination of Kane's (2013a, 2013b) IUA, Chapelle et al.'s (2008) expanded chain of inferences used for their TOEFL validation, and Bachman and Palmer's (2010) AUA. The aspects utilized from each of these frameworks, and the reasons for their use in this study are explained in the following paragraphs.

The overall theoretical framework for the argument-based approach employed in this study uses Kane's (2013a, 2013b) model, which consists of first articulating an IUA describing the claims to be made about the interpretations and uses of BET scores, and then building a validity argument which gathers and analyses evidence to evaluate the claims of the IUA. This study also focuses on the development stage of test validation as defined by Kane (2013b) and outlined in section 2.4.4.1. The reasons for this choice are that Kane’s IUA is the most up to date iteration of his influential work, and the IUA also addresses the flaw in previous iterations of Kane’s approach in which not enough attention was paid to the consequences of test use. In my opinion Kane’s IUA also seems to strike the right balance between emphasis on score-based interpretations and inferences, and justifications of test usage, by placing equal weight on each of these elements. I would argue, as does Kane (2013b), that Bachman and Palmer (2010) perhaps place too much emphasis on test usage in their framework at the expense of the essential foundation of building argument for score-based interpretations.

The structure of the particular IUA for this project draws heavily on the excellent example provided by Chapelle et al.’s (2008) validity argument for the TOEFL. One reason for structuring the IUA for this study around the chain of inferences detailed in Chapelle et al.’s study is that their work provides a transparent, coherent, and logically structured argument. In addition, the chain of inferences in their inferential argument expands on Kane’s work by articulating an extra two inferences of ‘domain definition’ and ‘explanation’. These two additional inferences are relevant both to the TOEFL iBT as a norm-referenced test, and also to the BETs as criterion-referenced tests. In addition, Chapelle et al.’s model provides clear examples of the types of evidence to be gathered to support an IUA for validating a standardised language proficiency test like the BETs.

Bachman and Palmer’s (2010) AUA, whilst detailed, comprehensive and insightful does not appear to be the ideal choice for use as the overarching framework for this validation study. One reason for this is that most of the final two claims of Bachman and Palmer’s AUA are adequately covered in Kane’s (2013a, 2013b) IUA and Chapelle et al.’s (2008) expanded chain of inferences. The warrant that interpretations about the test taker ability to be assessed are to be ‘meaningful’ is covered by Chapelle et al.’s explanation inference. The warrant about interpretations being ‘impartial’ is covered by the evaluation inference, and the warrant about interpretation being ‘generalizable’ is also covered by the domain and extrapolation inferences. Finally, the fourth claim in Bachman and Palmer’s AUA, that assessment records are consistent is equivalent to the evaluation and generalization inferences in Chapelle et al.’s interpretive argument and these same inferences in Kane’s IUA.

Bachman and Palmer’s ‘relevancy’ and ‘sufficiency’ warrants would seem to be, however, very useful conceptually for this study. Therefore, they are incorporated in the BET IUA utilization inference section, in assumptions assessing the efficacy of BET score-based course placement and class streaming decisions, and the use of the BETs as achievement tests. Relevancy and sufficiency are included only in the BET utilization inference, because in Bachman and Palmer’s AUA these two warrants are used to support decisions made based on test scores.

Another reason for not using Bachman and Palmer’s AUA entirely is that the second major claim in their chain of inferences, that decisions are values sensitive and equitable, while being an important area to consider for all tests, does not seem to be a major concern in the context of the BETs. One reason for this is that Japanese students are quite used to taking standardised tests, so using such tests as part of course grading, and to place students and measure their progress, aligns with the local testing culture. Indeed, Japan has a strong testing culture in which students are used to being graded and selected based on large-scale exams, with competitive entrance exams for private schools beginning right from kindergarten and continuing through to university (Coleman, 1996). In addition, Bachman and Palmer’s claim about tests being equitable, which aims to ensure that a test is not biased against any particular group of test takers would be very important in contexts in which the same test is taken by people from a variety of cultural and socioeconomic backgrounds, however, the BETs are taken by a very homogeneous group of students all of the same culture, sex, general age, and the same general socio-economic status. For the 2015–17 BETs which are covered by the span of this study, to the best of my knowledge, only Japanese students took tests as there were no international students enrolled in BECC courses. Therefore, in this context equitability as defined by Bachman and Palmer would not appear to be a major concern.

However, the great strength of Bachman and Palmer's AUA is the way in which it demands clear articulation of intended consequences of test use, and a thorough examination of actual test consequences. Bachman and Palmer's fourth claim in their AUA about a test being beneficial to all stakeholders seems particularly relevant to the context of the BECC at HBWU, because of the precarious situation of the university in an environment in which increasing competition for enrolments in a shrinking pool of students creates pressure on the BECC as an institution to demonstrate the effectiveness of its language programs to a range of stakeholders including the university administration, the university public relations department, students, and parents of students. In addition, one of the most important stakeholder groups is the BECC teachers, because if they are not convinced of the utility of the BETs the washback on the GE curriculum would be under question. Thus, the IUA for this study incorporates some of the warrants recommended by Bachman and Palmer for their 'consequences are beneficial' claim.

In summary, this study uses the latest version of Kane's (2013a, 2013b) argument-based validation framework to build an Interpretation Use Argument (IUA) which is structured around Chapelle et al.'s (2008) chain of inferences for their TOEFL validity study. The BET IUA also draws on selected aspects of Bachman and Palmer's (2010) AUA for a key warrant and assumptions.

The argument-based framework for the BETs consists of two major arguments an Interpretation/Use Argument (IUA), which outlines the major inferences in an argument for the proposed interpretations and uses of test scores, and a validity argument which analyses the evidence gathered in support of the IUA, to critically assess each of the warrants and backing for each inference in the IUA. Equal emphasis in this study is, therefore, placed on the final ‘utilization’ inference as is placed on the sum of the preceding five inferences of the BET IUA. This balance is what is recommended in Kane’s latest iteration of his argument-based approach, and it is also congruent with the importance placed on test usage in Bachman and Palmer’s (2010) Assessment Use Argument (AUA). In addition, it reflects the fact that the intended uses of the BETs examined in this study, for placement and streaming, measuring student progress in reading and listening ability, and the intended consequence of having positive impact on stakeholders, are central aspects of test validity to be investigated. For the final utilization inference, the warrant is modelled on that recommended by Bachman and Palmer for their ‘consequences are beneficial’ claim and the ‘relevancy’ and ‘sufficiency’ aspects of their third claim about interpretations based on test scores is incorporated into two assumptions beneath the utilization inference warrant.

Table 2.3 below outlines the major components of Kane’s (2013a, 2013b), Chapelle et al.’s (2008), and Bachman and Palmer’s (2010) argument-based validation frameworks. Where different terms for the same concepts are used in each framework the term used in this study is noted. More details are given for Bachman and Palmer’s framework, because the warrants and claims in their AUA do not fit neatly into Kane’s framework and Chapelle et al.’s adaption of it.

Table 2.3. *Aspects of Argument-based Frameworks Used in this Study*

Kane's Interpretation/Use Argument (2013a, 2013b)	**Chapelle Enright and Jamieson's TOEFL iBT Interpretive Argument (2008)**	**Bachman & Palmer's AUA (2010)**
	Domain definition inference	**Claim 3: interpretations are** ➢ **generalizable** ➢ **Warrant 1: The characteristics of the assessment tasks (i.e., setting, input, expected response, types of external interactions) correspond closely to those of TLU tasks.**
Scoring inference (labelled as the **Evaluation Inference** in this study and also in some of Kane's earlier work**)**	**Evaluation Inference**	**Claim 4: assessment records are consistent.** ➢ **Warrant 1: Administrative procedures are followed consistently across different occasions, and for all test taker groups.** ➢ **Warrant 2: Procedures for producing the assessment records are well specified and are adhered to.** **Claim 3: interpretations are** • **meaningful** ➢ **Warrant 2: The assessment task specifications clearly specify the conditions under which we will observe or elicit performance from which we can make inferences about the construct we intend to assess.** ➢ **Warrant 3: The procedures for administering the assessment enable test takers to perform at their highest level on the ability to be assessed.** *Claim 3: interpretations are • impartial
Generalization inference	**Generalization Inference**	**Claim 4: assessment records are consistent.** ➢ **Warrant 5: Scores on different tasks in the assessment are internally consistent (internal consistency reliability).** ➢ **Warrant 8: Scores from different forms of the test are consistent (equivalence, or equivalent forms reliability).** ➢ **Warrant 9: Scores from different administrations of the test are consistent (stability, or test-retest reliability).** ➢ **Warrant 10: Assessment records are of comparable consistency across different groups of test takers.**
***Theory-based Inferences**	**Explanation Inference**	**Claim 3: interpretations are** • **meaningful** ➢ **Warrant 4: the procedures for producing an assessment record focus on those aspects of the performance that are relevant to the construct we intend to assess.** ➢ **Warrant 5: Assessment tasks engage the ability defined in the construct definition.**

		➢ **Warrant 6: Assessment records can be interpreted as indicators of the ability to be assessed.** **Claim 3: interpretations are** ➢ **Meaningful** ➢ **Warrant 1: The definition of the construct is based on a frame of reference such as a course, syllabus, a needs analysis or a current research and/or theory, and clearly distinguishes the construct from other, related constructs.**
Extrapolation inference	**Extrapolation Inference**	**Claim 3: interpretations are generalizable** ➢ **Warrant 2: The criteria and procedures for recording the responses to the assessment tasks correspond closely to those that are typically used by language users in assessing performance in TLU tasks.**
Decision inference (labelled as the Utilization Inference in this study)	**Utilization Inference**	**Claim 1: consequences of the test are beneficial** **(This is the warrant for the BET IUA utilization inference)** *Claim 2: decisions are • values sensitive • equitable **Claim 3: interpretations are** • **relevant** • **sufficient** **Warrant 7: The test developer communicates the definition of the construct to be assessed in terms that are clearly understandable to all stakeholders. (These are incorporated in two of the assumptions underlying the BET Utilization Inference.)**

Note. The parts of each framework not used in this study are indicated by plain text and an asterisk.

CHAPTER THREE

Literature Review: Empirical Investigations Relevant to the Current Study

3.0 Introduction

The first section of this chapter reviews exemplary recent studies which used an argument-based framework for validation of large-scale, commercial language tests used for placement and/or certification. Due to limitations of space and relevancy, only studies of tests which include a listening and/or reading component are reviewed. In the second section of this chapter, validation studies which used an argument-based approach to validation of in-house placement tests are reviewed. The two sections in this chapter provide additional background on how other researchers in other contexts have validated their target language tests.

3.1 Exemplary argument-based validation studies of commercial language proficiency/placement tests with a focus on listening and reading

Due to the incentives involved, most detailed validation research has been conducted on large, high-stakes commercial language tests. These incentives are firstly the importance of validating the claims made based on these tests to ensure fair and equitable outcomes for all stakeholders, and secondly the availability of research grants provided by large testing organizations to validate their tests. For example, research grants are provided on a competitive basis for validation studies by ETS the provider of the Test of English as a Foreign Language (TOEFL), for IELTS by joint funding from the British Council, Cambridge English Language Assessment and IDP: IELTS Australia, by Pearson for their language proficiency tests, and for the Cambridge suite of exams by Cambridge English Language Assessment.

The first study to be reviewed in this section is Chapelle, Enright and Jamieson’s (2008) ground-breaking work, which details an interpretive argument and a validity argument for the new Internet-based version of the TOEFL, known as the TOEFL iBT. Before the TOEFL iBT was launched, many studies were done over a ten-year period to develop and evaluate it, and in their book titled *Building a Validity Argument for the Test of English as a Foreign Language* the authors organize and synthesize the evidence from these studies into a coherent validity argument to support an interpretive argument, which builds on the previous work of Kane (e.g., 2004, 2008) and others.

The structure of the TOEFL iBT interpretive argument has already been outlined in section 2.4.2, so in this section a brief summary of the contributions of this this book to the field of validation, and areas where TOEFL iBT interpretive argument validity argument could be improved are be given. The strengths of this book are a) that it focuses on what Kane (2013b) refers to as the development stage of test validation, b) that it exemplifies argument-based test validation through a clear chain of inferences with supporting evidence, and c) that it provides a practical path for test validation, which avoids focusing on controversial and incomplete theoretical constructs. Some possible weaknesses of the TOEFL validity argument presented in the book are a) that counter arguments or rebuttals are not presented, and b) that due to the timing of the volume the utilization inference had insufficient backing.

The first reason that Chapelle et al.'s study is an exemplar of an argument-based approach to validation is that, as Barkaoui (2009) points out, it focuses on the development stage of validation, in which the test and the Interpretation/Use Argument are developed and modified until satisfactory versions of both are produced. The vast majority of previous validation studies in the literature focus on the validation of tests that are already completed, so Chapelle et al.'s detailed account gives practitioners valuable insight into the process of test validation intertwined with the test development process.

Secondly, Chapelle et al.'s volume is an excellent example of the implementation of a chain of inferences within Kane's argument-based framework. It gives practitioners actual, clear examples of the kinds of evidence that can be used to support the assumptions for each inference in an interpretive argument. The current study has also used the same six inferences for the BET IUA as were used in Chapelle et al.'s study.

A final reason for the importance of Chapelle et al.'s validation of the TOEFL iBT is that it provides a way for practitioners to frame validity as an evidence-based argument to support the interpretations and uses of test scores, rather than placing the construct at the centre of validation (Chapelle, 2010). Thus, Chapelle et al.'s example shows a practical way to frame validation in terms of the real-world uses of a test, and it downplays the importance of clearly defining complex constructs such as language proficiency, which may lack an agreed-upon definition. In this approach, the construct forms one part of a chain of inferences in a validity argument, rather than taking centre stage.

One criticism that has been levelled at the validity argument detailed in Chapelle et al. is that it does not include rebuttals or counter claims (Barkaoui, 2009). However, Chapelle (2008) argues that this is because of the early stage of the TOEFL iBT, which focused on the development stage of validation. She states that "Such an argument might ultimately include rebuttals, which would weaken the strength of the inferences, but at this stage of design validity the emphasis is on the warrants supporting the inferences and their backing" (p. 321).

However, it is possible that there may be some confirmatory bias in the TOEFL iBT validity argument, so more research will need to be done by outside researchers focusing on possible evidence for rebuttals to the TOEFL iBT interpretive argument.

Another weakness of the TOEFL iBT validity argument as presented in the Chapelle et al. (2008) volume is that the utilization inference was not well supported. This is because Chapelle et al.'s study focused on the development stage of test validation. Therefore, all the evidence presented was gathered before the test was used commercially. This meant that the assumptions underlying the warrant for the utilization inference that "estimates of the quality of performance in English-medium institutions of higher education are useful for making decisions about admissions and appropriate curriculum for test takers" (p. 344) lacked sufficient supporting evidence. These assumptions included claims that all stakeholders are able to interpret test scores, that the test has positive backwash on how English is taught. N. Chen (2010) also points out that issues of test fairness, such as differential item functioning for various test taker groups had not yet been conducted for the TOEFL iBT, and that this is important for future validation research.

The next exemplary argument-based validation study of a major commercial test to be reviewed in this section is Aryadoust's (2013) validation study of the listening section of the IELTS test. In this study Aryadoust applied an interpretive argument based on Chapelle et al.'s (2008) example. He employed a chain of inferences which included all those used in Chapelle et al.'s study except for the domain inference. Aryadoust's study focused on addressing five gaps in the IELTS research, which informed his four research questions. His research questions were:

1) What listening sub-skills does the IELTS listening test assess? Is there evidence indicating that test items are tainted by construct-irrelevant factors?
2) Does the test method (in particular the item format) affect test performance?
3) Is the IELTS listening construct represented similarly across different gender, nationality, and other subgroups?
4) What are the construct-relevant factors that determine test item difficulty in the IELTS listening test? (Aryadoust, 2013, pp. 13–15)

The evidence to answer each of the research questions for this study also made up essential components of supporting evidence for the evaluation, generalization, and extrapolation inference warrant of his IELTS validity argument. Aryadoust found both support and rebuttal evidence for some of the inferences in his IELTS validity argument. The supporting and rebuttal evidence for each of the inferences is briefly summarized in the following paragraphs.

For the evaluation inference, Aryadoust presented backing that the test administrations of IELTS are strictly standardised and that both academic and non-academic texts are used. As rebuttal evidence, he found that some correct answers were not recognized by the rubric for the IELTS official sample test that he examined, and that the items did not cover a full range of academic listening skills. Aryadoust also argued that IELTS' claim to be able to predict both academic and general listening ability was not well supported.

For the generalization inference, Aryadoust gave backing evidence that IELTS reported high reliability for their tests and that he found that "As testified by Rasch person reliability analysis and low measurement error test items did stratify test takers into several ability levels" (p. 217).

For the explanation inference, Aryadoust presented rebuttal evidence from factor analysis, which he claimed showed that the construct of academic listening was underrepresented on the IELTS test he examined. He also gave rebuttal evidence from factor analysis which showed that IELTS test items in the test that he reviewed were contaminated by construct irrelevant variance from test method effects and students' backgrounds.

Finally, Aryadoust offered correlational evidence to support the extrapolation inference. He found a moderate correlation (r = .454; r^2 = .201) between the IELTS listening section sample test he examined and "an academic listening test developed by Educational Testing Service's (ETS) researchers" (p. 10), which he argues "somewhat supports" (p. 206) the extrapolation inference. In addition, a self-assessment questionnaire of English academic listening ability designed by the researcher, which elicited abilities on six academic listening subskills showed small-medium correlations (r = .144–.322) with the IELTs listening section, however these correlations were statistically significant for only three of the six academic subskills on the self-assessment survey, which Aryadoust claims may be further evidence of construct underrepresentation in the IELTS listening section, thus weakening the argument for the extrapolation inference.

Aryadoust's study is a significant contribution to the body of literature using an argument-based approach to language testing for several reasons. Firstly, Aryadoust's study serves as an excellent example of the types of backing that can be used for generalization, explanation and extrapolation inferences of a validity argument for a second or foreign language listening test. Secondly, his validation study of the IELTS listening test showed how Rasch analysis can be used to support the inferences in an argument-based approach. Thirdly, his research showed how rebuttal evidence can be organized in an argument-based validation framework, to suggest improvements to a test.

Some limitations of Aryadoust's study are firstly that the analysis of test items for his research questions 1 and 2 relied on three human raters, and as noted previously in section 2.1.1 above, human rating of test items has been criticized for a lack of reliability. Secondly, the utilization inference only cites anecdotal evidence as rebuttal evidence (p. 232), which is a major gap, as this is widely regarded as the most important inference in modern argument-based approaches (Bachman, 2010; Kane, 2013a, 2013b). Finally, Aryadoust's study was based on a single official sample IELTS test, so further analysis of more IELTS listening section samples is needed to confirm the results.

The final exemplary argument-based study to be reviewed in this section is Kumazawa, Shizuka, Mochizuki and Mizumoto's (2016) validation study of the Visualizing English Language Competency Test (VELC$^{®}$). This study is important to the current study, because it uses an argument-based to approach to validation for a placement test, which like the BETs, is designed specifically for the Japanese context. In Kumazawa et al.'s study the authors provide backing for four inferences in a validity argument. These inferences are scoring, generalization, extrapolation and decision. Firstly, the backing provided for each of these inferences in the VELC interpretive argument will be presented, followed by a brief analysis of the strengths and weakness of the VELC validity argument.

For the scoring inference Kumazawa et al. present backing for the assumption that “The VELC Test items work on test-takers to make placement decisions” (p. 14) in the form of a high Rasch item separation index of 31.07, and the fact that only 19 out of 120 (around 16%) items on the test had item discrimination values below .2.

For the generalization inference, they provide support for the assumption that “The test score is generalizable to other observations so as to reduce the measurement error” (p. 14) by presenting Cronbach’s reliability coefficients for the whole test of .93, and multivariate D study reliability of .89 for the whole test, and also a small standard error of measurement of 4.5 for a test with a total score of 120 points.

For the extrapolation inference the authors give backing for the assumption that “The test score indicates what test-takers can do with their English proficiency” (p. 14) by presenting the results of confirmatory factor analysis indicating three latent variables of vocabulary, listening and reading, along with correlational analysis with the TOEIC as a criterion. They also present the results of a self-assessment survey using can do statements given to 550 students in conjunction with the VELC test. Scores on the self-assessment survey were then used to indicate what students can do by comparing the participants’ self-assessment survey results to their VELC scores. In addition, the authors have reported elsewhere a coefficient of determination (R^2) of VELC scores with TOEIC scores of .68 (Shizuka & Mochizuki, 2014).

Finally, backing for the assumption underlying the decision inference that “The test score is appropriate for making placement decisions and useful for test-takers’ further learning” (p. 14) is provided in the form of Rasch person separation measures for the whole test in the range of 2.55–3.66, which the authors argue is evidence that the test is able to divide test takers into three levels, which in turn supports the decision inference, because most Japanese universities divide their English courses into three levels. The authors also present the score reports given to students as backing for the decision inference, but no warrant or assumption about score reporting is stated in their interpretive argument.

The strengths of Kumazawa et al.'s (2016) validity argument as presented in this paper are that they provide solid statistical evidence of the tests reliability and item properties. However, a major weakness is that the validity argument still lacks sufficient backing. The authors did not provide statistical correlations between their can do self-assessment survey and VELC test results for the extrapolation inference. Nor did they give details on how they linked self-assessment survey results to VELC scores. Furthermore, no detailed information is given about how the alternative VELC forms were equated to back warrants for the generalization inference. In addition, the backing for the VELC interpretive argument decision inference did not include any backing from the various VELC stake-holder groups that the test is fulfilling its intended uses. Overall, the VELC interpretive argument needs to be expressed in more detail, with more explicit statements for the assumptions underlying each warrant, and also more backing for the warrants and assumptions for each inference.

In conclusion, Chapelle et al.'s (2008) validation study of the TOEFL iBT remains the best example of a language test validation study using an argument-based approach, which includes a focus on reading and listening. This is because of the breadth, and detail of the interpretive argument, and the depth of the backing for the validity argument. Although argument-based validation studies of language tests are becoming more popular, most studies focus on presenting partial backing for a limited selection of inferences in the interpretive argument. This is likely due to limitations of time and resources for researchers, and to space limitations for journal articles. However, exemplary, innovative argument-based validation studies are beginning to appear two of which were outlined and examined above. Aryadoust's

(2013) argument-based validation study of the IELTS listening section provides excellent examples of the types of backing and rebuttal evidence that can be marshalled for a validity argument for a second/foreign language listening test, and Kumazawa et al.'s (2016) validation study of the VELC test provides a good example of the types of statistical backing that can be used in the validity argument for a placement test in the Japanese context.

3.2 Previous argument-based validation studies of in-house placement reading or listening tests

In this section, previous second/foreign language test validation studies relevant to the current study, because they included a reading or listening test component, and examined an in-house test are presented and critiqued. Such studies remain relatively few in the literature, so this review covers three PhD dissertations/theses, all of which also formed the basis for published studies. As in the previous section the relevant studies are presented chronologically.

Pardo Ballester (2007, 2010) employed Bachman's (2005) Assessment Use Argument (AUA) framework (see section 2.4.1) for the validation of the design and implementation of an online Spanish second language listening test called the Spanish Listening Exam (SLE), designed to place students into university Spanish classes. Pardo Ballester's AUA for the SLE consisted of six claims. The backing for each claim is briefly summarized in the following paragraphs.

The first claim of the AUA is that the "scores that were obtained from test-takers' performance and the ratings of different raters across different assessment tasks are consistent" (p. 196). Evidence to back this claim comes from presentation of the internal reliability of test scores in the form of Rasch reliability statistics which were .87 for the persons and .94 for the items. In addition, Rasch fit statistics were presented for persons and items.

Claim 2 was that “SLE scores were interpreted as indicators of Spanish Proficiency” (p. 196). Evidence to support this claim was sought from backing for a construct validity warrant that “SLE scores are valid indicators of the construct” (p. 196). Backing for this warrant came from expert opinion from four raters who rated test tasks by their content relevance and task difficulty, and whose ratings showed a high degree of consistency. For criterion-relatedness the class level that students were enrolled in prior to taking the test was compared to their Rasch ability with a one-way ANOVA which showed significant differences in Spanish language ability between the three groups. Evidence for content coverage came from a two-way ANOVA, and a Tukey post-hoc test which examined the relationship between the linguistic features of items and task difficulty levels, and which showed that lexical and syntactic features were useful for discriminating between test taker ability levels but phonological features were not.

Claim 3 was that “students were involved with the tool used for assessing listening” (p. 197). Backing came from evidence of test-takers correctly using the test software on which the test was administered, gathered from a post-test survey, which indicated that the overwhelming majority of test takers followed the test instructions correctly, and had no trouble understanding the test instructions.

Claim 4 was that “The SLE listening tasks were generalizable to the classroom domain” (p. 198). Backing for this claim came from test taker survey responses which indicated agreement that the test tasks were similar to classroom tasks. Backing was presented both from open answer questions and also from Likert scale items, which showed that a majority of test takers agreed that the test tasks were similar to classroom tasks.

Claim 5 was that “The consequences of using the SLE were beneficial to the stakeholders in order to assess Spanish comprehension for student placement” (p. 198). Supporting evidence for this claim came from backing for the impact warrant, which consisted of overwhelming agreement from test-takers that the test was a good technique for measuring their Spanish listening comprehension, and from instructor's perceptions that SLE cut-scores are reliable and can place learners into different proficiency levels.

Claim 6 was that the "SLE was a useful and practical web-based test" (p. 199). Evidence for the sixth claim also came from test taker survey responses, which indicated that nearly all test takers agreed that the tests web-based system was easy to use.

Pardo Ballester's study is particularly useful and relevant to the current study, because it validates the design and use of an in-house listening test. Overall, it is a sound example of an argument-based validity study, with thorough backing. The study is also an excellent example of the explicit statement of potential rebuttals in an AUA, which remains fairly uncommon in such studies. In addition, it provides a good example of the kinds of backing that can be used to support a validity argument for an in-house test, particularly backing arising from student survey data. Furthermore, this study provides an example of a standard setting attempt for an in-house test. This is the only example I know of such a standard setting or cut-score study for an in-house test within an argument-based framework.

One problem I see with the study, however, is that the way the argument and supporting evidence is organized is not as clear for the reader as with available examples of applying Kane's argument-based approach. Evidence to support claims in the SLE AUA is in some cases scattered between various warrants in the validity and utilization arguments, which makes it somewhat difficult for the reader to get a quick overview of the backing evidence. This seems to show that Kane's (1992, 1999, 2001, 2002, 2004, 2006) framework and Chapelle et al.'s (2008) adoption if it, with its linked chain of inferences, and also later examples of implementing it by other researchers (Aryadoust, 2013; Koizumi, In'nami, Asano, & Agawa, 2016; Kumazawa et al., 2016; Li, 2015a, 2015b) provide a more coherent and easily interpretable validity argument than Bachman's (2005) framework.

The second study, which used an argument-based approach for the validation of an in-house placement test is a validation study by Johnson (2012), and Johnson and Riazi (2015, 2017). This study used a combination of Kane's (2006, 2008) approach, Bachman's (2005) and Bachman and Palmer's (2010) AUA to validate the use of a standardised placement instrument, the Accuplacer Companion (AC), and a locally made graded essay or Writing Sample (WS) to place students into levelled classes in a developmental English program, or to exclude them from the program. Five claims were presented in the study. These were an evaluation claim, a generalizability claim, an extrapolation claim, a decision claim and a consequences claim. Johnson brought to bear a variety of backings and rebuttal evidence including student and instructor survey data, a faculty focus group interview, institutional procedures, multi-facet Rasch analysis of essay grading, test reliability statistics, and correlations between test scores and later course results. Overall Johnson found that rebuttal evidence gathered for his hybrid validation framework greatly outweighed the backing, thus indicating that neither the AC nor the WS were effective as placement tests in the local institutional context.

Strengths of Johnson's study are that it provides a useful example of how to apply an argument-based approach to validating a specific local use of a large-scale standardised test (for which there are few examples in the literature). It also shows the benefits of argument-based validation studies for suggesting changes and reforms to the use of placement tests in local contexts resulting from examining rebuttal evidence in a validity argument.

Limitations of the study are that individual student AC scores were not available, which limited the types of statistical analysis that could be done. In addition, no analysis was done to analyse the domain relevance of the AC and the WS, by examining actual test content in comparison to curriculum content. Such an analysis may have provided even stronger rebuttal evidence for the course placement function of these two tests.

The final study reviewed here, which used an argument-based approach for the validation of an in-house placement test was validation study of the English Placement Test (EPT) at Iowa State University by Li (2015a, 2015b). Li presented an IUA which used the same six inferences as used in Chapelle et al.'s (2008) study, except that the utilization inference was renamed as a ramification inference. His study focused on the last two extrapolation and ramification inferences. The backing Li marshalled for these two inferences in briefly summarized below, followed by a critique of the strengths and limitations of the study.

As backing for the extrapolation inference Li reported correlations between the EPT and the TOEFL iBT. Disattenuated correlations between the two measures ranged from .600–.684, and also disattenuated correlations between a student self-assessment and the subskills measured by the ETP were reported and were .288 for reading, .411 for listening and .440. for writing. Li judged this evidence to provide partial support for the extrapolation inference.

For the ramification inference Li presented backing for an assumption that "the decisions of ESL course placement are justifiable and comprehensible to test stakeholders" (p. 177) from the results of semi-structured interviews with three stakeholder groups of ESL students, academic advisors, and ESL instructors, and the results showed generally positive perceptions about placement decisions. The second assumption was "that that the decisions are beneficial for learners' improvement of academic English proficiency" (p. 177). Backing for this assumption was sought from pre and post scores in an English course which students took after taking the placement test, which showed that students made statistically significant score gains. However, the proportion of students who passed the post-test was quite low (26.5%), and students in two other courses who were placed by the EPT also showed a low hypothetical passing rate (50%) when their essays were rated by EP standards. Thus, this assumption was judged to lack sufficient support.

The third assumption beneath the warrant for the ramification inference was that "EPT performances are predictive of ESL learners' academic achievement at the university" (p. 178). Backing for this inference came from structural equation modelling which showed that EPT scores had an effect on the test takers first semester GPA (standardised regression coefficient = .306).

The strengths of Li's study are that like the previous studies it provides a rare example of applying an argument-based approach to validation to an in-house language proficiency placement test, and that it focuses on the impact of the test through an argument-based framework, which is also an understudied area. The study also provides a good example of how mixed methods such as Rasch analysis, self-assessment, factor analysis, and semi-structure interviews can be utilized as backing and rebuttal evidence in a validity argument.

Limitations of the study are that it focused only on the final two inferences in a chain of six inferences in the proposed IUA. This would seem to be a somewhat backward approach given that Kane views the inferences as a sequential series of bridges each which must be passed in turn by amassing sufficient backing for the assumptions beneath the warrant for each inference (Kane, 1999).

3.3 Closing comments

This chapter summarized and evaluated three exemlary argument-based validation studies of large scale commercial tests which included a reading or listening component. It then reviewed three exemplary argument-based validation studies of in-house tests with reading and/or listening components. From this brief review of relevant literature it is apparent that fully-elaborated argument-based validation studies of in-house reading and listening tests are few, and that more studies similar to the current study are needed to enrich the literature in this area. In the next chapter the BET Interpretation and Use Argument for the development phase of validation covered by this study is described in detail.

CHAPTER 4

The Bunkyo English Test Interpretation/Use Argument

4.0 Introduction

The argument-based validation framework used in this study, as outlined in section 2.6, is based on the latest iteration of Kane's (2013a, 2013b); argument-based framework and the inferences focused on in the BET IUA are modelled on those used in the interpretive argument for the TOEFL (Chapelle et al., 2008). In addition, some warrants and assumptions in the BET IUA are influenced by the work of Bachman and Palmer (2010).

This chapter outlines the BET IUA, which consists of a chain of six inferences, each supported by one or more warrants. The assumptions underlying each warrant are also detailed, along with evidence sought as backing for each assumption, and rebuttals which could arise resulting from counterevidence. Where necessary, additional literature specific to a warrant or assumption is reviewed to support decisions made in formulating the BET IUA.

4.1 Domain definition inference

This section describes the rationale for defining the domain of the BETs as the GE curriculum. As described in section 2.4.2, including a description of the domain or "the context of interest in which the ability test would be observed" (Chapelle, 2012, p. 20), was an inference first explicitly included in an application of Kane's argument-based approach in the validity argument for the TOEFL (Chapelle et al., 2008).

Bachman and Palmer (2010) point out that for language tests there are two general types of Target Language Use (TLU) domains. These are language teaching domains and real-life domains. When developing language tests, for a language course, the test developers must decide which of these two domains to focus on. Messick (1989) made a powerful argument for defining the domain of achievement tests in terms of lessons and curricula goals. He stated that “if the intended testing purpose is to certify minimal or desirable levels of knowledge and skill acquired through a course of instruction or to evaluate the effectiveness of the instruction, the nature of the curriculum or of actual classroom experiences similarly serve to delimit the domain” (p. 37).

In addition, researchers also argue that placement tests should match the curricula of the courses in which they place students (Brown & Hudson, 2002; Fujita, 2005, Pardo Ballester, 2007). This is important for three reasons:

1. it provides a baseline for measuring student progress through the curriculum
2. the test can be written to match the ability of the learners, so low-proficiency learners will not be discouraged by taking a commercial norm-referenced test in which most questions may be too difficult for them.
3. the test can be based on curriculum goals, which is likely to lead to more appropriate placement decisions than a commercial proficiency-based test with a very broad TLU domain.

As both placement and achievement tests, the domain of the BETs in the BET IUA is defined as a language teaching domain, which for the BET IUA is defined as the BECC General English curriculum. More specifically, as the BETs are tests of reading and listening, the domain definition for the BETs is the reading and listening skills needed to function effectively in GE classes.

Table 4.1 summarizes the warrants, assumptions, backing sought and potential rebuttals for the BET IUA domain definition inference. Whether the backing is relevant to the placement/streaming or achievement function of the BETs is noted in the far-right column.

Table 4.1. *The BET IUA Domain Definition Inference*

<table>
<tr><th colspan="4"><b><i>Domain Definition Inference</i></b></th></tr>
<tr><td colspan="4">Warrant for the BET Placement/Streaming Function IUA: <i>Observations of performance on the BETs reveal the level of reading and listening skills and abilities needed to function effectively in GE courses, and are representative of reading and listening performance in the General English curriculum.</i>
Warrant for the BET Achievement Function IUA: <i>Observations of performance on the BETs reveal achievement of the General English course goals for reading and listening.</i></td></tr>
<tr><th>Assumptions underlying the warrant</th><th>Backing sought to support assumption</th><th>Potential Rebuttals</th><th>Placement/Streaming or Achievement IUA</th></tr>
<tr><td rowspan="2">1. BET content is a representative sample of the GE curriculum.</td><td>1. Tables show a wide representation and even proportion of lesson materials from each GE unit on which items in each BET were based.</td><td>Tables of GE unit representation in the BETs show that some target GE units are not represented in some BETs.</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>2. An analysis of teacher and student survey responses shows that these stakeholders think BET tasks are representative of the targeted GE curriculum.</td><td>An analysis of teacher and student survey responses shows that these stakeholders view BET tasks as being unrepresentative of the targeted GE curriculum.</td><td>Placement/streaming and Achievement</td></tr>
<tr><td rowspan="3">2. BET task characteristics match the GE curriculum goals</td><td>1. Can do statements for testlets in the BET specifications accurately reflect the GE curriculum goals and testlet characteristics.</td><td>Analysis shows that several of the target can do statements for BET testlets in the BET specifications are not well-matched to the testlet tasks or curriculum goals.</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>2. Can do statements for BET testlets from the BET specifications cover a wide range of subskills at the curriculum target CEFR levels for reading and listening.</td><td>Only a narrow or incomplete range of subskills at the curriculum target CEFR levels for reading and listening is covered by the target can do statements from the BET specifications.</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>3. BERT and BELT lexical profiles are similar to other tests which target the same CEFR skills and levels. i.e. KET and PET reading and listening sections.</td><td>BERT and BELT lexical profiles are substantially different to KET and PET reading and listening section lexical profiles.</td><td>Placement/streaming and Achievement</td></tr>
<tr><td rowspan="2">3. BET task types are representative of reading and listening task types in the GE curriculum.</td><td>1. Tables analysing the representation of BET type tasks in the GE curriculum show a broad and even representation of BET type tasks in the curriculum.</td><td>Tables analysing the representation of BET type tasks in the GE curriculum show a lack of representation and/or an uneven representation BET-type tasks across the curriculum.</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>2. An analysis of teacher and student survey responses shows that these stakeholders view BET tasks as being broadly similar to reading and listening tasks in the GE curriculum.</td><td>An analysis of teacher and student survey responses shows that these stakeholders view BET tasks as being broadly dissimilar to reading and listening tasks in the GE curriculum.</td><td>Placement/streaming and Achievement</td></tr>
</table>

4.1.1 Domain definition warrants

As can be seen in Table 4.1, in order to reflect the two main uses of the BETs to be validated in this study there are two warrants for the domain definition inference in the BET IUA, one to reflect the placement/streaming function of the BETs, and the other to reflect the achievement function. These two warrants are supported by three assumptions.

4.1.2 Assumption 1 of the domain definition warrant

The first assumption is that *BET content is a representative sample of the GE curriculum.* This is based on criteria that tests should sample representatively from their domain of interest. As Kane asserts, agreeing with Guion (1977) "it is legitimate to take the observed performance as an estimate of overall performance in the domain, if … the observed performances can be considered a representative sample from the domain" (Kane, 2006, p. 19).

The methodology used to seek the first backing for assumption 1 of the domain definition inference warrant of creating tables showing GE course content represented in the BETs covered by this study, can be found in section 5.3.1.1, and an analysis of the results can be found in section 6.1.1.

The methodology for the teacher and student surveys from which questions were used to seek the second backing for assumption 1 of the warrant for the domain definition inference are explained in sections 5.3.5.1 and 5.3.5.2, and the results for the relevant questions are presented and discussed in section 6.1.2.

4.1.3 Assumption 2 of the domain definition warrant

The second assumption beneath the domain inference warrants is that *BET task characteristics match the GE curriculum goals*. As the domain of the BETs is defined as the GE curriculum, it is clearly important that BET tasks match the reading and listening goals of the curriculum, which are defined in the form of CEFR can do statements, in order for the tests to be able to fulfil both their placement/streaming and their achievement functions. The stated GE curriculum goals in the course outlines for students for the revised curriculum targeted by the BETs in this study were the overall listening comprehension and overall reading comprehension CEFR can do statements for level A2 for the lower stream course, and for level B1 for the higher stream course as follows:

Lower stream course reading goal: *Understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.*

Higher stream course reading goal: *Read straightforward factual texts on subjects related to your field and interest with a satisfactory level of comprehension.*

Lower stream course listening goal: *Understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated.*

Higher stream course listening goal: *Understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.*

Underlying these overall proficiency descriptors in the CEFR are further, more detailed can do statements for reading and listening sub-skills. Therefore, if the BETs are to claim to match the GE curriculum goals, they should test a wide range of the more detailed can do statements for reading and listening at CEFR levels A2 and B1. The methodology for analysing the can do statements from the BET specifications to assess their match to the task characteristics of BET testlets, and their match to the GE course goals is presented in section 5.3.1.3, and the results are presented and analysed in section 6.1.3.

Furthermore, if the BETs are to claim to be able to measure student achievement of the target CEFR goals for reading and listening, then it is clear that BET testlets should cover a good range of the subskills or tasks covered by the more detailed CEFR can do statements for those levels. Thus, the second backing sought for assumption 2 of the domain definition inference warrant comes from a comparison of the checked and revised can do statements covered by the BETs (created to seek the first backing for the second assumption of the domain inference warrant) with the complete list of CEFR can do statements to see if there are any gaps in the BET coverage of CEFR reading and listening subskills. A reasonable range of coverage of the more detailed can do statements for reading and listening at the curriculum target levels would provide backing, whereas, a poor range of coverage would provide rebuttal evidence for assumption 2 of the domain definition inference warrants.

The third backing sought for the second assumption underlying the warrant for the domain definition comes from a comparison of BERT and BELT lexical profiles with published results of KET and PET reading and listening section lexical profiles. One of the stated purposes of the BETs in the BET specifications is to assess achievement of the curriculum target proficiency levels of A2 and B1 for reading and listening. Therefore, the lexical profiles of the BETs should be similar to lexical profiles of other tests which claim to assess achievement of the target CEFR levels of A2 and B1 of the CEFR for reading and listening. Literature reporting lexical profiles for the KET and PET reading and listening sections is summarized in the following paragraphs, and the software used for creating the lexical profiles reported in this study is briefly described in section 5.3.10.4. The results of this analysis are presented, compared to the KET and PET results and discussed in section 6.1.5.

Lexical profiles, or lexical frequency profiles, are produced by software which analyses a text and places the words appearing in the text into frequency bands. The frequency bands into which words are placed have been calculated through computer analysis of a corpus of texts. Frequency bands are commonly divided into thousands, for example, the most commonly appearing 1000 words (first 1000), the second 1000 most commonly appearing words (second 1000) etc., Common frequency profiles are the Academic Word List (AWL) which comes from an analysis of academic texts (Coxhead, 2000), the General Service List (GSL) (West, 1953), and the British National Corpus (BNC) (University of Oxford, 2015). (See Laufer, 2012 for a succinct overview of lexical frequency profiles.)

Previous research has been conducted using lexical profiles as validity evidence for tests which claim CEFR A2 and B1 alignment within a socio-cognitive validation framework (Weir, 2005a). Specifically, these were two tests in the Cambridge suite of examinations, the Key English Test (University of Cambridge ESOL Examinations, 2012a), which aims to certify proficiency at the CEFR A2 level and the Preliminary English Test (University of Cambridge ESOL Examinations, 2012b), which aims to certify proficiency at the CEFR B1 level.

The lexical profiles for all of the reading sections of Cambridge suite of exams (i.e. KET, PET, FCE, CAE and CPE) reading sections were reported in Khalifa and Weir (2009) using the General Service List (GSL) and the Academic Word List (AWL), and the British National Corpus (BNC) 20 most frequent 1000 word bands. The analyses were run by Norbert Schmitt using Wordsmith Tools and Compleat Lexical Tutor (Cobb, 2018). Schmitt’s analysis was based on a corpus of 30 reading papers. However, the description does not make it clear if this was 30 reading papers for each level, or five reading papers for each level (i.e., 30 reading papers divided by six levels). The description also does not explain if the totals for the KET and PET are the averages for each set of papers, or the total for each set of papers. Based on an examination of total word counts of KET and PET reading sections from a couple of sample papers (a KET reading section was around 1200 words, and a PET reading section was around 2000 words) it is assumed that the numbers are averages.

The KET and PET analyses presented by Khalifa and Weir provide evidence supporting increasing difficulty from the KET (A2) to the PET (B1) as the proportion of words in higher bands clearly increases from the KET to the BET, as does the proportion of words from the academic word list.

Lexical profiles for all of the five Cambridge Main Suite Listening papers were also produced by Norbert Schmitt and were reported in chapter 5 of the book *Examining Listening: Research and Practice in Assessing Second Language Listening* (Eliot & Wilson, 2013). Tape scripts from five tests for each level were analysed using Lextutor Classic and BNC-20.

By using 95% lexical coverage as a general benchmark for comprehension, Eliot and Wilson (2013) point out that knowledge of the first 1000 words provides nearly 94% coverage of KET tapescripts, giving learners with such knowledge a good chance of understanding KET tapescripts. They also calculate that knowledge of approximately the first 1500 words allows for comprehension of the PET tapescripts.

4.1.4 Assumption 3 of the domain definition warrant

As the domain of the BETs is defined as the GE curriculum, it is important that BET task types are similar to reading and listening task types in GE lessons. This is important both for claiming that the BETs are achievement tests of the target curricula, and also for placing students in the appropriate course stream. This has been identified in the literature as an important aspect of domain validity (Kane 2006, 2013b; Guion, 1977). Bachman and Palmer (2010) also make a persuasive argument for why test tasks should be similar to language learning tasks in a curriculum.

> One way to minimize the potential for negative impact on instruction is to change the way we test so that the characteristics of the test and test tasks correspond to the characteristics of learning tasks in the instructional program. If the content of the

> assessment is thus aligned with the goals and objectives of instruction and with instructional activities, then "teaching to the test" may become an aspect of positive impact on instruction (p. 108).

To support assumption 3 of the domain definition warrant the first piece of backing sought consists of tables made by teachers in charge of planning and organizing the creation of the of the new GE first year materials in 2015, and also teachers involved in preparing revised GE second year materials in 2016 for 2017 curriculum. How these tables were put together is described in section 5.3.2.2. Tables showing an even representation and distribution of BET type tasks through the GE curriculum would provide strong backing for this assumption. On the other hand, tables showing an uneven representation of BET type tasks and/or an uneven distribution of BET type tasks in different units and years of the GE curriculum would provide rebuttal evidence for this assumption.

The second type of backing sought to support the third assumption of the warrant for the BET IUA domain definition inference comes from an analysis of answers to teacher and student surveys regarding their attitudes to the BETs. Three statements each from the student and teacher surveys administered at the end of the 2016/17 academic year were relevant and are used as backing for assumption 3. The relevant statements were adapted from Pardo Ballester's (2007, 2010) validation study. Unfortunately, these questions were not included in the earlier administrations of the teacher and student surveys, as the idea to include them only occurred to the researcher in 2016 after reviewing Pardo Ballester's work. The statements from the teacher and student surveys analysed as backing for assumption three of the domain inference warrant are listed in Table 4.2.

Table 4.2. *Survey Statements Used as Backing for Assumption 3 of the Domain Definition Inference Warrant*

Student Survey	Teacher and LA Survey
The BET included a wide range of content from what I studied in my BECC English classes.	BET content is representative of GE curriculum content.
BET reading tasks are similar to reading tasks in my BECC English classes.	BET reading tasks are similar to reading tasks in the GE curriculum.
BET listening tasks are similar to listening tasks in my BECC English classes.	BET listening tasks are similar to listening tasks in the GE curriculum.

The format and administration of the teacher and student surveys is described in section 5.3.5, and the results sought as the second backing for assumption three of the domain definition inference are analysed in section 6.1.7. Majority teacher and student agreement with these survey statements, along with a strong degree of agreement would provide firm backing for this assumption.

4.2 Evaluation inference

As explained in section 2.3.2 the evaluation inference links a test takers' performances to their observed scores. Kane et al. (1999) put forward the following assumptions underlying the evaluation inference: "The criteria used to score the performance are appropriate and have been applied as intended and second, that the performance occurred under conditions compatible with the intended score interpretation" (p. 9). Chapelle (2015) further explains that to "make evaluation inferences, backing is needed to support the quality of the sample of responses obtained from test takers" (p. 19). The warrant, assumptions, backing sought and potential rebuttals for the BET IUA evaluation inference are summarized in Table 4.3. Whether the backing is relevant to the streaming or achievement function of the BETs is noted in the far-right column.

Table 4.3. *The BET IUA Evaluation Inference*

Evaluation Inference			
Warrant: *BET tasks yield consistent observed scores, which are not contaminated by construct irrelevant variance.*			
Assumptions underlying the warrant	**Backing sought to support assumption**	**Potential Rebuttals**	**Placement/Streaming or Achievement IUA**
1. The BETs were administered and scored consistently.	Descriptions of the procedures followed for annual BET administrations show that the tests were administered and scored consistently.	Descriptions of the procedures followed for annual BET administrations show that the tests were administered and scored differently for some classes or students.	Placement/streaming and Achievement
2. Appropriate procedures were followed to develop BET testlets and items.	A description of procedures used for developing the BET testlets and items, shows that appropriate procedures were followed, for example, peer review of draft testlets, and standardised checking of item characteristics.	A description of procedures used for developing the BET testlets and items, shows that inappropriate or insufficient procedures were followed, for example, BET testlets were not piloted.	Placement/streaming and Achievement
3. BET task instructions were easily comprehensible for students.	Results of a student survey show that students thought that BET instructions were easy to understand.	Results of a student survey show that students think that BET instructions were unclear.	Placement/streaming and Achievement

4.2.1 Assumption 1 of the evaluation inference warrant

For the first backing to support the assumption that the BETs *are administered and scored consistently* (Kane, 2006), three sources of backing are presented. Firstly, procedures for the annual administration of the three BETs are described, secondly instructions for test administration from BET proctoring guidelines in the BET specifications are summarized, and finally a brief description of how the BETs are scored with a bubble sheet reader is presented. A presentation and analysis of these backings can be found in section 6.2.1.

4.2.2 Assumption 2 of the evaluation inference warrant

Backing for assumption 2 of the evaluation inference warrant was sought firstly from a description of the procedures followed for developing the BET testlets and items, and secondly from a checklist used for testlet/item review for making the 2016 and 2017 BETs. An analysis of these backings is presented in section 6.2.2.

4.2.3 Assumption 3 of the evaluation inference warrant

It is important that test-takers are able to understand test instructions, because if test instructions are not understood, this may introduce construct irrelevant variance into test scores. Backing for assumption 3, that *BET task instructions were easily comprehensible for students* comes from an analysis of the results of the student survey. Two questions relevant to this assumption that “BET DVD spoken instructions were easy to understand” and that “BET test paper instructions were easy to understand” are analysed in section 6.2.3. Results for a further two survey questions relevant to this inference which address instructions for the BELT and BERT sections specifically that “BET listening section instructions are easy to understand” and that “BET listening section instructions are easy to understand” are not presented due to considerations of space, and also because their results were consistent with the two survey questions presented.

4.3 Generalization inference

As explained in section 2.3.2 the generalization inference extends interpretations from those based on scores on a single test or the ‘observed score’ to interpretations about scores on other possible tests or tasks in the same universe of generalizations.

In this section the warrants, assumptions, backing sought and potential rebuttals for the BET IUA generalization inference are summarized in Table 4.4. Whether the backing is relevant to the streaming or achievement function of the BETs is noted in the rightmost column.

Table 4.4. *BET IUA Generalization Inference*

Generalization Inference			
Warrant: *Observed scores are estimates of expected scores on other versions of the BET (1, 2 & 3).*			
Assumptions underlying the warrant	**Backing sought to support assumption**	**Potential Rebuttals**	**Placement/Streaming or Achievement IUA**
1. Enough tasks are included to provide stable estimates of test taker performance.	Reliability, and dependability statistics for the BETs, BERTs and BELTs are appropriate for a moderate-stakes criterion-referenced test.	Low reliability and/or dependability statistics indicate poor test consistency for some BETs.	Placement/streaming (reliability) and Achievement (dependability)
2. Appropriate scaling and equating procedures are used to measure student achievement across BET forms.	Appropriate procedures are followed to equate the BETs to give information on student achievement of the curriculum.	The BET results presented to students in the frame of this study were not equated.	Achievement
3. Testlet specifications are well defined so that parallel tasks and test forms are created.	BET testlest specifications were sufficient for testlets of equivalent difficulty target equivalent language skills to be created.	The BET testlet specifications lacked sufficient clarity of detail in some areas for testlets of equivalent difficulty target equivalent language skills to be created.	Achievement

4.3.1 Assumption 1 of the generalization inference warrant

As the BETs within the frame of this study were intended to act as both achievement tests of the GE curriculum (BETs 1–3) and also placement/streaming tests (BETs 1 & 2) to place students into one of two courses, and into streams within those courses divided by English ability level, two estimates for the consistency of BET scores were calculated. Firstly, a norm-referenced measure of internal consistency known as the Kuder–Richardson Formula 20 (KR-20) which is based on a normal distribution, and which is commonly used for norm-referenced tests, or tests which are designed to spread test taker ability out along a normal or bell curve, was calculated as part of assessing the streaming function of the test as being able to separate test takers into separate groups. To assess the achievement function of the BETs a measure of internal consistency known as the phi dependability index or the generalizability coefficient for absolute error was also calculated for all of the BETs. Brown and Hudson (2002) use the term *dependability* for measures of internal consistency for criterion-referenced tests and their nomenclature is followed here. The phi dependability index is used for criterion-referenced tests on which absolute or achievement decisions are made, and it does not rely on a normal distribution.

There are no concrete rules about acceptable levels of reliability for tests in the literature, but some general guidelines can be found. For example, Weir (2005a) states that a reliability of .8 is generally considered minimally acceptable reliability. Reliabilities of .81 and .77 were reported as part of validity evidence for The City and Guilds Communicator exam and these were claimed to be “satisfactory” (O’Sullivan, 2010, p. 42).

Reliability also has implications for the number of levels into which a test is able to accurately divide test takers. An index of separation (or strata) can be used to estimate the number of statistically different performance strata into which a test can divide students (Wright, 1996). Kaftandjieva (2004) provides recommended numbers of cut points for a test based on test reliability and the resulting index of separation, which is reproduced as Table 4.5.

Table 4.5. *Reliability Recommendations Kaftandjieva (2004)*

Number of Levels	2	3	4	5	6
Number of Cut-off Points	1	2	3	4	5
Test Reliability	≥.61	≥.80	≥.88	≥.92	≥.95

Linacre (2017b) also provides the following guidelines for a tests ability to accurately divide learners into groups based on the person separation statistic and strata statistic. Linacre recommends using the person separation statistic if "the outlying measures are accidental" and strata statistic "if the outlying measures represent true performances" (p. 622). Table 4.6 shows Linacre's guidelines.

Table 4.6. *Person Separation and Strata Guidelines (Linacre, 2017b)*

Test Reliability	Person Separation	Strata
.5	1	1.64
.8	2	3
.9	3	4.33
.94	4	5.67

Based on the above recommendations from the literature a test reliability of ≥ .8 is seen to be sufficient evidence to support the placement function of the BETs during the timeframe of this study and a claim of the tests to be able to separate learners into CEFR A1 and A2 levels for course placement purposes. On the other hand, to support claims for the BERT and BELT to be able to place learners into CEFR levels A1, A2 and B1 for achievement purposes reliability of ≥ .8 would be needed for each of these BET sections using the more lenient strata index.

The phi coefficient is also known as the generalizability coefficient for absolute error and is a "general purpose estimate of the domain-referenced dependability of a test." (Brown, 1989, p. 102). According to Bachman (2004), it "estimates how dependable the test score is as an indicator of the test taker's level of ability or mastery of a particular content domain" (p. 194). Also, Bachman states that "the interpretation of the phi coefficient is directly analogous to that of the CTT internal consistency reliability estimates, as the proportion of total score variance that is domain score variance" (p. 195). I could not find any guidelines in the literature as to what constitutes an acceptable phi coefficient for domain referenced achievement tests such as the BETs, so the same general guideline as for assessing the streaming function of the bets with the KR-20 reliability index will be used. i.e. a phi coefficient of ≥ .8 or 80% of the variance in BERT and BELT scores being accounted for by ability within the domain of the GE curriculum is taken to be sufficient support for this assumption 1 of the BET IUA generalization inference warrant for the BET achievement test function. Results for the reliability and dependability of the BETs within the frame of this study are presented and discussed in section 6.3.1.

4.3.2 Assumption 2 of the generalization inference warrant

Assumption 2 of the generalization inference warrant is that *appropriate equating and scaling procedures are used to present students with their BET scores* (Chapelle et al., 2008). In order for the BETs to show students how they have progressed, and their achievement of the GE curriculum goals for reading and listening, students would need to be shown their BET, BERT and BELT scores for the BET 2 and BET 3, and to be given an indication of how their scores had improved relative to their previous BETs. This would require equating the BETs, which may be of differing difficulty in spite of being written to the same test specifications. Equating and scaling allow tests of the same format to be put on the scale so that scores between the tests are directly comparable (see Kolen & Brennan, 2014 for a detailed presentation of test equating issues and techniques.) Backing for this assumption would detail the procedures used to equate the BETs, BERTs and BELTs.

4.3.3 Assumption 3 of the generalization inference warrant

Backing for assumption 3 that *testlet specifications are well defined so that parallel tasks and test forms are created* was sought from an examination of the BET testlet specifications. The type of information presented in the specifications for BET testlets is presented and assessed as to whether it is sufficient for item writers to produce testlets across test forms which are of equivalent difficulty, and which target the same language skills, in section 6.3.3.

4.4 Explanation inference

As explained in section 2.4.2 the explanation inference was an inference first employed within Kane's argument-based approach to test validation by Chapelle et al. (2008) in their validation study of the new TOEFL iBT, and this inference seeks to link test scores to constructs of language proficiency. In the case of the BETs, which are intended to be achievement tests of the GE curriculum the test target constructs are intended to be reading and listening proficiency as defined by the CEFR scales for levels A2 and B1.

Alderson et al. (2006) stated that "the CEFR, being a comprehensive description of language use, can also be considered, implicitly at least, as a theory of language development" (p. 6). However, several writers have pointed out that the CEFR can do statements for reading and listening do not in themselves provide sufficient detail to allow for the construction of test items.

Weir (2005b) points out that the information given in the CEFR descriptors in not nearly detailed enough for the construction of valid test items, or to substantiate the claim that two different tests based on the CEFR are equivalent. He states that the CEFR

> was not designed specifically to meet the needs of language testers and that it will require considerable, long-term research, much reflective test development by

> providers, and prolonged critical interaction between stakeholders in the field to address these deficiencies. (p. 283)

Alderson et al. (2006) in relation to the development of the DIALANG project also stated that “the CEFR in its current form may not provide sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level of the CEFR” (p. 5).

In spite of the inherent difficulties of linking language tests to the underspecified CEFR statements, many studies that have sought to establish evidence of a link between reading and listening tests and the implicit CEFR reading and listening CEFR constructs.

Expert opinion is the most common evidence type used to link exams to the CEFR and its implicit language constructs. One form this expert opinion has taken is a review of the test specifications and test items using the procedures outlined in the *Manual Relating Language Examinations to the Common European Framework of Reference for Language* (Council of Europe, 2009), for example (Downey & Kollias, 2010; Elif, Thomas, O’Dwyer, & O’Sullivan, 2010; O’Sullivan, 2010; Szabo, 2010). Another form of expert review has been an analysis of test tasks using CEFR Grids made by Dutch CEFR Construct Project (Alderson et al., 2006), for example Kecker and Eckes, (2010), and Wu and Wu (2010).

Cut-score setting is also a common way of providing evidence for a link between an exam and the CEFR, and therefore its implied constructs (Brunfaut & Harding, 2014; Kecker & Eckes, 2010; O’Sullivan, 2010; Noijons & Kuijper, 2010). Details of CEFR cut-score setting projects are not summarized here, however, due to considerations of space, and because BET development had not yet reached the stage of effective CEFR-based cut-score setting during the time-frame of this study.

Relevant studies which have attempted to validate a link to the CEFR scales and their implied construct for a reading or listening test by providing evidence of correlation with other tests which already claim to be aligned to the CEFR, are briefly reviewed in the relevant sections for each of the assumptions of the BET IUA explanation inference warrant.

The warrants, assumptions, backing sought and potential rebuttals for the BET IUA explanation inference are summarized in Table 4.7. Whether the backing is relevant to the streaming or achievement function of the BETs is noted in the rightmost column.

Table 4.7. *BET IUA Explanation Inference*

Explanation Inference			
Warrant: *Observed scores are attributable to the constructs implicit in the CEFR levels A2-B1 for English reading and listening.*			
Assumptions underlying the warrant	**Backing sought to support assumption**	**Potential Rebuttals**	**Placement/Streaming or Achievement IUA**
1. Performances on BET measures are associated with performance on other test-based measures which claim CEFR alignment.	Large correlations are found between BET, BERT and BELT scores and Oxford Online Placement Test (OOPT) scores. Medium correlations are found between BET, BERT and BELT scores and TOEIC scores.	Small or non-existent correlations are found between BET, BERT and BELT scores and Oxford Online Placement Test (OOPT) scores and TOEIC scores.	Placement/streaming and Achievement
2. Performances on BET measures are associated with test taker self-assessments of CEFR can do statements for reading and listening.	Medium to large correlations are found between BERT and BELT scores and test taker self-assessments of CEFR reading and listening can do statements.	Small or non-existent correlations are found between BERT and BELT scores and test taker self-assessments of CEFR reading and listening can do statements.	Placement/streaming and Achievement
3. BET items effectively measure the constructs of interest.	1. Rasch item fit statistics show that most items function to effectively measure the constructs of reading and listening.	Rasch item fit statistics show that a large proportion of BET items do not function to measure the construct of interest.	Placement/streaming and Achievement
	2. Point-measure correlations indicate that most items function to measure the same construct.	A large proportion of items have point-measure correlations close to or below zero.	
4. The range of item difficulties on the test matches the range of abilities of the test takers.	BET Rasch person/item maps show an even spread of item difficulties across the test taker ability range.	A large proportion of BET items are too difficult or too easy for the range of test takers. There are large gaps in the difficulty range of items	Placement/streaming and Achievement

4.4.1 Assumption 1 of the explanation inference warrant

As backing for assumption 1 of the explanation inference warrant correlations between BET, BERT and BELT scores and the OOPT and TOEIC were sought.

For the BETs within the frame of this study BET, BERT and BELT scores were correlated with scores from representative samples of GE students who took the Oxford Online Placement Test (OOPT). Details of the OOPT test administrations are given in section 5.3.9.1. Reasons for choosing the OOPT as a criterion for the BETs in terms of measuring the CEFR constructs of reading and listening, are presented in the following paragraphs.

When planning this study, a criterion for the BETs was searched for which would be easy to administer to BECC students, which claimed to measure the constructs of reading and listening on the CEFR scales, and which had a strong existing validity argument as a measure of CEFR-defined constructs. Ease of test administration, meant the test should be able to be taken on iPads, which all Bunkyo students are given on entrance to the university. The DIALANG (Alderson, 2005; Alderson & Huhta, 2005) was considered as a criterion, however, at the time of project planning the DIALANG was not able to be administered on iPads due to flashplayer compatibility issues and also test instructions were not available in Japanese. These problems have since been rectified for the DIALANG, but unfortunately meant that it could not be used as a criterion for validity during the frame of this study.

The OOPT was chosen instead because it is relatively cheap to administer, students could take it on their iPads, and student results are conveniently aggregated online in a learning management system for later analysis. In addition, the OOPT is a computer adaptive test, which places learners at all CEFR levels, including the A1, A2 and B1 levels for English, which was thought to be the general range of ability of GE students. Finally, the OOPT was said to have a strong validity argument for its use as a CEFR-based placement test through “having a strong theoretical basis and having been through a rigorous test design, pretesting, and piloting stage” (Oxfordenglishtesting.com, 2017).

The OOPT consists of two parts. The first part is a Use of English section, and according to Purpura (2010):

> This section of the test includes three language knowledge tasks. The first task primarily aims to measure grammatical forms; the second mainly to measure semantic meaning; the third is a test of grammatical form and meaning, and a fourth (to be included in the test by early 2010) is designed to measure the students’ knowledge of the pragmatic (i.e., implied) meanings encoded in situated interactions. (p. 2)

According the Oxford testing website this section measures “vocabulary, grammar and the understanding of meaning in a conversation.” While the OOPT does not claim to be a direct test of reading comprehension, the well-established importance of vocabulary and grammar as components of reading comprehension (Shiotsu & Weir, 2007) makes this section testing these components a suitable criterion for the reading section of the BETs. This criterion is further relevant to the BERT as two BERT testlests focus on meaning and pragmatics in conversation, which the OOPT section 1 also claims to measure. In addition, the OOPT Use of English section claims to test meaning in conversations and two tasks in the BET similarly focus on understanding meaning in conversations.

The listening section of the OOPT claims to measure “understanding of meaning in a conversation” (Oxfordenglishtesting.com, 2017). According to Purpura (2010) the “Listening Section of the test is designed to present test takers with different types of listening passages from which they will need to identify the literal, intended, and implied meanings being communicated in what they hear” (p. 20). Similarly, in the BELTs the listening passages consist of dialogues and monologues, which mainly focus both on identifying specific information, in some cases inferring meaning such as the opinion and attitudes of the speaker.

Finally, the OOPT overall "claims to measure test takers' communicative language ability, so that the scores from the exam can be used to make relatively accurate placement decisions in a language program that is aligned with the CEFR" (Purpura, 2010, p. 21). This is also the goal of the BETs as placement tests, so the OOPT was chosen as a particularly suitable criterion for assumption 1 of the BET IUA explanation inference warrant. Results of correlational analyses of the BETs, BERTs and BELTs with OOPT overall scores, reading section scores and listening section scores are presented and discussed in section 6.4.1. Because of the close similarity between the constructs which the OOPT and the BETs claim to measure, "large" correlations (J. Cohen, 1988) between BET and OOPT overall scores are sought as sufficient backing for assumption 1 of the explanation inference warrant for the BET placement/streaming function. Also, large correlations between BERT scores and OOPT use of English section scores, and large correlations between BELT scores and OOPT listening section scores are sought as sufficient backing for the BET achievement function.

In addition, a second criterion of scores on the TOEIC, is used as backing for assumption 1 of the explanation inference warrant in this study. The TOEIC was chosen firstly because a convenience sample was available. Students in the Global Communication Department must take the TOEIC annually, and GCD students also took the BET1 2015, and the BETs 1 & 2 2016, so their scores were available for analysis. Secondly, a standard setting study has been done which links TOEIC scores to the CEFR scores (Tannenbaum & Wylie, 2008). Therefore, it can be argued that the BETs and the TOEIC overlap to at least some extent in measuring the constructs of reading and listening as implied by the CEFR scales. According to Powers (2010b), the TOEIC primarily claims to "measure a person's ability to communicate in English in the context of daily life and the global workplace environment using key expressions and common, everyday vocabulary" (p. 2), and the TOEIC is designed to help employers “make critical decisions concerning the English language skills of their employees and their prospective employees” (p. 2).

Given the emphasis on workplace communication in the TOEIC, which is different to the broader range of communicative situations in another country which the GE curriculum aims to cover, and also given that the CEFR was linked to the TOEIC post-hoc through a standardization panel, rather than CEFR-alignment being built into the test design phase, a priori, as was the case with the OOPT, “medium” correlations (Cohen, 1988) or correlations between .3 and .49 for the TOEIC overall, and reading and listening subsections with BET overall scores and reading and listening subsections are taken to be sufficient for this backing to support assumption 1 of the BET IUA explanation inference warrant.

4.4.2 Assumption 2 of the explanation inference warrant

Assumption 2 of the explanation inference warrant for the BET IUA is that *performances on BET measures are associated with test taker self-assessments of CEFR can do statements for reading and listening.* This section firstly reviews some previous literature on self-assessment in language learning, then studies which used student self-assessments as a criterion for test validation. Finally, the reasons for choosing CEFR can do statements as a criterion for the BETs are explained.

Self-assessment is defined as the “language learner’s evaluation of his or her own language skills” (Luoma, 2013). Blanche and Merino (1989) conducted a wide-ranging review of literature on self-assessment in language testing, and they found that correlations between self-assessments and a variety of criterion “ranged from .50 to .60”, and that higher correlations were “not uncommon” (p. 315). Ross (1998) performed a meta-analysis of self-assessment in foreign and second language testing, which included studies examining associations between self-assessment and language tests of reading, listening, reading and writing. Ross found a Pearson product-moment correlation of .61 for the 23 reading studies examined, and of .65 for listening from 18 correlations. From these two broad analyses of the literature, it can be seen that strong associations have been established between student self-assessments of reading and listening and standardised tests of these skills.

Two factors which have been reported to compromise the accuracy of self-assessments of second or foreign language ability are the respondents' relative familiarity with the situations or tasks in the can do statements through actual experience or lack of experience of the situations assessed by the can do statements, and differing interpretations of the Likert scale categories (Ross, 1998; Suzuki, 2015). In addition, North and Jones (2009) state that "It has often been observed that low-level learners tend to over-estimate themselves, while high-level learners tend to under-estimate" (p. 13). A more recent study by Suzuki (2015) also found that "less experienced second language speakers appeared to overestimate their ability, whereas those with more experience underestimated their language skills" (p. 64).

Studies have used self-assessment as backing in an argument-based approach to test validation. Li (2015b) compared the results of a standardised placement test called the MEPT with a self-assessment "which consisted of 54 statements on a six-point Likert scale in five sections: Self-assessment of English use (21 items), Academic self-efficacy (5 items), Learning motivation (8 items), Self-regulated learning strategies (10 items), and Anxiety about using English (10 items)". Li found weak correlations of .009–.19 for self-assessment and reading, and .142–.15 for listening, which did not provide backing for the extrapolation inference in the MEPT validity argument.

Aryadoust (2013) also sought correlational evidence between a self-assessment survey of academic listening ability and IELTs listening scores as backing for the extrapolation inference in the IELTs validity argument. He found small to medium correlations (*r* = .144–.322), which were only statistically significant for three out of six of the academic listening skills in his self-assessment questionnaire.

Wang, Eignor and Enright (2008) reported correlations between self-assessments of .47 for listening and .45 for reading and the new TOEFL reading and listening sections, which were the focus of their validation study. Chapelle (2008) also stated that “High correlations of assessments and instructor ratings with the TOEFL scores are not expected because use of the differences in measurement methods and differences in constructs of perceived ability and test performance (p. 343).”

Although the three above mentioned studies used correlational evidence of student self-assessments with test scores as backing for assumptions in extrapolation inferences, in the case of the BETs it is argued that correlations with self-assessments using CEFR statements are suitable as backing for the explanation inference, as the construct measured by the BETs is intended to be the reading and listening constructs implicit in the CEFR scales for these two skills.

As part of validating the DIALANG, Alderson (2005) correlated single overall can do statements for each CEFR level (A1–C2) with DIALANG test results for the skills of reading listening and writing. Alderson found Spearman’s correlations between overall can do statements and DIALANG results to be .54 for reading and .47 for listening. Alderson also correlated an IRT score for each skill that resulted from calibrating 18 more detailed can do statements for the three skills, with their DIALANG score, and these returned Spearman’s correlations of .49 for reading and .5 for listening.

Given that a previous study using self-assessment of CEFR-aligned can do statements found medium to large correlations (Alderson, 2005) between student self-assessments and test scores for reading and listening, and also that there are several factors that may moderate correlations such as unfamiliarity with the situations in the can do statements, differences between the constructs of reading and listening as measured by a multiple choice test and self-perceptions of listening skills, and overestimation of ability by low-proficiency learners. Medium correlations (.30–.49) are taken as moderately strong backing for assumption 2 of the BET IUA explanation inference warrant, and large correlations (> .5) are taken as strong backing.

CEFR self-assessment surveys for reading and listening at the GE curriculum target levels were chosen as a criterion for backing for assumption 2 of the explanation inference warrant, because the self-assessment scales represent the implicit construct of the CEFR, and are therefore perhaps the most direct means available to access this construct. Details of the administration and analysis of the CEFR self-assessment surveys administered during the frame of this study are given in section 5.3.5.4, and the results are presented and analysed in section 6.4.2.

4.4.3 Assumption 3 of the explanation inference warrant

Assumption 3 of the explanation inference warrant is that *BET items effectively measure the constructs of interest*. Backing sought for this assumption comes from item analysis using Rasch item fit statistics, and point-measure correlations. Specifically, items with mean squares between 0.5 and 1.5 are considered productive for measurement of the construct (R. Green, 2013; Wright & Linacre, 1994) so the proportion of items with mean squares less than 0.5 and greater than 1.5 for the 2016 BETs 1 & 2 and for the 2017 BETs 2 & 3 is examined. In addition, items with negative point measure correlations are identified, because “Negative … point-measure correlations … indicate that the responses to the item contradict the latent variable defined by the consensus of the items.” (Linacre, 2017b) Therefore, the proportion of items with negative point-measure correlations is also examined. Furthermore, items with very small positive point-measure correlations (between zero and point one) are identified and, because such items “may need further investigation” (R. Green, 2013).

For this study items with point-measure correlations of zero or below, and items with point-measure correlations of.1 or less were identified following R. Green’s guidelines. These items were then counted and their proportion of overall test items for the BERTs and BELTs was calculated. A low proportion of items with negative point measure correlations and of items with point-measure correlations less than 0.1 would constitute backing for this assumption, and high proportion of such items would be regarded as rebuttal evidence.

Further information about Rasch item fit statistics and point-measure correlations is given in section 5.3.10.3, and results and analysis for the backing sought for this assumption are given in section 6.4.3.

4.4.4 Assumption 4 of the explanation inference warrant

Assumption 4 of the explanation inference warrant is that *the range of item difficulties on the test matches the range of abilities of the test takers*. In Rasch analysis it is statistically important for the range of item difficulties of a test to be close to the range of person abilities of test takers. This correspondence is necessary for the abilities of the test takers to be measured precisely, and to minimize measurement error (Bond & Fox, 2007). As the BETs aim to measure reading and listening ability across the range of students in the GE curriculum it is important for the range of item difficulties on each BET to have a relatively even spread across the range of test taker ability. This is also important for the test to be able to separate test takers into separate ability groups for course placement (Linacre, 2017b). Backing for this assumption comes from an analysis of Rasch person/item maps for the BETs. (See section 5.3.10.2 for an explanation of Rash person/item or Wright maps.)

It is important to note that when planning the renewal of items from the 2015 and 2016 BETs for items to be recycled for the following year’s BETs the response probability for person/item maps was set at 50%, with the aim of revising items to produce a range of items that would maximize test reliability. However, for the analyses in this section the response probability is set at 80%, because according to Linacre (2017a) a response probability of 80% is “much less likely to provoke guessing or demotivate”. In addition, an 80% response probability has frequently been used as a definition of mastery (Jones, 2014), so setting the RP at 80% may be useful for BET planners in future for conceptualizing mastery when setting cut-score points at target CEFR levels for course streaming and/or certification, or for linking CEFR can do statements to BET items or testlets for giving diagnostic information to accompany BET scores. Therefore, it is recommended that an 80% RP be used for Wright maps when planning future BET revisions beyond the scope of this study.

4.5 Extrapolation inference

As cited in section 2.3.2, Kane, Crooks and Cohen define the target score as, a test taker’s “expected score over all possible performances in the target domain” (1999, p. 7). The target domain for the BETs is the GE curriculum, so the BETs should be predictive of student performance within the GE curriculum. Thus, the warrant for the BET IUA extrapolation inference is that *performances on the BERTs and BELTs account for the quality of linguistic performance in the domain of General English courses.* The warrants, assumptions, backing sought and potential rebuttals for the BET IUA explanation inference are outlined in Table 4.8. Whether the backing is relevant to the streaming or achievement function of the BETs is noted in the rightmost column.

Table 4.8. *BET IUA Extrapolation Inference*

Extrapolation Inference			
Warrant: *Performances on the BERTs and BELTs account for the quality of linguistic performance in the domain of General English courses.*			
Assumptions underlying the warrant	**Backing sought to support assumption**	**Potential Rebuttals**	**Placement/Streaming or Achievement IUA**
1. Performances on the BET measures are positively correlated with other criteria of language proficiency in General English courses.	1. Large correlations of BET scores with first semester GE grades	Small or non-existent correlations of BET scores with first semester GE grades	Placement/streaming and Achievement
	2. Large correlations of BET scores with speaking test scores	Small or non-existent correlations of BET scores with speaking test scores	
	3. Large correlations of BET scores with GE vocabulary quiz scores	Small or non-existent correlations of BET scores with GE vocabulary quiz scores	

4.5.1 Assumption 1 of the extrapolation inference warrant

The single assumption of the extrapolation inference warrant is that *performance on the BET measures are positively correlated to other criteria of language proficiency in General English courses*. The first backing to support this assumption was sought from correlations between total BET scores and student GE course grades. This section firstly briefly reviews previous literature from the broader field of education on the use of course grades as a criterion for tests. Secondly, literature on using course grades as a criterion for language proficiency tests is summarized. Finally, the methodology used to correlate BET overall grades to GE course grades for the BETs that function as placement/streaming tests within the frame of this study is described.

It is common practice in education to use course results as a criterion for validating placements tests through correlations of test scores with later course results. One of the challenges with the correlation approach is that course results are usually only available as letter grades and their numerical equivalent grade point average (GPA), which results in restricted range, which in turn leads to underestimation of actual correlations (Mattern & Packman, 2009).

Two notable examples of correlating academic skills placement tests with later course results to support test validity are the SAT, and the ACCUPLACER. These are both widely used tests for placement in university and college courses in the US. Kobrin, Patterson, Shaw, Mattern and Barbuti (2008) found an unadjusted correlation of .35 between combined SAT critical reading, mathematics and writing and first year GPA. The adjusted correlation was .53. When including high school GPA along with SAT scores as predictors the raw correlation rose to .46 and the adjusted correlation to .62. To avoid the problem of restricted range on correlations Mattern and Packman (2009) took a different approach to analysing the predictive validity of the ACCUPLACER by analysing the percentage of students achieving passing course grades, and the probability of success in a course for different ACCUPLACER scores. They found that the "mean operational validity for combinations of ACCUPLACER tests was 0.48 when success was defined as obtaining a "B or higher" and 0.40 when success was defined as obtaining a "C or higher," which supports a moderate-to-strong relationship between ACCUPLACER scores and course success" (p. 5).

Many studies have also been conducted, which examine the relationship between English language proficiency tests administered to non-native speakers and their later GPA in higher education institutions in which English is the medium of instruction. These have produced mixed results. For example, studies examining the TOEFL have reported correlations ranging from moderate (r = .40) (Ayers & Peters, 1977) to virtually no relationship (r = .05) (Ayers & Quattlebaum, 1992). Generally, studies report no, or small to medium correlations between commercial tests of academic English and later course GPA at

English language higher education (see Hajr, 2014; and Y. Lee & Greene, 2007 for useful summaries of studies in this area).

There are also some studies reporting none (Y. Lee & Greene, 2007), or small to medium correlations (Davies, 1990; Jochems, Snippe, Smid, & Verweij, 1996; T. Lynch, 2000) between in-house tests and later overall GPA in higher education courses in which that language is the medium of instruction. However, there seem no previous studies which have used correlations between scores on an in-house test of foreign or second language reading and listening ability and later second or foreign language course grades as backing for language test validation.

The BET2 2016 was used as a placement test for GE for students from all of the university's departments except for the Global Communication Department so all departments' results except for GCD were used for the analysis. In 2016 GCD students also took the FE A2–B1 GE course so their GE course results were also included in the correlational analysis for the BET1 2016.

Only first semester GE course grades were used to calculate correlations to seek backing for the first assumption of the extrapolation inference, because second semester GE course grades included a component based on students' end of year BET score. This backing seeks evidence of the relationship between BET scores and other measures of English proficiency in the GE course, so correlating BET scores with a course grade which also includes BET scores would not be appropriate. Details of how GE first semester grades were prepared for correlational analyses can be found in section 5.3.7. Results and discussion of correlations between GE grades and total BET grades are given in section 6.5.1.

The second backing sought for assumption 1 of the extrapolation inference comes from correlations between the BETs within this study used for course placement and class streaming, and the results of speaking tests administered at the ends of semesters 1 and 2.

Although the BETs are only tests of English reading and listening ability, they were used to place students into courses, which in addition to reading and listening goals, also had goals of improving students' productive skills of speaking and writing ability. Therefore, it is important that the BETS should have some predictive ability of students speaking ability, because they are used to place students into two levels of classes with speaking activities targeted at two separate ability levels.

The speaking tests used for these correlational analyses are explained in section 5.3.4. Large correlations between BET scores and later speaking test scores would be taken as strong backing for assumption 1 of the extrapolation inference, and small or non-existent correlations would be taken as rebuttal evidence. These correlations are presented and discussed in section 6.5.2.

The third backing for the assumption beneath the extrapolation inference was sought from correlations between scores for the BETs used for placement and streaming in this study and later vocabulary quiz scores in GE classes. GE vocabulary quiz results are considered a relevant measure of achievement in the GE curriculum with which to correlate BET scores for four reasons. Firstly, vocabulary quizzes were worth 10% of students' final grades in both semester 1 and semester 2 for all GE classes. Secondly, the GE vocabulary lists on which the vocabulary quizzes were based, were made by teachers in conjunction with making GE lessons, and by referring to the English Vocabulary Profile (Cambridge University Press, 2015) to choose words at the A1-B1 levels. Thus, the GE vocabulary lists represent words in the curriculum judged by the teachers who created the lessons to be important for students to learn, which makes the GE vocabulary lists an important part of the GE curriculum, and therefore of the BET domain. Thirdly, the BETs were deliberately written to include as many words as possible from the GE vocabulary lists in order to represent the GE domain, so high correlations with the GE vocabulary tests would provide good evidence of GE vocabulary domain coverage.

Finally, previous research has established a strong connection between L2 vocabulary knowledge and L2 reading and listening ability. For example, a meta-analysis which examined 10 predictor components of passage level reading comprehension by Jeon and Yamashita (2014) found that for 29 published studies which correlated vocabulary knowledge with reading comprehension, there was an overall large correlation ($r = .79$, $p < .01$). This indicated that measures of vocabulary knowledge in these studies accounted for 62% of the variance in reading comprehension scores.

As for the relationship between L2 vocabulary knowledge and listening comprehension, Stæhr (2009) found significant correlations between measures of breadth and depth of English vocabulary knowledge, and English listening ability of Danish advanced learners of English. In Stæhr’s study, vocabulary breadth was measured with Nation’s Vocabulary Levels Test (Nation, 1983) and vocabulary depth was measured with Read’s Word Associates Test (Read, 1993, 1998). Listening comprehension was measured using a listening section from the Cambridge Certificate of Proficiency in English. A multiple regression analysis revealed that together breadth and depth of vocabulary knowledge accounted for 51% of the variance in listening comprehension scores.

Given the intended overlap between the vocabulary covered by the BETs and the GE vocabulary quizzes, and also the strong relationship between L2 vocabulary knowledge and reading and listening ability established in the literature, large correlations between BET scores and later GE vocabulary quiz average grades are required for strong backing for the assumption underlying the BET IUA extrapolation inference warrant. The vocabulary quizzes and available vocabulary quiz data are explained in section 5.3.8. The results for this backing are presented in section 6.5.3.

4.6 Utilization inference

As outlined in section 2.4.2 the utilization inference focuses on whether or not the test effectively fulfils its stated uses. Another way of framing this is that the utilization inference involves an assessment of the extent to which the test has positive consequences for all stakeholders resulting from the decisions made using test scores. As Kane (2013a) states:

> Decision rules are evaluated in terms of their expected consequences, over some population of possible test takers. Decision programs that generally achieve their goals and do not have serious negative consequences are considered acceptable, or valid, and those that do not achieve their intended goals or have serious negative consequences are considered unacceptable, or invalid (p. 455).

Bachman and Palmer (2010) extend consequences further to include other stakeholders in the test. The first claim in Bachman and Palmer's AUA is that "The consequences of using an assessment and of the decisions that are made are beneficial to stakeholders" (p. 158). The warrant for the utilization inference of the BET IUA is drawn directly from Bachman and Palmer's AUA, and is that *uses of BET scores are beneficial to stakeholders*.

It is important to note here that much of the backing sought in the BET IUA argument which has been presented so far in the first five inferences of the BET IUA is also relevant as backing for assumptions in the utilization inference. Full evidence from this study for the course placement/streaming and achievement functions of the BETs from all inferences in the BET validity argument is discussed in sections 7.2.1 and 7.2.2. Additional backing to that sought for the first five inferences, which is relevant to the intended uses of the BETs, is presented in the utilization inference from three of the most important stakeholder groups in the BETs: students, teachers and university administrators. The BET utilization inference IUA is summarized in Table 4.9.

Table 4.9. *BET IUA Utilization Inference*

Utilization Inference			
Warrant: *Uses of BET scores are beneficial to stakeholders*			
Assumptions underlying the warrant	**Backing sought to support assumption**	**Potential Rebuttals**	**Placement/Streaming, Achievement IUA**
1. BET scores are sufficient and relevant for making decisions about GE course placement and class streaming.	1. Results of teacher attitudes surveys show that most teachers think that the BETs place/stream classes effectively	Results of teacher surveys show that most teachers think that the BETs do not place/stream classes effectively	Placement/streaming
	2. Results of a teacher perceptions of their class' ability surveys show clear differences between teacher perceptions of their class' ability between GE courses and class streams	Results of teacher perceptions of their class' ability surveys show little difference between teacher perceptions of their class' ability between GE courses and class streams	Placement/streaming
	3. Teacher comments in focus group interviews indicate that most teachers believe the BETs place/stream classes effectively	Teacher comments in focus group interviews indicate that some teachers don't believe the BETs place/stream classes effectively	Placement/streaming
	4. Rasch person separation measures indicate that the BETs can separate test takers into two distinct ability levels	Person separation measures indicate that the BETs cannot separate test takers into two distinct ability levels	Placement
	5. Results of student surveys show that a majority of students believe that their classmates' English ability is similar to their own and that their GE class is suitable for their level	Results of student surveys indicate that a large proportion of students believe that their classmates' English ability is dissimilar to their own and/or that their GE class is not suitable for their level	Placement / streaming
	6. Phi(lambda) statistics show that the BETs dependably classified students into the two course levels	Phi(lambda) statistics show that the BETs did not dependably classified students into the two course levels	Placement
2. BET scores are sufficient and relevant for assessing student achievement of the GE course goals.	1. Results of teacher surveys show that most teachers think that the BETs effectively measure student reading and listening proficiency	Results of teacher surveys show that most teachers think that the BETs do not effectively measure student reading and listening proficiency	Achievement

	2. Reliability statistics indicate that the BERTs and BELTs can separate test takers into three levels for assessment of the course goals for the two levels of GE courses	Reliability statistics indicate that the BERTs and BELTs cannot separate test takers into three levels for assessment of the course goals for the two levels of GE courses	
3. The BETs have beneficial impact on learning. 3.1. Students benefit from being placed in courses and streamed classes based on BET results.	1. Student survey responses show that most students prefer to be in streamed classes	Student survey responses show that most students prefer to be in mixed-ability classes	Placement/streaming and Impact
	2. Most teachers in focus group interviews agree that students benefit from course placement	Some teachers in the focus groups think that course placement is harmful for some students	
3.2. Students find BET scores to be informative	Student survey results indicate that a majority of students find BET scores to be useful indicators of their English ability	Student survey results indicate that a large proportion of students do not find BET scores to be useful indicators of their English ability	Impact
3.3. Students find BET scores to be motivating.	1. Student survey results indicate that most students are motived to study to improve their BET scores	Student survey results indicate that a large proportion of students are indifferent to improving their BET scores.	Impact
	2. Teachers in focus group interviews generally agree that their students are motivated to study hard to get good BET scores	Teachers in focus group interviews generally disagree that their students are motivated to study hard to get good BET scores	
4. The BETs have beneficial impact on teaching.	1. Teachers in focus group interviews indicate that the BETs have a positive influence on their teaching	Teachers in focus group interviews indicated that the BETs have a negative or little influence on their teaching	Impact
	2. Survey results indicate that teachers often thought about how their teaching would affect their students' BET scores when preparing classes	Survey results indicate that teachers rarely or never thought about how their teaching would affect their students' BET scores when preparing classes	
5. Uses of BET scores have beneficial impact on senior managers' attitudes to the BECC assessment system.	1. Senior managers in semi-structured interviews are in favour of GE course placement and class streaming	Senior managers would prefer mixed ability GE classes	Placement/streaming and Impact
	2. Senior managers in semi-structured interviews agree with presenting students with scores to show their reading and listening ability	Senior managers would prefer other evidence of learning over or in addition to test scores	Achievement and Impact

4.6.1 Assumption 1 of the utilization inference warrant

Assumption 1 of the utilization inference warrant is that *BET scores are sufficient and relevant for making decisions about GE course placement and class streaming.* That the BETs are able to stream students effectively into the A1–A2 and the A2–B1 GE courses, and also into class streams within those courses, is important in order to provide benefits to the most important stakeholder groups of students and teachers.

The first backing sought for this inference comes from results of teacher surveys in which teachers were asked to indicate their level of agreement or disagreement with two statements about the placement/streaming functions of the BETs. These statements were:

a. *The BETs do a good job of streaming students into classes by English language ability.*
b. *Students in the higher-streamed classes I teach clearly have higher overall English language proficiency than students in the lower-streamed classes I teach.*

Details about the administration of the teacher attitudes to the BETs surveys can be found in section 5.3.5.2, and survey results for these two items are discussed in section 6.6.1.

The second backing for this assumption was sought from the answers to another survey in which teachers were asked to estimate the percentage of students in their classes who could effectively perform each of CEFR can do statements from levels A1–B1. If the average proportions of students judged to be able perform can do statements are obviously higher for classes in the A2–B1 course than for classes in the A1–A2 course, this would provide backing for the course placement function of the BETs. Similarly, if the average proportions of students perceived to be able to perform can do statements are clearly higher for higher streamed classes within a GE course, this would provide solid backing for the BET streaming function. The teacher perceptions of GE class ability survey is explained in section 5.3.5.3, and the results ae presented in section 6.6.2.

The third form of backing for assumption 1 of the utilization warrant was sought from teacher comments in focus group interviews. Three questions in the teacher focus group interviews were relevant to assumption 1 of the utilization inference warrant. These were

a. *How effective is the use of BET scores for streaming students for GE classes?*

b. *What are the effects of current GE steaming policy for teacher classroom management, and class preparation?*

c. *How beneficial is the current GE streaming policy for students?*

Details about the administration of the teacher focus groups can be found in section 5.3.6.1, and results are presented and analysed in section 6.6.3.

The fourth backing for this assumption is that Rasch person separation and strata measures indicate that the BETs can separate test takers into two distinct ability levels for course placement, and three or four ability levels for class streaming. Person separation and strata indices are ways to "compute how many statistically different levels of performance can be identified" (Wright, 2001). Wright presents three different separation/strata indices which may be used depending on the situation. Person separation is the most conservative of the three measures and it assumes a normal distribution. Strata reliability assumes that the tales or extremes of the distribution are performance levels and thus results in higher values than the person separation index. Finally, for skewed distributions Wright provides a "Wright strata reliability", which shows "the maximum number of statistically different strata the test can identify." Wright recommends "to choose the one … that makes the most sense in your situation." In the case of the BETs it seems reasonable to assume that the tails of the distribution represent distinct performance levels. This is based on teacher comments which indicate that teachers believe that some of their learners are at the pre-A1 level and that some learners are at the B1 level. In addition, the results of the OOPT placed a few learners at the pre-A1 level and also a few learners at the B1 level. Therefore, Wright's second option of the strata statistic is used as the third backing for assumption 1 of the extrapolation inference warrant. Person separation statistics are also provided as a more conservative estimate. Person separation and strata statistics over two for the BETs used for course placement within this study would provide strong backing to support assumption 1 of the extrapolation inference warrant.

In 2015 the two GE courses were further streamed into a high and a low stream within each course. In 2016 the A1–A2 course was not subdivided, but the A2–B1 course was streamed into high and low classes. Therefore, person separation/strata statistics of 4 or over would be required for the BETs as a whole to support the 2015 streaming policy, and person separation/strata statistics of 3 or over would be required to support the 2016 streaming policy. Person separation and strata statistics for the BETs in this study which were used for course placement and streaming are provided and discussed in section 6.6.4.

The fifth backing for the first assumption of the BET IUA utilization was sought from the results of five items in the student attitudes to the BETs surveys. The five Likert scale items were:

a) I think that the other students in my BECC English class have similar English language ability to myself.

b) In my class, most students’ English skills are similar to mine.

c) I feel that the level of my class is appropriate for my level of English.

d) The level of the classroom handouts (materials I download) is appropriate for my English level.

e) I can clearly understand what my classmates are saying in English.

Results relevant to assumption 1 of the utilization inference warrant are presented and discussed in section 6.6.5. These questions were not included in the April/May 2016 survey administered to recently entered first-year students on the assumption that the students had not yet had enough time in class (around one month) to answer the questions in an informed way.

The final backing sought for assumption 1 of the utilization inference warrant comes from phi(lambda) statistics calculated for the BETs within the frame of this study that were used for course placement purposes. Phi(lambda) is a squared-error loss agreement approach which provides an indication of the dependability of cut scores (Brown, 2005). According to Boyle and Rahman (2013) “Squared-error loss agreement indices are – as a matter of principle – the most appropriate indices for ‘internal reliability’ analyses of CRTs with cut scores” (p. 30). Phi(lambda) can be interpreted as the percentage of correct classification decisions made. For example, a phi(lambda) of .9 would be interpreted as meaning that 90% of test takers had been classified correctly. I could not find any rules of thumb in the literature for what constitutes an acceptable percentage of phi(lambda) for an in-house placement test. Phi(lambda) is also dependent on the distance of the cut-point from the mean test score, with cut points set further from the mean being more dependable and thus resulting in higher phi(lambda) statistics, and cut-points set close to the mean being less dependable and consequently giving lower phi(lambda) statistics. Given that the ability range of GE students is relatively narrow, and the moderate stakes nature of classification decisions based on the test phi(lambda) statistics of around .8, meaning that 80% of students are correctly classified into either the A1–A2 course or the A2–B1 course would be taken as sufficient backing for this assumption. These results are presented in section 6.6.6.

4.6.2 Assumption 2 of the utilization inference warrant

Assumption 2 of the utilization inference warrant is that *BET scores are sufficient and relevant for assessing student achievement of the GE course goals*. The majority of the backing and rebuttal evidence relevant to this assumption is presented and discussed in sections covering the previous five inferences in the BET IUA, and an overall evaluation and discussion of the achievement function of the BETs within the frame of this study is presented in section 7.2.2. In this section, two additional backings sought, which are related to the achievement function of the BETs, are presented. This first backing comes from answers to two questions on the teacher attitudes to the BETs surveys. The two questions are:

a) I think the BETs are an effective way to measure GE students' English reading proficiency.

b) I think the BETs are an effective way to measure GE students' English listening proficiency.

These two questions are both directly relevant to assessing the ability of the BETs within the frame of this study to act as effective achievement tests of the GE course reading and listening goals. The results for these two Likert scale items are presented and discussed in section 6.6.7.

The second backing for this assumption was sought from the reliability statistics, and resulting strata statistics, for the BERTs and BELTs. The BERTs and BELTs aim to assess achievement of the GE course goals for reading and listening for both the higher (A2–B1) stream and for the lower (A1–A2) stream. In order to separate test takers into groups who had achieved the B1 goal of the higher GE course, the A2 goal of the lower GE course, and those who had not achieved the course goals the BERTs and BELTs would have to be able to separate test takers into three groups. As explained in section 4.3.1 this would require reliability statistics for the BERTs and BELTs of at least .8. Reliability statistics as backing for this assumption are discussed in section 6.6.8.

4.6.3 Assumption 3 of the utilization inference warrant

Assumption 3 of the BET utilization inference warrant is that *The BETs have beneficial impact on learning*. This assumption is broken down into three sub-assumptions.

4.6.3.1 Sub-assumption 3.1

The first sub-assumption is that *students benefit from being placed in courses and streamed classes based on BET results*. Backing sought for this sub-assumption comes from student and teacher opinions.

The first backing sought for this assumption comes from answers to a question on the student attitudes to the BETs survey which asked students whether they would prefer to study in a class with a wide range of English ability or a class of students with English ability similar to their own ability.

The item in English and Japanese follows.

この 17 問目の質問には、自分の意見に当てはまるものにチェックをいれてください

1. もし自分で選ぶことができるのなら、私は…

a) 英語のレベルが自分とほぼ同じ学生のクラス

b) どちらでもよい

c) 上級、中級、また初級の学生がそれぞれいるレベルが混ざっているクラス

For question 17, choose the answer which matches your opinion.

If I could choose, I would prefer to be …

a) in a class with students whose English skills are about the same level as mine

b) I don’t have a preference either way

c) in a mixed-level class, with some students having more advanced level English skills and others having intermediate or beginner level English skills.

This item was adopted from previous research on student streaming preferences by Joyce and McMillan (2010). A large majority of students expressing a preference for being in a class with students of similar ability to their own would provide strong backing for sub-assumption 3.1 of the utilization inference on the assumption that meeting students preferences for course streaming is beneficial for student motivation and learning. Proportions of students who chose each of the three options are presented and discussed in section 6.6.9.

The second backing sought for sub-assumption 3.1 of the utilization inference warrant comes from the responses of teachers in the focus group interviews. Two questions in the focus group interviews were relevant as backing. These were:

a) What are the effects of current GE steaming policy for teacher classroom management, and class preparation?

b) How beneficial is the current GE streaming policy for students?

Comments from teachers in the focus group interviews which generally see the GE course placement/streaming policy as beneficial would provide backing for this assumption. These results are presented and discussed in section 6.6.10.

4.6.3.2 Sub-assumption 3.2

The second sub-assumption related to the positive effect on learning is that *students find BET scores to be informative*. Backing to support this assumption was sought from an analysis of answers to the following five statements in the student attitudes to the BETs and class streaming surveys.

a) I think my BET results help me to know what I can do in English.

b) I think my BET scores help me to know my English proficiency level in reading, listening (and grammar).

c) I think my BET scores help me to know my weak points in English.

d) I think my BET scores help me to know my strong points in English.

e) My BET scores are useful to help me to plan my English study

A high proportion of survey takers agreeing to these statements and a generally strong level of agreement would provide solid backing for a claim that the BET scores provide students with useful information about their English ability. The results for these survey statements are presented and discussed in section 6.6.11.

4.6.3.3 Sub-assumption 3.3

The third sub-assumption for assumption 3 of the utilization inference is that *students find BET scores to be motivating*. The first backing for this sub-assumption was sought from the answers to two Likert scale statements in the student attitudes to the BETs and class streaming surveys, as follow.

a) I want to improve my BET score on my next BET.

b) I'm motivated to study harder to improve my BET score.

A high proportion of students agreeing to these statements, along with a high degree of agreement would provide solid backing for this assumption. The relevant results are discussed in section 6.6.12.

The second backing for this assumption was sought from teacher answers in the focus groups. The question relevant to this assumption was "To what extent do you think BETs affect students' motivation for learning English language?" Relevant answers from the focus groups are presented and discussed in section 6.6.13.

4.6.4 Assumption 4 of the utilization inference warrant

Assumption 4 of the warrant for the BET utilization inference is that *the BETs have beneficial impact on teaching*. The first backing for this assumption was sought from answers elicited from teachers in the two focus groups conducted in 2015 and 2016. The questions which aimed to elicit information about how the BETs affected teaching practice in the GE courses was *How do you try to prepare your GE students for the BETs?* The 2015 focus group interview also had a second question related to this assumption which was *How much class time do you spend on preparing your students for the BETs*?

If teacher responses in the focus groups indicate that the BETs have a positive influence on their teaching, this would provide backing for this assumption. On the other hand, if teachers indicate that the BETs have a negative influence on the teaching or little or no influence this would provide rebuttal evidence for this assumption. An analysis of teachers' answers to these questions in the focus groups is presented in section 6.6.14.

The second backing for assumption 4 of the utilization inference warrant was sought from the results of the anonymous teacher surveys described in section 5.3.5.2. The survey question relevant as backing for this assumption was: *I think (thought) about how my teaching will (would) affect my students' BET scores when preparing my GE classes.* (the wording was slightly different depending on when the survey was administered). If teachers indicate that they generally did not think about the BETs when preparing their classes, this would provide rebuttal evidence for assumption 6 of the utilization warrant. These results are presented and discussed in section 6.6.15.

4.6.5 Assumption 5 of the utilization inference warrant

The fifth assumption of the utilization warrant is that *uses of BET scores have beneficial impact on senior managers' attitudes to the BECC assessment system*. The first backing for assumption 5 of the utilization warrant was sought from the answers to a question in semi-structured interviews with HBWU senior managers. The question was: "What do you think about streaming English classes for GE courses?". If managers show general agreement with the placement/streaming policy for GE classes this would provide backing for this assumption, and consequently for the placement/streaming function of the BETs. Ambivalent or negative opinions toward course placement and class streaming from senior managers would provide rebuttal evidence for this assumption. The methodology of the management interviews is explained in section 5.3.6.2, and the results relevant to assumption 5 are presented in section 6.6.16.

The second backing for assumption 6 of the utilization inference warrant backing was sought from answers to the question in the management semi-structured interviews: "How important is it to give students numerical scores to show their English ability in reading, listening and speaking?" If interviewees agree on the value of presenting students with numerical scores representing their reading and listening ability, this would provide backing for this assumption and therefore backing for the achievement function of the BETs from the perspective of this important group of stakeholders. Results from these interviews are presented in section 6.6.17.

CHAPTER 5

Materials and Methods

5.0 Introduction

This chapter explains the materials and methods used for this study. Firstly, the study participants and ethical consent procedures are described. Then, the materials used to seek backings for the BET IUA are described, and the methods used to analyse the data are explained.

5.1 Participants

Potential students to participate in this study were all first year GE students in the 2015/16 academic year, and all first and second year GE students in 2016/17 academic year. From examining GE class lists it was found that in the 2015/16 academic year, 269 students studied the first year of the new General English curriculum. For the first time in the 2016/17 academic year English language majors in the Global Communication Department also took the new GE curriculum resulting in a total of 274 students. 274 students also took the second year of the new GE curriculum in the 2016/17 academic year. The number of students studying the second-year GE curriculum in 2016–2017 is slightly higher than the number of first year students studying GE in 2015–2016 due to some repeating students. All students in the frame of this study were female Japanese nationals around the ages of 18 or 19, as no foreign students were enrolled in the university over this time.

Table 5.1. *Numbers of Students Taking the GE Curriculum Within the Frame of this Study*

Department	**2015 GE Freshman Curriculum**	**2016 GE Freshman Curriculum**	**2016 GE Sophomore Curriculum**
Early Childhood Education	125	111	127
Welfare	52	39	54
Psychology	22	25	22
Nutrition	70	78	71
Global Communication	NA	21	NA
Total	269	274	274

BECC teachers and learning advisors also contributed their opinions to this study through surveys. At the start of the 2015/16 academic year there were 13 full-time teachers and learning advisors (LAs) working at the BECC. Specifically, there was a BECC Director, a BECC Assistant Director (the researcher of this study), nine teachers and two full-time learning advisors. The Assistant Director also taught BECC classes, and the BECC Director taught BECC classes and also did some learning advising. 12 of the teachers and learning advisors had MAs in applied linguistics, education or TESOL, and one had an MA in Japanese Language and Society. Of these 13 full-time staff, five were female and seven were male. The staff were a mixture of nationalities consisting of five Americans, one Canadian, two British, two Japanese, two New Zealanders, and one Australian.

In the second semester of the 2015/16 academic year the BECC Director and one learning advisor left the BECC, one of these staff was replaced by a new male American teacher with a Masters in TESOL and a new American LA with a Masters in Business Administration, bringing the total number of American staff to seven out of 13.

Finally, in 2016 one American teacher left the BECC and two new teachers of Filipino nationality one with a PhD in English language, and one with an MA in applied linguistics joined the BECC, bringing the total number of teachers and learning advisors to 14. For this study, these BECC teachers and LAs are considered to be experts on students who took the GE curriculum, and those teachers who taught the GE curriculum are considered to be experts on the GE curriculum.

More details about the participants for each of the data collection methods used are presented later in this chapter, when explaining the collection methods.

5.1.1 Ethical issues

All students who took the 2015 BET1 and the 2016 BET1 were requested to fill in and sign an informed consent form after the test, which gave permission to use their BET results for research. The informed consent form was in the students' native language of Japanese, and a Japanese native speaker was present to answer any student questions. Copies of the student informed consent form in both English and Japanese are attached as Appendix B.

All BECC teachers who took surveys related to the BETs, and/or who participated in the focus groups were also asked to fill in informed consent forms. Copies of the informed consent forms for instructors are attached as Appendix C.

In addition, HBWU senior managers who agreed to participate in semi-structured interviews completed informed consent forms. Copies of the informed consent forms for university senior managers are attached as Appendix D.

Approval for the study was granted by both the Hiroshima Bunkyo Women's University Ethics Committee and the Human Research Ethics Committee of Macquarie University (attached as Appendix E.)

5.2 Mixed-methods approach

This study employs a mixed-methods research (MMR) approach to gathering and analysing data (See Riazi and Candlin, 2014 for a thorough overview of MMR in linguistics and language education research). The MMR approach taken in this study is what Greene, Caracelli & Graham (1989) define as a *complementarity* approach. In a complementarity MMR approach a combination of qualitative and quantitative research methods is used to gain insight into different levels and aspects of a phenomenon. Caracelli & Graham explain that the different quantitative and qualitative data sought in a complimantry MMR aproach are best gathered concurrently, as was the case in this study. MMR is used in the BET validity argument to seek backing for varied inferences and assumptions.

5.3 Materials and data collection

This section firstly presents a table which summarizes the sources of evidence to be used in the BET validity argument. Secondly a brief outline of each source of evidence is presented.

Table 5.2 provides a summary of the types of evidence analysed in this study.

Table 5.2. *Sources of Evidence Used in the BET Validity Argument*

Data Source		**Description**
GE and BET Administrative Documents	BET domain coverage documents	Documents made during BET design, which show the units from which lessons were chosen, on which to base BET item and testlet content
	The BET specifications	A document describing the structure and purpose of the BETs, BET administrative procedures, and giving specifications for how to write BET testlets. Three iterations of the BET specifications from the 2014/15, 2015/16 and 2016/17 academic years were examined
	GE course outlines	Separate documents for each of the GE first-year course and second-year course which describe the course goals, timeline, assessment schedule, homework and attendance policy etc
	Tables showing BET task type representation in the GE curriculum	Documents made as part of the new GE curriculum design and revision process in 2015 and 2016, which show how BET type tasks are represented across the GE curriculum
BET Documents	BET question booklets	Copies of the BET question booklets for the BET1 2015, the BETs 1 & 2 2016, and the BETs 1–3 2017
	BELT tapescripts	Tapescripts for listening passages from the BELTs within the frame of this study

Language Proficiency Tests	BET results	Results of the BET1 2015, the BETs 1 & 2 2016 and the BETs 2 & 3 2017
	BEST results	Results of in-house speaking tests known as the BESTS, specifically 2015 BEST2, and the 2016 BESTs 1–4
	Oxford Online Placement Test	The OOPT, a standardised, computer adaptive, CEFR-aligned test, was administered to a representative sample of GE students who took each of the BETs within the frame of this study
	TOEIC	TOEIC scores were available for a convenience sample of students in the Global Communication Department
Surveys	Student Attitudes to the BETs and Class Streaming Survey	Surveys of students' opinions about the BETs and GE class streaming were administered around the time of each BET within the frame of this study
	Teacher Attitudes to the BETs survey	Surveys of teacher and LAs opinions of the BETs and class streaming, which were administered three times over the time period covered by this study
Interviews	Focus Group Interviews	Focus groups interviews about the BETs were conducted in May 2015 and again in August 2016
	Semi-structured Interviews	Semi-structured interviews were conducted with key senior managers at HBWU
Course Grades	Student grades for the GE course	Grades for the first semester of the FE and GE courses were used for correlational analyses
Vocabulary Quizzes	Scores from GE unit vocabulary quizzes.	Results of vocabulary quizzes on important vocabulary encountered in lessons, which GE students took at the end of each GE unit
Statistical Analyses	Correlational analyses	Spearman and Pearson correlations of BET scores with several criteria
	Rasch Analysis	Rasch analysis was used for item analysis, separation statistics, and for creating Wright maps
Vocabulary Profiling	Vocabulary frequency profiles	Vocabulary frequency profiles of BET question booklets and tapescripts were produced using tools from the Compleat Lexical Tutor website

5.3.1 BECC administrative documents

Several BECC administrative documents are drawn on for this study. These documents are briefly described in the following sections. As Bachman and Palmer (2010) explain, such administrative documents can be an important form of backing to support warrants in an argument-based approach to validation.

5.3.1.1 Tables showing the representation of GE curriculum content in BETs

Google sheets were used to enter which unit and lesson titles were used as the basis for designing BET items and testlets. The work of writing/rewriting the BETs was divided amongst writers by testlet, and testlet writers were responsible for entering the lesson on which they based their new or recycled testlet in the document. The sheets for the BETs 1 & 2 2016 listed which unit of the six units in the 2015–16 FE curriculum each item in the test was based on, and a formula was used to add up the totals and to calculate a percentage of representation of each unit in the BET1 and BET2. As only the first year of the new curriculum was available for designing the 2016 BETs, only the FE material is covered by the BET1 and BET2 2016. Items that drew on content from more than one unit were entered as "multiple."

The same process was followed in the planning and item writing stages of the 2017 BETs. As the newly designed second-year or Sophomore English (SE) materials for the new GE curriculum were used in the 2016/17 academic year, all three BETs 1, 2 and 3 were included in the analysis. The longer-term plan for the BETs was that the BET1 will contain 50% FE content and 50% SE content to act as a baseline achievement test of the whole GE curriculum, so that student progress can be measured across two years of study. The BET2 would contain 100% FE content to act as an achievement test of students first year of study, and the BET3 would contain 50% FE and 50% SE content to act as a final achievement test of the whole GE curriculum. Due to limited teacher hours available for revising the 2017 BETs it was decided that the BET1 2017 would be based only on the FE curriculum to limit the amount of new items that had to be written, the BET2 would also be based on the FE curriculum to act as an achievement test of the first-year of students' studies, and the BET3 2017 would contain 50% FE and 50% SE materials to act as an achievement test for the first cohort to complete the new GE curriculum.

For the BET1 2015, a table was created showing the units and lessons on which each testlet was created post hoc by the researcher as no such table was available from the creation process. While the BET1 2015 was being made the first semester of materials for the new curriculum was also being made, so it was not possible to base the BET1 2015 on the whole of the new curriculum. Instead the BET1 2015 was based on the first semester of materials that was undergoing creation, as well as those lessons that it was thought were likely to remain in the new curriculum, based on the curriculum design plan. Therefore the researcher analysed the BET1 2015 content post hoc against the actual GE curriculum completed by the students who took the BET1 2015 over two years, to see the extent to which the items in the BET1 2015 actually represented the themes and content of the revised curriculum that the students went on to study. Summary tables for the BETs within this study are attached as Appendix F. As noted in section 4.1.2 for the BETs to act as effective placement and achievement tests, it is very important that the BETs cover a broad and representative sample of the GE curriculum. Therefore these tables were used to seek backing for the first assumption in the BET IUA domain definition inference.

5.3.1.2 The BET specifications

The BET specifications are a document which outlines the BET purpose, structure and content, as well as including proctoring guidelines for the administration of the tests. Test specifications are important to define the overall purpose of a test, to give clear instructions for standardised test administration, and to describe how tasks should be written (Bachman & Palmer, 1996, 2010; Carr, 2011). Test specifications thus form the bedrock of test development, and as such are an essential component of backing for some assumptions in the BET IUA. The first version of the BET specifications was made in 2014 as part of the process of making the 2015 BETs. The specifications were also updated periodically over the span of BET development covered by this study. Three versions of the BET specifications were available for analysis in this study. These are BET specifications 2014–15 dated as being updated on August 23, 2015, and the BET Specs 2015–16, dated as being updated on January 15, 2016, and the BET specifications 2016–17 dated as being updated on September 22, 2016.

5.3.1.3 Can do statements for testlets in the BET specifications

Analysis of the can do statements, which were listed in the BET specifications for each testlet type to show the reading or listening CEFR can do descriptor targeted by each testlet were analysed to produce backing sought for assumption 1 of the domain definition (presented in section 4.1). For this backing, the can do statements that BET testlets claim to test for three versions of BET specifications from 2014–2016 were listed and analysed in a table. The can do statements were then analysed by the researcher for their fit to the curriculum goals, and their match to the testlet task. Can do statements that did not appear to match the curriculum goals of CEFR levels A2 and B1, or that did not appear to match the task requirement of a testlet were identified and suggestions for more appropriate can do statements in terms of matching the testlet task requirements and the curriculum goals of CEFR A2 and B1 proficiency were suggested as needed. In some cases, both in the existing BET specifications, and in the suggestions for improved target can do statements which were made by the researcher, EAQUALS can do statements were used instead of CEFR can do statements if an EAQUALS can do statement was seen to provide a better description of a task requirements than a CEFR descriptor. EAQUALS can do statements are revised versions of the CEFR can do statements developed for the European Language Portfolio project (EAQUALS, 2017). The results of this analysis are attached as Appendix G and are discussed in section 6.1.3.

In addition, these checked can do statements along with the suggested revisions were compared to the detailed CEFR reading and listening can do statements at the curriculum target levels of A2 and B1 to determine the actual BET coverage of the can do statements at the target CEFR levels. This analysis was done to seek the second backing for assumption 2 of the BET domain definition inference warrant. Results of this analysis are presented in Appendix H and discussed in section 6.1.4. The can do statements in the BET specifications were analysed, because they form a key link between BET testlets and the aspects of the CEFR which the BETs claim to test. Thus, the can do statement for BET testlets are an essential component of arguing that the BETs are able to place student proficiency at the target CEFR levels.

5.3.2 BET Documents

5.3.2.1 BET question booklets and tapescripts

BET question booklets, which contain test instructions, reading passages, items, keys and distractors for the BETs within the frame of this study were used for vocabulary analysis. Tapescripts, which have transcripts of the listening passages used in the BELTs within the frame of this study, were also used for vocabulary analysis. For comprehension of reading and listening texts it is important that the majority of vocabulary be comprehensible to readers, therefore the vocabulary in the BET question booklets and tapescripts was analysed as explained in section 5.3.10.4, to assess if the level of vocabulary was appropriate to the test target CEFR levels.

5.3.2.2 BET task and curriculum task comparison tables

As explained in section 4.1.4, it is important for both achievement and placement tests that test task types are similar to tasks in the target domain. Thus tables showing the number of tasks in the GE curriculum which were similar to BET tasks were analysed to seek support for assumption 3 of the BET IUA domain definition inference.

Tables were made to count the number and dispersion of tasks across the GE curriculum which were similar to BET tasks as part of the curriculum design and review process in 2015 and 2016. To make the tables the teachers responsible examined the listening and reading tasks in the classroom materials and compared them to the testlet types specified in the BET specifications. Differences between the lesson tasks and BET testlet types were noted in the table, and if a lesson task was judged to be sufficiently similar to a BET testlet type it was counted. The analysis focused on features such as the question response type (i.e., the BETs have multiple choice questions, so only multiple choice questions were counted), the amount of words in a text, the number of texts for a task, and the format of the task, for example, if it included an example question.

However, BET review lessons which were given to students before their BETs, were not included in the analysis, so they were added post hoc by the researcher. In addition, periodic listening, reading, vocabulary and grammar assessments in the curriculum were not included in the counts, so these have also been added by the researcher. For example, GE students also take three vocabulary quizzes per semester, and these quizzes each contain 5 items very similar to BERT Part 2. Therefore, BERT Part 2 counts for these vocabulary quizzes were also added to the tables in the assessments category by the researcher.

Tables showing the representation of BET testlet type tasks in the 2016 GE materials were made as part of preparation for revising the GE materials for implementation in the first semester of 2017. As a consequence, the counts for first semester materials were based on revised materials to be used from 2017, and only the semester 2–4 materials counts represented the GE 2016 curriculum. As such, it was necessary for the researcher to check and enter tasks which were similar to BET testlets for FE 2016 semester 1 only. The analysis for semester 2–4 lessons was left as is, because it represents expert teacher opinion of the 2016 GE materials. The researcher also checked that the interpretation of whether lesson tasks sufficiently matched BET testlet specifications to be counted was consistent across the two years of analysis for lessons which were unchanged, in cases in which there was a differing analysis, the second-year interpretation took precedent. The tables are attached as Appendix I.

5.3.3 BET results

The results of the 2015 BET1, the 2016 BETs 1 & 2 and the 2017 BETs 2 & 3 are analysed in this study. The format of the tests is summarized in sections 1.5.4–1.5.6. The number of students who gave permission for their results to be used for this research is summarized in Table 5.3. The BET1s, and the BET2 2017 were taken by students from all five of the university's departments, but the BET2 2016 and the BET3 2017 were not taken by students in the GCD department, which reduced the overall number of students who took these tests.

Table 5.3. *BET Results Available for Analysis*

BET	Number of Test takers giving informed consent
BET1 2015	261
BET1 2016	246
BET2 2016	222
BET2 2017	239
BET3 2017	217

BET results were analysed using Rasch analysis, as explained in section 5.3.10.2. Reliability and dependability statistics for the BETs examined in this study are give in table 6.1 in section 6.3.1.

5.3.4 Bunkyo English Speaking Test results

As explained in section 4.5.1, results of speaking tests aligned to the curriculum target CEFR levels were correlated with BET scores to seek backing for the assumption beneath the BET IUA extrapolation inference warrant. This backing was sought because the BETs 1 and 2 in this study were used for placing and streaming students into courses and classes which had speaking proficiency goals, and which also had a large spoken communication component. Therefore, to justify the placement function of the BETs, it is important to establish a strong relationship between BET scores and English speaking ability, as speaking ability is a major part of language performance in the BET domain.

In conjunction with the BETs speaking tests aligned to the new GE curriculum were introduced in 2015 and 2016. These tests were labelled as the Bunkyo English Speaking Tests (BESTs) and form part of the BET suit of exams in addition to the BERTs and BELTs. As with the BERTs and BELTs, BEST tasks were designed by adapting the style of tasks in the KET and PET, in this case the speaking sections of these tests. The BEST is a standardised test with a paired format in which a pair of two students is double graded by two examiners. One examiner employs a holistic rubric, and the other examiner employs an analytic rubric. The holistic rubric is graded out of 5 and the analytic rubric has three categories each graded out of five. The final test score is weighted by doubling the holistic rating to a score out of ten, adding it to the analytic score out of fifteen for a total out of 25. A final score is then calculated out of 15, by dividing scores by 25 and then multiplying them by 15. This means that the holistic score has a weighting of 40% and the analytic score has a weighting of 60%. Each BEST is administered at the end of a semester, and is designed to represent tasks from that semester’s GE materials. BEST speaking tasks are intended to represent a wide sample of topics and tasks from the target semester of the GE curriculum.

The BEST1 2015 was not used for analysis as student numbers were not recorded with test grades, which complicated matching BET scores to BEST scores in this case. All other BESTs within the frame of this study were used for analysis.

Space restriction do not permit a detailed validity argument for the BESTs to be presented here. However, the facts that BEST rubrics were base on CEFR can do statements at the A1-B1 levels, that teachers were given standardised rater training before each BEST administration, that BEST tasks were thoroughly edited during the creation process to ensure maximum coverage of the target curriculum and consistency between tasks, and that the BESTs have reasonable Chronbach's alpha reliabilities as shown in table 5.4 indicate that the BESTs are a reasonably reliable measure of speaking ability in the BET target domain of the GE curriculum. The reliabilities of the BESTs 1 and 3, and the BESTs 2 and 4 are very similar, and this is likely due to these tests having the same examiners, who received the same training, and that the tests were the same format and were administered around the same time.

Table 5.4. *BEST Reliabilities*

BEST	Chronbach's Alpha
BEST2 2015 (*n* = 223)	.87
BEST1 2016 (*n* = 245)	.78
BEST2 2016 (*n* = 242)	.87
BEST3 2016 (*n* = 217)	.78
BEST4 2016 (*n* = 215)	.88

5.3.5 Surveys

5.3.5.1 Student attitudes to the BETs and class streaming surveys

Student attitudes to the BETs and class streaming surveys were administered five times across the two-year timeframe of this study, soon after BET administrations. The survey was administered using the online survey platform SurveyMonkey, and was taken by students in class on their iPads when BECC teachers were able to spare the lesson time. Students were asked to indicate how strongly they agree with a set of statements on a scale from 1–6, in which 1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = somewhat agree, 5 = agree, 6 = strongly agree. The survey was administered in Japanese. Statements from the surveys are attached as Appendix J. The questions are presented in both their Japanese translation for the survey administration, and in the original English. There were a few differences in questions across administrations, and some additional questions were added for the 2017 administration. Therefore, the survey administration(s) for which statements were used, are indicated with check marks in the columns to the right of the table.

To analyse the survey data, results for students who opened the survey but who did not respond to the items were first removed. Responses which consisted of a single numeral for all answers (e.g. all fives) were also removed, as these respondents had probably not thought about their answers and simply entered the same number to complete the survey as quickly as possible. Data for respondents who had missed some questions was left in the analysis. Responses for those students who had not given permission for their survey results to be used for this research were also removed. The dates of the survey administrations for which data was used for this study, and the total number of useable responses available are given in Table 5.5, along with Chronbach’s alpha for the usable survey results, including all answers with a 1-6 scale.

Table 5.5. *Number of Usable Responses to the Student Attitudes to the BETs and Class Streaming Survey*

Survey administration dates	Number of usable responses	Chronbach's Alpha
April 30–May 7, 2015	175 first-year students	.95
January 21–February 2, 2016	206 first-year students	.93
April 18–May 1, 2016	197 first-year students	.93
January 20–January 23, 2017	215 first-year students	.94
	193 second-year students	.95

5.3.5.2 Teacher attitudes to the BETs surveys

Teacher attitudes to the BETs surveys were administered three times across the period covered by this study. As with the student surveys, the survey was administered using SurveyMonkey. Teachers and LAs were asked to indicate how strongly they agreed with a series of statements which were later converted to a scale from 1–6 for analysis. The statements were: 1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = somewhat agree, 5 = agree, 6 = strongly agree. Survey administration dates and the numbers of respondents are listed in Table 5.6. For the final question on the survey respondents were asked to indicate how often they thought about how their teaching would affect their students' BET scores when preparing their GE classes. These statements were also later converted to a scale from 1–5 for analysis. The statements were 1 = never, 2 = rarely, 3 = occasionally, 4 = usually, and 5 = always. To avoid any possible bias or conflict of interest the researcher did not take any of these surveys. It was important to gather anonymous teacher attitudes to the BETs data, as teachers are a key BET stakeholder group.

Table 5.6. *Teacher Attitudes to the BETs Survey Administration: Dates and Numbers of Respondents*

Survey administration dates	**Number of respondents**
June 29 – July 10, 2015	9
July 4 – July 29, 2016	12
January 10 – February 6, 2017	11

Statements used in the teacher attitudes to the BETs surveys are attached as Appendix K.

5.3.5.3 Teacher perceptions of GE class ability survey

This survey asked teachers to estimate the proportion of students in their GE classes who could perform each of the CEFR can do statements from levels A1, A2 and B1 by choosing a percentage from 0% to 100%, with choices rising in increments of 10%, (i.e. 0%, 10%, 20%, 30%, ... 100%). If teachers felt that they did not have enough information to judge their class' ability on a can do statement, they were instructed to choose an N/A option. All can do statements from the CEFR self-assessment grid for levels A1–B1, for reception, interaction and production were included in the survey for a total of 27 can do statements. Due to limitations of space these can do statements are not included here but they can be found online in the *Structured Overview of all CEFR Scales* (Council of Europe, 2001b, p. 6).

The survey was administered online through SurveyMonkey three times, firstly in June/August 2015, secondly in July 2016 and finally in January 2017. Each survey was analysed separately by administration date for the FE and SE courses. Any can do statement that was entered as N/A was removed from the analysis, to ensure that the same can do statements were consistent for each analysis. Estimated proportions for the remaining can do statement were then averaged for each class, and lastly combined averages were made for each GE course and class stream. The results are presented in section 6.6.2. This survey was considered important to include, because teachers perceptions of their students' ability provides a further avenue to triangulate teacher perceptions of the effectiveness of the BETs for class streaming/placement purposes.

5.3.5.4 CEFR self-assessment survey

A self-assessment survey using CEFR can do statements for the levels A1–B1 was administered to students within a month of each of the BETs examined in this study. Can do statements for the survey were taken from the CEFR self-assessment grid for these three levels for the skills of listening, speaking, reading and writing. The survey was administered online using the online survey platform SurveyMonkey, and was taken by students on their iPads. All BECC teachers were requested to run the survey with their GE classes. The entire survey was delivered in Japanese including the instructions and the questions. Respondents were asked to indicate the extent to which they agreed that they were able perform the can do statements when communicating in the English language on a Likert scale, which were later assigned numerical values for statistical analysis. The Likert scale categories were strongly agree = 6, agree = 5, somewhat agree = 4, somewhat disagree = 3, disagree = 2 and strongly disagree = 1. The translated CEFR can do statements were taken from an official translation of the CEFR into Japanese (Council of Europe 2004/2010). Minor adjustments were made to expressions in four of the translated can do statements, to make them easier for Japanese university students to understand.

Only results for the can do statements for reading and listening are used for this study, as they are directly relevant as evidence for the BERTs and BELTs as tests of reading and listening. These can do statements in their original English and also in their translated Japanese versions are given in Appendix L. The survey can do statements for speaking and writing are not included due to considerations of relevancy and space.

To prepare the survey data for analysis entries with no answers, or with most answers missing were removed. Data for respondents that had chosen the same response for all answers (e.g. all ‘somewhat agree’) or for more than 10 answers in a row were also removed, as these respondents had most likely not thought carefully about their survey answers and had just rushed through the survey in order to complete it as soon as possible. However, responses from survey takers who had entered all ‘strongly agree’ or all ‘strongly disagree’ were left in, as these responses might indicate students who perceived their ability to be higher than B1 or lower than A1. Responses from survey takers who did not give permission for their survey results to be used for this research were also removed. Several respondents seemed to have taken the survey twice, so in these cases the second response for each survey taker was removed, unless the respondent had only partially completed the survey the first time then fully completed the survey the second time, then in this case the second response set was kept.

After cleaning the data as outlined above, responses for writing and speaking can do statements were removed. This left a total of six can do statements for reading and five can do statements for listening. The reading and listening can do statements are each treated as a separate survey for this analysis, and Cronbach’s alpha reliability for the total reading survey and the total listening survey were calculated using Winsteps. The number of usable survey responses, as well as Cronbach’s alpha for each of the surveys relevant to this study are given in Appendix M. All Cronbach’s alpha statistics were above .8 for the reading and listening can do statement sections of the surveys, which is within the range (.7–.9) commonly considered a sufficient to indicate enough internal consistency for survey instruments. Before running the correlations BERT and BELT items with point measure correlations of .05 for below were removed as these items may not be effectively measuring the construct of interest. Also, an attempt to correct to measurement error, disattenuated correlations were calculated for correlations between BERT, BELT and self-assessment survey results for reading and listening. Reasons for choosing student CEFR self-assessment data as a criterion for this study were presented in section 4.4.2.

5.3.6 Interviews

5.3.6.1 Teacher focus group interviews

Focus groups are a qualitative research method in which a group of participants is encouraged to discuss a topic under investigation and to talk freely about the topic (Giliflores & Alonso, 1995). Two focus group interviews were conducted in the course of this study. The first focus group took place on May 15, 2015 shortly after the BET1 2015 was administered. The second focus group took place on August 2, 2016. The same moderator led both focus group discussions. An outside moderator was chosen in order that participants would feel free to discuss the topics without judgement. The moderator was an experienced researcher, who worked as an English language instructor at another university. She had a PhD in linguistics, and an MA in TESOL. 13 questions were discussed by the two focus groups, and three questions changed for the 2016 focus group in an attempt to elicit better backing for assumptions in the BET IUA.

There were six teacher participants in the 2015 focus group, all of whom taught the GE curriculum that year. No teachers with managerial roles were invited to participate in the group, in order to keep it as homogeneous as possible and to avoid any power differential between participants which may impede a free exchange of opinions (Krueger & Casey, 2009). Members of the GEAC were also not invited to participate in the 2015 group to keep it homogeneous, which is thought to promote better discussion (Krueger & Casey, 2009). The 2016 focus group also had six participants, however this time GEAC members were invited because with the expanded GEAC committee size in 2016, there would not have been enough members to meet the recommended minimum focus group size without including GEAC members. The focus group in 2016 consisted of six BECC teachers all of whom taught the GE curriculum that year. Four of the teachers were on the GEAC and two were not. Once again teachers in management positions were not invited to participate in order to avoid a power differential between participants.

All focus group participants were compensated for their participation with a 3000 yen iTunes voucher. The informed consent form for focus group participants is attached as Appendix C. The focus group questions are attached as Appendix N.

Focus groups were chosen as qualitative data source for this research, because this technique provides a variety of perspectives on a topic (Ho, 2010). Also, due to the somewhat somewhat freeform nature of focus groups, they can shed light on issues which may be missed in a survey, and can also provide more in-depth results than a survey (Morgan, 1996).

5.3.6.2 Senior manager semi-structured interviews

Five senior managers at HBWU were contacted to participate in semi-structured interviews regarding their knowledge and opinions of the BETs over the course of this study. All five managers agreed to be interviewed. The exact administrative positions of these administrators within the university are not revealed in this study to maintain the participants' anonymity. However, it can be revealed that all of the managers had positions in the top one or two management tiers within the university.

The interviews were semi-structured. Semi-structured interviews are interviews in which fixed questions are asked of interviewees, and the interviewer also asks free form follow-up questions, depending on an interviewee's responses. (see Ho, 2012 for a succinct introduction to semi-structured interviews.) After the first interview, the questions were reconsidered in order to make them more of a discussion nature, rather than direct questions, in an effort to make the interviewees feel more at ease, rather than feeling challenged about their level of knowledge of the GE curriculum. Questions which were asked in the first semi-structured interviews, and the revised questions for the following four interviews are attached as Appendix O. A semi-structured format was chosen for the senior management interviews, because it allows the interviewer to ask for clarification and to elicit more detail on areas of interest (Ho, 2012).

5.3.7 GE course grades

There were six grade classifications for GE course results awarded during the frame of this study. These grads were E, D, C, B, A and S. In grade explanation material provided to BECC teachers these grades were describe as S = superb (90–100 points), A = excellent (80–90 points), B = good (70–79 points) and C = pass (60–69 points.) For reasons described in section 4.5.1 only first semester grades were used for correlational analyses. To prepare GE first semester grades for correlational analysis an S grade was converted to a score of four, an A grade was converted to a score of three, a B grade was converted to a score of 2 and a C grade was converted to a score of 1. D grades for students who had not completed sufficient assessments to pass the course, and E grades for students who had not attended the minimum of 67% of classes were removed from the correlation, because these students' grades were likely affected by factors other than English language proficiency, such as poor motivation, or problems in their personal lives. The reasons for using course grades as a criterion in this study are presented in section 4.5.1.

5.3.8 GE vocabulary quiz grades

As part of GE course assessment, students were required to take a vocabulary quiz after each unit of study. The quizzes had a standardised format of 25 questions, with five questions focusing on Japanese to English translation, five questions focusing on English to Japanese translation, five questions on parts of speech, five multiple choice gap-fill questions in which the best word from a selection of words on the vocabulary list had to be chosen to complete a sentence, and five listening questions in which, quiz takers had to listen to a word and then spell it. The quizzes could either be administered online through the Moodle course management system, or as paper-based tests. Teachers were free to choose which medium they preferred. Data are available for those classes whose teachers chose to administer the vocabulary tests through Moodle. Data was only used from students who took all six vocabulary quizzes across a year of study, in order not to negatively affect averages from students who missed a quiz, which would lower their average.

The numbers of student participants who had available both BET scores and quiz scores for all six quizzes across an academic year are given in Table 5.7.

Table 5.7. *Available GE Vocabulary Quiz Results*

Course Quizzes	Number of usable results
FE 2015	177
FE2016	179
SE 2016	203

KR-20 reliability statistics for the vocabulary quizzes used in this study are presented in Appendix P. The individual unit vocabulary quizzes have reasonable KR-20 values for low-stakes, in-house quizzes, ranging from .66 to .84. The reasons that GE vocabulary quiz grades were considered a suitable criterion for use in this study are outlined in section 4.5.1.

5.3.9 English proficiency tests

Results of two standardised English proficiency tests were used in this study. These were the Oxford Online Placement Test (OOPT) and the Test of English for International Communication (TOEIC). Results available for these two tests are briefly described in the following two sections. Reasons for using the OOPT and the TOEIC as a criteria are given in section 4.4.1

5.3.9.1 Oxford Online Placement Test

The OOPT was administered to representative samples covering the ability range of the student population who took the BET1 2015, the BET1 2016, the BET2 2016, and the BET3 2017. The test was given to a representative sample rather than to the whole population of test takers due to considerations of cost and practicality (i.e. each test credit cost four British pounds, and the test took up to a full period of class time). The representative sample was chosen by having one class from each stream of GE classes take the test. However, for the BET2 2017 only data from a convenience sample of 15 students in the Global Communication Department (who took the GE curriculum A2–B1 course) is available, so this cannot be said to be representative sample of the whole BET test taker population. For all of the BETs examined the OOPT was administered within one month of taking the BET, which is assumed not to be enough time for significant changes in reading and listening ability to take place, given that GE classes were only taken twice a week.

The OOPT was administered to students three months after taking BET2 2016, after the students returned from their spring vacation. In this case it is also assumed that no major changes to English reading and listening ability occurred over these three months, as most students probably did not study any English over their two month break, and it is also unlikely that a single month of twice per week GE classes resulted in significant changes to English reading or listening ability. For example, the Cambridge English Assessment website suggests that approximately 200 hours of guided study are needed to increase proficiency by one CEFR level (Desveaux, 2018), and a study by Saegusa 1985 (as cited in Trew, 2007) showed that over 200 hours of intensive study were needed to gain approximately 100 points on the TOEIC.

In 2015 GE students were divided into a higher and lower stream for the A1–A2 course and also a higher and lower stream for A2–B1 course, so students from four classes took the OOPT. In 2016 there was just one stream for the A1–A2 course, and two streams for the A2–B1 course, so the OOPT was administered to three classes. Results for students who did not give permission for their BET or OOPT results to be used were removed from the analysis. A summary of the number of students from each GE course and stream whose OOPT results were able to be used for analysis is attached as Appendix Q.

5.3.9.2 TOEIC

The TOEIC is a widely accepted and used standardised, norm-referenced test targeted at assessing communicative English for the workplace and everyday life (Powers, 2010b). TOEIC results are presented for reading, listening, and an overall score is also given.

For this study a convenience sample of TOEIC scores was available for students in the Global Communication Department who also took the BET1 2015, the BET1 2016 and the BET2 2017. All of these students took the TOEIC within a month of taking the relevant BET.

5.3.10 Data analysis procedures

In this section data analysis procedures and statistical methods used in this study are briefly described.

5.3.10.1 Correlational analyses

Both Pearson and Spearman correlations are provided in this study, so that their results may be compared. Spearman’s rho is a non-parametric test, which has less stringent assumptions than the Pearson product-moment correlation (i.e. a linear association and homoscedasticity are not assumed, and it is robust to outliers). Spearman’s rho has just two assumptions. The first assumption is that the data is ordinal and the second assumption is that one variable is monotonically related to the other variable. All data used for Spearman’s rho correlation in this study was ordinal, and a monotonic relationship between each pair of variables was checked through examining scatterplots of the two variables, which showed that as the value of one variable increased, so did the value of the other variable.

Pearson product-moment correlation coefficients are parametric and have assumptions that each variable should be continuous, that there are no outliers, and that the relationship between the two variables is linear and homoscedastic. Scatterplots for each pair of variables correlated were examined and seemed to show a sufficiently linear relationship and homoscedasticity. All variables used for Pearson correlations in this study were continuous, except for Likert scale data. However, Likert scale data is commonly used for Pearson correlations, and this use is considered acceptable by many researchers (Norman, 2010). Therefore, Pearson correlations were also calculated for Likert scale data in this study. Apparent outliers were not removed from Pearson correlations, as Spearman's rho correlations are provided for comparison.

For classifying the strength of correlations in this study the commonly cited rule of thumb conventions established by J. Cohen (1998) are used, i.e. *r* between .10 and .29 is a "small" association, .30–.49 is a "medium" association, and .5 or greater is a "large" association.

5.3.10.2 Rasch analysis

Rasch analysis is a probabilistic model commonly employed for language test analysis (See McNamara, 1996 for a detailed description of the Rasch model and its applications in language testing). Rasch analysis was chosen for analysing BET items and the tests as a whole because Rasch analysis is the only IRT-based approach that is usable with relatively small sample sizes like the BET taker population (Kolen & Brennan, 2014). Linacre (1994) provides the guidelines in the table below for minimum sample sizes needed to provide stable Rasch estimates. It can be seen from the table that for the available data for the BETs within the frame of this study only the BET1 2015 meets the criteria for a definitive and high stakes test. The other four BETs examined should provide stable Rasch item calibrations and person measures within plus or minus one half of a logit with 99% confidence, which seems sufficient for the purposes of this study examining the BETs as moderate stakes tests.

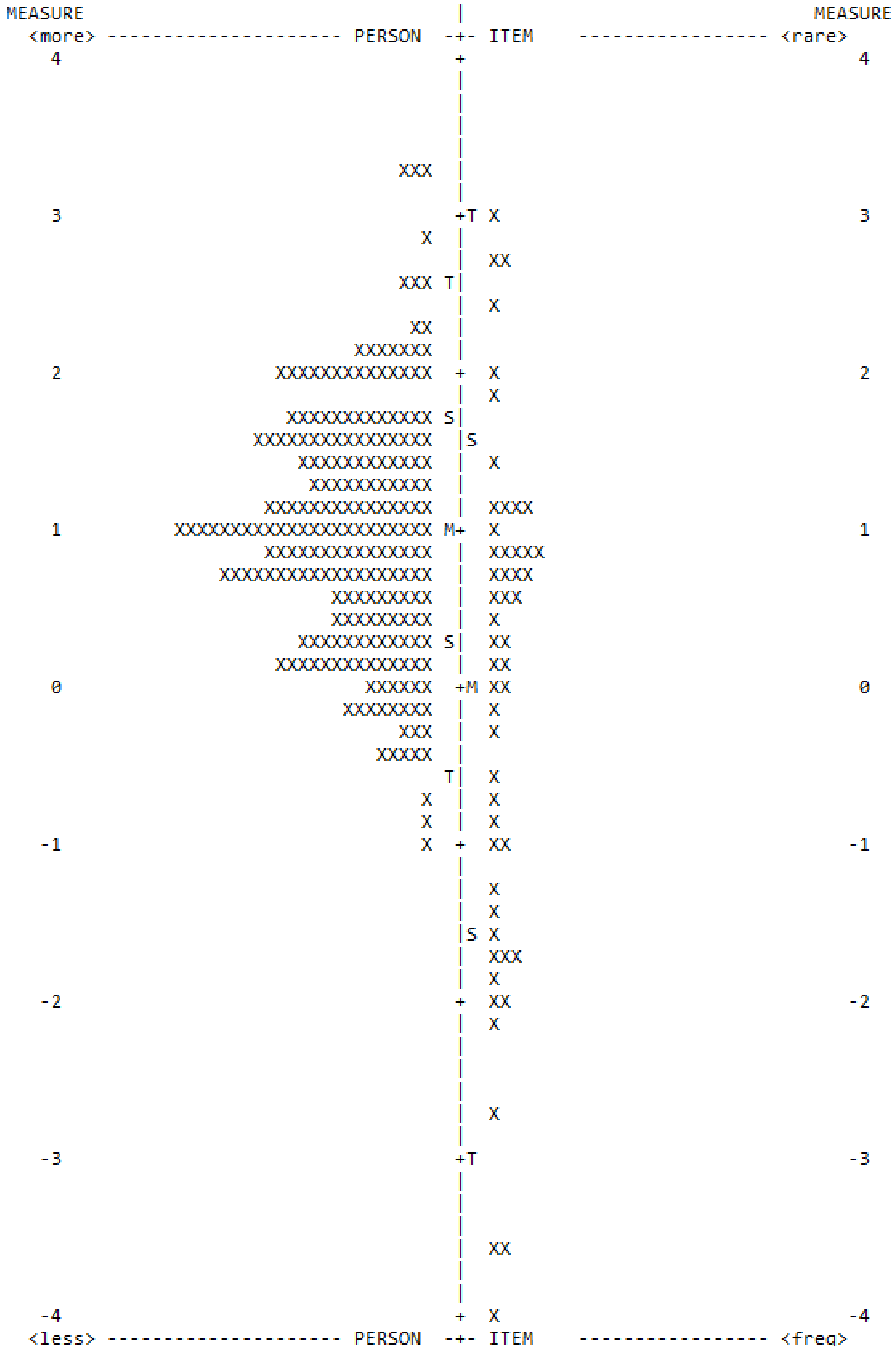
Table 5.8. *Recommended Minimum Sample Sizes for Rasch Analysis*

Item Calibrations or Person Measures Stable Within	**Confidence**	**Minimum Sample Size Range (Best to Poor Targeting)**	**Size for Most Purposes**
± 1 logit	95%	16–36	30 (minimum for dichotomies)
± 1 logit	99%	27–61	50 (minimum for polytomies)
± ½ logit	95%	64–144	100
± ½ logit	99%	108–243	150
Definitive or High Stakes	99%+ (Items)	250–20*test length	250
Adverse Circumstances	Robust	450 upwards	500

Note. Adapted from “Sample Size and Item Calibration Stability” by M. J. Linacre, 1994, *Rasch Measurement Transactions 7*(4), p. 328. Adapted with permission.

Rasch analysis is commonly used for the validation of both norm-referenced and criterion-referenced tests (Gochyyev & Sabers, 2012; Smith & Stone, 2009), and is a widely accepted method of test analysis in educational testing (McNamara & Knoch, 2012), therefore it was deemed suitable for use in this study.

The Rasch model is used for item analysis in this study, as described in the following section 5.3.10.3. In addition, person-item maps, or Wright Maps are used in this study. Rasch analysis allows for test item difficulty and learner ability to be placed on the same scale and represented graphically. Such pictures are known as person-item maps or Wright Maps. An example of a Wright map for the BERT2 2016 is given as Figure 5.1.



```
MEASURE                                 |                               MEASURE
  <more> --------------------- PERSON  -+- ITEM    ----------------- <rare>
    4                                   +                                   4
                                        |
                                        |
                                        |
                                        |
                                   XXX  |
                                        |
    3                                   +T X                                3
                                     X  |
                                        |  XX
                                   XXX T|
                                        |  X
                                    XX  |
                               XXXXXXX  |
    2                   XXXXXXXXXXXXXX  +  X                                2
                                        |  X
                         XXXXXXXXXXXXX S|
                      XXXXXXXXXXXXXXXX  |S
                          XXXXXXXXXXXX  |  X
                           XXXXXXXXXXX  |
                       XXXXXXXXXXXXXXX  |  XXXX
    1          XXXXXXXXXXXXXXXXXXXXXXX M+  X                                1
                       XXXXXXXXXXXXXXX  |  XXXXX
                   XXXXXXXXXXXXXXXXXXX  |  XXXX
                             XXXXXXXXX  |  XXX
                             XXXXXXXXX  |  X
                          XXXXXXXXXXXX S|  XX
                        XXXXXXXXXXXXXX  |  XX
    0                           XXXXXX  +M XX                               0
                              XXXXXXXX  |  X
                                   XXX  |  X
                                 XXXXX  |
                                       T|  X
                                     X  |  X
                                     X  |  X
   -1                                X  +  XX                              -1
                                        |
                                        |  X
                                        |  X
                                        |S X
                                        |  XXX
                                        |  X
   -2                                   +  XX                              -2
                                        |  X
                                        |
                                        |
                                        |
                                        |  X
                                        |
   -3                                   +T                                 -3
                                        |
                                        |
                                        |
                                        |  XX
                                        |
                                        |
   -4                                   +  X                               -4
  <less> --------------------- PERSON  -+- ITEM    ----------------- <freq>
```

Figure 5.1 Wright Map of the BERT2 2016

Test takers are represented on the left side of the Wright Map ordered by their ability on the BERT reading construct, and and test items are represented on the right side ordered by difficulty. Both test takers and items are measured on the same scale in logits. For this Wright Map , the response probability is set at 80%, so a test taker with the same ability level as an item’s difficulty level on this map (i.e. at the same horizontal level on the map) has an 80% chance of answering that item correctly.

In order for a test to measure test taker ability effectively, it is important for the item difficulty to be distributed relatively evenly along the spectrum of test taker ability. It can be seen in Figure 5.1 that even though the response probability has been set at an 80% chance of answering an item correctly for the BET2 2016, there are several item below the ability of the lowest proficiency test taker. These items would need to be revised or replaced in order to more precisely measure test taker reading ability with the BERT 2. (See Bond & Fox, 2007 and Boone, Staver & Yale, 2014 and for thorough and accessible explanations of Wright Maps.)

5.3.10.3 Item analysis

Two criteria for item analysis are used as backings in this study. The first is point-measure correlations, and the second is Rasch fit statistics. These two statistics are used as backing for assumption 3 of the warrant for the BET IUA explanation inference as they can be interpreted as evidence of unidimensionality, or evidence that all items on a test measure the same construct. Unidimensionality is a somewhat controversial topic in testing with there being no broadly agreed upon definition or universally accepted criteria for assessing unidimensionality (Smith, 2002).

The point-measure correlation is similar to the point biserial correlation. Point-biserial correlation is the Pearson correlation between responses to an item and scores on the whole test. Point-measure correlation is the correlation between Rasch person measures and persons responses to an item (Linacre, 2017b). Zero or negative point measure correlations can indicate items that do not function effectively to measure the construct of interest (Linacre, 2017b), and very low point-measure correlations indicate that an item may be in need of further investigation, and by inference not measuring the construct well (R. Green, 2013, Sick, 2010). For this study items with point-measure correlations of zero or below, and of .1 or below are identified and discussed.

Rasch fit statistics can give a further indication that items that may not measure the construct of interest (Sick, 2010; Smith, 2002), and Rasch fit statistics falling within acceptable boundaries have been used in previous studies as evidence of unidimensionality of a test, or in other words, as evidence that the items in a test contribute to measurement the same construct (Dunlea, 2015; Mohsen & Dennick, 2013; McCreary et al., 2013).

As mentioned in section 4.4.3, in this study criterion of MNSQ values for BERT and BELT items of between .5 and 1.5 is considered to indicate an acceptable level of contribution to measurement of the construct in question.

As one of the functions of the BETs is as criterion-referenced achievement tests, ideally criterion-referenced item analysis statistics such as the difference index (DI) and the *B*-index (Brown, 2005; Brown & Hudson, 2002) should also be used to inform item revisions. However, these two item statistics were not used in this study. DI was not used, because it requires the same item to be administered on both the pre and post test. Due to the difficulty of creating multiple alternative test forms and a concern that students may remember test items students were administered different items each time they took a BET, so DI could not be used. The B-index, on the other hand, requires that cut scores be set, so that test takers can be split into master and non-master groups before it can be used. For the

BETs within the frame of this study, cut scores to define CEFR achievement levels had not yet been set, meaning that *B*-index scores also could not be used.

5.3.10.4 Lexical profile analysis

Lexical frequency profiles were created for the BERTs and the BELT tapescripts within the frame of this study using the ‘Vocabprofile’ tool available on the Compleat Lexical Tutor website (Cobb, 2018). BERT and BELT-tapescript lexical profiles were produced for each of the BETs within the frame of this study based on the General Service List first 1000 and second 1000 word lists, the Academic Word List and the British National Corpus. Results comparing the BET lexical profiles to KET and PET reading and listening section lexical profiles are attached as Appendices R, S and T, and are discussed in section 6.1.5.

It is widely accepted that one factor affecting the level of comprehension difficulty of a text is the difficulty of the words it contains (Laufer, 1992; Nation, 2006). In addition, the difficulty of words is highly correlated with their frequency of appearance in corpora (Breland, 1996), which makes vocabulary profiles one of the important tools for measuring the difficulty of text comprehensibility. In addition, vocabulary profiles generated by tools on the Compleat Lexical Tutor website have been used as backing in validation studies of other criterion-referenced tests, which claim CEFR alignment (Eliot & Wilson, 2013; Khalifa & Weir, 2009; Khalifa & Schmitt, 2010). Therefore, Vocabprofile was considered to be a suitable and valid tool to use in this study, in order to seek backing in the BET validity argument.

CHAPTER 6

Results

6.0 Introduction

This chapter presents the results of the backings sought for the BET IUA presented in Chapter 4. It follows the structure of the BET IUA to present backing or rebuttal evidence for each inference's warrants and their related assumptions. The chapter is thus divided into six sections, one for each inference in the BET validity argument. Within each inference section, the evidence gathered for each of the assumptions beneath the warrant is presented and evaluated.

The results are presented as backing for each inference in the IUA, rather than as direct support addressing each of the research questions in this study. This is because within Kane's argument-based validation framework, test score interpretations and uses are judged based on the sufficiency of support for each inference in a chain of inferences. Therefore, examining the backing for each inference in the BET IUA is essential to evaluate the validity of the BETs as placement/streaming tests (research question 1), and as achievement tests (research question 2).

An overall assessment as to the extent to which each assumption is supported or rebutted is made based on the evidence. Evaluations of the overall evidence for the warrant for inference are presented in the following Chapter 7.

6.1 Domain definition inference

As shown in Table 4.1 in Chapter 4 the BET IUA domain definition inference has two warrants. The first warrant for the streaming function if the BETs is *observations of performance on the BETs reveal the level of reading and listening skills and abilities needed to function effectively in GE courses, and are representative of reading and listening* *performance in the General English curriculum.* The second warrant for the achievement function of the BETs is *observations of performance on the BETs reveal achievement of the General English course goals for reading and listening.* These warrants are supported by three assumptions, for which the evidence is assessed in the following sections.

6.1.1 Backing/rebuttal 1 for assumption 1 of the domain definition warrant

As described in section 5.3.1.1 backing for the first assumption of the domain definition warrant *that BET content is a representative sample of the GE curriculum*, comes from tables made during the 2016 and 2017 BET planning and creation process, which show the proportions of items from each unit in the GE curriculum on which BET items were based. Such a table was not available for the BET1 2015, so it was made post hoc by the researcher. These tables are attached as Appendix F.

It can be seen from the tables that only the BET2 2016, the BET2 2017 and the BET3 2017 provided a wide and even representation of the target units across the GE curriculum (FE for the BET2 and FE and SE for the BET3). In spite of the test writers' best efforts to predict material that would be in the new GE curriculum for the BET1 2015, it had a narrow coverage of the new two-year curriculum that the students actually studied. Also, due to only one year of curriculum material having been completed for the BET1 2016 and due to revision priorities decided for the BET1 2017 both of these tests only covered the first year of the curriculum, not both the first and second year, as they were intended to do in the long term. Overall, for the BETs examined in the first two years of test development covered by this study, backing for assumption 1 of the BET IUA domain definition inference warrant is weak due to insufficient curriculum representation in three of the BETs examined. This provides strong rebuttal evidence for this assumption for the BET1 2015. However, strong backing evidence was found for this assumption for the 2016 BET2, and also the 2017 BETs

2 & 3 as a result of updates made to these tests in 2016 to match them to the new curriculum. Overall this is judged to represent moderately strong rebuttal evidence for this assumption.

6.1.2 Backing/rebuttal 2 for assumption 1 of the domain definition warrant

Descriptive statistics for Likert scale answers to the BET teacher and student survey questions used as backing for assumption 1 of the domain definition warrant are attached as Appendix U. The survey results show that at the end of the 2016/17 academic year all teachers who took the survey agree that “BET content is representative of GE curriculum content.” The level of agreement is also quite strong, as shown by an average Likert scale response of 5.1 out of 6. Also, most first and second year students agreed that the 2017 BET2 and BET3 “included a wide range of content from what I studied in my BECC English classes” (>90%), with a moderately high level of agreement indicated by an average Likert scale response of 4.6 out of 6. This provides moderately strong backing for assumption 1 of the domain definition warrant for the 2017 BETs 2 and 3, probably as a result of the updates and changes made to these tests in 2016.

6.1.3 Backing/rebuttal 1 for assumption 2 of the domain definition warrant

An analysis of the can do statements that the BET specifications for each BET testlet claimed to measure within the frame of this study is attached as Appendix G. The analysis shows that many of the can do statements selected to represent the testlet tasks in the 2014/15 BET specifications, which were entered when the goals of the curriculum were defined in terms of the CEFR-J levels A1 and A2, were not updated to match the new curriculum goals of level A2 for the new lower course stream and B1 for the new higher course stream. These can do statements were also not updated in the 2015/16 iteration of the specifications. For the 2016/17 BET specifications the can do statements for the reading section were updated to better match the new curriculum goals, however, the majority of the listening section can do statements still remained to be updated. Given the importance of test specifications as backing for warrants in a validity argument (Bachman & Palmer, 2010) this mismatch between the actual BET testlet task requirements, the course goals, and the can do statements for each BET testlet in the BET specifications represents strong rebuttal evidence for the second assumption of the BET IUA domain definition warrant.

6.1.4 Backing/rebuttal 2 for assumption 2 of the domain definition warrant

As described in section 4.1.3, backing 2 for assumption 2 of the domain definition warrant was sought from a post hoc analysis of the can do statements for each of the BET testlets in the BET testlet specifications. As problems with the match between the can do statements listed for the BET testlets and the course goals and testlet task requirements were revealed by the analysis, better-matching can do statements were suggested by the researcher (Appendix G). These can do statements were then compared to the detailed can do statement for the target CEFR goals of the GE course to get an idea of the actual coverage of the CEFR reading and listening skills at the course target levels, that were covered by the BETs within the frame of this study. The analysis is attached as Appendix H. After the review of can do statements in the BET testlet specifications the BERT appears to cover 7/15 or 47% of the CEFR reading subscales at levels A2 and B1, and the BELT appears to cover 7/16 or 44% of the listening subscales at levels A2 and B1. The CEFR reading subscales of reading correspondence, and reading instructions are not represented, and the CEFR listening subscale of watching TV and film is also not represented. On the other hand, skills relevant to reading, but not directly covered by the CEFR reading subscales seem to be tested by the BERT. These are vocabulary range, grammatical accuracy, sociolinguistic awareness, understanding a native speaker interlocutor and information exchange. Vocabulary range and grammatical accuracy seem appropriate for inclusion in a reading test as vocabulary and grammar have been shown to be important components of reading comprehension ability. It is debatable, however, whether reading tasks based around conversation turns are appropriate for a reading test, and whether aspects of conversation ability can be assessed through a printed format, so this is an area for more research.

Overall this analysis represents strong rebuttal evidence for a claim that the BETs test a broad range of subskills represented by the CEFR subscales for reading and listening, which would seem to be implicit in the course goals set by the overall CEFR can do statements. Partial coverage of the CEFR subskills for the target reading and listening levels is not in itself a problem as a curriculum may prioritize certain CEFR subskills, and test space is also limited. However, this lack of coverage points to a need to define clearly for stakeholders which aspects of reading and listening ability within the CEFR, the GE program aims to improve in students, so that the match between the BETs and the curriculum goals is clear for all. Therefore, it is suggested in section 7.3.2 that more specific reading and listening goals be included in the course outline for students, and in the curriculum overview for BECC teachers and learning advisors in addition to the overall reading and listening CEFR proficiency statements.

6.1.5 Backing/rebuttal 3 for assumption 2 of the domain definition warrant

The first piece of backing sought for assumption 3 of the domain definition warrant is a table showing the lexical profiles of the BERTs within the frame of this study and KET and PET reading sections using the GSL first 1000 and second 1000 word lists, and the AWL, which is attached as Appendix R. The average proportion of words from the first 1000 words of the GSL for the BERTs within this study is 81.03%, and from the second 1000 words of the GSL is 6.07% for a total of 87.10% of words coming from the first 2000 words of the GSL. This compares to 86.95% of words coming from the first 1000 words GSL list, and 5.04% from the GSL second 1000 word list for the KET (91.99%), and for the PET 81.22% of words being from the GSL first 1000 word list, and 8.81% of words being from the GSL second 1000 word list (total 90.03%). While the proportion of words from the first two 1000 word lists of the GSL for the BERTs examined is not greatly different to the KET and PET papers examined in Khalifa and Weir (2009). It is notable that the BERTs have a similar proportion of words from the first 1000 word list as the PET examples, and 5.92% less words from the first 1000 word list than the KET examples. This may indicate that the vocabulary level of the BERTs is a little too high for a test targeted at both the A2 and B1 levels and as such represents weak rebuttal evidence for assumption 2 of the domain definition warrant.

Secondly, a table displaying lexical profiles of the BERTs within the frame of this study and the KET and PET (Khalifa & Weir, 2009) using the first 20,000 words of the British National Corpus is attached as Appendix S. A similar concern arises from an examination of this table as arose for the above analysis using the GSL. It can be seen from the table that the proportion of words from the K1 or first 1000 words from the BNC in the BERT1 2015 (83.52%) and the 2017 BERTs (83.25%) is less than in the KET (89.3%) and PET (84.73%) papers examined in Khalifa and Weirs study. The 2016 BERTs, on the other hand, have proportion of K1 words midway between the KET and PET proportions revealed by Khalifa & Weir. Overall, this is another possible indication that the vocabulary used in the BERTs within the frame of this study is overall a litter higher than is suitable for reading tests at the CEFR A2 and B1 levels.

Finally, a table presenting lexical analyses of tapescripts of the BELTs within the frame of this study and KET and PET tapescripts (Eliot & Wilson, 2013) using the BNC is attached as Appendix T. Similar to the BERT lexical analysis outlined above, the analysis of the tapescripts shows that the BELTs have a significantly lower proportion (81.39–88.43%) of words from the BNC first 1000 word list than the KET (93.72%) and PET (92.38%) from Eliot and Wilson’s study. The difference between the proportion of words from the 2017 BELTs and the KET and PET is 5.29% and 3.95% respectively. This is evidence that the level of vocabulary in the BELTs may also need to be simplified somewhat to better target the A2 and B1 CEFR levels, and as such is some further weak rebuttal evidence for assumption 2 of the domain definition warrant.

6.1.6 Backing/rebuttal 1 for assumption 3 of the domain definition warrant

The first type of backing sought for assumption 3 of the warrant for the BET domain definition inference is presented in the form of tables attached as Appendix I, showing the distribution of listening and reading tasks in both the 2015 FE curriculum and in the 2016 overall GE curriculum, which were analysed to be similar to the BET testlet specifications. After examining the tables, it is clear that both the representation, and distribution of BET-like tasks in the 2015 FE and 2016 overall GE curriculum was rather poor. Tasks similar to BERT Part 6 are greatly over represented with 18 BERT 1 Part 6 type tasks in the FE 2015 materials, and 21 in the 2016 GE curriculum. The balance of BET-like tasks in the materials is also a problem. For example, there were no BERT Part 1–4 or BERT Part 8 type tasks in the main FE lessons in 2015 or 2016, and there were also no BERT Parts 2, 3, 7 and 8 type tasks in the main SE lessons in 2016. In addition, there were no BELT Part 1, 5 or 6 type tasks in the 2015 and 2016 FE main lessons, and no BELT part 2, 4 or 5 type tasks in the SE main lessons in the 2016 GE materials.

BECC teachers in charge of designing and revising the GE curriculum for 2017 were aware of this mismatch between curriculum task types and BET testlet types. Indeed, the BET review lessons, which are included in the table, were made toward the end of the 2015/16 and 2016/17 academic years to make sure that students had at least some exposure to BET type tasks before taking the BETs 2 and 3, and these review lessons did go some way toward improving the backing for assumption 3. For example, the 2016 SE BET review lesson has examples of all of the BELT testlet types, and examples of all of BERT testlet types except for 2, 5 and 6. Also, there is some backing for assumption 3 to be found in the tables as it can be seen that lesson tasks similar to all BET testlet types were presented at least twice by the end of 2016 GE and at least once by FE students by the end of 2016 FE (except for BELT Part 6). Further work is obviously needed, however, to increase the amount of those BET task types that are shown to be underrepresented in the GE curriculum, and to better balance the representation of BET-type tasks across the GE curriculum, for example by reducing the amount of BERT part 6 type tasks, and replacing them with task types the that are similar to other BERT testlets. Indeed, these goals were part of the revisions to the GE curriculum undertaken at the end of the 2016/17 academic year for FE first semester 2017 materials, and this work continued in 2017. I would suggest that having students encounter a lesson task similar to each BET testlet type at least once a semester would provide sufficient backing for assumption 4 of the BET IUA domain inference warrant.

However, it is important to note that it would not be desirable for all curriculum reading and listening task to be similar to BET testlet task types, as it is useful for learners to experience other types of reading and listening lesson tasks, such as open response comprehension items, responses to reading and listening passages that involve reflection and personalization, sentence ordering activities, and picture ordering activities etc. Such learning tasks can teach learners reading and listening strategies at the CEFR A2 and B1 levels, in ways which are not limited by multiple choice responses as are the BETs.

The mismatch between the types of tasks in BET testlets and the types of language learning tasks in the GE curriculum revealed in the tables is moderately strong rebuttal evidence for assumption 3 of the warrant for the BET IUA domain definition inference. However, the GE curriculum designers at the end of the 2016/17 academic year viewed this as a problem to be addressed through continued revisions to the GE curriculum, rather than through revisions to the BETs. As such, this is a case of washback from the test into the curriculum, Curriculum designers saw greater representation of BET style tasks in the curriculum as desirable, because the BET testlet tasks were viewed as being appropriate to the curriculum target A2 and B1 levels, as they were derived from KET and PET tasks, which were designed by language testing experts to assess skills at the CEFR A1 and B1 levels. A long-term aim for GE curriculum revisions should not be for all tasks in the GE curriculum to be like BET testlets. Rather, it should be to increase the representation of BET testlet type tasks in the lessons, and to better balance the representation of different BET testlet type tasks across units and years.

6.1.7 Backing/rebuttal 2 for assumption 3 of the domain definition warrant

An analysis of the two teacher survey statements and two student survey statements from surveys administered at the end of the 2016/17 academic year, which were used to seek backing for assumption 3 of the BET IUA domain definition warrant is attached as part of Appendix U. Most teachers who took the survey (10/11 or 91%) agreed that "BET reading tasks are similar to reading tasks in the GE curriculum.", and also that "BET listening tasks are similar to reading tasks in the GE curriculum." The average level of agreement is also high at 4.9 and 5.0 respectively on a 1–6 scale. However, one teacher slightly disagreed with both of these statements and commented in the free response section that:

> I feel the topics of BET 2 and 3 are much more aligned to our current GE course, however the different task types used in the BETs are still lacking in the GE lessons. For example, Reading part 1 (matching statements with signs) does not occur anywhere in the FE curriculum, nor does the longer text in Reading part 8. Therefore the GE curriculum needs to be revised further to incorporate these tasks into lessons so students can practice prior to being given the "BET Review" lesson at the end of the year.

This teacher's comments are congruent with the analysis from backing 1 of assumption 3 of the domain definition warrant, that more work needs to be done to include examples of BET task types in the GE curriculum.

Majorities of FE and SE students also agreed that “BET reading tasks are similar to reading tasks in my BECC English classes” (>89%), and that “BET listening tasks are similar to listening tasks in my BECC English classes” (>80%). However, the somewhat low average level of agreement for these two statements of 4.32–4.46, is further evidence that more BET type tasks need to be included in the GE curriculum to improve the alignment between the test tasks and curriculum tasks.

Overall the analysis of teacher and student survey questions which were used to seek backing for assumption 3 of the warrant for the domain definition inference provide moderate backing for this assumption.

6.2 Evaluation inference

The warrant for the evaluation inference in the BET IUA is that *BET tasks yield consistent observed scores, which are not contaminated by construct irrelevant variance*. Evidence for the three assumptions supporting this warrant is evaluated in the following sections.

6.2.1 Backing/rebuttal for assumption 1 of the evaluation warrant

To evaluate the first assumption of the warrant for the evaluation inference of the BET IUA that *the BETs are administered and scored consistently* the procedures for administering the annual BETs are described here.

For all the BETs within the scope of this study, proctoring guidelines were issued to the test proctors, who were BECC teachers and learning advisors, before the test, and a short meeting was also held before each test administration to answer any questions from proctors.

The BET1 was administered in early April to entering students, in two large lecture theatres. Firstly, test papers, note papers and bubble sheets for answering questions were distributed and the sound quality of the test video was checked. Simple instructions such as to turn off mobile phones, and that dictionary use is not permitted were either written on a whiteboard, or projected where they were visible to all test takers. After that, students were allowed to enter the exam rooms. The test was automatically timed with the test DVD, and verbal test instructions for all test sections were given in Japanese.

40 minutes was given to answer the reading section, and a countdown of remaining test time was presented on the screen for the reading section, so that test takers they could see how much time they had remaining. The listening section was also played from the test DVD, which made the timings and instructions of the listening section completely standardised.

The BET2s and the BET3 within the scope of this study were administered at the end of semester 2 of the academic year in January. These tests were administered in classrooms, and the administration procedures were the same as for the BET1. In this case the students' classroom teacher acted as the proctor for their class.

After each test, all question booklets and note papers were collected to ensure test item security, and bubble sheets were also collected. The bubble sheets with test taker answers were then run through an optical scanner, which automatically read student answers and entered them into an Excel spreadsheet.

Overall the BET administrations within the scope of this study were well standardised. Clear instructions were given to test proctors in the proctoring guidelines document, and the opportunity to clarify any concerns was given in a pre-test meeting. The use of a DVD for all test timing and instructions meant that all test takers took the test in the same way, under the same conditions, which minimized the risk of construct irrelevant variance contaminating test scores. The use of an optical scanner to read test responses, also reduced the risk of human error in scoring test sheets. Overall, these backings provide strong support for assumption 1 of the evaluation inference warrant.

6.2.2 Backing/rebuttal for assumption 2 of the evaluation warrant

The second assumption underlying the warrant for the evaluation inference of the BET IUA is that *appropriate procedures were followed to develop BET testlets and items*.

The first backing for this assumption comes from a brief description of procedures followed for developing, giving peer-feedback on and revising new or recycled testlets for the 2015, 2016 and 2017 BETs. To ensure item quality it is important that new test items be subject to review by other item writers, and then revised based on peer feedback, before the test items are trialled with a target population or used in an actual test (ALTE, 2011; Carr, 2011; Fulcher, 2013; Rodriguez, 2016).

For the 2015 BET1, testlets were each written by one member of the then three-member assessment committee, and subsequently checked and given feedback by the other two committee members. The feedback was presented and discussed in a meeting, appropriate changes were agreed on in the meeting, and then revisions were made for the final test versions.

Draft testlets for the 2016 BETs were reviewed by the three other item writers on the General English Assessment Committee (GEAC), and then revised according to feedback. Also, for creation of the 2017 BETs, draft testlets were reviewed by at least three other item writers and revised testlets were uploaded for further review until a consensus was reached on the quality of a testlet and its items.

The second backing for this assumption comes from checklists used by item writers when making and reviewing BET testlets and items in 2015 for the 2016 BETs, and in 2016 for the 2017 BETs. The checklists were provided to testlet writers/reviewers, and GEAC members were encouraged to use the lists when making lesson assessments of reading and listening for GE lessons, and also when making and revising BERT and BELT testlets. The checklist used for the 2016 BETs contained 8 points, seven of which were relevant to BET items, and the checklist used for making the 2017 BETs had eleven points, ten of which were relevant to BET testlets and items. The questions on the checklist relevant to the BETs are presented below. Points 1–7 were on the checklists for making both the 2016 and 2017 BETs, and points 8–10 were added before making the 2017 BETs.

Listening and Reading Assessment Checklist

1. Does the vocabulary in the text mostly come from the appropriate GE vocabulary lists?
2. Does the length of the text match the BET specifications?
3. Is there a single clear answer (key) for each multiple choice item? (Rodriguez, 2016)
4. Are all of the distractors plausible? (i.e., there are no obviously incorrect or silly distractors.) (Rodriguez, 2016)
5. Are the answer choices all about the same length? (Rodriguez, 2016)
6. Is the item independent of other items? (i.e. there are no hints to answer this item in other items. The answer to this item does not depend on having correctly answered another item.) (Rodriguez, 2016)
7. Can test takers guess the answer from their general knowledge? (The item should test test takers language ability, not their subject knowledge.)
8. Are the questions are in the same order as the information appears in the text? (Khalifa & Weir, 2009)
9. Are the distractors with numbers in numerical order? e.g. a) $3 b) $13 c) $30 (Rodriguez, 2016)
10. Are the stem (i.e. the question) and the choices positively worded? (i.e. the word ‘not’ isn’t used) (Rodriguez, 2016)

Rebuttal evidence for assumption 2 of the evaluation inference comes from the fact that new and revised BET testlets and items during the two-year span of this study were not piloted before each annual administration. Piloting of test items is standard practice in test development. However, the decision not to pilot items was a practical decision based on the limited man hours available for item writing, and the impracticality of obtaining a student sample big enough for Rasch analysis outside of set exam periods. The lack of piloting of BET items significantly weakens the support for this assumption, because it meant that psychometric properties of test items could not be checked before the items were used in actual test administrations.

The testlet review procedures outlined above in which draft BET testlets received feedback from two peers when designing the BET 2015, and from at least three peers when designing the BET 2016 and BET 2017 testlets, along with use of the reading and listening assessment checklist provides strong supporting evidence for assumption 3 of the explanation inference warrant. However, this backing is also strongly attenuated by rebuttal evidence that new BET testlets and items were not piloted before being included in the final test. Overall the backing for assumption 3 of the evaluation inference warrant is judged to be moderately strong.

6.2.3 Backing/rebuttal for assumption 3 of the evaluation warrant

The final backing for assumption 3 of the evaluation inference warrant, *that BET task instructions are easily comprehensible for students,* comes from the fact that all task instructions on the BETs are delivered in the test takers' native language of Japanese. It is paramount that test takers are able to easily understand task instructions, because misunderstanding the nature of the task could lead to construct irrelevant variance entering test scores. Therefore, to facilitate easy understanding of task instructions it is widely recommended that where possible the instructions be put in the test takers' first language Bachman and Palmer (2010) state that "if there is any doubt that test takers might misunderstand it is best to present the instructions in the native language" (p. 385). The fact that all BET task instructions are in Japanese provides strong backing for this assumption.

The second form of backing for assumption 3 of the evaluation inference warrant comes from results of surveys given to the test takers shortly after they took the BETs examined in this study. From Appendix V it can be seen that test takers consistently agreed that the BET DVD spoken instructions and the task instruction on the test papers were easy to understand, with over 92% of students agreeing across all of the BETs focused on here. This represents further strong backing for assumption 3 of the evaluation inference warrant.

6.3 Generalization inference

6.3.1 Backing/rebuttal for assumption 1 of the generalization warrant

Reliability and dependability statistics for the whole BETs as well as for the BERTs and BELTs within the frame of this study are given in table 6.1. The table provides KR-20 reliability stastics, and Phi dependability statistics. Descriptions of these statistics and the assumptions are provided in section 4.3.1.

For the placement/streaming function of the BETs it can be seen that all of the BETs within the frame of this study that were used for placement/streaming have acceptable KR-20 reliability statistics of greater than .8. This provides strong backing for this assumption for the placement function of the BETs.

On the other hand, Phi dependability statistics for all of the BETs, BERTs and BELTs within the frame of this study are all below .8 and are particularly concerning for the BELTs, which have Phi values ranging from .49–.61 indicating that 39–51% of the variance in BELT scores is either caused by error, random variance, or is measuring something outside of the BELT domain.

Table 6.1. *Reliability and Dependability Statistics for BETs, BERTs and BELTs*

BETS	Reliability KR-20	Dependability Phi	BERTs	Reliability K-R20	Dependability Phi	BELTs	Reliability KR-20	Dependability Phi
BET1 2015	.83	.79	BERT1 2015	.78	.74	BELT1 2015	.59	.49
BET1 2016	.81	.78	BERT1 2016	.76	.72	BELT1 2016	.59	.55
BET2 2016	.81	.77	BERT2 2016	.76	.71	BELT2 2016	.56	.51
BET2 2017	.80	.77	BERT2 2017	.76	.72	BELT2 2017	.59	.53
BET3 2017	.84	.81	BERT3 2017	.78	.75	BELT3 2017	.67	.61

In addition, reliability statistics for the separate BERTs and BELTs are all below of .8, which is the approximate threshold that would indicate that these sub-tests would be able to divide the test taker population into three groups (A1, A2, and B1 levels) in order to assess achievement or non-achievement of the GE goals of CEFR A2 and B1 ability, based on strata statistics (see table 4.6). This provides strong rebuttal evidence for this assumption as it relates to the BET achievement test function.

In summary, reliability and dependability statistics for the BETs, BERTs and BELTs within the frame of this study provide strong rebuttal evidence for the second assumption of the generalization warrant for the BET achievement test function, but strong backing for the BET placement test function (see section 6.6.4 for an assessment of the BET streaming function based on separation and strata statistics).

As a caveat, it is worth noting that the reliability of the BETs, BERTs and BELTs may be curtailed by two factors. One factor is that the tests are made up of testlets aimed at testing different reading and listening subskills or different aspects of the same construct, which may reduce reliability estimates (Bachman, 2004; Brown & Hudson, 2002; Khalifa & Weir, 2009; Jones, 2001; Saville, 2003; Weir, 2005a). The other factor is that the students who take the BETs have a relatively narrow English language ability range, which may also lower reliability statistics (Bachman, 2004; Jones, 2001; Saville, 2003; Weir, 2005a).

6.3.2 Backing/rebuttal for assumption 2 of the generalization warrant

Assumption 2 of the generalization warrant is that *appropriate equating and scaling procedures are used to present students with their BET scores*. However, students were not presented with equated and scaled BET scores for the BETs within the frame of this study. Anchor items were placed within the BERTs and BELTs with the intention of linking the tests, however, the tight timeline between test administration and the requirement to present students with their test scores, as well as a lack of technical equating expertise on the GEAC prevented the calculation of equated BET scores within the frame of this study. This represents strong rebuttal evidence for assumption 2 of the generalization warrant.

6.3.3 Backing/rebuttal for assumption 3 of the generalization warrant

To assess the third assumption of the warrant for the generalization inference of the BET IUA that *testlet specifications are well defined so that parallel tasks and test forms are created* a short analysis of the 2015/16 and 2016/17 BET specifications is presented here.

The first version of the BET specifications was written in 2014 in conjunction with making the 2015 BETs. Updates to the BET specifications were then made through 2015 and 2016 to match changes to the format of the BETs, and based on issues which arose and feedback given during the testlet and item writing process. This section examines the BET specification sections which are relevant to writing the 2016 and 2017 BET reading and listening sections. Firstly, backing evidence is given, and secondly rebuttal evidence is presented.

The overall domain of the BETs was clearly stated in the BET specifications "As far as possible the content of each of the BETs will represent a wide, and even sample of the language use tasks for reading, grammar and listening across its target units of the GE curriculum." The specifications also stated "As far as possible the vocabulary in BETs 1 and 3 will be limited to the vocabulary on the FE and SE word lists. Vocabulary for BET 2 will as far as possible be limited to the FE word list." Finally, the BET specifications provided fairly detailed information about each BET testlet to assist item writers, along with an example of each testlet type. The specifications for each BET testlet include information under five categories as shown in Table 6.2. Examples from the specifications are given to illustrate the type of information provided in each category. Unfortunately, actual examples of the testlets presented in the BET specifications cannot be presented here due to test security concerns, as the examples in the BET specifications may be recycled in future BETs. It is important to provide task specifications in addition to can do statements, because as noted by critics of the CEFR (Alderson et al., 2004; Weir, 2005b), CEFR do statements alone are too ambiguous to be used for designing actual test tasks.

Table 6.2. *Types of Information Presented in the Specifications for BET Testlets*

Categories	Descriptions
Task Type and Format	These two categories describe the task response format. E.g. “Three option multiple choice sentence gap fill”, “Matching five prompt sentences to eight notices, plus one example”.
Task Focus	Describes the language skill which the testlet claims to test. E.g. “Reading for detailed understanding and main idea(s)”.
Can Dos Targeted	Lists the CEFR or related (EAQUALS, CEFR-J) can do statement(s) that the testlet claims to assess.
Number of Items	The number of items in the testlet.
Task Specifications	Gives details such as text length, the text type, the order of questions.

An examination of the 2015/16 and 2016/17 BET specifications also reveals two sections of the specifications with out of date information, which may be interpreted as rebuttal evidence for this assumption. The 2015/16 and 2016/17 BET specifications contained an inventory of functions, notions and communicative tasks, and an inventory of grammatical areas, which were found in the 2014 BET curriculum. The language functions listed were for example “asking people to do something”, and the grammatical areas listed are for example “present continuous tense”. These lists were produced by a semester-long exhaustive examination of all the lessons in the 2014 curriculum, and the lists were left in the specifications in the following two years, even though the curriculum had been renewed, with the intention that the lists would be updated to match the new curriculum when the human resources became available. It was not feasible to update these sections during 2015 and 2016, because the priority for the General English Assessment Committee was on updating the BERTs, BELTs and also the Bunkyo English Speaking Tests (BESTs), along with lesson assessments and vocabulary tests. On the other hand, BET testlet reviewers and developers were instructed not to use these lists, but instead to draw on lessons from the new curriculum when choosing content and topics for testlets for the 2016 and 2017 BETs. Thus, the presence of these outdated sections in the 2015/16 and 2016/17 BET specifications does not significantly weaken assumption 2 of the explanation inference warrant.

Overall the BET testlet specifications seem sufficiently detailed for testlet writers to write testlets for alternate test forms that target the same language abilities, particularly considering that example testlets are included in the specifications. However, rebuttal evidence presented for assumption 2 of the domain definition warrant in section 6.1.3 shows that many of the target can do statements for BET testlets over the frame of this study did not match the course goals or the testlet task characteristics well, which also provides rebuttal evidence for this assumption. To know if BET teslets on alternate BET forms are actually equivalent in difficulty would either require piloting the alternate testlets with the same test taker population, or equating the BET forms from different administrations. However, piloting of the alternate testlets was not possible due to time and resource constraints, and the BETs within the frame of this study have not been equated. Taking both the available backing and rebuttal evidence for this assumption into consideration, overall the backing for assumption 3 of the generalization inference is assessed as weak.

6.4 Explanation inference

The warrant for the BET explanation inference is that *observed scores are attributable to the constructs implicit in the CEFR levels A2–B1 for English reading and listening*. This warrant is support by four assumptions, which are evaluated according to the evidence found in the following sections.

6.4.1 Backing/rebuttal for assumption 1 of the explanation warrant

Correlations between BET, BERT and BELT scores and OOPT overall, use of English and listening scores are given in Table 6.3. Before calculating these correlations BET, BERT and BELT items with point measure correlation equal to or below .05 were removed from the analysis as a conservative estimate of items that may not be measuring the intended construct effectively.

Table 6.3. *Correlations between BETs, BERTs, BELTs and the Oxford Online Placement Test (All* p *values are less than .001 except as indicated in the table)*

BET	**BET Overall and OOPT Overall**		**BERT and OOPT Use of English**		**BELT and OOPT Listening**	
	Spearman	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**
BET1 2015	.70	.69	.64	.64	.64	.60
BET1 2016	.62	.65	.65	.62	.50	.48
BET2 2016	.75	.73	.72	.72	.38 (*p* = .003)	.35 (*p* = .008)
BET2 2017	.69	.82	.75 (*p* = .001)	.77 (*p* = .001)	.61 (*p* = .017)	.67 (*p* = .006)
BET3 2017	.62	.62	.61	.59	.52	.52

From the table it can be seen that all correlations between BET total scores and OOPT overall scores are greater than .5 or are “large” according to the widely cited rules of thumb given by J. Cohen (1998). This represents strong backing for assumption 1 of the BET IUA explanation warrant. Correlations between BERT and OOPT use of English scores are also all large, however two correlations between the BELT2 2016 and the OOPT listening section were only medium according to J. Cohen’s rules of thumb.

Correlations between BET, BERT and BELT scores and TOEIC overall, reading and listening sections are given in Table 6.4. As expected, overall correlations between TOEIC and BET scores, were less significant than correlations between OOPT scores and BET scores, so in this case a one tailed test was used rather than a two tailed test, as only a positive correlation was expected between the two variables. As with the OOPT scores, before calculating these correlations BET, BERT and BELT items with point measure correlation below .05 were removed from the analysis as a conservative estimate of items that may not be measuring the intended construct effectively.

Table 6.4. *Correlations between BETs and the TOEIC (One-tailed tests)*

BET	**BET and TOEIC Total Scores**		**BERT and TOEIC Reading Scores**		**BELT and TOEIC Listening Scores**	
	Spearman	**Pearson**	**Spearman**	**Pearson**	**Spearman**	**Pearson**
BET1 2015 (*n* = 29)	.46 (*p* = .006)	.50(*p* = .003)	.46 (*p* = .006)	.46 (*p* = .006)	.35 (*p* = .030)	.38 (*p* = .020)
BET1 2016 (*n* = 16)	.75 (*p* < .001)	.79 (*p* < .001)	.68 (*p* = .002)	.75 (*p* < .001)	.67 (*p* = .002)	.70 (*p* = .001)
BET2 2017 (*n* = 14)	. 84 (*p* < .001)	.75 (*p* = .002)	.34 (*p* = .114)	.46 (*p* = .049)	.81 (*p* < .001)	.83 (p < .001)

As can be seen from Table 6.3 significant correlations were found between BET, BERT and BELT scores and available TOEIC overall scores, reading section scores and TOEIC listening scores respectively, except for the BERT2 2017 and the TOEIC reading section. All correlations were medium to large according to J. Cohen's (1998) criteria, however, there was a rather wide range of correlation strength ranging from .46–.84 for overall scores, .34–.75 for the reading section and .35–.83 for the listening section. This wide variation is probably due to the small sample size available for these analyses, as small samples, or underpowered studies, tend to produce varied results (Reinhart, 2015).

Given the small sample sizes involved, as well as the smaller and in one case non-significant p values returned when compared to the OOPT correlational analyses, backing for assumption 1 of the BET IUA explanation inference from BET and TOEIC correlations is assessed to be moderately strong.

Criterion related evidence (i.e. evidence of correlations between the focus test and external measures of the same construct the test claims to measure, or evidence of equivalent difficulty of items between the focus test and the criterion) have previously been used to validate a link to the CEFR and standardised tests. For example, Kecker and Eckes (2010), correlated scores on the reading and listening sections of the TESDAF Institute, a test of German, and the German section of the DIALANG. Kecker and Eckes found a Spearman's rho correlation of .43 with the DIALANG reading section (which may have been compromised by problems with some distractors not appearing for some candidates) and the TESDAF Institute reading section and .6 with the DIALANG listening section and the TESDAF Institute listening section.

Overall, the correlations found between BET scores and the OOPT and TOEIC were comparable to that obtained in another CEFR alignment study (Kecker & Eckes, 2010). Given that most correlations between the BET, BERT, BELT and OOPT sections were large, backing for assumption 1 of the explanation warrant is judged to be strong overall.

6.4.2 Backing/rebuttal for assumption 2 of the explanation warrant

Backing for assumption 2 of the explanation warrant was sought from correlations between BERT scores and reading self-assessment data and BELT scores and listening self-assessment data. As can be seen from the table of correlations between CEFR self-assessment survey results for reading and listening and BERT, BELT scores in Appendix M, all correlations were statistically significant at the alpha < .05 threshold and disattenuated correlations ranged from .30–.65. It is possible that student unfamiliarity with some of the situations in the can do statements may have served to lower the accuracy of students' self-assessment ratings. Specifically, "work" and "job" and "employment" is mentioned in three of the can do statements, but it is most likely that none, or almost none of the students who took the survey had actual experience working abroad.

Given factors which may have mitigated the correlations between student self-assessments and BERT and BELT results, such as the relatively low reliability of the BELTS, students' lack of experience with work situations mentioned in some of the can do statements, a tendency for low proficiency language learners to overestimate their ability (North & Jones, 2009), and likely differences in the constructs of reading and listening ability as measured by a multiple choice test, and student self-perceptions of their reading and listening ability (Wang, Eignor, & Enright, 2008), the medium to strong correlations between student CEFR self-assessments and BERT and BELT scores are taken to be moderately strong backing for assumption 2 of the warrant for the BET IUA explanation inference warrant.

6.4.3 Backing/rebuttal for assumption 3 of the explanation warrant

The first form of backing for the third assumption underlying the warrant for the explanation inference of the BET IUA *that items effectively measure the construct of interest* comes from an analysis of the proportion of items from Rasch analysis with mean squares less than 0.5 and greater than 1.5. Items should fall within this range to be considered to effectively measure the construct of the test.

There were no items in any of the BETs within the timeframe of this study with Rasch mean square values outside the acceptable range of 0.5 and 1.5, which provides strong backing evidence that the items on the BELTs and BERTs examined in this study are measuring the same construct.

The second form of backing for assumption 3 of the evaluation inference comes from further item analysis. The proportions of items with negative point measure correlations, and point measure correlations between zero and .1 are presented in Table 6.5.

Table 6.5. *Proportions of BET Items with Negative, Zero or Low Point-Measure Correlations*

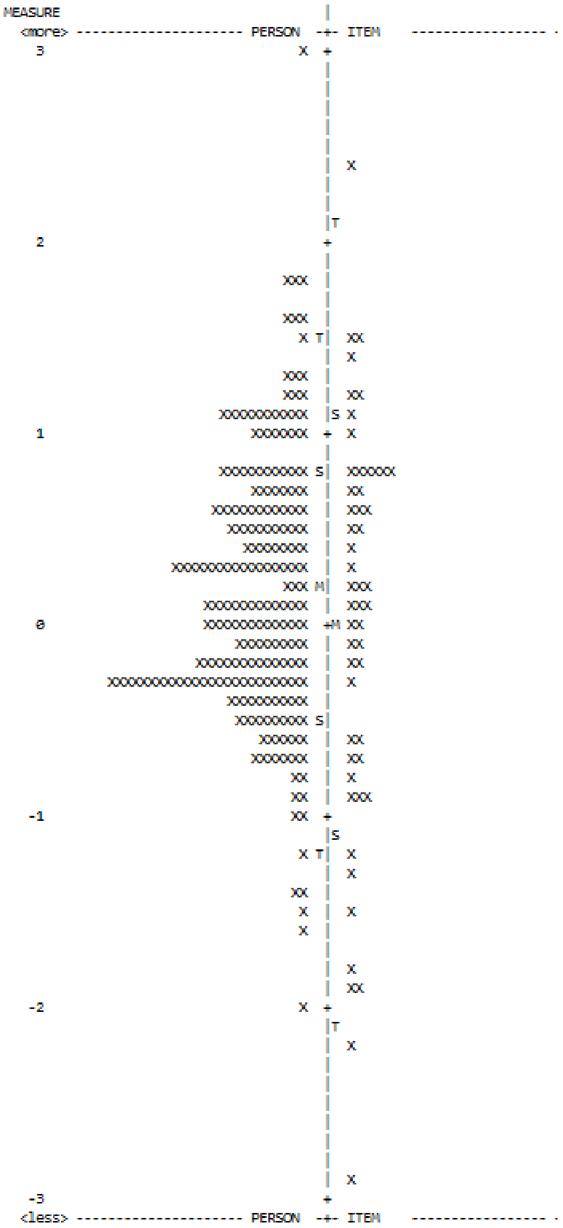
	BERTs					BELTs				
	2015	2016		2017		2015	2016		2017	
	1	1	2	2	3	1	1	2	2	3
Proportion of items with zero or negative point measure correlations.	1/42 2.4%	2/52 3.8%	2/52 3.8%	3/52 5.8%	0%	0/26 0%	1/34 2.9%	2/34 5.9%	0/37 0%	0/37 0%
Proportion of items with point measure correlations between zero and .1.	1/42 2.4%	3/52 5.8%	4/52 7.7%	5/52 9.6%	3/52 5.8%	2/26 7.7%	4/34 11.8%	2/34 5.9%	4/37 10.8%	2/37 5.4%
Total	2/42 4.8%	5/52 9.6%	6/52 11.5%	8/52 15.4%	3/52 5.8%	2/26 7.7%	5/34 14.7%	4/34 11.8%	4/37 10.8%	2/37 5.4%

As can be seen from Table 6.5, there were a few items on all of the BETs during the timeframe of this study which did not effectively measure the construct of interest either reading or listening, as indicated by zero or negative point measure correlations. The proportion of items was not great however, ranging from 0–5.9% across the BETs. Perhaps, of greater concern is the proportion of items with point measure correlations between zero and .1, which ranges from 0–11.8% across the BETs. According to R. Green (2013) such items with small point measure correlations may need further investigation, so these percentages, as well as the total of percentages of both categories (negative/zero and low point measure correlations) clearly indicate that further item revision is needed to improve the quality of BET items for future administrations in order to improve measurement of the constructs of reading and listening implied by the A1–B1 CEFR scales. Overall, item analysis statistics across the BETs examined in this study provide moderate backing for the assumption that items effectively measure the construct of interest, but also show that further item revision for future test administrations is needed to strengthen the backing for this assumption.

6.4.4 Backing/rebuttal for assumption 4 of the explanation warrant

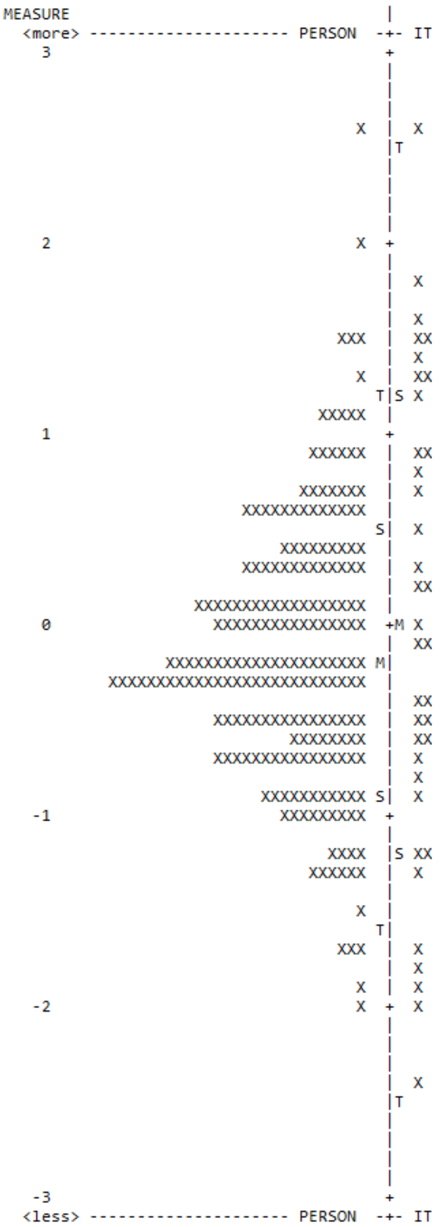
The final form of backing sought for assumption 4 of the BET explanation warrant comes from Wright maps of BERT and BELT test takers and items, which should show a good match between test taker ability and item difficulty, in order to measure the targeted construct accurately (Bond & Fox, 2007). Wright maps were made for all of the BERTs and BELTs within the timeframe of this study that were targeted at the new GE curriculum, and for which test data is available (i.e. BET1 2015, BETs 1 & 2 2016, and BETs 1 & 2 2017). The response probability was set at 80% for the reasons indicated in section 4.4.4. Due to limitations of space, only the Wright maps for the 2017 BET3 are presented as an example in Figure 6.1. It can be seen from these maps that a few more items at the high end of the difficulty range are needed for the BERT3 to effectively measure the ability of the higher-level test takers. Important points from the Wright maps for the BERTs within the frame of this study are summarized in Table 6.6, and the BELTs in table 6.7. In cases where it was difficult to judge a difference between the level of adjacent person ability and item difficulty from a Wright map for judging the number of items above the ability of the highest test taker or below the ability of the lowest test taker (as in the BELT3 2017 example below for the highest person and item) the actual Rasch person ability and item difficulty statistics were directly checked.

Wright map of the BERT3 2017



```
MEASURE                             |
  <more> ------------------ PERSON -+- ITEM    ----------------- .
    3                            X  +
                                    |
                                    |
                                    |
                                    |
                                    |
                                    |  X
                                    |
                                    |
                                    |T
    2                               +
                                    |
                               XXX  |
                                    |
                               XXX  |
                                 X T|  XX
                                    |  X
                               XXX  |
                               XXX  |  XX
                       XXXXXXXXXXX  |S X
    1                      XXXXXXX  +  X
                                    |
                       XXXXXXXXXXX S|  XXXXXX
                           XXXXXXX  |  XX
                      XXXXXXXXXXXX  |  XXX
                        XXXXXXXXXX  |  XX
                          XXXXXXXX  |  X
                 XXXXXXXXXXXXXXXXX  |  X
                               XXX M|  XXX
                     XXXXXXXXXXXXX  |  XXX
    0                XXXXXXXXXXXXX  +M XX
                         XXXXXXXXX  |  XX
                    XXXXXXXXXXXXXX  |  XX
         XXXXXXXXXXXXXXXXXXXXXXXXX  |  X
                        XXXXXXXXXX  |
                         XXXXXXXXX S|
                            XXXXXX  |  XX
                           XXXXXXX  |  XX
                                XX  |  X
                                XX  |  XXX
   -1                           XX  +
                                    |S
                                 X T|  X
                                    |  X
                                XX  |
                                 X  |  X
                                 X  |
                                    |
                                    |  X
                                    |  XX
   -2                            X  +
                                    |T
                                    |  X
                                    |
                                    |
                                    |
                                    |
                                    |
                                    |
                                    |  X
   -3                               +
  <less> ------------------ PERSON -+- ITEM    ----------------- .
```

Wright map of the BELT3 2017



```
MEASURE                               |
  <more> -------------------- PERSON -+- IT
    3                                 +
                                      |
                                      |
                                      |
                                   X  |  X
                                      |T
                                      |
                                      |
                                      |
                                      |
    2                              X  +
                                      |
                                      |  X
                                      |
                                      |  X
                                 XXX  |  XX
                                      |  X
                                   X  |  XX
                                     T|S X
                               XXXXX  |
    1                                 +
                              XXXXXX  |  XX
                                      |  X
                             XXXXXXX  |  X
                       XXXXXXXXXXXXX  |
                                     S|  X
                           XXXXXXXXX  |
                       XXXXXXXXXXXXX  |  X
                                      |  XX
                  XXXXXXXXXXXXXXXXXX  |
    0               XXXXXXXXXXXXXXXX  +M X
                                      |  XX
               XXXXXXXXXXXXXXXXXXXXX M|
         XXXXXXXXXXXXXXXXXXXXXXXXXXX  |
                                      |  XX
                    XXXXXXXXXXXXXXXX  |  XX
                            XXXXXXXX  |  XX
                    XXXXXXXXXXXXXXXX  |  X
                                      |  X
                         XXXXXXXXXXX S|  X
   -1                      XXXXXXXXX  +
                                      |
                                XXXX  |S XX
                              XXXXXX  |  X
                                      |
                                   X  |
                                     T|
                                 XXX  |  X
                                      |  X
                                   X  |  X
   -2                              X  +  X
                                      |
                                      |
                                      |
                                      |  X
                                      |T
                                      |
                                      |
                                      |
                                      |
   -3                                 +
  <less> -------------------- PERSON -+- IT
```

Figure 6.1 Wright Maps of the BERT and BELT 3 2017

Table 6.6. *Wright Map Analysis of the BERTs*

	BERT1 2015	**BERT1 2016**	**BERT2 2016**	**BERT2 2017**	**BERT3 2017**
Wright map shows a good spread of item difficulty across the test taker ability range.	YES, but few more items needed in the high range	YES, but few more items needed in the high range.	YES, but few more items needed in the high range.	YES	YES, but two or three more items needed in the high range
Number of items above the ability of the most proficient test taker.	0	1	0	1	0
Number of items below the ability of the lease proficient test taker.	6	9	15	3	2

Table 6.7. *Wright Map Analysis of the BELTs*

	BELT1 2015	**BELT1 2016**	**BELT2 2016**	**BELT2 2017**	**BELT3 2017**
Wright map shows a reasonably good spread of item difficulty across the test taker ability range.	A few gaps in items matching test taker ability.	YES	YES, but a few more items needed in the high range.	YES	YES
Number of items above the ability of the most proficient test taker.	1	0	0	1	1
Number of items below the ability of the lease proficient test taker.	5	1	5	3	1

Generally, the Wright maps for the BETs under investigation show a reasonable match between the spread of item difficulty and the spread of test taker ability, although in several cases a few more items need to be targeted at the higher test taker ability range across the tests. The proportion of items below the ability of the least proficient test taker was of concern for the 2015–2016 BERTs, and the BELT 1 2015 and BELT2 2016, however, this seems to have been largely corrected in the 2017 BERTs through revisions made in 2016, which aimed to increase the difficulty of the easiest items to better target the tests to the learners. Although more work still needs to be done to better match the range of item difficulty to the range of test taker ability, this backing evidence provides moderately strong support for assumption 4 of the BET IUA explanation inference warrant.

6.5 Extrapolation inference

The warrant for the extrapolation inference in the BET IUA is that *performances on the BERTs and BELTs account for the quality of linguistic performance in the domain of General English courses*. This warrant is supported by a single assumption, for which the evidence is evaluated in the following section.

6.5.1 Backing/rebuttal 1 for assumption 1 of the extrapolation warrant

As explained in section 4.5.1 the first backing sought for the assumption of the extrapolation inference warrant, that that *performance on the BET measures are positively correlated to other criteria of language proficiency in General English courses*, was sought from correlations between BET scores and semester 1 course grades for the BETs in this study used for course placement and streaming. Correlations between the BETs within the frame of this study which were used for course streaming and GE course grades for the first semester after taking each BET are shown in Table 6.8. All correlations are significant at the *p* < .01 level.

Table 6.8. *Correlations between BET grades and Semester 1 Course Grades*

BET	Spearman Correlations	Pearson Correlations
BET1 2015 (*n* = 224)	.50	.53
BET1 2016 (*n* = 243)	.51	.52
BET2 2016 (*n* = 217)	.61	.57

It can be seen from Table 6.8 that large correlations (.50–.61) were found between BET total scores, which were used for course placement and class streaming, and FE and SE semester 1 course grades. These correlations have not been adjusted for restriction of range (which results from using letter grades rather than total course assessment scores,), so it is safe to assume that all correlations would be greater when corrected for restricted range.

Considering that “course performance is dependent on not only a student’s ability, but also other factors not measured by placement tests such as motivation, perseverance, and attendance.” (Mattern & Packman, 2009, p. 1), and also considering that the correlations between overall BET scores and GE first-semester course grades are likely reduced by range restriction, large correlations found between BET total scores and GE semester 1 course grades are strong backing for assumption 1 of the warrant for the extrapolation inference.

6.5.2 Backing/rebuttal 2 for assumption 1 of the extrapolation warrant

As described in section 4.5.1 the second backing for assumption 1 of the extrapolation warrant was sought from correlations between the BETs in this study used for course and class streaming and later speaking test scores. Course placement and streaming was solely based on the combined raw scores of the BERT and BELT, however students were placed into two courses with CEFR-level goals for all four macro skills. Therefore, it is important to provide evidence that raw BET scores are able to predict later speaking test scores in GE courses. Spearman correlations for the BETs are provided in Table 6.8. All correlations are significant at the p <.01 level.

Table 6.9. *Correlations Between BET and BEST Scores*

BET	**BEST**	**Spearman's Correlations**	**Pearson Correlation**
BET1 2015	BEST2 2015 (*n* = 223)	.66	.63
BET1 2016	BEST1 2016 (*n* = 245)	.44	.45
BET1 2016	BEST2 2016(*n* = 242)	.62	.62
BET2 2016	BEST3 2016(*n* = 217)	.72	.70
BET2 2016	BEST4 2016 (*n* = 215)	.66	.68

It can be seen from Table 6.9 that all correlations between raw BET (combined BERT and BELT) scores used for class streaming within the frame of this study and subsequent BEST scores have large correlations, except for the BET1 2016. The medium correlation between the BET1 2016 and the BEST1 2016, which was administered at the end GE students' first semester in the course may be due to these students taking the BEST for the first time, and therefore lacking familiarity with the test format. Overall this evidence can be interpreted as providing strong backing for assumption 1 of the extrapolation warrant.

6.5.3 Backing/rebuttal 3 for assumption 1 of the extrapolation warrant

As also explained in section 4.5.1, the third backing for the assumption underlying the extrapolation warrant inference was sought from correlations between grades for the BETs used for class streaming and average vocabulary quiz scores across the year of the GE curriculum into which students were placed into courses. Results for these correlations are presented in table 6.10.

Table 6.10. *Correlations Between BET Scores and GE Vocabulary Quiz Scores (all correlations significant at p < .001)*

BET	Vocabulary Quizzes	Spearman Correlations	Pearson Correlations
BET1 2015	FE Vocabulary Quizzes 1–6	.71	.77
BET1 2016	FE Vocabulary Quizzes 1–6	.57	.60
BET2 2016	SE Vocabulary Quizzes 7–12	.67	.68

As can be seen from table 6.10 large correlations were found between all of the BETs used for placement and streaming in this study, and subsequent GE vocabulary quiz average scores. This evidence is taken as further strong backing for the assumption beneath the extrapolation inference warrant.

6.6 Utilization inference

The warrant for the BET IUA utilization inference is that *uses of BET scores are beneficial to stakeholders*. This is judged to be the most important inference, and it rests on six assumptions. The evidence gathered to seek backing for these assumptions is analysed and evaluated in the following sections.

6.6.1 Backing/rebuttal 1 for assumption 1 of the utilization warrant

The first backing sought for assumption 1 of the utilization inference warrant, that *BET scores are sufficient and relevant for making decisions about streaming GE classes* comes from answers to two items on a teacher attitudes to the BET survey related to teacher attitudes to the efficacy of class streaming based on BET scores. Answers to the two questions for each of the three administrations of the survey across the time of this study are attached as Part A of Appendix W. The results show that all teachers who were surveyed agreed with the statements (except for one teacher somewhat disagreeing with one item in the 2017 survey), and that the overall strength of agreement as indicated by the average of a Likert scale from 1–6 was over 5 for all except for one of the statements in the 2015 survey.

Overall these results from the opinions of this key stakeholder group provide strong backing for assumption 1 of the BET IUA utilization inference warrant.

6.6.2 Backing/rebuttal 2 for assumption 1 of the utilization warrant

The second backing sought for this assumption comes from answers to surveys eliciting teacher perceptions of the proportion their class' able to perform A1–B1 CEFR can do statements from the CEFR self-assessment grid. After removing can do statements for which teachers felt they did not have enough information about their students to judge, the resulting averages for the remaining can do statements for each class and stream are presented in Table 6.11.

Table 6.11. *Averaged Teacher Perceptions of Proportions of their Classes who can Perform A1–B1 CEFR Statements*

June–August 2015 Survey (13 can do statements) FE		
Class/Stream	**Number of classes**	**Average**
A1–A2 Course	5	54%
A2–B1 Course	5	84%
A1–A2 Low Stream	3	42%
A1–A2 High Stream	2	72%
A2–B1 Low	3	79%
A2–B1 High	2	91%
July 2016 Survey (12 can do statements) FE		
Class/Stream	**Number of classes**	**Average**
A1–A2 Course	6	68%
A2–B1 Course	5	80%
A2–B1 Low Steam	3	74%
A2–B1 High Stream	2	88%
January 2017 Survey (17 can do statements) FE		
Class/Stream	**Number of classes**	**Average**
A1–A2 Course	6	74%
A2–B1 Course	5	81%
A2–B1 Low Steam	3	76%
A2–B1 High Stream	2	88%
July 2016 Survey (19 can do statements) SE		
Class/Stream	**Number of classes**	**Average**
A1–A2 Course	5	58%
A2–B1 Course	6	86%
A2–B1 Low Steam	3	85%
A2–B1 High Stream	3	86%
January 2017 Survey (24 can do statements) SE		
Class/Stream	**Number of classes**	
A1–A2 Course	5	57%
A2–B1 Course	6	83%
A2–B1 Low Steam	3	82%
A2–B1 High Stream	3	86%

The results show clear differences in the average percentages between the A1–A2 course and the A2–B1 course, with the A2–B1 course ranging from 7–30% higher than the A1–A2 course across the surveys and FE and SE courses. Unfortunately, it is not possible to run inferential statistical texts on the mean difference between the ratings for the two courses due to the small sample size of classes. Nevertheless, this provides some moderate backing for the first assumption of the utilization warrant for the BET course placement function.

On the other hand, the differences between the class streams within the GE courses is less well defined, ranging from just 1% to 30%. Because of the very low difference in some cases between the two streams within the A2–B1 course, i.e. SE 2016 (1% difference) and SE 2017 classes (4% difference) this is judged to be overall weak rebuttal evidence for the class streaming function of the BETs.

6.6.3 Backing/rebuttal 3 for assumption 1 of the utilization warrant

Teachers in both the 2015 and 2016 focus group interviews generally expressed agreement that the BETs were able to place students effectively in courses and stream students effectively into classes, aside from the occasional student who teachers noted seemed to have been misplaced. One teacher in the 2015 group stated that:

> the classes are composed of relatively equal levels. So that we can move at a pace that is appropriate to them and we can supplement it as needed. As opposed to having so many high-level girls and then low-level girls all in the same class …”

Another teacher in the 2016 focus group stated that “I think everybody [in my class] is almost at the same level.” In addition, a teacher in the 2015 interview stated that:

> I think it’s really good. I think it’s good especially in the low-level classes, you get the ones that pop out as kind of a bit of a leader and that gives them extra confidence that they might not have if they were in a mixed stream class.

Another participant in the 2015 focus group interview stated that “I think it’s better for everybody, students and teachers.”

On the other hand, there were contrasting opinions about the value of placing students into two streams in the lower A1–A2 course, as had been the streaming policy for 2015–2016 academic year, in which there were two streams in the A1–A2 course and two streams in the A2–B1 course. One teacher in the 2016 group expressed that if all students in one class had very low ability it could be demotivating:

> When you don’t have anyone that is anywhere in the ballpark of where they could go, then it just reinforces this kind of bottom of the pool, like scum, leftover. They just ... they know that ... they just feel that they’re stupid. They look around, they know that they’re stupid. They feel that because there’s no one else around them. So they’re like well, why am I even trying.

In contrast, another teacher found that her class in the lowest A1–A2 stream had been quite motivated:

> I had a really low-level last year but it was actually quite good. I didn’t have ... test scores, they were kind of the same as the other low-level classes, but they just seemed to just be more motivated in general and more enthusiastic students in general.

Teachers in the 2015 focus group interview also expressed a desire to have more measures upon which to base student placement and streaming, in order to reduce the chance of placement errors. One teacher stated:

> It gets difficult. There are people, like I’m horrible at tests, which is why I ended up pursuing a degree in history. So, I can write papers. Because every time I sat down to take a test, even multiple choice, I don’t know. I’m just really bad at it. So, some people just aren’t good test takers, but how do you account for that?

Although there was some disagreement in the focus group interviews about the value of streaming students into classes within the lower GE course, there seemed to be fairly strong general agreement that the BETs did a good job overall of dividing students into courses. Thus, evidence from the focus group interviews constitutes fairly-strong backing for the first assumption of the BET utilization warrant.

6.6.4 Backing/rebuttal 4 for assumption 1 of the utilization warrant

The fourth backing sought for assumption 1 of the utilization inference comes from person separation and strata statistics as explained in section 4.3.1. The person separation and strata statistics for the BETs which were used for course placement and streaming within the frame of this study are presented in Table 6.12. Strata statistics are calculated by multiplying the person separation statistic by four, adding one, and then dividing this total by 3 (Wright & Masters, 2002).

Table 6.12. *BET Person Separation and Strata*

BET	Person Separation	Strata
BET1 2015	2.11	3.15
BET1 2016	2.04	3.05
BET2 2016	1.97	2.96

Table 6.11 shows that the three BETs used for course placement within the frame of this study all had person separation statistics above or very close to 2, and strata statistics around 3. This provides strong backing for assumption 1 of the utilization inference warrant by showing the BETs used for course placement within the frame of this study were able to divide test takers into two distinct ability groups for course placement purposes. On the other, hand this provides strong rebuttal evidence for the 2015 policy of streaming both the A1–A2 course and the A2–B1 course into two streams each, resulting in four levels of classes, as the strata statics indicate that these tests could at most divide the test takers into three distinct levels. The 2016 streaming policy of dividing students into two levels of courses, and then streaming the higher course into two levels of classes, however, has supporting evidence from the strata statistics of around three.

6.6.5 Backing/rebuttal 5 for assumption 1 of the utilization warrant

The fifth backing sought for assumption 1 of the warrant for the utilization inference comes from the results for Likert scale statements relevant to class streaming from student surveys as explained in section 4.6.1. The results for the five statements relevant to the BETs placement and streaming functions are presented in Part A of Appendix X.

It can be seen from the results that a majority of students agreed with all five statements relevant to class streaming, across all of the survey administrations. The level of agreement was above 80% for all except one statement for the January, 2017 second year responses, which was the statement that “In my class, most students’ English skills are similar to mine.”, which had a proportion of 76.88% of respondents agreeing. However, the strength of agreement is only overall only moderate as indicated by most averages for the 1–6 Likert scale being just or somewhat above four and also by large proportions of students who only agreed “somewhat” to the statements (29.13%–54.02%). Overall backing for assumption 1 of the utilization warrant for the BETs within the frame of this study is judged to be moderately strong.

6.6.6 Backing/rebuttal 6 for assumption 1 of the utilization warrant

Phi(lambda) statistics for the BETs used for course placement in this study, i.e. BET1 2015, BET1 2016 and BET 2 2016 are reported in Table 6.13 along with mean BET scores for these tests, and the cut-score used to separate students into the A1–A2 and A2–B1 courses.

Table 6.13. *Course Placement BET Phi(lambda), Mean and Cut Scores*

BET	**Phi(lambda)**	**BET Mean Score**	**BET Cut Score**
BET1 2015	.78	44.77	45
BET1 2016	.77	51.24	51
BET2 2016	.77	60.01	58

All three phi(lambda) statistics for the BETs used for course placement are very close to the criteria set of .8. Given that the test taker population has relatively homogenous English language proficiency, meaning that there are likely to be many test takers around the A1/A2 ability borderline, and also taking into consideration that the cut scores are quite close to the test mean scores (which lowers phi(lambda)) the reasonably high dependability indicated by these phi(lambda) statistics is taken to be moderately strong backing for assumption 5 of the utilization warrant, and therefore for the placement function of the BETs.

In order to inform future revisions of BET cut scores phi(lambda) statistics for cut scores further from the mean for the BETs in this study are presented in Appendix Y. Cut-scores and the resulting proportions of students who would have been placed in the A1-A2 and the A2-B1 streams for phi(lambda) statistics of .8, .85 and .9 are given. It can bee seen from the results that obtaining the target phi(lambda) of .8 in order to have fewer lower ability students mis-classified into the higher stream, would result in approximately 40% of students going into the higher stream in the years covered by this study. Similarly, to have fewer higher level students mis-classified into the lower stream, lower cut scores needed for a phi(lambda) of .8 would result in around 29-34% of students going into the A1-A2 stream. Given that approximately 50% of GE students were placed into each GE course based on BET results over the time of this study, it can can be seen that in order to move the phi(lambda) by two or three percentage points to reach the target of .8 set in the BET IUA, the proportion of students who would change their GE course placement is around 10-16% depending on the year and direction of the cut-score change. Given that evidence from teachers and students presented elsewhere in the BET validity argument shows that both students and teachers are largely in agreement with course placement decisions, and that the the BETs are only a moderate stakes test. It does not seem worth changing course placement decisions for such a large proportion of students in exchange for such a relatively small gain in dependability.

6.6.7 Backing/rebuttal 1 for assumption 2 of the utilization warrant

As explained in section 4.6.2, in addition to the evidence presented in other warrants of the BET validity argument, additional backing for the achievement function of the BETs was sought as backing for assumption 2 of the utilization warrant that *BET scores are* *sufficient and relevant for assessing student achievement of the GE course goals*. The first backing was sought from the results of two items on the teacher attitudes to the BETs surveys which elicited teachers' opinions of the efficacy of the BETs for measuring students' reading and listening proficiency. These results are attached as Part B of Appendix W. The results show fairly high agreement (means between 4.75 and 5.27 on a scale of 1–6) from teachers and LAs who responded to the survey that the BETs were effective at measuring both students' reading and listening proficiency. Generally, the results cluster at "agree" with a few respondents indicating that they strongly agree. However, it is interesting to note that there is one outlier who indicated somewhat disagreeing that the BET effectively measured both reading and listening in the 2015 survey, and also one outlier who somewhat disagreed that the BETs effectively measured listening in the Jan/Feb 2017 survey. It would be valuable to find out why this respondent or these respondents disagreed, however, as the survey was anonymous this is not possible. The proportion of survey respondents agreeing with these two statements ranged from 87.5% to 100% across the three surveys. Overall, these results provide moderately strong backing for assumption 2 of the utilization warrant. A full discussion of the effectiveness of the BETS within the frame of this study to function as achievement tests of the new GE curriculum is given in section 7.2.2 in answer to research question 2.

6.6.8 Backing/rebuttal 2 for assumption 2 of the utilization warrant

The second backing for assumption 2 of the utilization warrant was sought from reliability statistics of the BERTs and BELTs in this study. As explained in section 4.6.2 reliability statistics of at least .8 would be needed as backing to support an assumption that BET scores are sufficient for assessing student achievement of the GE course goals. As can be seen in Table 6.1 in section 6.3.1, the BERTs and BELTs all have reliability statistics below .8. The BERTs have KR-20 reliability between .76 and .78 and the BELTs have KR-

20 reliabilities between .56 and .67. Overall, this represents moderate rebuttal evidence for assumption 2 of the utilization warrant.

6.6.9 Backing/rebuttal 1 for sub-assumption 3.1 of the utilization warrant

Assumption 3 of the BET IUA is that *the BETs have beneficial impact on learning*. This assumption is subdivided into three sub-assumptions, each of which addresses a different beneficial aspect of the uses of BET scores on learning.

The first sub-assumption is that *students benefit from being placed in courses and streamed classes based on BET results.* Backing for this sub-assumption 3.1 comes from analysis of an item on the student attitudes to the BET survey, which addressed student preferences to being in classes divided by ability, i.e. a class of similar English ability to their own or in a class of mixed ability. The survey item is given in Japanese and in English in section 4.6.3.1 and the results for the students who were streamed based on BET results in this study are presented in Table 6.14.

Table 6.14. *Student Class Streaming Preference from Surveys*

Survey Administration	**Same level class preference**	**Mixed level class preference**	**Whichever is okay**
April/May 2015 (*n* = 172)	133 (77.23%)	21 (10.47%)	18 (12.21%)
Jan 2016 (*n* = 205)	157 (76.59%)	18 (8.78%)	30 (14.63%)
Jan 2017 FE (*n* = 214)	163 (76.17%)	19 (8.88%)	32 (14.95%)
Jan 2017 SE (*n* = 193)	146 (75.65%)	24 (12.44%)	23 (11.92%)

It can be seen from Table 6.14 that a large majority of students for all of the student attitudes to the BETS and class streaming surveys administered during this study preferred to be in English classes of students made up of a similar ability to their own and this proportion is quite consistent across all of the surveys at just over 75% of survey takers. In addition, a fair proportion of survey takers appear to be indifferent to course streaming with around 12–15% of students choosing this option in the surveys. There is also a reasonable proportion of students who would prefer to be in non-streamed classes or classes of mixed ability with around 9–12% of respondents choosing this option.

Overall these survey results provide fairly strong backing for assumption 1 of the third assumption of the warrant for the BET IUA as more than 87% of respondents across all survey administrations either had their steaming preference met, or were indifferent to streaming policy.

6.6.10 Backing/rebuttal 2 for sub-assumption 3.1 of the utilization warrant

The second backing for sub-assumption 3.1 of the BET utilization inference warrant was sought from teacher responses to questions 8 and 9 in the focus group interviews, which were: "What are the effects of current GE steaming policy for teacher classroom management, and class preparation?" and "How beneficial is the current GE streaming policy for students?" Teachers in the 2015 focus group interview all agreed that course placement and class streaming based on BET scores was beneficial for students.

Some representative comments from the 2015 focus group interview are:

> I think it's really good. I think it's good especially in the low-level classes, you get the ones that pop out as kind of a bit of a leader and that gives them extra confidence that they might not have if they were in a mixed stream class.

Another participant stated: "I think it's better for everybody, students and teachers."

Focus group participants in the 2015 interview also noted that lower level students would be more likely to participate in a class of similar ability to their own, because they might be intimidated and reluctant to participate if students in their class had a much higher ability than them. "Probably the low-level students would be more inhibited by giving their presentation after seeing a really awesome, high-level presentation. Better to keep it as separated a bit."

One participant in the 2016 focus group stated “[It’s] beneficial to students because they can participate more knowing that the other students are on the same level so they’re not pressured probably.” Aside from this teacher, however, in 2016 focus group, other teachers did not directly express agreement that streaming was beneficial for students as they did in the 2015 group, because the conversation wandered to the topic of dividing the A1–A2 course into two streams. Thereafter, a teacher expressed that it was not beneficial to have a very low class stream as was the case with the 2015 streaming, stating that “The danger is if you have everybody in that class who’s super low motivated it can be even more deadening for everyone”. Another teacher stated the opinion that:

> I would say that in terms of the low-level streaming class, it’s important to understand what your goal is for the course. Is your goal to extend these students who can do as much possible or is your focus more on supporting the lowest level learners? And if your focus is on the extension on where ever student can go, pushing them as far as possible, it would make sense to sort the stream a lowest of the low and then a middle-low. But I don’t think GE is at the point where we’re looking to be such an intensive, pushing course and we’re more on the supporting and nursing our students who are at the lower end of the stream. And so I think this current set up makes more sense for our students and for our university as a whole for what we are intending.

Overall, there seemed to be some agreement among the participants that it was better not to divide the A1–A2 course into a high and low stream as was the 2015 policy, however not all participants expressed their opinion on this issue.

Overall, evidence from the 2015 and 2016 focus groups provides moderate backing to support the claim that course placement and class streaming based on BET scores is beneficial for GE students, as available comments indicated that teachers generally agreed that course placement and class streaming was beneficial for students. In addition, teacher opinions expressed in the 2016 focus group generally supported the policy change implemented in 2016, of not streaming the A1–A2 course into a high and low stream.

6.6.11 Backing/rebuttal for sub-assumption 3.2 of the utilization warrant

Backing for sub-assumption 3.2 of the utilization warrant that *students find BET scores to be informative* was sought from responses to Likert scale items on student attitudes to the BETs and class streaming surveys as outlined in section 4.6.3.2. These results are presented in Appendix X, Part B. The results show that a large majority of students for each of the survey administrations believed that the BETs were useful to show them their English ability level for reading and listening with over 90% of students agreeing that the BETs helped them to know their strong and weak points in English and their reading and listening proficiency level. Over 90% of the respondents on all of the survey administrations also agreed that the BETs were useful for them to plan their English study.

However, one caveat is that a fair proportion of students disagreed with these statements (around 5–11% across the statements and surveys), and also a large proportion of students (29–40%) only somewhat agreed with the statements across the survey administrations. This indicates that further improvements are needed in the way that BET scores are presented to students. Some suggestions for how this may be achieved are presented in sections 7.3.6 and 7.3.7

Overall, results from the relevant items in student surveys provide moderate backing for sub-assumption 3.2 of the utilization warrant.

6.6.12 Backing/rebuttal 1 for sub-assumption 3.3 of the utilization warrant

As described in section 4.6.3.3 the first backing for sub-assumption 3.3 of the utilization warrant that *students find BET scores to be motivating* was sought from answers to two Likert scale statements on the student surveys. Descriptive statistics for the responses to these two statements for students who had recently taken the BET1 or the BET2 are provided in Appendix X, Part C. Students who had recently taken the BET3 were not asked to respond to these survey items, because they had finished the GE curriculum and would therefore not take another BET.

The results show that a large majority of respondents (over 95%) in all survey administrations reported that they were motivated to improve their scores on their next BET, and a large majority of respondents across all surveys felt motivated to study hard in order to improve their BET score (over 94%), and overall the degree of agreement was fairly high with averages of 4.72–5.04 on a scale of 1–6. On the other hand, here was a significant proportion of students who either disagreed that they wanted to study to improve their BET scores or only somewhat agreed, with around 30% of respondents in all three surveys either disagreeing that they wanted to improve their next BET score or only somewhat agreeing, and around 40% either disagreeing or just somewhat agreeing that they were motivated to study hard to improve their next BET score.

Overall, results of the student survey provide fairly-strong backing for sub-assumption 3.3 of the BET utilization warrant.

6.6.13 Backing/rebuttal 2 for sub-assumption 3.3 of the utilization warrant

In contrast to the results from the student surveys, teachers in both the 2015 and 2016 focus group interviews generally expressed beliefs that GE students were not motivated to study hard to get a good score on the BETs. In the 2015 interview teachers agreed that the BETs were probably not a motivational factor for those GE students who had low motivation to study English. One teacher in the 2015 group commented that “All they care about is passing and not having to repeat.” Reasons stated by teachers in the focus groups for the BETS not being a strong motivational factor for students were that the BETs form only 15% of students’ final grade for semester 2, and other components such as participation (20%) are of more immediate concern and carry more weight. Teachers also pointed out that students were only likely to think about studying for the BET right before the test, as students were busy and generally not concerned with long term planning. One teacher in the 2015 focus group stated that

> I think the BET test, they'll be really concerned about it right before the test, they're very short-sighted. You could give a writing assignment three weeks ahead and then last day, they'll come up to you all panicked. So I think it could be motivating but they really tend to be short-sighted …

On the other hand, one teacher in both the 2015 and 2016 interviews thought that the BET was useful extrinsic motivation, commenting that he often reminded his students about the BETs and BESTs to attempt to motive them. "I just noticed that as I've been teaching here, I've been having to constantly remind them of those extrinsic motivators at the end and I have noticed a difference." Teachers also stated that the lower level students probably don't think about their final test at all, but the higher-level students are likely to be more motivated to study hard to do well on the test.

Taken as a whole, teacher opinions expressed in the two focus groups provide moderately strong rebuttal evidence for sub-assumption 3.3 of the utilization warrant.

6.6.14 Backing/rebuttal 1 for assumption 4 of the utilization warrant

As described in section 4.6.4, the first backing for assumption 4 of the utilization warrant that *the BETs have beneficial impact on teaching* was sought from teacher comments in the focus group interviews. Two teachers in the 2015 focus group interview indicated that they taught simple test taking strategies with the BET in mind, and two teachers also indicated that they would tell their students when a classroom task was similar to a BET task to help them to prepare for their next BET. Another two teachers in the 2015 interview indicated that they simply told their students to review their lesson materials and vocabulary lists for the tests, because the BETs were achievement tests of the curriculum, therefore no special preparation was required.

In the 2016 focus group interview one teacher indicated that he would try to prepare his students for the BET by giving them tips on test taking strategies when they do reading and listening tasks in class with taking the BET in mind. These included reading the questions before reading the text, or giving students a time limit for answering a reading question to imitate test conditions. Another teacher indicated that he would have students highlight where they found the answer in reading text and discuss this with a partner, to improve their ability to answer test-like questions. One participant indicated that as a new teacher he had not yet had much time to think about preparing his students for the BETs stating:

> A lot of us are still trying to learn the content of the lessons … and then to facilitate for the first time, and, possibly as we see the BET, as we get a more substantial, um, better understanding of the BET as we go on we'll be able to then tailor our specific activities in the lessons to the, to, as [Participant 4] was saying, what skills students are going to need when they take the BET.

Another teacher mentioned that that he tells students that a task type will be on the BET after he has taught that task type in a lesson.

Overall results from the teacher focus group interviews provide weak rebuttal evidence for assumption 4 of the utilization warrant, with teachers indicating that the BETs caused them to teach some test taking strategies, and to raise students' awareness of the types of tasks on the BETs when a similar task appeared in the classroom materials. There is no indication of the BETs having a negative influence on teaching practice, nor is there any real evidence of the BETs having a positive influence.

6.6.15 Backing/rebuttal 2 for assumption 4 of the utilization warrant

The second backing for assumption 4 of the utilization warrant was sought from answers to teacher surveys as described in section 4.6.4. One teacher stated in the open response section of the January 2017 teacher survey that "Sometimes I informed students that a particular style of questioning would be on their exam, but I don't feel like BET performance was constantly in the back of my mind while guiding them through the tasks". A teacher in the Jan 2017 survey stated that:

> Maybe it's a wording thing, but if I read the final question as 'am I preparing for the BET test in classes', my answer would be 'no'. I'm not 'teaching for the exam' because that rationale has clearly failed our students in the past, and clearly doesn't work now. I think about how my teaching will affect my students' desire to study, and abilities to learn and use the materials in our lessons. IF I do this correctly / effectively, then my teaching will affect my students' BEST scores. However, I do take the time to point out if a reading/ listening style activity is the same style as one that will be in our tests, and how learning vocabulary will help them in their tests ... hence the 'occasionally' reply."

Results of the Likert scale item in the teacher attitudes to the BETs and class streaming survey are attached as Appendix X, Part C. It can be seen from the survey results that a majority of survey respondents on all three survey administrations (55.55%–63.63%) indicated that they thought about how their teaching would affect their students BET score when preparing for classes only rarely or occasionally. This presents further moderately strong rebuttal evidence for a positive influence of the BETs on GE teaching, and further research would be needed on the ways in which teacher class preparation is influenced by the BETs to draw any firm conclusions.

6.6.16 Backing/rebuttal 1 for assumption 5 of the utilization warrant

Assumption 5 of the utilization warrant is that *uses of BET scores have beneficial impact on senior managers' attitudes to the BECC assessment system*. As described in section 4.6.5 two backings were sought for this assumption from semi-structured interviews with university senior managers.

The first backing was sought from answers to the question "What do you think about streaming English classes for GE courses?" All of the managers interviewed were strongly in favour of dividing students into GE classes by their English ability based on BET scores. Stated reasons were to avoid lower ability students becoming frustrated, to challenge higher level students to further improve their ability, and to avoid teachers having to adjust their class to their lowest proficiency learners in a mixed ability class.

Some representative comments follow from the interviews.

Interview 1: Students were enrolled by different entrance exams. Some of them studied English very well and they were placed in a high-level class originally. Some of them were not good at English. Therefore, I think it is very important to grasp the characteristics and abilities of each student properly and to classify them according to their abilities. … This is a very good system.

これについてはですね、やはり学生さんが入ってきたときに様々な入試を通して入ってくる学生さんがいますので、また中には非常に英語をよく勉強し中には英語の特殊クラス、あの特定のクラスから入ってくる学生もいれば、英語を得意としない学生もいますので、それをきちんとそれぞれの学生さんの特徴や能力を把握して、その能力に応じてクラス分けをしていくというのは非常に大事なことだと思っています 。。。とてもいい制度だと思っています。

Interview 2: I totally agree with it. It’s really necessary to separate classes depending on the level, especially for the classes to acquire skills.

これはもう絶対必要なことだと思ってます。やっぱりグレード分けないと特にスキルを身に付けるための授業ではやっぱりグレードをきちんと分けていくっていうのは絶対必要なことだと思います.

Interview 3: That’s very necessary. It’s important to make a class that enables high performers to become motivated to aim for further levels. They are willing to study even outside of the BECC to get language training, like studying abroad.

それは大事な必要なことだと思いますね。出来る子たちがもっと自分たちで意欲的に学ぼう、もう少し上のレベルまで行こう、っていう、そういう学生たちもいますけれども、そういう学生たちが留学しようとかですね、語学研修の機会を求めてベック以外のところでまた勉強しようっていう意欲を持っていますから、そういう学生たちの更に上のレベルを目指していく意欲を引き出すようなハイレベルのクラスを作ることは大事なことだと思います。

Interview 4: I think there is a big gap in the students’ English levels even when they are in the same department, so it’s natural that classes should be divided … if we don’t do this, teachers will adjust the level of the class to the lowest proficiency learners, so if high and low performers are mixed together in the class, it won’t work out. This level classification is a very good idea and it’s necessary.

あの、英語のレベルというのは、同じ学科でもかなり学力に差があるんじゃないかと思いますので、当然クラス分けをしていかなくてはならないと思いますし…やはりある程度のレベルでクラス分けをしませんと、どうしても分からない学生に授業は合わせてしまいますので、ハイレベルな子と英語がさっぱり分からない子が一緒

にいたのでは授業が成り立たないと思いますので、このクラス分けはとてもいいと思いますし、当然やって行かなければいけないと思いますし、

Interview 5: It is common for teachers to focus on the average or just a little lower level in class when students with the highest level and students who are not high in level are in the same class. In that case, high performers won’t be able acquire new skills and it’s a problem. Classification by level is a necessary measure for the development of lessons and also to improve students’ skills.

もともと持っているレベルが高い学生があまりレベルの高くない学生と同じクラスになると、どうしても授業的には平均的なところかもうちょっと下のあたりに焦点を絞るっていうのが一般的だと思いますので。そうするとレベルの高い学生っていうのは、新しく何も身に付けることが出来ない、となりますから、それは非常にまずいと思いますので、そういった意味ではレベルによるクラス分けをしていくっていうのは、授業の展開上、それから学生さんのスキルアップさせる上でも必要な措置だと思いますので、それはそれでいいんじゃないかと思っているんです。

Of course, I think it has worth and meaning. That is to say, if low performers take the same class as high performers, they will stop studying because they get panicked. Probably they are the people who were not good at English in high school and don’t like English. Therefore, first of all we need to give them some experiences with the can do approach.

逆に、意味、価値はあると思ってるんです。というのは、下のクラスの子たちに上のクラスの子たちと同じ内容をしてしまうと、やっぱりもう分からないっていうのが先に立って、学生さんたちがもうそれ以上勉強しようとしないと思うんです。恐らく下のクラスの子たちは高校までのところで英語が苦手、あるいは嫌いという人

たちなので、まずはやればできるんだよっていう経験をさせてやる必要があると思うんですよね。

Overall, responses from the senior management interviews provide strong backing for assumption 5 of the utilization inference.

6.6.17 Backing/rebuttal 2 for assumption 5 of the utilization warrant

All five senior-managers who were interviewed also agreed with the value of giving students numerical scores representing their reading and listening ability. Some representative comments follow.

Interview 1: If they can see their result, like the score at the beginning of first semester, the one at the end of first semester, then the one in the second semester, it would be good. It will be very useful and meaningful.

やはりこれはどれくらい出来るようになったかという材料として与えられるのであれば、例えば前期の最初これくらいだった、それが前期の終わりにはこれくらいになり、そして後期これくらいになり、と、こういうふうにわかっていくような形であれば、こういうことがあってもいいんじゃないかと思います。それで出来れば非常に有用なんじゃないかな、意味があることなんじゃないかなというふうには思います。

Interview 2: When talking about analyzing connections with each department’s education, having this score is very important. It’s also important to show students’ scores to them. They should know how much skill they have now.

さっきの、その後の各学科の教育ともつながりを分析する意味では、このスコアがあるっていうのはとっても重要だと思っているんですけど、学生に示すっていう時

に、学生に示すことも大事です。彼女たちは自分の力がどれくらいあるかってことを知るべきですから。

Interview 3: Of course, it is very important to grasp the result by themselves of what they have studied and made efforts for, and I think it will also motivate them to move on to the next step. Therefore, this is very important.

これもまあ、当然自分が学んで努力をした成果っていうか結果を自分自身で把握するっていうことは大事なことですし、それが次のステップに進んでいく際のモチベーションにもなると思いますから、それはとても大事なことだと思います。

Interview 4: If you keep taking the tests, you can compare the results with the previous ones. Once again, it will motivate you. You can't know how much you can do, or how far you developed by a single test. But students take this test at the enrollment, at the end of first-year and so on, so they can receive the results constantly. To show the results to students has plenty of meaning.

継続してそのテストを受けていたら、前のテストと比較出来ますので、何度も言いますけれどもモチベーションにつながっていくと思います。単発で受けただけだったら自分がどれくらいなのか、伸びているのか分かりませんけれども、このテストの場合でしたら、入った時に受けて、１年生の終わりで受けて、っていうふうにちゃんと定点観察、定期的に受けるわけですから、学生に示すことは私は意味があるとそういう意味では十分思います。

Interview 5: I think this is very important. We separate English ability into four skills in Japan. The fact that skill scores are shown separately is very meaningful. Quite often Japanese universities do not return a test score report. Some students told me about this the other day, and they feel uncomfortable about no feedback, because they had it when they were in high school. I think we all should think about it not only at BECC. Your current method is very good.

これ多分すごい重要だと思うんです。で、英語の場合は四技能とかいうふうに日本では言ってますけど、そういうふうに分けて示せるようになってるから余計やり易いところがあるのかもしれないと思っているんですけど、先程お話したことと関係しますが、そういった意味では分けられるところは分けて示されているというのはとっても意味があると思いますし、なかなか日本の大学ではテストをしても返さないっていうケースが多いんですね。この前も学生が言っていたんですけど、そういったフィードバックがないっていうのは、あんまり彼女たちに言わせると、高等学校まではフィードバックが一応あったので、ちょっと違和感があるらしいんですね。ベックの英語だけじゃなくて全体的に今後考えていかなければならない問題なんだと思いますけど、現状こうされていることは非常に素晴らしいんじゃないかなと思います。

However, one interviewee suggested exercising caution when giving students test scores as it could be demotivating to some already unmotivated students with low confidence in their English ability.

Interview 2: However, I suggest we should be careful when showing test scores to students who found English hard … maybe students from some departments don’t care about the results, for example, students from the Nutrition course, Human Welfare course, Psychology course. But the students who are in Global Communication course or Elementary Education course may suffer greatly. We should be careful and consider how to treat these students. It’s important that students know about their own skill but if each student gets her scores, it will be a problem. I mentioned earlier but education results in rank-ordering students. We consider it important to take some special care for students who are sensitive about their rankings.

ただ、ちょっとだけ気になるのは、英語でつらい思いをした子たちにとってはどんなものになるかっていうことをちょっと考えておかないといけないかもしれない。そうです。もしかすると例えば人間栄養学科とか人間福祉学科の子たちは、いいや、心理学科の子たちもいいや、って思うかもしれない。だけど例えばグローバルコミュニケーションの子とか、それから初等教育の子たちはすごく悩んでしまうかもしれない。ここのところはそういう子たちが出る可能性がある時にどういうふうに対処するかって考えて置かないといけないと思います。だから、本人たちがどれくらい力を持っているかって知ること自体大事だと思うんですけど、全体がそれを知ることになると。一人ひとりみんなが知ることになると、何ていうんだろう、そこでさっき言いましたけど、教育って順番がついてしまう。順番に対して敏感な反応をする学生へのケアが必要になるかなって思います。でも大切なことだと思います、これは。

Responses in the management semi-structured interviews show overall strong agreement with the value of presenting students with numerical grades to represent their English reading and listening ability, and thus provide strong backing for assumption 5 of the utilization inference warrant.

6.7 Closing comments

This chapter presented and analysed the backing/rebuttal evidence gathered to assess each of the assumptions beneath each warrant for the six inferences in the BET IUA. An overall summary and evaluation of this backing/rebuttal evidence is presented in the following Chapter 7, which is drawn upon to answer the research questions. Table 7.7 in Chapter 7 summarizes overall backing/rebuttal evidence for the BET validity argument.

CHAPTER 7

Summary, Conclusions and Reflections

7.0 Introduction

This chapter firstly gives a summary of the obtained backing and/or rebuttal evidence for each inference in the BET IUA, to form a judgement of whether each warrant related to the inferences is sufficiently supported. Secondly, each of the research questions articulated in Chapter 1 is addressed, in light of the evidence presented in this study. This chapter then presents limitations of this study, followed by reflections on using Kane’s argument-based approach to validation in the context of in-house tests, and a summary of this study’s contributions to the field of language test validation. Finally, suggestions for future research are made.

7.1 Summary of the results for the BET validity argument

The main purpose of this study was to build an Interpretation/Use Argument for the BETs and to apply it through a validity argument to the development stage of test validation, which took place over the first two years of implementing a new English language curriculum. The following sections evaluate the backing for each of the inferences of the BET validity argument for the first two years of test development.

7.1.1 Domain definition

The validity argument for the BET domain definition inference for BETs within the frame of this study is summarized in Table 7.1.

Table 7.1. *BET Domain Definition Validity Argument*

<table>
<tr><th colspan="4"><b><i>Domain Definition Inference</i></b></th></tr>
<tr><td colspan="4">Warrant for the BET Placement/Streaming Function IUA: <i>Observations of performance on the BETs reveal the level of reading and listening skills and abilities needed to function effectively in GE courses, and are representative of reading and listening performance in the General English curriculum.</i>
Warrant for the BET Achievement Function IUA: <i>Observations of performance on the BETs reveal achievement of the General English course goals for reading and listening.</i></td></tr>
<tr><th>Assumptions underlying the warrant</th><th>Backing/Rebuttal Evidence</th><th>Analysis of Evidence</th><th>Placement/Streaming or Achievement IUA</th></tr>
<tr><td rowspan="2">1. BET content is a representative sample of the GE curriculum.</td><td>1. Tables analysing the representation of GE unit content on the BETs, show good representation for BET2 2016 and BETs 2 & 3 2017, but poor representation for BETs1 2015, 2016 and 2017.</td><td>Overall, moderately strong rebuttal evidence for the BETs within the frame of this study, which is somewhat moderated by evidence of improving representation within each cycle of BET revision</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>2. An analysis of teacher and student survey responses shows that overall these stakeholders think that BET tasks are representative of the targeted GE curriculum.</td><td>Moderately strong backing evidence</td><td>Placement/streaming and Achievement</td></tr>
<tr><td rowspan="3">2. BET task characteristics match the GE curriculum goals</td><td>1. Analysis shows that several of the target can do statements in the BET specifications are not well-matched to the testlet content and/or curriculum goals.</td><td>Strong rebuttal evidence for this assumption</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>2. An incomplete range of subskills at the curriculum target CEFR levels for reading and listening is covered by the target can do statements from the BET specifications.</td><td>Strong rebuttal evidence for this assumption</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>3. BERT and BELT lexical profiles are somewhat different to KET and PET reading and listening section lexical profiles particularly at the 1000-word level.</td><td>Weak to moderate rebuttal evidence for this assumption</td><td>Placement/streaming and Achievement</td></tr>
<tr><td rowspan="2">3. BET task types are representative of reading and listening task types in the GE curriculum.</td><td>1. Tables analysing the representation of BET type tasks in the GE curriculum show a lack of representation for some task types, and an uneven representation BET-type tasks across the target curriculum.</td><td>Strong rebuttal evidence for this assumption</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>2. An analysis of teacher and student survey responses shows that these stakeholders view BET tasks as being similar to reading and listening tasks in the GE curriculum.</td><td>Moderate backing evidence for this assumption</td><td>Placement/streaming and Achievement</td></tr>
<tr><td>Intermediate Conclusion</td><td colspan="3">Overall backing is insufficient to support the assumptions for BET IUA domain definition warrants for both placement/streaming and achievement.</td></tr>
</table>

As can be seen from Table 7.1, there is some evidence of improving backing for the domain inference during each of the three BET creation and renewal cycles covered by this study. This includes improved BET domain coverage, better matching of CEFR can do statements to testlet tasks in the BET specifications, and slightly improved task representation in the curriculum. There is also fairly strong backing from surveys that the two most important stakeholder groups for the BETs, teachers and students, agree with assumptions beneath the domain definition inference warrant that BETs are representative of GE curriculum content, and that BET tasks are similar to reading and listening tasks in the GE curriculum.

However, overall it is judged that the backing for some of the assumptions in the BET IUA domain definition warrant across the timeframe of this study, is outweighed by the rebuttal evidence, which shows that three of the six BETs available for analysis in this study did not sufficiently represent their target units of the GE curriculum, that the target can do statements for BET testlets in the BET specifications in several cases did not accurately reflect testlet tasks and/or the course goals, that only around half of the possible target CEFR detailed can do statements for each target level were covered by the BETs, that the vocabulary in the BETs was somewhat more difficult than the KET and PET, and finally that there was poor and uneven representation of BET task types in the target GE curriculum. These results points to a clear need for further revisions to the BETs, the BET specifications, and the GE curriculum which are dealt with in section 7.3.

7.1.2 Evaluation

The validity argument for the BET evaluation inference for BETs within the frame of this study is summarized in Table 7.2.

Table 7.2. *BET Evaluation Validity Argument*

Evaluation Inference			
Warrant for the BET IUA: *BET tasks yield consistent observed scores, which are not contaminated by construct irrelevant variance*.			
Assumptions underlying the warrant	**Backing/Rebuttal Evidence**	**Analysis of Evidence**	**Placement/Streaming or Achievement IUA**
1. The BETs were administered and scored consistently.	Descriptions of the procedures followed for annual BET administrations show that the tests were administered and scored consistently.	Strong backing	Both placement / streaming and Achievement
2. Appropriate procedures were followed to develop BET testlets and items.	A description of procedures used for developing the BET testlets and items, shows that appropriate procedures were followed, for example, peer review of draft testlets, and standardised checking of item characteristics. BET items were not piloted.	Overall moderately strong backing	Both placement / streaming and Achievement
3. BET task instructions were easily comprehensible for students.	Results of a student survey show that students thought that BET instructions were easy to understand.	Strong backing	Both placement / streaming and Achievement
Intermediate Conclusion	Overall backing is sufficient to support the assumptions for BET IUA evaluation warrants for both the streaming and achievement functions.		

Table 7.2 shows that the weight of evidence for the BET validity argument supports the assumptions beneath the warrant for the evaluation inference. On the one hand, strong rebuttal evidence for assumption three of the evaluation inference came from the fact that BET testlets were not able to be piloted, which is a consequence of designing the tests with limited resources. On the other hand, backing for assumption three of the evaluation

inference will improve as 2018 the BETs should have full domain coverage, meaning that the focus of test design can move to improving existing testlets and items, rather than designing new items, meaning that all testlets will have been previously used on the test taker population. In addition, strong backing was found from the consistent way in which the BETs were administered, from BET development procedures, and from student survey results showing that a great majority of students found the test instructions easy to understand, which is judged to be sufficient to support the evaluation inference warrant.

7.1.3 Generalization

Table 7.3 summarizes the validity argument for the BET generalization inference for BETs within the frame of this study.

Table 7.3. *BET Generalization Validity Argument*

Generalization Inference			
Warrant for the BET IUA: *Observed scores are estimates of expected scores on other versions of the BET (1, 2 & 3).*			
Assumptions underlying the warrant	**Backing/Rebuttal Evidence**	**Analysis of Evidence**	**Placement/Streaming or Achievement IUA**
1. Enough tasks are included to provide stable estimates of test taker performance.	Reliability statistics for the BETs 1 and 2 as a whole are sufficient for moderate-stakes in-house placement tests. Dependability statistics indicate that a large proportion of BELT variance is not based on ability in the curriculum domain. Reliability statistics, separation and strata statistics show that the BERTs and BELTs in this study could not divide test takers into three groups for A1, A2, and B1 levelling.	Strong backing evidence for the course placement function of the BETs Moderately-strong backing for the 2016 class streaming policy Overall strong rebuttal evidence for the achievement function of the BETs	Placement / streaming and Achievement
2. Appropriate scaling and equating procedures are used to measure student achievement across BET forms.	The BETs were not equated during the span of this study.	Strong rebuttal evidence	Achievement
3. Testlet specifications are well defined so that parallel tasks and test forms are created.	BET testlest specifications seem to be sufficient for testlets of equivalent difficulty target equivalent language skills to be created. Some testlet can do statements to not match the course goals, or task requirements well. Without equating the BETs it is not possible to compare testlet difficulty across BETs.	Overall weak backing evidence	Achievement
Intermediate Conclusion	Backing evidence is sufficient to support the placement/streaming function of the BETs, but insufficient to support the achievement function.		

Table 7.3 shows that strong supporting evidence for the placement function of the BETs was found from all BETs within the frame of the study having KR-20 reliabilities above .8, which is within the range generally found to be acceptable for a moderate-stakes test (Weir 2005a) and also indicates that the BET 1 and 2s were able to separate the learners into two groups statistically for placement into two courses. (Statistical evidence relating to the class streaming function of the BETs is assessed in the utilization inference.)

On the other hand, fairly strong rebuttal evidence for the assumptions beneath the warrant for the generalization inference was found in the BET validity argument. Specifically, KR-20 reliabilities for the BERTs and BELTs indicate that person separation and strata statistics for all of the tests would not be sufficient to divide learners into three groups (i.e. CEFR A1, A2 and B1) for course achievement purposes, that is to say, in order to measure course achievement in terms of separate reading and listening skills at levels A2 and B1 of the CEFR. In addition, it is not possible to assess the equivalency of testlet difficulty across BETs without equating the tests. Overall, the generalization inference for the achievement function of the BETs within the frame of this study is found to be insufficiently supported.

7.1.4 Explanation

The validity argument for the BET explanation inference for BETs within the frame of this study is summarized in Table 7.4.

Table 7.4. *BET Explanation validity argument*

Explanation Inference			
Warrant: *Observed scores are attributable to the constructs implicit in the CEFR levels A2–B1 for English reading and listening.*			
Assumptions underlying the warrant	**Backing/Rebuttal Evidence**	**Analysis of Evidence**	**Placement/Streaming or Achievement IUA**
1. Performances on BET measures are associated with performance on other test-based measures which claim CEFR alignment.	Large correlations were found between BET scores and OOPT scores. Large correlations were found between BERT and OOPT Use of English scores. Medium to large correlations were found between BELT and OOPT listening section scores. Medium to large correlations were found between BET and TOEIC overall scores. Small to large correlatiosn were found between BERT and TOEIC reading scores, and small to large correlations were found between BELT and TOEIC listening scores.	Overall strong backing evidence	Placement/streaming and Achievement
2. Performances on BET measures are associated with test taker self-assessments of CEFR can do statements for reading and listening.	Medium to large disattenuated correlations were found between BERT and reading self-assessment survey results, and between BELT and listening self-assessment survey results.	Fairly-strong backing	Placement/streaming and Achievement
3. BET items effectively measure the constructs of interest	Rasch item fit statistics show that BET items function to effectively measure the constructs of reading and listening. Point-measure correlations indicate several items may be off-construct and need revision or replacement	Overall moderate backing	Placement/streaming and Achievement
4. The range of item difficulties on the test matches the range of abilities of the test takers.	The range of item difficulties to test taker abilities shows improved matching across time of this study, and overall the matching is reasonably good. A few more items are needed at the more difficult end of the range to challenge and measure the most able test takers.	Overall moderate backing	Placement/streaming and Achievement
Intermediate Conclusion	Backing evidence for the evaluation warrant is sufficient to support the placement/streaming and achievement functions of the BETs.		

Overall there seems to be enough evidence in the BET validity argument at this stage to support the assumptions underlying the warrant for the explanation inference. The strongest backing comes from large correlations between BET overall scores and OOPT overall scores, BERT and OOPT use of English scores, and medium to large correlations between BELT and OOPT listing scores. This is because as explained in section 4.4.1 the OOPT is a test which had CEFR alignment build into it at all stages of its development and validation, and similar to the BETs, it aims to assess communicative ability against the CEFR, and to place students into CEFR levels (Pollit, 2009; Purpura, 2010).

Medium to large correlations between BERT scores and reading items in a CEFR self-assessment survey, and between BELT scores and listening items in the same self-assessment survey also provide moderately strong backing for this assumption 2 of this the warrant for the explanation inference. The correlations for the BERTs and reading self-assessment statements are similar to those obtained in Alderson's validation of the DIALANG (Alderson, 2006) for overall reading self-assessment and DIALANG reading scores, but are weaker for the BELTs than for the DIALANG listening results. Correlations may have been moderated by test-taker unfamiliarity with some of the actual situations, which the CEFR self-assessment statements referred (Ross, 1988), which would seem to be an unavoidable when using the CEFR can do statements with young learners with no actual experience in a country where the target language is spoken. The lower overall correlations between CEFR listening self-assessment statements and the BELTs for three of the BELTs (see Appendix M), may due at least in part to the lower reliability of the BELTs than the BERTs. As described in section 7.3.3 further work to improve BELT items and therefore the reliability of the BELTs should result in as least slightly stronger correlations, between BELT scores and CEFR listening self-assessment survey results, which would provide improved backing for assumption 2 of the explanation inference warrant.

Similarly, Wright maps for the BERTs and BELTs within the frame of this study show that there is a fairly good match of the range of test taker ability to test item difficulty, and thanks to targeted revisions this match improved across the BETs examined in this study, with the 2017 BERTs and BELTs examined, having better matches between tester ability spread and item difficulty spread than the 2015 and 2016 BERTs and BELTs examined. The match can be further improved, however, so this another area in which future BET revisions should lead to strengthening backing for an assumption beneath a BET validity argument warrant.

In contrast to backing for the first three assumptions of the warrant for the explanation inference, assumptions 4 of the explanation inference has weaker backing, because a small proportion of correlations between test taker Rasch ability scores and answers to some test items (point-measure correlations) are negative, indicating that these items do not measure the same construct as the other test items, or may have problems which are causing the items not to function effectively. In addition, a significant proportion of items have low point-measure correlations indicating that they may need revision to better assess the target constructs of reading and listening.

This problem is in part an unavoidable consequence of the lack of resources for BET development during the frame of this study, which meant that new and revised test items could not be piloted. In future it is important that such problematic items are identified, examined and revised which, should eventually eliminate items with negative or very low point-measure correlations, and therefore improve the backing of assumption 4 of the BET validity argument.

7.1.5 Extrapolation

Table 7.5 summarizes the argument for the BET extrapolation inference for BETs within the frame of this study.

Table 7.5. *BET Extrapolation Validity Argument*

Extrapolation Inference			
Warrant: *The constructs of reading and listening proficiency on the BETs account for the quality of linguistic performance in the domain of General English courses.*			
Assumptions underlying the warrant	**Backing/Rebuttal Evidence**	**Analysis of Evidence**	**Placement/Streaming or Achievement IUA**
Performances on the BET measures are positively correlated with other criteria of language proficiency in General English courses.	Backing: Medium to large correlations of BET scores with first semester GE grades. Medium to large correlations between BET scores and speaking test scores. Large correlations between BET scores and GE vocabulary quiz average scores.	Overall strong backing evidence	Placement/streaming and Achievement
Intermediate Conclusion	Backing evidence for the extrapolation warrant is sufficient to support the placement and achievement functions of the BETs.		

Strong backing for the single assumption beneath the warrant for the BET IUA extrapolation inference comes from large correlations between overall scores BET scores used for class placement and streaming, and later GE course grades at the end of students' first semester. Further fairly-strong backing comes from medium to large correlations between those same BET scores and students later speaking test grades. Final strong backing comes from large correlations between BET scores used for placement and streaming and later GE vocabulary test average scores. Overall it is judged that this provides sufficient backing for the extrapolation inference warrant that the BETs are able to predict students' linguistic performance in English in the domain of GE courses.

7.1.6 Utilization

The utilization argument for the BETs during their first two years of development is summarized in Table 7.6.

Table 7.6. *The BET Utilization Inference Validity Argument*

Utilization Inference			
Warrant: *Uses of BET scores are beneficial to stakeholders*			
Assumptions underlying the warrant	**Backing/Rebuttal Evidence**	**Analysis of Evidence**	**Placement/Streaming or Achievement IUA**
1. BET scores are sufficient and relevant for making decisions about GE course placement and class streaming.	1. Results of teacher surveys show that all teachers surveyed agreed that the BETs place/stream classes effectively.	Strong backing	Placement/streaming
	2. Results of a teacher perceptions of their class' ability survey, show clear differences between teacher perceptions of their class' ability between GE courses. Results of teacher perceptions of their class' ability surveys do not show clear differences between teacher perceptions of some class streams.	Moderate backing for the BET placement function and weak rebuttal evidence for the streaming function	Placement/streaming
	3. Teachers in the focus groups generally agreed that the BETs placed students effectively into courses and streamed them effectively into classes.	Moderately strong backing	Placement/streaming
	4. Rasch person separation measures indicate that the BETs can separate test takers into two distinct ability levels for course placement, and strata statistics indicate that the BETs can separate learners into 3 streams for class streaming.	Moderately strong backing for the BET course placement function and 2016/17 streaming policy. Strong rebuttal evidence for the BET 2015/16 streaming policy	Placement/streaming
	5. Results of student surveys show that majorities of students believe that their classmates' English ability is similar to their own and that their GE class is suitable for their level.	Moderately strong backing	Placement/streaming
	6. Phi(lambda) statistics show that the BETs classified students into the two course levels with reasonable dependability.	Moderately strong backing	Placement
2. BET scores are sufficient and relevant for assessing student	1. Results of teacher surveys show that most teachers think that the BETs effectively measure student reading and listening proficiency.	Moderately strong backing	Achievement

achievement of the GE course goals.	2. Low reliability statistics indicate that the BERTs and BELTs cannot separate test takers into three levels for assessment of the course goals for the two levels of GE courses.	Strong rebuttal evidence	
3. The BETs have beneficial impact on learning. 3.1. Students benefit from being placed in courses and streamed classes based on BET results.	1. Student survey responses show that a large majority of students prefer to be in classes divided by ability.	Moderately strong backing	Placement/streaming and impact
	2. Teachers in focus groups generally agreed that course placement and streaming is beneficial for students.		
3.2. Students find BET scores to be informative	Survey results indicate that a great majority of students find their BET scores to be useful indicators of their English ability	Moderately strong backing	Impact
3.3. Students find BET scores to be motivating.	1. Student survey results indicate that a great majority of students are motived to study to improve their BET scores.	Overall weak backing	Impact
	2. Teachers in focus group interviews generally think that many of their students are not motivated to study hard to get good BET scores.		
4. The BETs have beneficial impact on teaching.	1. Teachers in focus group interviews indicated that the BETs have little or no influence on their teaching.	Overall moderate rebuttal evidence	Impact
	2. Survey results indicated that a majority of teachers thought about how their teaching would affect their students' BET scores when preparing classes only occasionally or rarely.		
5. Uses of BET scores have beneficial impact on senior managers' attitudes to the BECC assessment system.	1. Senior managers in semi-structured interviews are in favour of GE course placement and class streaming.	Strong backing evidence	Placement/streaming and Impact
	2. Senior managers in semi-structured interviews agree with presenting students with scores to show their reading and listening ability.	Strong backing evidence	Achievement and Impact
Intermediate Conclusions	Backing for the utilization warrant and the rest of the BET validity argument is sufficient to support the placement function of the BETs. Backing for the utilization warrant and the rest of the BET validity argument is insufficient to support the achievement function of the BETs. Backing for the utilization warrant and the rest of the BET validity argument is sufficient to support positive impact from the BETs on learning, and on managers' attitudes to the BECC assessment system but insufficient to support positive impact on teaching .		

Assessing the utilization inference is somewhat complex as the assumptions beneath the inference relate to three intended functions of the BETs. These functions are course placement/streaming, achievement of the GE reading and listening course goals, and having beneficial impact on learning, teaching and senior managers' attitudes to the BECC assessment system. The efficacy of the BETs for their placement/streaming and achievement functions during the timeframe of this study are major research questions and therefore are evaluated in the sections answering these questions. The placement/streaming function of the BETs is evaluated in section 7.2.1, and the achievement function is addressed in section 7.2.2. Other beneficial impacts of the BETs on stakeholders are addressed in section 7.1.6.1. Before assessing the utilization inference assumptions in detail, it is important to summarize the whole BET IUA for their first two years of development. This summary is presented in table 7.7.

Table 7.7. Summary of support for BET IUA Assumptions and Warrants

Inference	Assumptions	Sufficient Support?	Warrant(s)	Sufficient Support?
Domain definition	1. BET content is a representative sample of the GE curriculum.	NO	Placement/streaming: *Observations of performance on the BETs reveal the level of reading and listening skills and abilities needed to function effectively in GE courses, and are representative of reading and listening performance in the General English curriculum.* Achievement: *Observations of performance on the BETs reveal achievement of the General English course goals for reading and listening.*	Placement/ Streaming NO Achievement NO
	2. BET task characteristics match the GE curriculum goals	NO		
	3. BET task types are representative of reading and listening task types in the GE curriculum.	NO		
Evaluation	1. The BETs were administered and scored consistently	YES	*BET tasks yield consistent observed scores, and are not contaminated by construct irrelevant variance*	Placement/ Streaming YES Achievement YES
	2. Appropriate procedures were followed to develop BET testlets and items	YES		
	3. BET task instructions were easily comprehensible for students.	YES		
Generalization	1. Enough tasks are included to provide stable estimates of test taker performance.	YES	*Observed scores are estimates of expected scores on other versions of the BET (1, 2 & 3).*	Placement/ Streaming YES Achievement NO
	2. Appropriate scaling and equating procedures are used to measure student achievement across BET forms.	NO		
	3. Testlet specifications are well defined so that parallel tasks and test forms are created.	YES		
Explanation	1. Performances on BET measures are associated with performance on other test-based measures which claim CEFR alignment.	YES	*Observed scores are attributable to the constructs implicit in the CEFR levels A2–B1 for English reading and listening.*	Placement/ Streaming YES Achievement

	2. Performances on BET measures are associated with test taker self-assessments of CEFR can do statements for reading and listening.	YES		YES
	3. BET items effectively measure the constructs of interest	YES		
	4. The range of item difficulties on the test matches the range of abilities of the test takers.	YES		
Extrapolation	Performances on the BET measures are positively correlated with other criteria of language proficiency in General English courses.	YES	*The constructs of reading and listening proficiency on the BETs account for the quality of linguistic performance in the domain of General English courses.*	Placement/ Streaming YES Achievement YES
Utilization	1. BET scores are sufficient and relevant for making decisions about GE course placement and class streaming.	YES	*Uses of BET scores are beneficial to stakeholders*	Placement/ Streaming YES Achievement NO Impact YES
	2. BET scores are sufficient and relevant for assessing student achievement of the GE course goals.	NO		
	3. The BETs have beneficial impact on learning.	YES		
	4. The BETs have beneficial impact on teaching.	NO		
	5. Uses of BET scores have beneficial impact on senior managers’ attitudes to the BECC assessment system.	YES		

7.1.6.1 Assessment of beneficial impacts of the BETs

In this section the beneficial impact function of the BETs is assessed. In the BET validity argument utilization inference presented and analysed in this study, backing was sought for three aspects of beneficial impact of the BETs on students' learning. These were: students benefiting from course placement/streaming, students finding BET scores to be informative and students finding BET scores to be motivating. Evidence for each of these aspects of positive impact was presented and evaluated in sections 6.6.9–6.6.13. Student survey results indicated that large majorities of students were in favour of being placed in classes divided by ability level, and teachers in the focus groups also generally agreed that course placement and class streaming were beneficial for students. Student survey results also consistently showed that a majority of students found their BET scores to be informative about their English language ability.

Regarding BET impact on student motivation, vast majorities of students indicated that they were motivated to improve their BET scores, and to study harder to improve their BET scores, which is taken to be fairly strong backing to support this aspect of positive test impact. On the other hand, as presented in section 6.6.10, most teachers in the focus group interviews agreed that the BETs were probably not a major motivating factor for many GE students, as a lot of students had generally low motivation to study English, the BETs formed just a small proportion (15%) of students final grade in only semester 2, and most students were unlikely to think forward to a test given at the end of the year.

Thus, evidence for the motivating influence of the BETs for students to study is mixed. While it is important to listen to the opinions of students, as they are the most aware of their individual motivations, the opinions of teachers are also important, because teachers have regular contact with the learners, and are able to make inferences about student's motivation from their actions and attitude toward homework and in-class tasks. It is also important to note that student surveys were given within a few weeks of taking the BETs, so students' motivation as a result of recently receiving their test scores is likely to be higher with test fresh in their minds, than it would be several months later, after the impression of receiving their test scores had faded. Weighting this backing from student survey results, and rebuttal evidence from teacher focus groups, and giving student perceptions a little more weight than teacher perceptions, overall there seems be weak backing evidence for the claim that BETs have a positive washback on student motivation to study.

Evaluating the overall body of evidence for the positive impact of the BETs on student learning, it is judged that there is sufficient evidence to support this assumption based on evidence from the three sub-assumptions.

The second impact claim for which evidence was sought in the BET IUA utilization inference was assessed through the utilization inference warrant assumption 4, which is that *the BETs have a beneficial impact on teaching.* Backing was sought from answers to a teacher survey and answers in focus groups. As explained in section 6.6.14 teachers in the focus groups reported that the BETs had little influence on their teaching except for attempting to impart their students with some test taking strategies and to raise their students' awareness of the types of tasks in the BETs. Teachers did not report that the BETs had a positive influence on their teaching beyond this, and thus results from the teacher focus groups can be interpreted as weak rebuttal evidence for this assumption. In addition, backing sought from teacher surveys for this assumption, attached as Appendix W, Part C, showed that large proportions of teachers on each survey thought about how their teaching would affect their students' BET scores only rarely or occasionally when preparing lessons. Overall, evidence from both teacher focus groups and surveys provides moderately strong rebuttal evidence for a claim that the BETs have a beneficial impact on teaching in the GE curriculum.

Finally, evidence from semi-structured interviews was sought as backing to support an assumption that *uses of BET scores have beneficial impact on senior managers' attitudes to the BECC assessment system.* Comments from senior managers indicate strong overall agreement with the BECC policy of course placement and class streaming based on BET scores, and also with the policy of presenting students with their overall BET grades, and BERT and BELT subsection scores. This indicates that the use of BET scores is in alignment with the expectations of this key stakeholder group, and thus has a positive impact on this groups attitudes to BECC assessment policies.

In conclusion, the available evidence, for the development stage of the BETs covered in this study, provides sufficient support for the BETs having a positive influence on student learning, and on senior management attitudes to the uses of BET scores, but insufficient evidence for the BETs having a positive influence on teaching practice.

7.2 Answers to research questions

In the following sections overall results of the BET validity argument are used to answer each of the three research questions. Answers to the research questions draw on evidence gathered to support warrants for inferences in a chain of six inferences for the interpretations and uses of BET scores outlined in the BET IUA (Chapter 4), and analyzed in the BET validity argument (Chapter 6). Different types of evidence are drawn upon in order to support the separate intended functions of the BETs as placement/streaming (norm-referenced) tests and achievement (criterion-referenced) tests.

7.2.1 Answer to research question 1

The first research question in this study is *to what extent did the BETs within the frame of this study fulfil their functions as course placement and streaming tests in terms of* *inferences in the validity framework?* This section answers this research question drawing on the evidence presented and the conclusion drawn in the BET validity argument for the BET development phase for the two academic years from 2015–2017. Firstly, the course placement function is examined, and secondly, the function of using BETs to stream classes within the two GE courses is discussed.

This section first reviews the evidence for and against the efficacy of the BETS in this study as course placement tests to place students into a CEFR A1–A2 course and a CEFR A2–B1 course. Then a conclusion is presented that minimally sufficient evidence exists in the BET validity argument to support the efficacy of the BETs as placement tests in their first two years of development, but not for the BET class streaming function.

Overall, fairly strong supporting evidence for the efficacy of the placement function of the BETs was found in the BET validity argument. Strong backing was found to support the assumption of the efficacy of the BETs as placement tests in the utilization inference, coming from the opinions of the two most important stakeholder groups of students and teachers. Student survey results as presented in Part A of Appendix X, and analysed in section 6.6.5 show that most students think that the ability of other students in their class is similar to their own and that the level of their class was appropriate to their English level. All teachers also agreed in both the anonymous survey (Appendix W, Part A) that The BETs do a good job of streaming students into classes by English language ability. In the teacher focus groups, comments also indicated that teachers generally thought the BETs place students effectively and that this placement is beneficial for both students and teachers. Further backing comes from Rasch person separation statistics which show the BETs are able to separate the test taker population into at least two distinct groups (Section 6.6.4). Additional evidence comes from the BET validity argument for the explanation inference in which medium to large correlations between scores on the BETs used for course placement and course grades in the semester following the test, and also later speaking, and vocabulary test grades, which shows that the BETs have reasonably strong relationship to other assessments of student English language proficiency after course placement. Furthermore, in the BET validity argument generalization inference, the BETs used for course streaming were shown to have reasonable KR-20 reliability statistics, showing that they had sufficient internal consistency for course placement purposes. In addition, evidence presented in the BET validity argument for the evaluation inference showed that appropriate test administration procedures were followed to minimize construct irrelevant variance. Backing gathered from the two key stakeholder groups of teachers and students for the domain inference in the BET validity argument (section 6.1.7) also provided support for the course placement function of the BETs showing that large majorities of both teachers and students regarded BET task as being similar to reading and listening tasks in the GE curriculum. In addition, strong support for the placement and streaming functions of the BETs was found in the opinions of university senior managers in semi-structured interviews.

On the other hand, a significant amount of rebuttal evidence for the placement function of the BETs also arose in the BET validity argument. This rebuttal evidence was found in the explanation inference in which it was shown that several BET items had point-measure correlations below or near zero, and particularly in the domain inference in which the evidence showed that half of the BETs within the frame of this study did not contain a representative sample of material based on the target units of the GE curriculum, that many target can do statements for BET tasks in the specifications matched poorly with the GE curriculum goals or with the actual task characteristics, and that expert opinion from teachers involved in curriculum renewal showed that BET task types were actually poorly represented in the GE curriculum.

Another argument that can be made against the effective placement function of the BETs in the 2015/16 and 2016/17 academic years, is that the cut scores were based on the approximate proportions of students shown to be at the A1 and A2 levels by the OOPT, rather than on the results of a cut-score panel as recommended by manual for *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009). When the OOPT was administered to a representative sample of entering freshmen in 2015 (see section 5.3.9.1 for a full description of OOPT administrations). OOPT results showed that around 47% were A0 or A1, and around 53% were A2 or higher. Extrapolating these proportions from the representative samples given the OOPT to the whole cohort, cut scores were set which divided the cohorts approximately in half. While this method of setting cut scores does have some empirical backing from the OOPT results and from the large correlations between overall BET scores and OOPT scores presented in section 6.4.1, a properly conducted cut-score panel (as suggested in section 7.3.6) would provide a stronger argument that the BETs are placing students at their appropriate CEFR levels for the two courses in alignment with the curriculum goals.

Strictly speaking, in Kane’s conception of validity if any span in the chain in inferences lacks sufficient support, or metaphorically speaking, if any span of the bridge is not sufficiently supported, it disallows all subsequent inferences. However, overall, in spite of the strong rebuttal evidence found for the placement function of the BETs in the domain and explanation inference in the BET validity argument, and the fact that course placement cut scores were based on approximate proportions derived from 2015 OOPT results, rather than from a cut-score panel using the actual BET tests, there does seems to be minimally sufficient backing found for the other inferences to indicate that the BETs did sufficiently achieve their placement function in 2015 and 2016 for practical purposes.

All of the evidence presented on the placement function of the BETs is also relevant to the streaming function of the BETs for placing students into either a high or a low stream within the two GE courses. For the streaming function, however, two additional pieces of evidence need to be considered. Firstly, the Rasch person separation and strata statistics for the BETs used for class streaming with values around 2 and 3 respectively as shown in Table 6.12 in section 6.6.4 are rebuttal evidence for the function of the BETs in streaming both the A1–A2 and the A2–B1 courses into two streams each in 2015, for a total of four ability levels.

Secondly, in the focus group interviews some teachers expressed opposition to the creation of a very low stream within the A1–A2 course for classroom management and student motivation reasons, and no teachers expressed an opinion on streaming the A2–B1 course. Thirdly, there were not clear differences between teachers' perceived ability of their classes in two A2–B1 streams for the 2016 SE course from the teacher perceptions of GE class ability surveys (section 6.6.2). Therefore, it is judged that there is insufficient evidence found in the BET validity argument to support both the 2015/16 class streaming policy, and the 2016/17 class streaming policy.

Given the lack of evidence of a clear teacher preference for streaming the A2–B1 course, it is recommended that teachers be directly asked their opinion of streaming the A2–B1 course, then if no clear preference for this policy is emerges, perhaps a simpler policy of dividing students into the two courses with no further subdivision into streams could be implemented.

7.2.2 Answer to research question 2

This section firstly reviews the evidence for, and secondly the evidence against, the efficacy of the BETS in this study as GE course achievement tests in order to answer research question 2, *to what extent did the BETs within the frame of this study fulfil their function as* *achievement tests in terms of inferences in the validity framework*? Then, after weighting the evidence, a conclusion is presented that the BETs did not function as effective course achievement tests within the frame of this study.

An examination of the BET validity argument for the development phase of validation covered by this study provides some backing for the BETs acting as effective achievement tests. In the domain definition inference section of the BET validity argument opinions elicited from both teachers and students in surveys showed that vast majorities of these stakeholder groups believed that the BETs contained a good representation of curriculum content, and that BET tasks were similar to reading and listening tasks in the GE curriculum. Furthermore, backing presented in the evaluation inference showed that the BETs were written and administered using sufficiently standardised procedures to minimize the chance of construct irrelevant variance contaminating test scores. Backing presented in the extrapolation inference also showed that BET scores had strong associations with other measures of language proficiency in BET courses, such as overall course grades and speaking test grades, indicating that BET scores were able to account for/predict the quality of linguistic performance in the domain of GE courses to a sufficient extent. Additional evidence was presented in the explanation inference, in which medium to large correlations between BET scores and other measures of the constructs of reading and listening, provided backing for a claim that the BETs measured the constructs of reading and listening implicit in the CEFR statements, which were used as GE course goals. Furthermore, the opinions of teachers elicited from surveys presented in the BET validity argument show that most teachers thought that the BETs were effective measures of student reading and listening proficiency, and that the level of agreement on the Likert scale was reasonably high, all of which supports the claim of the BETs being able to act as achievement tests of the GE course goals. Finally, strong support for the BECC policy of providing GE students with numerical scores representing their achievement of reading and listen ability in the GE curriculum was found from the opinions of university senior managers in semi-structured interviews.

On the other hand, strong rebuttal evidence was revealed for several of the assumptions underlying the warrant for the domain definition inference. Rebuttal evidence was found showing that several of the BETs within this study did not contain tasks based on a representative sample of units from their target years of the GE curriculum, that the target can do statements for BET testlets in the BET specifications were in many cases unrepresentative of the curriculum goals, or not very accurate to the actual task requirements, and that the spread and representation of BET type listening and reading tasks across the GE curriculum, although improving, was not sufficient in the first two years of the BETs and the new curriculum. Further strong rebuttal evidence comes from the low phi dependability statistics calculated for the BELTs within the frame of this study within the generalization inference. In order to effectively measure achievement of the curriculum reading and listening goals reasonably high phi dependability statistics would be required. However, the phi dependability statistics are quite low for the BELTs (although close to a sufficiency criterion of .8 for the BERTs). This indicates that the amount of variability in BELT scores which does not represent knowledge of the GE curriculum domain was too high for these separate tests to act as effective achievement tests. Additional strong statistical rebuttal evidence was presented in in the BET validity argument for the generalization and utilization inferences. For the BETs to be able to act as effective achievement tests for both the lower GE course stream, which has a CEFR A2 achievement goal, and also for the higher course stream, which has a CEFR B1 achievement goal, both the BERT and the BELT would have to be able to split students into three achievement groups, B1 for those who achieve the higher stream course goal, A2 for those who achieve the lower stream course goal, and A1 for those who have not yet achieved the course goals. However as discussed in section 6.6.8, reliabilities, and resulting person separation and strata statistics for the BERTs and BELTs within the frame of this study indicated that they were not able to separate students into three groups.

Final and decisive rebuttal evidence for the achievement function of the BETs was found in the generalization inference validity argument, coming from the fact that the BETs in this study were not equated and students were not presented with comparable scores from their BET1, BET2 and BET3. In order for stakeholders to know students' actual progress toward the curriculum reading and listening goals, directly comparable scores between tests would have to be generated through proper test equating procedures.

Overall, in spite of some solid backing for the achievement function of the BETs found in the domain, evaluation, explanation, extrapolation and utilization inferences of the BET validity argument in this study, decisive rebuttal evidence emerged from the backing sought for the domain and generalization inferences which overall outweighs the backing and clearly shows that the achievement function of the BETs was not supported by the evidence within the frame of this study.

7.2.3 Reflections on answers to research questions 1 and 2

In the previous two sections it was found that after evaluating the evidence in the BET validity argument the BETs fulfilled their function as placement tests during the timeframe covered by theis research, but did not yet fullfill their function as achievement tests. This indicates that the BETs were able to separate students by ability sufficiently well for norm-referenced decisions of course placement, (which rely on separating test takers along a normal distribution), but that they could not yet be used for achievement decisions or absolute decision based on the criterion of CEFR levels. Suggestions for improving the BETs and BET administrative procedures and documents to provide stronger backing for the BET placement/streaming function, and especially for the BET achievement function are provided in section 7.3.

7.2.4 Answer to research question 3

The third research question in this study asks: *based on any weaknesses found in backing for inferences in the BET validity argument across two academic years from 2015–17, what recommendations can be made for changes to the BETs and their accompanying documents, procedures and policies?*

The following section 7.3 recommends ten improvements, which if made, would strengthen the BET validity argument. Firstly, the test specifications must continue to be updated to match changes to the curriculum, and to better specify which CEFR can do statements each testlet targets. More detailed can do statements beyond the general CEFR reading and listening can statements should also be specified in the GE course outline and GE curriculum overview documents. Further work is also needed to improve the quality of BET items, and the BET developers may also consider rebalancing the number of listening and reading items and testlets. In addition, methods for equating the BETs should be investigated, with a view to providing comparable test scores across administrations. To truly achieve the achievement function of the BETs it will also be necessary to set cut scores at the goal CEFR levels for reading and listening for the BERT and BELT (and also the BEST), which would allow for more individualized and informative feedback to be given to test takers. It is also recommended that further measures be used for course placement beyond the BETs, at least for SE streaming. Finally, recommendations are made that the BET specifications be given a little more prominence in the lesson review process, that a review of the BETs and the BET specifications be included in future orientations for new GE teachers, and that BET validity evidence be shared with university senior managers.

This study demonstrates that a solid foundation for the BETs has been built, but the first iteration of the BET validity argument presented here has also shown that a good deal of further work remains to be done. Language testing experts point out that the development and validation of a test and interpretive and validity arguments is an iterative and ongoing process with no clear end. As Bachman and Palmer (2010, p. 430) describe:

> We have ... described the entire process of assessment development and use as iterative, with the possibility of making changes in the warrants in the AUA, the Design Statement, the Blueprint, and the assessment itself, at any stage. Thus, information that is collected during assessment use will also be used to provide backing for warrants and to improve the assessment itself. This information is also likely to present the test developer and user with evidence that may weaken some of the warrants of the AUA, so that they will need to go back and reconsider these, make changes in them, or make changes to the blueprint and assessment.

7.3 Conclusions and recommendations for reforming BET systems

In this section, firstly changes made to the BETs over the course of this study with a view to strengthening the BET validity argument are presented, followed by recommendations to further improve the BET system. Table 7.8 summarizes changes made to the BETs, the GE curriculum and associated documents and procedures in attempts to strengthen the BET validity argument over the timeframe of this study.

Table 7.8. *Revisions made to the BETs, the GE curriculum and associated documents and procedures to strengthen the BET validity argument*

BETs	Domain	Evaluation	Generalization	Explanation	Extrapolation	Utilization
2016	• FE course domain representation much improved from BET1 2015 to BET1 2016 • BET2 introduced with balanced coverage of topics from all FE units	• Draft testlets reviewed by at least one more peer, for the 2016 and 2017 BETs than the two reviewers for the 2015 BET1	• A ten item testlet added to the BERTs, and a six item testlet added to the BELTs in an attempt to increase reliability and dependability	• Recycled items with Rasch difficulty below the ability of the weakest test taker, with very low or negative point-measure correlations or with too strong distractors were revised.	N/A	• Streaming policy changed to three class streams in two courses, rather than four class streams, which is supported by strata statistics of around 3 for the BETs 1 & 2 as a whole.
2017	• BET3 introduced with fairly-balanced coverage of topics from all GE units • BERT 2016/2017 testlet specification can do statements updated to better match the curriculum and testlet task requirements • Slight improvement in the representation of BET testlet style tasks in the 2016 FE curriculum.	• Three points added to the listening and reading assessments checklist for making the 2017 BETs • Testlet peer review of the 2017 BETs done online, which allowed for more efficient asynchronous feedback	• Three items added to BELT testlet 1 in an attempt to improve reliability and dependability	• Recycled items with Rasch difficulty below the ability of the weakest test taker, with very low or negative point-measure correlations or with too strong distractors were revised.	N/A	N/A

In the following sections, ten recommendations for the improvement of BET system drawn from the results of the current study are provided. These suggestions relate to strengthing the BET validity argument for the three BET functions examined in this study of placement/streaming function, achievement function, and positive impact. The majority fo these suggestions, however, focus on providing stronger backing for the BET achievement test function, which was found to have insufficient backing. The achievement function of the BETs would involve making absolute criterion-referenced decisions about test taker's English language proficiency based on the CEFR scales.

7.3.1 Reviewing and revising the BET specifications

It is recommended that the BET specifications be given a thorough review and update taking into consideration the weaknesses in the specifications revealed in this study.

A useful method to facilitate an analysis of the test specifications and test items in relation to the CEFR would be to follow the specifications process outlined in the manual for *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009) using the tables and forms in the Manual appendices. Users in pilot studies "commented that completing these forms was a very good way to review and evaluate the coverage of the examination and to re-assess its fitness for its stated purpose (p. 29)" as well as resulting in improvements to the test specifications (Szabo, 2010). The manual states that:

> This is an awareness-raising process which cannot be undertaken by a single researcher or team member. Sometimes, this exercise throws up a lack of coherence between official test specifications, which may not have been revised for some years, and the test in practice – as represented by forms of the test administered in recent sessions. The exercise is certainly easier to complete if formal test specifications exist. If they do not exist, the

> process of completing the forms associated with this chapter will help the user to consider aspects that should be included in such a specification (p. 27).

An alternative would be to use the CEFR Grids developed by the Dutch CEF Construct Project to analyse BET tasks (Alderson, 2006), and then to use the results to update the task specifications for each BET testlet, including the target can do statements. Wu (2010) recommends using the Dutch CEF Construct Project over using the specifications in the Manual because “specifications of item-level comprehension operations, which should be equally important when test constructs are examined and compared, are overlooked” (p. 210).

Revising the BET specification based on the problems identified in this study as well as based on an expert review by teachers who are stakeholders in the test using either of the above resources would almost certainly lead to better backing for assumptions beneath the warrants for the BET IUA domain and generalization inferences, which would in turn greatly strengthen the BET validity argument.

7.3.2 Providing more detailed GE course proficiency goals

Results from the BET domain definition inference validity argument indicated that the BETs in this study tested less than half of the possible CEFR subscales for reading and listening at the target A2 and B1 levels. As noted in section 6.1.4, this is not a problem in itself, as the length of the BETs is time limited, which therefore limits the amount of testable reading and listening skills. It is also up to curriculum designers to decide on which aspects of the CEFR they will focus in their courses. However, these results do point to a need to clarify for all stakeholders, which specific reading and listening skills represented in the CEFR reading and listening subscales the GE curriculum aims to improve in GE students. Clearly stating these more detailed goals would be beneficial for students, teachers, and BET designers on the GEAC, and would also strengthen the backing for the BET domain definition inference validity argument.

7.3.3 Increasing BET reliability and dependability

Attempts were made during the timeframe of this study to increase the reliability and thus person separation of the BERTs and BELTs in order to bring the tests to a stage where they are able to statistically divide the learners into three groups for course achievement purposes. This was attempted in three ways. The first was to increase the length of both the BERTs and the BELTs, which is well known as the easiest way to increase test reliability (Bachman, 2004; Carr, 2011; Brown, 2005). The second was to attempt to better match the abilities of the test takers to the test items, which may also improve test reliability (Linacre, 2017b). This was attempted by rewriting/replacing items to increase the difficulty of items with Rasch item difficulty below the Rasch person ability of the lowest ability test taker. The third way of attempting to increase test reliability, was by revising recycled test items with negative or very low point-measure correlations, and/or too strong distractors.

However, as can be seen from Table 6.1 in section 6.3.1, this did not result in sufficient increase in test reliability to be able to claim that the BETs, BERTs and BELTs are able to split learners into three groups (i.e. A1, A2 and B1) for achievement classification purposes. Therefore, in addition to using low point-measure correlation to identify poorly functioning BET items it is also recommended that in future another measure from Classical Testing Theory item analysis, known as item discrimination, be used as a criterion for identifying items for revision. This is because increasing item discrimination increases test variance, which leads to increased test reliability. (Brown, 2005; Ebel, 1967; R. Green, 2013).

Item discrimination was not used as a criterion for BET item revision during the frame of this study, because it is generally used for item analysis for norm-referenced tests, in which it is important to discriminate learner ability as much as possible. As the BETs are classified as criterion-referenced tests, it was thus thought not to be appropriate to use item discrimination as a criterion for BET item revision (Bachman 2004; Brown & Hudson, 2002;

Fulcher, 2013). However, upon further reflection, if the BETs are to fulfil their function as achievement tests for two course streams with different levels of proficiency goals (A2 and B1) it is important for the BETs to be able to discriminate between learners, therefore it will be necessary for items which are not contributing to test reliability because poor item discrimination values to be revised.

As such, if the GEAC and BECC management wish to pursue a goal of CEFR certification for levels A2 and B1 based on BET scores it is recommended that criteria of item discrimination be used as part of item analysis for future BET revisions, using common NRT based thresholds for item revision or replacement. For example, a threshold of .25 for point biserial correlations (Khalifa & Weir, 2009; Programme for International Student Assessment, 2006), or a threshold for item discrimination of .3 (Ebel, 1979). Revising BET items using item discrimination as a criterion will likely lead to increased BERT and BELT reliability, which in turn will provide stronger backing for assumption 1 of the generalization inference warrant in the BET IUA for the achievement function of the BETs.

7.3.4 Changing the BET format

BET developers in future may also consider rebalancing the proportion of items in the BET listening and reading sections to reflect the relative importance of these goals in the GE curriculum. In the GE course outlines for the time period covered by this study there was no indication that reading skills are more important than the listening skills, so it seems a little incongruous that the BET reading sections were so much longer than the BET listening sections for the BETs over this period (see Appendix A for a summary of the BET formats). If reading and listening are to be given equal weight it in terms of importance as learning targets, then it would seem to make more sense for these test sections to be of equal length, so that they can be measured with equal accuracy. One step to move toward this could be to remove the BET testlet which targets vocabulary, as students are already tested regularly on the GE vocabulary lists as another component of GE assessment. Such decisions would of course depend on the current GE goals and learning priorities at the time of future BET revisions.

7.3.5 Providing stakeholders with test scores representing curriculum achievement

For the BETs to fulfil their function as achievement tests it will be necessary to make tests scores across the BETs directly comparable. Although the BETs are written from the same test specifications, and therefore should be of similar difficulty, it is possible that BET tests may vary to a significant degree in difficulty. Differences between the difficulty of individual BETs would lead to inaccurate comparisons between test scores for the purpose of assessing student development of reading and listening ability, and presenting indicators of progress in these skills to stakeholders. As Dorans, Moses and Eignor (2010) state "In reality, it is virtually impossible to construct multiple forms of a test that are strictly parallel …" (p. 4).

One way to make BET, BERT and BELT scores directly comparable across years to measure student progress, would be to equate the tests. Test equating attempts to put separate tests of the same construct on the same scale through statistical procedures, resulting in scaled scores for both tests which are directly comparable. Usually for test equating large samples of examinees are required (Kolen & Brennan, 2014). However, there are also several equating methods which are recommended for relatively small sample sizes such as the BETs using a common-item, non-equivalent groups design, in which a set of common items is used on both of the tests to be equated (see Kurtz & Dwyer, 2013 for a brief overview of these methods). A recent study by LaFlair, Isbell, May, Arvizu and Jamieson (2017) indicated that circle-arc equating may be the most practical and accurate method for test equating for relatively small-scale tests like the BETs.

However, test equating also has several challenges for a small, in-house programmes like the BETs. First test equating procedures are rather difficult for those who do not have a strong back ground in statistics. Secondly, the criteria for anchor items are quite stringent. Anchor items, or the common items between tests, should be a mini test version of the test as a whole, in that it should represent the total test both in content and statistical properties (Kolen & Brennan, 2014, von Davier Holland, & Thayer, 2004). The requirement for the anchor items to be representative of the content of the whole test is impractical for the BETs, which consist of testlets, each with a different question format, many of which are a set of questions based around a single text. As such, it would not be possible to make a representative mini test without including most of the actual test.

In addition, creating suitable anchor testlets which meet the requirements of having a mean difficulty similar to the total test (Petersen, Kolen & Hoover, 1989), a spread of item difficulty across the test item difficulty range (Linacre, 2017b), and a standard deviation similar to the whole test, is also quite difficult, given the limited resources available for BET development. (On the other hand, it should be noted that Sinharay and Holland (2007) have presented research suggesting that the requirement for a spread of anchor item difficulty similar to the overall test may be relaxed.)

Making test scores directly comparable between the BETs is essential to provide strong backing for the BET IUA generalization inference. In addition, equating the BETs would likely provide further evidence to support the explanation inference of the BET IUA by producing additional evidence of BET alignment to the implied CEFR constructs of reading and listening.

Given in the level of technical expertise required for test equating it may be advisable for those involved in future development of the BETs, and also for other test developers in similar situations to hire an external consultant to train those involved in test development in the appropriate procedures, if the resources for this are available. In addition, given the importance of test equating for achievement tests, from a broader perspective on the education of language teachers, it would be valuable for those in the field of assessment teacher training to consider including test equating, particularly for small sample sizes, in in their published textbooks and in courses on language testing, as I could not find simple step-by-step instructions for test equating in any of the available language assessment textbooks.

7.3.6 Setting CEFR-based cut scores

In order to make a strong claim that the BETs are able to place students into their CEFR levels for reading and listening to assess achievement of the GE course goals it will be necessary to set cut-scores. This would require following the procedures presented in the *Manual Relating Language Examinations to the CEFR* (2009) and preparing a cut-score panel. There are various procedures for running cut-score panels (see Kaftandjieva, 2004 for an overview), and in the case of the BETs the panel would most conveniently consist of teachers and learning advisors in the BECC who are familiar with the GE curriculum and students. Initial attempts were made in 2016 to set cut-scores for the BERTs, however, errors made in the preparation of panel materials led to the results being unusable. Before conducting such a further cut-score panel it will be advisable to improve the reliability and person separation of the BERTs and BELTs (see section 7.3.3). In addition, it would be useful to provide phi(lambda) statistics to the cut-score panel for different possible cut scores, and also the proportion of students who would fall into each category for differen cut-scores, as part of the cut-score setting procedure. This would be similar to the information in Appendix Y, for the BETs within the frame of this study that were used for streaming/placement purposes.

7.3.7 Presenting BET results along with can do statements and study suggestions

Once the BETs have been equated and cut-score panels successfully run, it would then be possible to provide students with CEFR or localized can do statements showing what they are able to do in terms of reading and listening in English. Presenting these can do statements, along with study suggestions to improve students' English in each area covered by a can do statement, which are targeted at the students current level, would provide invaluable diagnostic information for students.

As one of the senior managers stated in his interview:

… students are curious about their progress. It's not easy to understand it unless if you show them the details very specifically. It's better to give them detailed feedback, in a manner, for example, a part of your result is this, another part is this, this part is difficult for you, etc. In that sense I think it is very meaningful to divide details whenever possible and show them separately.

学生さんって自分がどれくらい出来たかやっぱり知りたがっていますので、一つの授業全体としてどうかよりも、細分化出来るものは細分化して、この部分はあなたはこうですよ、別の部分はこうですよ、ここはちょっと難しかったね、とかいうようなのを具体的に示してあげないと、多分自分が本当にどうだったかっていうのは理解しにくいですよね。

7.3.8 Using further measures of student ability for placement/streaming decisions

It is well known and widely advocated that using multiple methods and formats to test language proficiency is likely to lead to fairer and more accurate placement decisions (Powers, 2010a). It is not possible to use further inputs for FE streaming decisions because the BET1 is administered only a few days before classes commence. For SE streaming decisions to place students into GE second-year classes, however, it would be possible to use further sources of information about student English language ability such as BEST (speaking test) grades and overall GE course grades. Indeed, progress was made on this front as BEST scores were used just after the frame of this study as part of SE streaming for the 2017/18 academic year. Looking further into the future it is recommended that FE course grades also form a portion of the score used for placing students into the two GE courses.

7.3.9 Increasing teacher familiarity with the BET specifications

In both focus groups in 2015 and 2016 participants commented that it would be beneficial to make the BET specification more prominent, and to increase teacher familiarity with the specifications. Teacher survey results indicate that the specifications were being more actively referred to in the lesson revisions done in the second semester of 2016 than in previous rounds of design of the new GE materials, and the increased number of teachers on the GEAC also undoubtedly increased general familiarity with the specifications. However, it is suggested that steps be taken to further familiarize teachers and LAs with the BET specifications. One teacher in the 2016 focus group stated “So I would like to make a recommendation of the BET specs if they are ... I mean they’re much more important now, they need to be updated more often, and maybe they need to be more publicly distributed to teachers, maybe as more of a reminder to reference it. So maybe once every, at the end of every semester, the BET specs need to be updated.” One further suggestion for this is to give an annual presentation on the BETs which would be attended by teachers and also open to other stakeholders, such as university administrators, staff and teachers from other departments. This presentation would outline the uses and future goals of the BETs, the BET development procedures, and also direct attendees to where the BET specs can be found. As suggested by one of the teacher focus group participants, it would also be useful to share annual or semester updates of the BETs with all relevant stakeholders via email and a download link.

7.3.10 Sharing BET validity evidence with senior managers

From the management interviews it became apparent that although all of the interviewed senior managers agreed with the proposed interpretations of BET scores as placement/streaming and achievement tests, they were unaware of how well the BETs were achieving their intended purposes. Therefore, it would be beneficial to share BET validity evidence with this key stakeholder group, both the results of this study, and also further validity evidence as it emerges from future research, which strengthens the BET validity argument. To share BET validity evidence with senior managers it would be useful to write a concise summary in Japanese to share with these stakeholders. It would also be beneficial to hold a brief annual meeting or presentation to present and discuss the findings of validity research with senior managers. A brief meeting would allow for clarification of any uncertainties, and it would also give an opportunity for feedback into the BET validation process and associated issues.

7.4 Limitations of the current study

This study was limited by its nature as a validation study focusing on the development phase. As noted elsewhere in this thesis there were limited institutional resources available for BET development, which meant that new and revised test items could not be piloted, while the General English Assessment Committee (GEAC) prioritized test writing and revisions. Technical expertise on the GEAC was also limited, meaning that techniques such as Rasch analysis had to be learnt in conjunction with BET development. This lack of technical expertise placed some restrictions on the types of backing obtained in the BET validity argument. Incorporating further statistical analyses in future iterations of the BET IUA and validity argument, such as test equating, generalizability studies, and factor analysis, may better support some assumptions in the BET validity argument.

In addition, sample sizes for some criterion measures to support the BET IUA explanation inference were limited due to the cost of administering these tests. Larger sample sizes would strengthen confidence in the correlations with other standardised English language proficiency criterion. Also, individual student scores for other important criteria of curriculum achievement, such as writing assignment grades, and presentation grades were not aggregated and were thus not available for this study. Correlational evidence of BET grades with such criteria would further strengthen backing for the extrapolation inference in the BET validity argument. Furthermore, correlations between teacher ratings of individual student reading and listening ability with BET scores may provide strengthened backing for the evaluation inference.

7.5 Reflections

7.5.1 Reflections on utilizing Kane's argument-based framework

In this section, I reflect on my experiences as a researcher in utilizing Kane's argument-based approach to test validation for the ongoing validation of in-house language tests in the development phase. Firstly, benefits found in using this approach, and secondly the challenges encountered are identified and discussed. Finally, suggestions are made for test developers and researchers who are thinking about implementing an argument-based approach to validation in similar contexts.

Based on the experience of this study, Kane's argument-based approach can be a viable framework for the validation of in-house placement/streaming and achievement tests. Kane's approach was applied to create an Interpretation/Use Argument, which enabled the proposed interpretations and uses of the BETs to be evaluated through seeking suitable backing evidence. The subsequent presentation and evaluation of evidence gathered in the BET validity argument clearly exposed areas in which the BETs, the BET specifications, BET-related policies and also integration between the BETs and the GE curriculum must be improved in order for the BETs to fulfil their proposed interpretations and uses (see the previous section 7.3 for the specific suggestions). This study is an example of how Kane’s argument-based approach is feasible for in-house language tests in the development phase, showing that it can help test developers to clarify what kind of evidence need to be gathered and examined to support inferences that test makers seek to make from test scores, and in turn, to expose places in the validity argument where the evidence is found to be wanting. This process thus enables test developers to set priorities for the next iteration of test development in order to strengthen the test validity argument.

However, although Kane’s framework was ultimately found to be practical and effective in this context, a few caveats and cautions are needed, as some difficulties were encountered in implementing this approach. The first difficulty found in implementing this approach, is that it is not a simple matter to craft an IUA, nor is it easy to identify the kinds of evidence that are needed as backing for its assumptions. Argument-based approaches allow for great flexibility in the structure of an IUA depending of the test and its interpretation and uses, and are thus argued to be superior to checklist approaches to validation (Bachman, 2005). However, the downside of this open and flexible nature, is that considerable time and effort is needed to craft an IUA for a specific test, with specific interpretations and uses in a specific context. Construction of an IUA requires wide reading first of theoretical work in the field, then of previous argument-validation studies of similar tests in similar contexts in order to build the IUA and to choose suitable backing to seek. Researchers must be prepared to devote a good deal of work to this process.

Another challenge in utilizing Kane’s argument-based approach is summarizing and weighing whether sometimes contradictory evidence is sufficient to support inferences made from test scores. For example, in this study it was judged that sufficient evidence was marshalled to at least minimally support a course placement function based on BET scores.

However, this judgement was no easy matter, and is inevitably subjective. Another difficulty found was that judging whether weak evidence should be interpreted as backing or rebuttal evidence can be challenging.

More guidance is needed in the literature, and more examples like this study are needed to provide precedents of validating test score uses in specific contexts, at specific stages of test development, to further discussion, and move toward an evidence-based consensus on these kinds of difficult judgements in the field.

Bachman and Palmer (2010), draw an analogy between an argument-based approach to test validation and building a legal case:

> This process is analogous to that of building a legal case to convince a judge or a jury. A lawyer presents her "case" to the court that, let us say, her client is innocent of any wrongdoing. This case consists of a clearly articulated argument and evidence that the lawyer submits to the court to support this argument. The lawyer's purpose in presenting her case is to convince the judge or jury that her client is innocent. Similarly, the process of assessment justification consists of building a "case" that the intended uses of the assessment are justified. (p. 95)

It would be useful for the language teaching/testing profession to expand the scope of Bachman and Palmer's analogy further, in order to view available argument-based language test validation studies in the literature as being similar to precedents under a common law system. Common law is a legal system which "is largely based on *precedent*, meaning the judicial decisions that have already been made in similar cases." (The Regents of the University of California, 2018, p. 1), as opposed to a civil law system, which consists of "continuously updated legal codes that specify all matters capable of being brought before a court, the applicable procedure, and the appropriate punishment for each offense." (p. 1). Thus, more argument-based studies of more types of language tests in more contexts are needed in order to set precedents for what constitutes appropriate and sufficient backing for inferences made from test scores. This will make it easier for practitioners to formulate their Interpretation/Use Arguments, and to judge their IUA's and the backing and rebuttal evidence found in validity arguments against suitable precedents. This study represents one such precedent, for a thus far unique context of a placement/streaming and achievement test in the development phase, aiming for CEFR alignment, at a Japanese tertiary institution.

Furthermore, caution needs to be exercised in choosing the types of inferences that test makers claim to make from test scores in an IUA. In hindsight, the scope of this validation study was rather ambitious. Creating an almost entirely new curriculum and accompanying in-class assessments, while at the same time developing a new set of standardised tests of reading, listening and speaking, and simultaneously aiming to align both the curriculum and the standardised tests to an outside standard such as the CEFR within two years, was an enormous undertaking for a group of teachers who were also teaching a full-time load along with all of its additional requirements such as class preparation, grading, and student pastoral care. Given the many other responsibilities born by those in the BET project, it is perhaps not surprising that the tests were only partially able to fulfil their intended purposes in the first two years of their development covered by this study. While this project was certainly worthwhile, and continues to move toward its goals, such endeavours should not be undertaken lightly.

To make language test validation using Kane's framework more easily viable for other test developers in similar contexts, with limited resources, I would suggest doing a series of validation studies, each focusing on one inference in an IUA, rather than a single large validation study as was the case with the BETs. Such studies should move from the domain inference, then through each succeeding inference until the utilization inference. This approach makes sense from a theoretical perspective, as Kane (2013b) suggests that the inferences in an IUA "can be envisioned as the spans of a bridge leading from test performances to the conclusions and decisions include in the proposed interpretation and use: if one span falls, the bridge is out, even if the other spans are strongly supported" (p. 13). Therefore, just as when building a bridge, the first span should be built first, it would make sense to run a validation study on the domain inference first, and to move onto later inferences only after a solid foundation for the first span, and then for each later span had been established.

In addition, running a single validation study for each inference in the IUA at a time would allow those responsible for the test design and validation project, to familiarize themselves with the research methodologies and statistical procedures needed for the backing sought for each warrant. As Kane (2013a), points out "developing the evidence to support the claims being made typically requires technical skill and ingenuity" (p. 456), and as most graduate programs in TESOL do not teach statistics such as correlation, regression, and test equating, those involved in the test program will need time to develop such skills in practice. Therefore, if there is not urgent external pressure on the test development program to produce a complete validity argument, doing a series of step-wise validation studies, one for each inference, each with a timeline of 6 months to a year may be the most practical way to gradually build a complete validity argument.

7.5.2 Contributions of the current study to the field of language testing

In this section, significant contributions of this study to the field of language test validation (which were briefly outlined in section 1.2) are firstly summarized in point form, then each point is expanded upon in the following paragraphs. Finally, the section ends with a short summary of this study's contributions to the field of language test validation.

7.5.2.1 The significance of this study

1. Perhaps the first example of presenting a domain definition inference within Kane's IUA for in-house reading and listening placement/achievement tests, with corresponding backing and rebuttal evidence presented in a validity argument.
2. A novel and practical example of how elements of Bachman and Palmer's (2010) AUA can be applied to utilization inferences within Kane's (2013a, 2013b) IUA.
3. A rare example of defining both backing and rebuttal evidence in an IUA, and showing how rebuttal evidence suggests clear avenues for reforms to tests, test specifications, and administrative procedures.
4. A useful example for practitioners of the challenges faced in the development phase of test validation for in-house tests focusing on a local in-house curriculum which aims for CEFR alignment.
5. Perhaps the first example of a detailed IUA and corresponding validity argument covering all of the six inferences exemplified by Chapelle et al. (2008) for in-house reading and listening tests in the development phase of validation, with a more comprehensive coverage of assumptions for each warrant than given in preceding similar studies.

The first useful contribution of this study to the field of language test validation is to present a rare example of a detailed IUA and supporting validity argument using Kane's argument-based approach with backings for the domain definition inference of an in-house criterion-referenced language test. To the best of my knowledge, there are very few other studies to date which have included a domain definition inference for a similar testing context. Most of the examples of argument-based validation studies of in-house placement or achievement tests for language programs that I was able to find (Fujita, 2005; Johnson, 2012;

Kumazawa, 2013; Li, 2015a) did not give detailed interpretive or validity arguments for a domain definition inference.

The only example I could find of an argument-based validation study of an in-house language test, which included the equivalent of a domain definition inference with details of backing was Pardo Ballester's (2007, 2010) study validation study of a web-based Spanish listening placement exam (SLE), which utilized the earlier (2005) version of Bachman's argument-based approach.

Argument-based validity studies of commercial language studies show a similar lack of focus on a domain definition inference. Of the exemplary argument-based validity studies of commercial reading and listening language tests that I was able to find (Aryadoust, 2013; Chapelle et al., 2008; Kumazawa et al., 2016) only Chapelle et al.'s study included details of a domain inference in the interpretive argument, and gave supporting evidence in a validity argument.

The detailed IUA domain definition inference, and corresponding validity argument presented for the BETs in this study seems to represent the first attempt to examine a domain inference in Kane's argument-based framework for an in-house test targeted at the domain of a local curriculum. It is thus a useful example and precedent for other practitioners. More such studies are needed, however, in order to further explore the types of evidence that can be brought to bear to support the domain in inference in such situations.

The second valuable contribution of this study to the field is to give a worked example of combining the latest iteration of Kane's argument-based approach (2013a, 2013b) with a warrant and some assumptions derived from Bachman and Palmer's AUA (2010) to form an IUA for the validation of in-house placement/streaming and achievement tests. Johnson (2012) and Chapelle et al. (2008) adapted earlier versions of Kane's approach and also incorporated Bachman's (2005, 2010) work for their utilization inference, but as far as I know this is the first study to combine aspects of the most up-to-date versions of these two argument-based validation frameworks for in-house tests. As Kumazawa (2013) pointed out "not many studies have been conducted to validate in-house placement test score interpretation and uses, and no study has evaluated the validity of such low stakes tests using Kane's validity framework" (p. 73). Further validation studies utilizing the most recent versions of these two argument-based approaches are needed to shed further light on the pros and cons of using each approach, or of combining them.

The third contribution of this study to the field of language test validation is as a relatively rare example of incorporating potential rebuttals in an IUA and weighing rebuttal evidence in a validity argument. Chappelle et al.'s (2008) study has been criticized for not explicitly incorporating rebuttals, and validation arguments for commercial language tests can tend to focus solely on backing (e.g. Kumazawa et al., 2016) rather than rebuttal evidence. Like Johnson's (2012) and Johnson and Riazi's (2015, 2017) studies, this study shows how rebuttal evidence can be analysed to create suggestions for reforms to tests, test specifications, and institutional policies, and may thus serve as a useful reference and precedent for practitioners who are considering using an argument-based validation framework for in-house test validation.

This leads to the fourth contribution of this study to the field, which is as an example of the challenges of implementing and validating multi-purpose institutional criterion-referenced tests. It is hoped that this example may prove of assistance to others who aim to implement similar tests for similar interpretations and uses. To my knowledge, no other validation studies exist of in-house tests, designed in conjunction with an in-house curriculum, with the broad aims of both course placement, and testing achievement of A2 and B1 CEFR-based goals. Evidence from the domain inference in this study highlights the importance clearly defining the test domain, and aligning curriculum content and curriculum goals to test task types and content from the beginning of test and curriculum development. Evidence from the generalization inference also points to the importance of not underestimating the challenges involved in developing the technical, and statistical expertise necessary to support claims of achievement tests.

The fifth way in which this study makes a unique contribution to the field of language testing, is through the presentation of a complete IUA and validity argument, covering all of the inferences set forth in Chapelle et al.'s (2008) study, with a relatively full set of assumptions stated for the warrant for each inference. To my knowledge, no other such examples of a full IUA for in-house reading and listening tests exists. Other such studies have focused on a more limited range of inferences for a validity argument.

One benefit of including all six of Chapelle et al.'s inferences in a validation study is that it requires researchers to examine all aspects of test validity, and thus it may overcome a potential temptation to focus only on the well-supported inferences. The wide focus gained from examining a full chain of inferences in an IUA and supporting validity argument reduces the chance of confirmatory bias identified by Kane (2012) as a danger when validating tests in the development phase of validation. Examining the full chain of inferences also increases the chance of identifying important weaknesses in a validity argument, which might be overlooked when focusing on fewer inferences. The relative comprehensiveness of the validity argument presented in this study makes it a valuable reference for others who may seek to thoroughly validate an in-house test, in order to communicate the value of the test to stakeholders, or to build an evidence base for test renewal.

In summary, this study provides a small, but uniquely useful contribution to the field of validation in second language assessment, through serving as an example of applying a combination of recent argument-based approaches to test validation to build a relatively complete and thorough IUA, and subsequent validity argument, for in-house reading and listening tests aimed at CEFR alignment. This study also exemplifies how rebuttal evidence found in a validity argument, in this case in the context of a Japanese university English language program, clearly indicates areas for improvement for the next cycle of test development.

7.6 Suggestions for further research

Several suggestions for areas of future research arise from weakness found in the BET validity argument. Suggestions for future research which would strengthen the BET validity argument, and which would also make useful contributions to the field of language test validation are outlined below.

1) It would be beneficial to research the process of reviewing and revising the BET specifications as suggested in section 7.3.1. This would serve both as a means to strengthen backing for the BET domain definition inference in the BET validity argument, and would also provide a useful example of how in-house test specifications can be systematically reviewed and revised to provide stronger evidence of representation of a CEFR-based curriculum domain within an argument-based framework.
2) A further useful area of future research would be to run a cut-score panel on the BETs, and to report on the methodology and results. This is because, to the best of my knowledge, while there are many studies reporting on cut-score setting for large commercial or state tests, there are few studies examining cut-score setting for small-scale in-house tests. In particular, there seem to be no cut-score setting studies aimed at alignment to an external standard such as the CEFR, which have been conducted on relatively small-scale in-house reading and listening tests. Conducting and reporting such a study for the BETs would make a valuable

contribution to the literature, and could also provide a useful reference for other practitioners. In addition, a well-run cut-score panel with resulting cut scores could furnish stronger backing for the BET IUA explanation inference. Finally, such research could also strengthen the utilization inference warrants that BETs are placing students into their correct CEFR levels for reading and listening, and are able to measure student progress against CEFR reading and listening goals.

3) Another valuable research project would be to present research on equating the BETs. Equating the BETs would firstly serve to solidify backing for the generalization inference of the BET validity argument. Secondly, once the GEAC establishes procedures for test equating, it would be a great service to the testing community to share the equating methodology in an easily accessible form. This is because the specifics of test equating (as noted in section 7.3.5) are not included in language test courses for most EFL teacher preparation programs, nor in available textbooks on language testing, and it can be quite difficult for the uninitiated, non-psychometrician to decipher the literature on test equating. Given the importance of test equating for achievement tests, which have an indispensable place in language learning programs, more effort needs to be made to make test equating accessible to language teachers. A simple step-by-step guide on test equating for small-scale programs would be invaluable for practitioners, and once the BETs have been successfully equated, it may be possible for BECC researchers to contribute to such a guide.

4) Another area for further research is to investigate how teachers use BET scores as a diagnostic indicator of their students' ability. The issue of how informative students find BET scores is addressed in assumption 4 of the utilization inference of the BET IUA. However, the issue of how informative teachers find BET scores,

has yet to be addressed. This area was not covered in the teacher survey or focus groups, and it would be a useful avenue for further research on the BETs. Such research would also strengthen backing for the BET utilization inference validity argument.

5) A further productive area of research would be to look in detail at how students feel about the way their BET scores are presented to them, and how students actually use their BET scores. This is also an area which seems to be under researched in the literature on language testing. Some backing for the utilization inference assumptions that BET scores are informative and motivating for students was gathered in this study from the answers to survey Likert scale items attached as Appendix X, Part B. However, more detailed answers from students could provide richer and stronger backing for these assumptions. This backing could take the form of transcripts of focus group interviews with representative groups of BET takers from the higher and lower stream classes. Such research could also provide the GEAC with information about how to better present and explain the meaning of BET scores to students. As Brown and Hudson (2002) state:

> … handing those score reports to the teachers and students should not be the end of the process. The views of students and teachers should be systematically gathered not only on what the scores mean personally to them and to the curriculum … but also on how the tests, administrations procedures, and score report strategies themselves can be improved. Gathering feedback from teachers and students in this way can prove informative, but also it may increase their motivation in the class and during the testing session. (p. 286).

6) This project also raises the need for further research on teacher and administrator attitudes to placement and achievement tests in various contexts. This seems to be an under researched area, and such research could address questions such as what are teachers' attitudes to the placement/achievement tests at their institutions? What obstacles stand in the way of teachers developing and validating their own placement/achievement tests? To what extent do teachers view the placement/achievement tests at their institutions to be aligned to their language course content? Such research would be illuminating for the field both in terms of shedding light on this issue in varied contexts, and also to provide useful needs analysis for language teacher training programs.
7) Finally, as advocated in section 7.5.1, more validation studies such as this one, utilizing argument-based frameworks, are needed across further contexts in the field of language testing. This is necessary to build a rich bank of precedents for test interpretation and use arguments in various contexts, and also to provide a fuller scope of examples of the types of evidence suitable for backing in different validity arguments, and of criteria for judging sufficiency of backing for assumptions, warrants and inferences.

7.7 Closing words

This study is a pioneering example of an ambitious attempt to design and validate small-scale, localized, curriculum-based, CEFR-aligned, placement/streaming and achievement tests in Japan. As such, it serves as a valuable example for others who may attempt similar projects. Results of this research also highlight the need to commence test validation from the earliest stages of test development, and the iterative nature of validation research, which is needed to support inferences made from test scores, even in local contexts for small-scale, in-house tests.

This study also demonstrated that the process of creating an IUA, and gathering and examining backing and rebuttal evidence in a validity argument, is invaluable for clarifying the aims of an in-house language test and its corresponding language programme. Finally, this research has exhibited how an argument-based validation process can reveal aspects of tests, and their accompanying policies, procedures and documents which need to be revised in order to better support claims made based on test scores.

REFERENCES

A mobile first, embeddable collaboration platform. (n.d.). Retrieved from http://moxtra.com/

Alderson, J. C. (1990). Testing reading comprehension skills (part 1). *Reading in a Foreign Language, 6*(2), 425–438. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl62anderson.pdf

Alderson, J. C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46–57). Alexandria, VA: TESOL.

Alderson J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (Ed.). (2017). Special issue on the Common European Framework of Reference for Languages (CEFR) for English language assessment in China [Special issue]. *Language. Language Testing in Asia*, *7*(20). doi:10.1186/s40468-017-0049-9

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Sauli, T., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project, *Language Assessment Quarterly, 3*(1), 3–30. doi:10.1207/s15434311laq0301_2

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). The development of specifications for item development and classification within the Common European Framework of Reference for Languages: Learning, teaching, assessment. Reading and listening. Final report of the Dutch CEF construct project. Unpublished document. Available from http://www.research.lancs.ac.uk/portal/en/publications/the-development-of-specifications-for-item-development-and-classification-within-the-common-european-framework-of-reference-for-languages-learning-teaching-assessment-reading-and-listening-final-report-of-the-dutch-cef-construct-project-(c8b8d75a-e71e-461a-8e26-75000bb6353b).html

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320. doi:10.1191/0265532205lt310oa

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing, 30*(4), 535–556. doi:10.1177/0265532213489568

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5*(2), 253–270. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl52alderson.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing* (6th edition). New York, NY: Macmillan.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle: Cambridge Scholars Publishing.

Association of Language Testers in Europe, (2011). *Manual for language test development and examining: For use with the CEFR*. Strasbourg, Council of Europe.

Ayers, J. B., & Peters, R. M. (1977). Predictive validity of the Test of English as a Foreign Language for Asian graduate students in engineering, chemistry or mathematics. *Educational and Psychological Measurement*, *37*(2), 461–463. doi:10.1177/001316447703700221

Ayers, J. B., & Quattlebaum, R. F. (1992). TOEFL performance and success in a master's program in engineering. *Educational and Psychological Measurement, 52*(4), 973–975. doi:10.1177/0013164492052004021

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34. doi:10.1207/s15434311laq0201_1

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Barkaoui, K. (2009). Building a validity argument for the Test of English as a Foreign Language. [Review of the book *Building a validity argument for the test of English as a foreign language*, by Carol A. Chapelle, Mary K. Enright, and Joan M. Jamieson (Eds.)]. *The Canadian Modern Language Review, 65(4),* 657– 659.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning, 39*(3), 313–338. doi:10.1111/j.1467-1770.1989.tb00595.x

Bond, T., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York, NY: Springer.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061

Bower, J., Runnels, J., Rutson-Griffiths, A., Schmidt, R., Cook, G., Lusk Lehde, L., & Kodate, A. (2017). Aligning a Japanese university's English language curriculum and lesson plans to the CEFR-J. In F. O'Dwyer, M, Hunke, A. Imig, N. Nagai, N. Naganuma, & M. G. Schmidt (Eds.), *Critical, constructive assessment of CEFR-informed language teaching in Japan and beyond* (pp. 176–225). Cambridge, England: Cambridge University Press.

Bower, J., Rutson-Griffiths, A., & Sugg, R. (2014). Setting and raising standards: the rationale for, and the structure of the Bunkyo English Tests. *Hiroshima Bunkyo Women's University Bulletin, 49,* 65–78. Retrieved from http://harp.lib.hiroshima-u.ac.jp/h-bunkyo/detail/1216920150204165253;jsessionid=BF59C15C5D3D975C501417C4C89 1F2F5

Boyle, A., & Rahman, Z. (2013). The internal reliability of some City and Guilds tests. Retrieved from https://www.gov.uk/government/publications/the-internal-reliability-of-some-city-and-guilds-tests

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science, 7*(2), 96-99. doi:10.1111/j.1467-9280.1996.tb00336.x

Brown, J. D. (1989). Criterion-referenced test reliability*. University of Hawai'I Working Papers in ESL, 8*(1)*,* 79–113. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/38543/1/Brown%20(1989)_WP 8(1).pdf

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment.* Singapore: McGraw-Hill.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge, England: Cambridge University Press.

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of References*. Taipei: The Language Training and Testing Center. Retrieved from http://eprints.lancs.ac.uk/69811/1/Brunfaut_Harding2014_GEPT_listening_test_linking_study.pdf

Cambridge University Press. (2015). *English Vocabulary Profile*. Retrieved from http://englishprofile.org/wordlists

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19,* 254–72. doi:10.1017/S0267190599190135

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K., Enright, & J, M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 319–354). New York, NY: Routledge.

Chapelle, C. A. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple …. *Language Testing*, *29*(1), 19–27. doi:10.1177/0265532211417211

Chapelle, C. A. (2015). *Validity arguments for four-skill language tests*. Paper presented at the 17th Academic Forum on English Language Testing in Asia, Tokyo, Japan.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. III, pp. 1079–1097). Hoboken, NJ: Wiley-Blackwell. doi:10.1002/9781118411360.wbcla110

Chen, N. (2010, November 20). Building a validity argument for the Test of English as a Foreign Language, by Carol A. Chapelle, Mary K. Enright, and Joan M. Jamieson (Eds.) [Review of the book *Building a validity argument for the test of English as a foreign language*, by Carol A. Chapelle, Mary K. Enright, and Joan M. Jamieson (Eds.)]. *Language Assessment Quarterly, 7*(4), 377–382. doi:10.1080/15434303.2010.519418

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155–163. doi:10.1177/026553228500200204

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436–470. doi:10.1017/S0261444815000233

Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In S. Jones and B. Spolsky (Eds.), *Language testing proficiency* (pp. 10–24). Arlington, VA: Center for Applied Linguistics.

Cobb, T. (2018). *Compleat lexical tutor*. Retrieved from https://www.lextutor.ca/

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing. 1*(1), 70–81. doi:10.1177/026553228400100106

Cohen. J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coleman, J. (1996, December 1). Japan toddlers cram for kindergarten. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1996-12-01/news/mn-4579_1_cram-schools

Council of Europe (Ed.). (2001a). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

Council of Europe *(Ed.). (*2001b*). Common European framework of reference for languages: Learning, teaching, assessment: structured overview of all CEFR scales*. Retrieved from http://ebcl.eu.com/wp-content/uploads/2011/11/CEFR-all-scales-and-all-skills.pdf

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR).* Strasbourg, France: Council of Europe.

Council of Europe. (2010). *外国語教育〈2〉外国語の学習、教授、評価のためのヨーロッパ共通参照枠*. (Yoshijima, S., & Ohashi, N, Trans.). Tokyo: 朝日出版社. (Original work published 2004.)

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly, 34*(*2*), *213–38.* doi:10.2307/3587951

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. doi:10.1037/h0040957

Davies, A. (1990). *Principles of language testing*. Oxford, England: Blackwell.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing, 14*(3), 328–339. doi:10.1177/026553229701400309

Desveaux, S. (2018). *How many hours do I need to prepare for my exam?* Retrieved from https://support.cambridgeenglish.org/hc/en-gb/articles/202838506-Guided-learning-hours

Dorans, N. J., Moses, T. P., & Eignor, D. (2010). Principles and practices of test score equating (ETS Research Report No. RR-10–29). Princeton, NJ: ETS. Retrieved from https://www.ets.org/Media/Research/pdf/RR-10-29.pdf

Downey, N., & Kollias. C. (2010). Mapping the Advanced Level Certificate in English (ALCE) examination onto the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 119–130). Cambridge, England: Cambridge University Press.

Dunlea, J. (2015). *Validating a set of Japanese EFL proficiency tests: Demonstrating locally designed tests meet international standards*. (Doctoral thesis, University of Bedfordshire). Retrieved from http://uobrep.openrepository.com/uobrep/bitstream/10547/618581/1/Binder1.pdf

EAQUALS. (2017). *Revision and refinement of CEFR descriptors*. Retrieved from https://www.eaquals.org/resources/revision-and-refinement-of-cefr-descriptors/

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4*(3), 125–128. doi:10.1111/j.1745-3984.1967.tb00579.x

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Elif, K., Thomas, C., O'Dwyer, J., & O'Sullivan, B. (2010). Benchmarking a high-stakes proficiency exam: The COPE linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 102–118). Cambridge, England: Cambridge University Press.

Eliot, M., & Wilson, J. (2013). Context validity. In Geranpayeh, A., & Taylor, L. (Eds.), *Examining listening: research and practice in assessing second language listening* (pp. 152–241). Cambridge: CUP.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485. doi:10.1093/elt276et/ccs037

Fujita, T. (2005). *Validation of a Japanese university English language placement test*. (Doctoral dissertation). Temple University, Japan.

Fulcher, G. (2000). The ‘communicative’ legacy in language testing. *System, 28*(4), 483–497. doi:10.1016/S0346-251X(00)00033-6

Fulcher, G. (2013). *Practical Language Testing*. London, England: Routledge:

Giliflores, J., & Alonso, C. G. (1995). Using focus groups in educational research: Exploring teachers’ perspectives on educational change. *Evaluation Review, 19*(1), 84–101. doi:10.1177/0193841X9501900104

Gochyyev, P., & Sabers, D. (2012). Item analysis. In N.J. Salkind (Ed.), *Encyclopedia of research design* (pp. 642-645). Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781412961288

Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, *13*(2), 39–51. doi:10.6018/ijes.13.2.185891

Green, R. (2013). *Statistical analyses for language testers*. New York, NY: Palgrave Macmillan.

Greene, J. C., Caracelli, V. J., &. Graham, W. F. (1989). Toward a conceptual framework for mixed method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255–274. doi:10.2307/1163620

Guion, R. M. (1977). Content validity–The source of my discontent. *Applied Psychological Measurement, 1*(1), 1–10. doi:10.1177/014662167700100103

Hajr, F. A. (2014). The predictive validity of language assessment in a pre-university programme in two colleges in Oman: a mixed-methods analysis. *International Multilingual Journal of Contemporary Research, 2*(2), 121–147. Retrieved from http://imjcr.com/journals/imjcr/Vol_2_No_2_June_2014/8.pdf

Ho, D. G. E. (2012). Interviews. In C. A. Chapelle (Ed), *The encyclopedia of applied linguistics*. Blackwell. doi:10.1002/9781405198431.wbeal0571

Huy, N, V., & Hamid, O. M. (2015). Educational policy borrowing in a globalized world: A case study of Common European Framework of Reference for languages in a Vietnamese university. *English Teaching: Practice & Critique, 14*(1), 60–74. doi:10.1108/ETPC-02-2015-0014

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. doi:10.1111/lang.12034

Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China Standards of English: Challenges at macropolitical and micropolitical levels. *Language Testing in Asia, 7*(1). doi:10.1186/s40468-017-0032-5

Jochems, W., Snippe, J., Smid, J. H., & Verweij, A. (1996). The academic progress of foreign students: Study achievement and study behaviour. *Higher Education*, *31*(3), 325–340. doi:10.1007/BF00128435

Johnson, R. C. (2012). *Assessing the assessments: Using and argument-based validity framework to assess the validity and use of an English placement system in a foreign language context.* (Doctoral thesis, Macquarie University, Sydney, Australia). Retrieved from http://www.researchonline.mq.edu.au/vital/access/manager/Repository/mq:30434

Johnson, R. C., & Riazi, M. (2015). An argument-based validation of both test score meaning and impact. *Papers in language testing and assessment, 4*(1), 31–58. Retrieved from http://www.altaanz.org/uploads/5/9/0/8/5908292/johnson_riazi.pdf

Johnson, R. C., & Riazi, M. (2017). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing, 32,* 85–104. doi:10.1016/j.asw.2016.09.002

Jones, N. (2001). Reliability as one aspect of test quality. *Research Notes, 4,* 2–5. Retrieved https://www.cambridgeenglish.org/Images/23115-research-notes-04.pdf

Jones, N. (2014). *Multilingual frameworks: The construction and use of multilingual proficiency frameworks*. Cambridge: Cambridge University Press.

Joyce, P., & McMillan, B. (2010). Student perceptions of their learning experience in streamed and mixed-ability classes. *Language Education in Asia*, *1*(1), 215–227. doi:10.5746/LEiA/10/V1/A18/Joyce_Mcmillan

Kaftandjieva, F. (2004). Standard setting. In *Council of Europe, Reference supplement to the pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF)*. Strasbourg: Language Policy Division.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41. doi:10.1111/j.1745-3992.2002.tb00083.x

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135–170. doi:10.1207/s15366359mea0203_1

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2008). Validating the interpretation and uses of test scores. In Lissitz, R. (Ed.), *The concept of validity* (pp. 39–64). Charlotte, NC: Information Age Publishing.

Kane, M. (2012). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3–17. doi:10.1177/0265532211417210

Kane, M. (2013a). The Argument-Based Approach to Validation. *School Psychology Review, 42*(4), 448–457.

Kane, M. (2013b). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x

Kecker, G., & Eckes, T. (2010). Putting the Manual to the test: the TestDaF-CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 50–79). Cambridge: Cambridge University Press.

Khalifa, H., &Weir, C. J. (2009), *Examining reading*. Cambridge: Cambridge University Press.

Khalifa, H., & Schmitt, N. (2010). A mixed-method approach towards investigating lexical progression in Main Suit Reading test papers. *Cambridge ESOL: Research Notes, 41*, 19-25. Available from https://docs.wixstatic.com/ugd/5f2482_e1f018c567854243a076d58f981cbd81.pdf

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). Validity of the SAT for predicting first-year college grade point average. (College Board Research Report, No.2008-5). New York, NY: The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED563202.pdf

Koizumi, R., In'nami, Y., Asano, K., & Agawa, T. (2016). Validity evidence of Criterion for assessing L2 writing proficiency in a Japanese university context. *Language Testing in Asia*, *6*(5). doi:10.1186/s40468-016-0027-7

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices* (3nd ed.). New York, NY: Springer-Verlag.

Krueger, R. A., & Casey, M. A. (2009). *Focus groups: a practical guide for applied research*. Thousand Oaks, CA: Sage Publications.

Kumazawa, T. (2013). Evaluating validity for in-house placement test score interpretations and uses. *JALT Journal, 35*(1), 73–100. Retrieved from http://www.jalt-publications.org/files/pdf-article/jj2013a_art4.pdf

Kumazawa, T., Shizuka, T., Mochizuki, M., & Mizumoto, A. (2016). Validity argument for the VELC test score interpretations and uses. *Language Testing in Asia, 6*(2). doi:10.1186/s40468-015-0023-3

Kunnan, A. J. (2003). Test fairness. In M. Milanovic & C. Weir (Eds.), *Select papers from the European year of languages conference, Barcelona* (pp. 27–48). Cambridge, England: Cambridge University Press.

Kunnan, A. J. (2008). Towards a model of test evaluation: using the test fairness and wider context frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: achieving transparency, assuring quality, sustaining diversity. Papers from the ALTE conference in Berlin, Germany* (pp. 229–251). Cambridge, England: Cambridge University Press.

Kunnan, A. J. (2018). *Evaluating Language Assessments*. New York, NY: Routledge.

Kurtz, A. M., & Dwyer, A. C. (2013). Small sample equating: Best practices using a SAS Macro. Retrieved from http://analytics.ncsu.edu/sesug/2013/BtB-11.pdf

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* New York, NY: McGraw-Hill.

LaFlair, G. T., Isbell, D., May, L. D. N., Arivzu, M. N. G, & Jamieson, J. (2017). Equating in small-scale language testing programs. *Language Testing, 34*(1), 127–144. doi:10.1177/0265532215620825

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). Basingstoke, England: Macmillan

Laufer, B. (2012). Lexical frequency profiles. In C. A. Chapelle (Ed), *The encyclopedia of applied linguistics*. Blackwell. doi:10.1002/9781405198431.wbeal0692

Lee, A. (2011). Reflexivity. In R. Thorpe & R. Holt (Eds.), *The SAGE dictionary of qualitative management research* (pp. 184–185). London, England: Sage Publications. doi:10.4135/9780857020109.n86

Lee, H. (2011). Investigating the applicability of the CEFR to a placement test for an English language program in Korea. English Language and Linguistics, 17(3), 29-60. Retrieved from http://www.elsok.org/xe/index.php?page=8&document_srl=143387&mid=ell2

Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*(4), 366–389. doi:10.1177/1558689807306148

Li, Z. (2015a). *An argument-based validation study of the English Placement Test (EPT) – Focusing on the inferences of extrapolation and ramification*. (Doctoral thesis, Iowa State University, Ames, USA). Retrieved from http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=5545&context=etd

Li, Z. (2015b). Using an English self-assessment tool to validate an English placement test. *Papers in language testing and assessment, 4*(1), 59–96. Retrieved from http://www.altaanz.org/uploads/5/9/0/8/5908292/li.pdf

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328. Retrieved from https://files.eric.ed.gov/fulltext/ED453662.pdf

Linacre, J. M. (2017a, August 8). Setting a response probability for educational tests. Message posted to http://raschforum.boards.net/thread/753/setting-response-probability-educational-tests

Linacre, J. M. (2007b). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com

Little, D. (2006). The Common European Framework of Reference for Languages: content, purpose, origin, reception and impact. *Language Teaching*, *39*(3), 167–190. doi:10.1017/S0261444806003557

Liu, L., & Jia, G. (2017). Looking beyond scores: validating a CEFR-based university speaking assessment in mainland China. *Language Testing in Asia*, *7*(2). doi:10.1186/s40468-017-0034-3

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement*, *3*, 635–694. doi:10.2466/pr0.1957.3.3.635

Luoma, S. (2013). Self-assessment. In C. A. Chapelle (Ed), *The encyclopedia of applied linguistics*. Blackwell. doi:10.1002/9781405198431.wbeal1060

Lynch, T. (2000). An evaluation of the revised test of English at matriculation at the University of Edinburgh. *Edinburgh Working Papers in Applied Linguistics*, 10, 61–71. Retrieved from https://eric.ed.gov/?id=ED373551

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351–372. doi:10.1177/026553220101800403

Matsutani, M. (2012, January, 10). Student count, knowledge sliding. *Japan Times*. Retrieved from http://www.japantimes.co.jp/news/2012/01/10/reference/student-count-knowledge-sliding/#.VAAyJ2POuBJ

Mattern, K. D., & Packman, S. (2009). *Predictive validity of ACCUPLACER scores for course placement: A meta-analysis* (Research Report No. 2009-2). New York, NY: The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED561046.pdf

McCreary, L. L., Conrad, K. M., Conrad, K. J., Scott, C. K., Funk, R. R. & Dennis, M. L. (2013). Using the Rasch Measurement Model in Psychometric Analysis of the Family Effectiveness Measure. *Nursing Research*, *62*(3), 149–159. doi:10.1097/NNR.0b013e31828eafe6

McNamara, T. (1996). *Measuring second language performance*. London, England: Longman.

McNamara, T. (2003, October 1). Fundamental Considerations in Language Testing. Oxford: Oxford University Press, Language testing in practice: designing and developing useful language tests. [Review of the book Language testing in practice: designing and developing useful language tests]. *Language Testing*, *20*(4), 466–473. doi:10.1191/0265532203lt268xx

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555-576. doi:10.1177/0265532211430367

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. doi:10.1177/026553229601300302

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496. doi:10.1191/0265532202lt241oa

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. doi:10.1207/S15366359MEA0101_02

Mohsen, T., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher, 35*(1), 838–848. doi:10.3109/0142159X.2012.737488

Morgan, D. L. (1996). Focus groups. *Annual Review of Sociology, 22*, 129-152. doi:10.1146/annurev.soc.22.1.129

Morrow, K. (1979). Communicative language testing: revolution or evolution? In: Brumfit, C.K., Johnson, K. (Eds.), *The communicative approach to language teaching* (pp. 143–159). Oxford, England: Oxford University Press.

Morrow, K. (2012). Communicative language testing. In: C, Combe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 140–146). Cambridge, England: Cambridge University Press.

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines*, 5(1), 12–25. Retrieved from https://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/1983-Testing-and-teaching.pdf

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82. doi:10.3138/cmlr.63.1.59

Negishi, M., Takada, T., & Tono, Y. (2012) A progress report on the development of the CEFR-J. In: E, D. Galaczi, & C. J. Weir (Eds), *Studies in Language Testing 36* (pp. 137–165). Cambridge, England: Cambridge University Press.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 1–29. doi:10.1080/15366367.2012.669666

Noijons, J., & Kuijper, H. (2010). Mapping the Dutch foreign language state examinations onto the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 247–265). Cambridge, England: Cambridge University Press.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5), 625-632. doi:10.1007/s10459-010-9222-y

North, B., & Jones, N. (2009). Further material on *maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling.* Retrieved from Council of Europe website: https://rm.coe.int/1680459fa0

North, B., Ortega, A., & Sheehan, S. (2010). *British Council – EAQUALS Core Inventory of English*. British Council & EAQUALS. Retrieved from https://englishagenda.britishcouncil.org/sites/default/files/attachments/books-british-council-eaquals-core-inventory.pdf

O'Sullivan, B. (2010). The City & Guilds Communicator examination linking project: a brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 33–49). Cambridge, England: Cambridge University Press.

Oxfordenglishtesting.com. (2017). *Oxford Online Placement Test – features*. Retrieved from https://www.oxfordenglishtesting.com/defaultmrarticle.aspx?id=3074

Palmer, A. S., Groot, P. J. M., & Trosper, G. A. (Eds.). (1981). *The construct validation of Tests of communicative competence*, TESOL, Washington, DC. Retrieved from https://files.eric.ed.gov/fulltext/ED223103.pdf

Pardo Ballester, C, M. (2007). *The Development of a Web-Based Spanish Listening Placement Exam*. (Unpublished doctoral dissertation). University of California Davis, Davis, California.

Pardo Ballester, C. (2010). The Validity Argument of a Web-Based Spanish Listening Exam: Test Usefulness Evaluation. *Language Assessment Quarterly, 7*(2), 137–159. doi:10.1080/15434301003664188

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (ed.). *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.

Pollit, A. (2009). *The Oxford Online Placement Test: The meaning of OOPT scores*. Oxford, England: Oxford University Press. Retrieved from https://www.oxfordenglishtesting.com/uploadedFiles/Buy_tests/oopt_meaning.pdf

Powers, D. E. (2010a). The case for a comprehensive, four-skills assessment of English-language proficiency. *R & D Connections*, *14*, 1–12. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections14.pdf

Powers, D. E. (2010b). Validity: What does it mean for the TOEIC tests? In D. Powers (Ed.), *TOEIC compendium* (1st ed., pp. 1.1–1.11). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/TC-10-01.pdf

Programme for International Student Assessment. (2006). *PISA 2006 Technical Report*. Retrieved from https://www.oecd.org/pisa/data/42025182.pdf

Purpura, J. (2010). *The Oxford Online Placement Test: What does it measure and how?* Oxford: Oxford University Press. Retrieved from https://www.oxfordenglishtesting.com/uploadedfiles/6_New_Look_and_Feel/Content/oopt_measure.pdf

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. doi:10.1177/026553229301000308

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Erlbaum.

Regents of the University of California. (2018). *The common law and civil law traditions*. Retrieved from https://www.law.berkeley.edu/library/robbins/pdf/CommonLawCivilLawTraditions.pdf

Reinhart, A. (2015). *Statistics Done Wrong: The woefully complete guide*. San Francisco: No Starch Press.

Riazi, A. M., & Candlin, C. N. (2014). Mixed-methods research in language teaching and learning: Opportunities, issues and challenges. *Language Teaching, 47*(2), 135-173. doi:10.1017/S0261444813000505

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., 259–273). New York, NY: Routledge.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1–20. doi:10.1177/026553229801500101

Saville, N (2003). The process of test development and revision within UCLES EFL.in C. Weir, and M. Milanovic (Eds.), *Continuity and innovation: revising the Cambridge Proficiency in English Examination 1913–2002* (57–120). Cambridge, England: Cambridge University Press.

Shephard, L. A. 1993. Evaluating test validity. *Review of Research in Education* 19, 405–450. doi:10.2307/1167347

Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. Cambridge: Cambridge University Press.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99–128. doi:10.1177/0265532207071513

Shizuka, T., & Mochizuki, M. (2014). VELC Test for testing competency: Verification of reliability and validity: Retrieved from http://www.velctest.org/contact/VelcTest-for-TestingCompetency.pdf

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147–170. doi:10.1177/026553228400100203

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests.* Harlow, England: Pearson Education Limited.

Sick, J. (2010). Rasch measurement in language education part 5: Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing and Evaluation Newsletter, 14*(2), 23-29. Retrieved from http://hosted.jalt.org/test/sic_5.htm

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249–275. doi:10.1111/j.1745-3984.2007.00037.x

Skaggs, G. & Lissitz, R. W. (1986). An Exploration of the Robustness of Four Test Equating Models. *Applied Psychological Measurement, 10*(3), 303–317. doi:10.1177/014662168601000308

Smith, E. V. (2002). Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Smith, E. V., & Stone, G.E. (Eds.). (2009). *Criterion referenced testing: practice analysis to score reporting using Rasch measurement models*. Maple Grove, Minnesota: JAM Press.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577–607. doi:10.1017/S0272263109990039

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing*, *32*(1), 63–81. doi:10.1177/0265532214541885

Szabo, G. (2010). Relating language examinations to the CEFR: ECL as a case study. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 133–144). Cambridge, England: Cambridge University Press.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (TOEFL iBT Research Report RR-08-34). Princeton, NJ: Educational Testing. Retrieved from https://www.ets.org/Media/Research/pdf/RR-08-34.pdf

Thompson, G. & Foale, C. (2008). Adapting the BEPP model: The BECC curriculum project. *Studies in Linguistics & Language Teaching*, 19, 253–289.

Tono, Y., & Negishi, M. (2012). The CEFR-J: Adapting the CEFR for English teaching in Japan. *Framework Language & Portfolio SIG Newsletter, 8*, 5–12. Retrieved from https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxmbHBzaWd8Z3g6NGQ5ZDM1MTg3NTg3MTY1NQ

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, England: Cambridge University Press.

Trew, G. (2007). A teacher's guide to TOEIC listening and reading test preparing your students for success. Retrieved from https://elt.oup.com/elt/students/exams/pdf/elt/toeic_teachers_guide_international.pdf

Trim, J. L. M., (2012). The Common European Framework of Reference: Learning, teaching, assessment. In: M, Byram., & L, Parmenter (Eds.), *The Common European Framework of Reference: The globalization of language policy* (pp. 14–35). Bristol, England: Multilingual Matters.

University of Cambridge ESOL Examinations (2012a). *Cambridge English Key English Test (KET) CEFR Level A2 handbook for teachers.* Cambridge, England: Cambridge English Language Assessment:

University of Cambridge ESOL Examinations (2012b). *Cambridge English Preliminary: Preliminary English Test (PET) CEFR Level B1 handbook for teachers.* Cambridge, England: Cambridge English Language Assessment:

University of Oxford. (2015). *British National Corpus*. Retrieved from http://www.natcorp.ox.ac.uk/

van Ek, J. A. (1976). *The threshold level for modern language learning in schools.* London, England: Longman.

van Ek, J, A. (1998a). *Threshold 1990* (2nd ed.). Cambridge: Cambridge University Press.

van Ek, J. A., and Trim, J. L. M. (1998b). *Waystage 1990* (2nd ed.). Cambridge: Cambridge University Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel Method of Equating*. New York, NY: Springer.

Wang, W., Eignor, D., & Enright, M. (2008). A final analysis. In C. A. Chapelle, M. K., Enright, & J, M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 259–318). NY: Routledge.

Weir, C. J. (2005a). *Language testing and validation: An evidence-based approach.* Hampshire: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the common European framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300. doi:10.1191/0265532205lt309oa

West, M. (1953). *A general service list of English words*. Longman, London

Wright, B. D. (1996). Reliability and Separation. *Rasch Measurement Transactions, 9*(4), 472. Retrieved from https://www.rasch.org/rmt/rmt94n.htm

Wright, B. D. (2001). Separation, reliability and skewed distributions: statistically different levels of performance. *Rasch Measurement Transactions, 14*(4), 786. Retrieved from https://www.rasch.org/rmt/rmt144k.htm

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. Retrieved from https://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., & Masters, G, N. (2002). Number of Person or Item Strata: (4*Separation + 1)/3. Rasch *Measurement Transactions, 16*(3), 888. Retrieved from https://www.rasch.org/rmt/rmt163f.htm

Wu, J., & Wu, R. (2010). Relating the GEPT reading comprehension tests to the CEFR. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of*

Europe's draft manual (pp. 204–244). Cambridge, England: Cambridge University Press.

Xi, X. (2008). Methods of test validation. In E. Shohamy and N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed,Vol 7, 177–196). Springer Science.

Xi, X. (2010). How do we go about integrating test fairness? *Language Testing, 27*(2), 147–170. doi:10.1177/0265532209349465

APPENDICES

APPENDIX A
BET Formats

BET 2015 Format

Test Section and timing	Testlet Name	Testlet Type	Target CEFR Level
Reading (30 minutes)	Reading Part 1	*Matching* Match five options to the appropriate signs with the same meaning from eight sign options.	A2
	Reading Part 2	*Multiple Choice* Five three-option multiple choice items, matching the appropriate answer to a question for a verbal exchange.	A2
	Reading Part 3	*Gap fill* Choose the five appropriate phrases to complete a dialogue from eight options.	A2
	Reading Part 4	*Multiple Choice Reading Comprehension* Seven three-option multiple choice items focusing on comprehension of a short text.	A2
	Reading Part 5	*Multiple Choice Reading Comprehension* Five three-option multiple choice items focusing on comprehension of a short text.	B1
Vocabulary & Grammar (10 minutes)	Vocabulary	*Multiple Choice Gap fill* Five three-option multiple choice items choosing the most appropriate words to complete five sentences in a short passage.	A2
	Grammar	*Multiple Choice Gap fill* Ten three-option multiple choice items choosing the most appropriate words to complete five sentences in a short passage.	A2
Listening (approximately 30 minutes)	Listening Part 1	*Choose the Correct Picture* Five three-option multiple items choosing the correct picture to match the listening.	A2
	Listening Part 2	*Matching*	A2

		Choose the correct options from eight choices to answer five items based on the information in a dialogue.	
	Listening Part 3	*Multiple Choice Listening Comprehension* Five three-option multiple choice items focusing on comprehension of a short dialogue.	A2
	Listening Part 4	*Multiple Choice Listening Comprehension* Five three-option multiple choice items focusing on comprehension of a short monologue.	A2
	Listening Part 5	*Multiple Choice Listening Comprehension* Six three-option option multiple choice items focusing on comprehension of a longer monologue or interview.	B1

BET 2016 Format

Test Section and timing	Testlet Name	Testlet Type	Target CEFR Level
Reading (45 minutes)	Reading Part 1	*Matching* Match five options to the appropriate signs with the same meaning from eight sign options.	A2
	Reading Part 2	*Multiple Choice Gap fill* Five three-option multiple choice items choosing the most appropriate words to complete five sentences in a short passage.	A2
	Reading Part 3	*Multiple Choice* Five three-option multiple choice items, matching the appropriate answer to a question for a verbal exchange.	A2
	Reading Part 4	*Gap fill* Choose the five appropriate phrases to complete a dialogue from eight options.	A2
	Reading Part 5	*Multiple Choice Gap fill* Ten three-option multiple choice items choosing the most appropriate words to complete five sentences in a short passage.	A2

	Reading Part 6	*Multiple Choice Reading Comprehension* Seven three-option multiple choice items focusing on comprehension of a short text.	A2
	Reading Part 7	*Multiple Choice Reading Comprehension* Five four-option multiple choice items focusing on comprehension of a short text.	B1
	Reading Part 8	*True / False Reading Comprehension* Ten true / false items focusing on comprehension of a longer text.	B1
Listening (approximately 30 minutes)	Listening Part 1	*Choose the Correct Picture* Seven three-option multiple items choosing the correct picture to match the listening.	B1
	Listening Part 2	*Matching* Choose the correct options from eight choices to answer five items based on the information in a dialogue.	A2
	Listening Part 3	*Multiple Choice Listening Comprehension* Five three-option multiple choice items focusing on comprehension of a short dialogue.	A2
	Listening Part 4	*Multiple Choice Listening Comprehension* Five three-option multiple choice items focusing on comprehension of a short monologue.	A2
	Listening Part 5	*Multiple Choice Listening Comprehension* Six three-option option multiple choice items focusing on comprehension of a longer monologue or interview.	B1
	Listening Part 6	*True / False Listening Comprehension* Six true / false items focusing on comprehension of a dialogue.	B1

APPENDIX B
Student Information and Consent Forms in Japanese and English

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: 082-218-0810
Email:jack.bower@students.mq.edu.au

研究責任者/ スーパーバイザー　氏名：メディ　リアージ

研究責任者/ スーパーバイザー　肩書き：准教授

Student Information and Consent Form

十分な説明を受け、理解したうえでの同意書

プロジェクト名：文教イングリッシュテストの評価

あなたは文教イングリッシュテスト評価研究の参加へ招待されています。
この研究の目的は文教イングリッシュテストが英語の読解力とリスニング力を効果的に測定し定められた目的を達成しているかどうかを調査することです。

この研究はオーストラリアにあるマッコーリー大学で博士課程の学生であるジャック バウワーによって行われています。(電話番号: 082-814-3191 ; email: jbower@h-bunkyo.ac.jp) また この研究は言語学の博士号要件を満たすためのもので、マッコーリー大学　言語学科メディ リアージ教授の監督の下で行われています。(電話番号: +61-406682439; email: mehdi.riazi@mq.edu.au)

協力していただけるのであれば、この研究のために以下のデータを使用することを認めてください。

- BET(Bunkyo English Test)のリスニングとリーディングの結果
- 授業内での語彙クイズ・文法クイズの結果と、学期末のスピーキングテストの結果
- BECC英語クラスの成績評価
- BETに関する意見とBET結果の使用法についての短いオンラインアンケート
- 自身の英語能力の評価についてのオンラインアンケート
- オックスフォードオンラインレベル分けテストの結果
- TOEICテストの結果

研究の過程で集められる情報や詳細な個人情報はすべて機密です。データ結果を使って発行されるどの出版物においても個人が特定されることはありません。提

供していただいた情報はジャック　バウワーと彼のスーパーバイザー、また場合によっては翻訳会社のみが見ることができます。
オンラインアンケートの集計結果を希望される場合は、下のボックスにチェックマークを入れて、あなたのe-mailアドレスを記入してください。

□はい、私はオンラインアンケートの集計結果を希望します。

研究が修了した際に研究結果の要約を送ることができます。希望する場合は下のボックスにチェックマークを入れてください。

はい、□私は研究結果の要約を希望します。

私のe-mailアドレスは
____

この研究への参加は任意です。参加する義務はなく、参加を決めた後でもいつでも参加を止めることができます。またその際に理由を求められることなく、不利益を受けることもありません。あなたがこの研究に参加する/参加しない__の決定は広島文教女子大学やマッコーリー大学とあなたとの関係に何ら影響を与えることはありません。

私（ローマ字ですべて**大文字**で記入）___________________________________,　　は以上の情報を読み理解し、質問した場合にはそれについての満足な回答を得られました。
私はこの研究に参加することに同意し、今後不利益なしにいつでも研究の参加を止めることができます。私は保管用にこの同意書の控えを渡されています。

参加者　氏名　________________________________
（ローマ字ですべて**大文字**で記入）

学籍番号: ____________________________________

参加者　署名: ____________________________年月日:　____________________________

研究者　氏名:____
（ローマ字ですべて**大文字**で記入））

研究者　署名:________________________年月日:　________________________

この研究の倫理面についてはマッコーリー大学ヒューマン研究倫理委員会と広島文教女子大学ヒューマン研究倫理委員会によって承認されています。　また、この研究に関して研究者の研究倫理等に関する苦情や疑問が発生した場合は、マッコーリー大学ヒューマン研究倫理委員会のディレクターに連絡をしてください。(電話番号 +61 (02) 9850 7854; email ethics@mq.edu.au)
いかなる苦情であってもマッコーリー大学ヒューマン研究倫理委員会が外部に漏れることなく調査をし、その結果をご報告いたします。

Women's University Hiroshima Bunkyo

MACQUARIE UNIVERSITY

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: 082-218-0810
Email: jack.bower@students.mq.edu.au

Chief Investigator's / Supervisor's Name: Mehdi Riazi
Chief Investigator's / Supervisor's Title: Associate Professor

Student Information and Consent Form

Name of Project: *Assessing the Bunkyo English Tests*

You are invited to participate in a study of Assessing the Bunkyo English Tests. The purpose of the study is investigate if the Bunkyo English Tests are achieving their stated purpose of measuring student English reading and listening ability effectively.

The study is being conducted by Jack Bower, a doctoral student at Macquarie University, Australia (tel.: 082-814-3191; email: jbower@h-bunkyo.ac.jp). This study is conducted to meet the requirements for the degree of Doctor of Philosophy (in Applied Linguistics) under the supervision of A/Prof. Mehdi Riazi, Department of Linguistics, Macquarie University (tel.: +61-406682439; email: mehdi.riazi@mq.edu.au).

If you decide to participate, you will give permission for any of the following data to be used for this research.

- Test results from the Bunkyo English Tests of listening, and reading
- Results of in-class vocabulary quizzes, grammar quizzes and end of semester speaking tests
- Your General English Course grades
- Results of an online survey about your opinions of the BETs and how they are used
- An online self-assessment survey of your English language ability
- Results of the Oxford Online Placement Test
- TOEIC test results

Any information or personal details gathered in the course of this study are confidential. No individual will be identified in any publication of the results. Only Jack Bower, his supervisors, and a local translation company will have access to the information you provide.
If you would like to obtain a copy of the overall survey results, please put a tick (√) in the box below and provide an email address.

□ Yes, I would like to receive a copy of the overall survey results.

If you would like a summary of the research results sent to you when the study is completed, please check the box below.

□ Yes, I would like to receive a summary of the results of this study.
My email address is
____

Participation in this study is entirely voluntary: you are not obliged to participate and if you decide to participate, you are free to withdraw at any time without having to give a reason and without consequence. Your decision to participate or not to participate in this study will not impact on your relationship with the Hiroshima Bunkyo Women's University or Macquarie University.

I, ____________________________________, have read 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.

Participant's Name:____
(Block letters)

Student Number: _________________

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 and the Hiroshima Bunkyo Women's University Ethics Committee. If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Macquarie University Ethics Committee through the Director, Research Ethics (telephone +61 (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.

Appendix C
Teacher and Learning Advisor Information and Consent Form for Online Surveys and Focus Groups

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: +81-090-8500-1088
Email: jack.bower@students.mq.edu.au

Chief Investigator's / Supervisor's Name: Mehdi Riazi
Chief Investigator's / Supervisor's Title: Associate Professor

Teacher Information and Consent Form

Name of Project: *Assessing the Bunkyo English Tests*

You are invited to participate in a study of Assessing the Bunkyo English Tests. The purpose of the study is to investigate if the Bunkyo English Tests are achieving their stated purpose of measuring student English reading and listening ability effectively, and having a positive influence on the General English curriculum, lesson content and pedagogy.

The study is being conducted by Jack Bower, a doctoral student at Macquarie University, Australia (tel.: 082-814-3191; email: jbower@h-bunkyo.ac.jp). This study is conducted to meet the requirements for the degree of Doctor of Philosophy (in Applied Linguistics) under the supervision of A/Prof. Mehdi Riazi, Department of Linguistics, Macquarie University (tel.: +61-406682439; email: mehdi.riazi@mq.edu.au).

If you decide to participate, you will be asked to attend a focus group, and to complete two online surveys.

If you decide to participate in the focus group, you will be asked to share and discuss your opinion about various aspects of the Bunkyo English Tests. Participation in a focus group will take between one and a half, and two hours. The focus group sessions will be video and audio recorded, and the resulting data will be transcribed and analyzed for research purposes. Each participant will receive a 3000 yen iTunes voucher. The focus group will be run by Fuyuko Takita of Hiroshima University. The identity of focus group participants will be kept anonymous in all reports of the results through the use of pseudonmyms for any quotes. All participants will be given access to a summary of the final thesis of this research via a Dropbox link.

If you decide to participate in the online surveys, you will be asked to answer a few questions about your opinions of the Bunkyo English Tests (BETs), and to estimate the percentage of students in your classes who can perform CEFR can do statements to a satisfactory level.

Any information or personal details gathered in the course of this study are confidential. No individual will be identified in any publication of the results. Only Jack Bower, and his supervisor will have access to the information you provide. If you would like to obtain a copy of the focus group discussion and/or the overall online survey results, please put a tick (√) in the box below and provide an email address.

□ Yes, I would like to receive a copy of my focus group discussion, and the overall online survey results.

If you would like a summary of the final research results of this study sent to you when the study is completed, please check the box below and provide an email address.

□ Yes, I would like to receive a summary of the results of this study.

Please check the box below if you agree to have anonymous extracts from the transcript of your interview used in publications and presentations.

□ I consent to have anonymous transcript extracts of my video/audio recording used in oral presentations and dissemination of results.

My email address is:____

Participation in this study is entirely voluntary: you are not obliged to participate and if you decide to participate, you are free to withdraw at any time without having to give a reason and without consequence.

I, _____________________________________, have read 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.

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 and the Hiroshima Bunkyo Women's University Ethics Committee. If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Macquarie University Ethics Committee through the Director, Research Ethics (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.

APPENDIX D
Senior Manager Information and Consent Forms in Japanese and English

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: +81-090-8500-1088
Email: jack.bower@students.mq.edu.au

研究責任者/ スーパーバイザー　氏名：メディ　リアージ
研究責任者/ スーパーバイザー　肩書き：准教授

十分な説明を受け、理解したうえでの同意書（大学のシニアマネージメントへのインタビューについて）

プロジェクト名：文教イングリッシュテストの評価

あなたは文教イングリッシュテスト評価研究の参加へ招待されています。

この研究の目的は文教イングリッシュテストが英語の読解力とリスニング力を効果的に測定し定められた目的を達成しているかどうかを調査することです。また BECC 英語クラスのカリキュラム、レッスン内容そして教え方への波及効果のねらいも含まれています。

この研究はオーストラリアにあるマッコーリー大学で博士課程の学生であるジャック バウワーによって行われています。(電話番号: 082-814-319 ; email: jbower@h-bunkyo.ac.jp) またこの研究は言語学の博士号要件を満たすためのもので、マッコーリー大学　言語学科メディ リアージ教授の監督の下で行われています。(電話番号: +61-406682439; email: mehdi.riazi@mq.edu.au)

参加協力してくださる場合、BET(文教イングリッシュテスト)についてのあなたの知識と意見を伺います。インタビューの所要時間は 1 時間もかかりません。

インタビューは音声録音され、研究目的のために翻訳された後、分析されます。

□ 出版物や学会での発表の為に、あなたの氏名を匿名にし、行ったインタビューの逐語記録を抜粋したものを使用してもよい場合は下のボックスにチェックマークを入れてください。

あなたは本研究の最終論文、またデータに基づき出版された全ての出版物を読む権利があります。またこの研究の過程で収集されたすべての情報、個人情報は機密で扱われます。データ結果を使って発行されるどの出版物においても許可のないものに関しては個人が特定されることはありません。

提供していただいた情報はジャック バウワーと彼のスーパーバイザーのみが見ることができます。

自分の逐語記録を希望される場合は、下のボックスにチェックマークを入れて、あなたの e-mail アドレスを記入してください。

□ はい、私は自分のインタビューの逐語記録を希望します。

□ 研究が修了した際に研究結果の要約を送ることができます。希望する場合は下のボックスにチェックマークを入れてください。

私の e-mail アドレスは

____

この研究への参加は任意です。参加する義務はなく、参加を決めた後でも、いつでも参加を止めることができます。またその際に理由を求められることなく、不利益を受けることもありません。

私（氏名を英字ブロック体で記入）________________________________，は以上の情報を読み理解し、質問した場合にはそれについての満足な回答を得られました。

私はこの研究に参加することに同意し、今後不利益なしにいつでも研究の参加を止めることができます。私は保管用にこの同意書の控えを渡されています。

参加者　氏名 ____
（ローマ字ですべて**大文字**で記入）

参加者　署名: ____________________ 年月日: ____________

研究者　氏名:____
（ローマ字ですべて**大文字**で記入）

研究者　署名: __________________ __ 年月日: ____________

この研究の倫理面についてはマッコーリー大学ヒューマン研究倫理委員会と広島文教女子大学**研究倫理委員会**によって承認されています。　また、この研究に関して研究者の研究倫理等に関する苦情や疑問が発生した場合は、ヒューマン研究倫理委員会のディレクターに連絡をしてください。(電話番号 (02) 9850 7854; email ethics@mq.edu.au)

いかなる苦情であっても**マッコーリー大学**ヒューマン研究倫理委員会が外部に漏れることなく調査をし、その結果をご報告いたします。

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: +81-090-8500-1088
Email: jack.bower@students.mq.edu.au

Chief Investigator's / Supervisor's Name: Mehdi Riazi
Chief Investigator's / Supervisor's Title: Associate Professor

University Senior Management Information and Consent Form

Name of Project: *Assessing the Bunkyo English Tests*

You are invited to participate in a study of Assessing the Bunkyo English Tests. The purpose of the study is investigate if the Bunkyo English Tests are achieving their stated purpose of measuring student English reading and listening ability effectively, and having a positive influence on the General English curriculum, lesson content and pedagogy.

The study is being conducted by Jack Bower, a doctoral student at Macquarie University, Australia (tel.: 082-814-319 ; email: jbower@h-bunkyo.ac.jp). This study is conducted to meet the requirements for the degree of Doctor of Philosophy (in Applied Linguistics) under the supervision of A/Prof. Mehdi Riazi, Department of Linguistics, Macquarie University (tel.: +61-406682439; email: mehdi.riazi@mq.edu.au).

If you decide to participate, you will be asked about your knowledge and opinions of the Bunkyo English Tests. Interviews should take less than one hour. The interviews will be audio recorded, and the resulting data will be transcribed and analyzed for research purposes. Please check the box below if you agree to have anonymous extracts from the transcript of your interview used in publications and presentations.

□ I consent to have anonymous transcript extracts of my video/audio recording used in oral presentations and dissemination of results.

Any information or personal details gathered in the course of this study are confidential. No individual will be identified in any publication of the results without prior permission. Only Jack Bower, and his supervisor will have access to the information you provide. If you would like to obtain a copy of your interview, please put a tick (√) in the box below and provide an email address.

□ Yes, I would like to receive a copy of my individual interview.

If you would like a summary of the research results sent to you when the study is completed, please check the box below.

□ Yes, I would like to receive a summary of the results of this study.

My email address is

____

Participation in this study is entirely voluntary: you are not obliged to participate and if you decide to participate, you are free to withdraw at any time without having to give a reason and without consequence.

I, ____, have read 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.

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 and the Hiroshima Bunkyo Women's University Ethics Committee. If you have any complaints or reservations about any ethical aspect of your participation in this research, you may contact the Macquarie University Ethics Committee through the Director, Research Ethics (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.

Appendix E
Ethics Review Committee (Human Research) Approval

Dear Associate Professor Riazi,

Re: "Setting and raising standards: Creating and validating institutional standardized reading and listening test aligned to the Common European Framework levels A2-B1 at a Japanese university"(5201401091)

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 December 2014. This email constitutes ethical approval only. This approval is subject to the following condition/s:

1. Please forward the official permission document from Hiroshima Bunkyo Women's University for records when this is available.
2. Please forward the advertising emails to the Ethics Sub-Committee should they be changed (currently indicated as drafts).

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/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:
Associate Professor Mehdi Riazi
Mr Jack Bower
NB. STUDENTS: IT IS YOUR RESPONSIBILITY TO KEEP A COPY OF THIS APPROVAL EMAIL TO SUBMIT WITH YOUR THESIS.

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 December 2015
Progress Report 2 Due: 16th December 2016
Progress Report 3 Due: 16th December 2017
Progress Report 4 Due: 16th December 2018
Final Report Due: 16th December 2019

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. Progress reports and Final Reports are available at the following website:
http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

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 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 Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:
http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

5. Please notify the 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/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

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 FHS Ethics at the address below.

Please retain a copy of this email as this is your official notification of ethics approval. Yours sincerely,

Dr Anthony Miller
Chair
Faculty of Human Sciences
Human Research Ethics Sub-Committee

Appendix F
GE Curriculum Content Represented in the BETs

Curriculum Content Represented in BET1 2015		
Unit	Items	Percentage
Introduction Unit	13	19.12%
Everyday Life Unit	0	0.00
My Home Unit	0	0.00
Travel Unit	24	35.29%
Relationships Unit	0	0.00
Leisure Time Unit	13	19.12%
Orientation Unit		0.00
Health Unit		0.00
Services Unit		0.00
Food & Drink Unit	1	1.47%
Shopping Unit	1	1.47%
Places Unit		0.00
Multiple Units		0.00
N/A	16	23.53%
Total	68	100%

Curriculum Content Represented in BET1 and BET2 2016				
	BET 1 2016		BET 2 2016	
Unit	Items	Percentage	Items	Percentage
Introduction Unit	11	13%	12	13.95%
Everyday Life Unit	10	12%	12	13.95%
My Home Unit	6	7%	12	13.95%
Travel Unit	21	24%	12	13.95%
Relationships Unit	14	16%	13	15.12%
Leisure Time Unit	13	15%	14	16.28%
Multiple	11	13%	11	12.79%
Total	86	100%	86	100.00%

Curriculum Content Represented in BET2 and BET3 2017						
	BET1 2017		**BET2 2017**		**BET3 2017**	
Unit	**Items**	**Percentage**	**Items**	**Percentage**	**Items**	**Percentage**
Introduction Unit	8	9%	13	15%	10	11%
Everyday Life Unit	15	17%	15	17%	8	9%
My Home Unit	11	12%	15	17%	6	7%
Travel Unit	15	17%	16	18%	8	9%
Relationships Unit	9	10%	11	12%	6	7%
Leisure Time Unit	19	21%	18	20%	11	12%
Orientation Unit	0	0%	0	0%	10	11%
Health Unit	0	0%	0	0%	7	8%
Services Unit	0	0%	0	0%	6	7%
Food & Drink Unit	0	0%	0	0%	5	6%
Shopping Unit	1	1%	0	0%	6	7%
Places Unit	1	1%	0	0%	6	7%
Multiple Units	10	11%	1	1%	0	0%
Total	89	100%	89	100%	89	100%

Appendix G
Analysis of BET Testlet Specifications Can Do Statements

Testlet	2014/15 BET Specs Can dos targeted	Analysis	BET Specs 2015/16 Can dos targeted	Analysis	BET Specs 2016/17 Can dos targeted	Analysis	Suggestion for Improvement
BERT 1	Same as BET Specs 2015/16		**CEFR-J A1.3 Reading** Be able to understand short narratives with illustrations and pictures written in simple words	This can do statement does not match the task well as task involves reading signs not short narratives. The can do statement level is also below the new course goals of A2 and B1.	**CEFR A2 Reading for Orientation** Can understand everyday signs and notices: in public places, such as streets, restaurants, railway stations; in workplaces, such as directions, instructions, hazard warnings.	This can do statement matches the task and the course goals well.	None.
BERT 2	2015 BET Vocabulary Section Same as BET Specs 2015/16	Same as BET Specs 2015/16	**CEFR-J A1.3 Reading** Be able to understand texts of general interests (e.g. articles about sports, music, travel, etc.) written with simple words supported by illustrations and pictures. **CEFR-J A2.2 Reading** Be able to understand the main points of texts dealing with everyday topics (e.g. personal information, goals and wishes, hobbies) and obtain the information they need.	This task does not have supporting pictures and the minimum goal of the course is CEFR A2. Therefore, this can do statement does not seem appropriate for the task. This task involves test takers choosing the best vocabulary to fill a gap, so the CEFR-J A2.2 statement does not seem appropriate to the task.	**CEFR Vocabulary Range** **A2** Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. **A2** Has a sufficient vocabulary for the expression of basic communicative needs. **A2** Has a sufficient vocabulary for coping with simple survival needs.	These seem to be the most appropriate CEFR can do statements for this vocabulary focused task at the target A2 level.	None
BERT 3	Same as BET Specs 2015/16	Same as BET Specs 2015/16	**CEFR-J A1.3 Spoken Interaction** Be able to ask and answer simple questions about familiar topics such as hobbies, university life, weekend activities provided people speak clearly. **Modified CEFR-J A1.3 Spoken Interaction** Be able to carry out simple classroom activities (checking answers,	This can do statement refers to spoken clarity, so it does not seem appropriate for a reading task. This can to statements' target level is below the lowest new course goal level of A2. This can to statements' target level is also below the lowest new course goal level of A2.	**CEFR Sociolinguistic Appropriateness** **A2** Can perform and respond to basic language functions, such as information exchange and requests and express opinions and attitudes in a simple way. **A2** Can handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, apologies etc.	These can do statements are at the appropriate level, and they seem broad enough to cover the possible range of dialogues in the GE curriculum. They also seem appropriate to a task in which test takers must choose the appropriate sentence or phase to respond to a conversational turn.	None

			brainstorming, borrowing items). **Modified CEFR-J A1.3 Spoken Interaction** Be able to exchange basic information and simple opinions about familiar topics (e.g. travel destinations, music, relationships and advice), using simple words a limited range of expressions. **CEFR-J Spoken Interaction A2.2** Be able to exchange opinions and feelings, express agreement and disagreement, and compare things and people using simple English. **CEFR-J A2.2 Spoken Interaction** Be able to interact in predictable everyday situations (e.g. a restaurant, a shop), using a wide range of words and expressions.	This descriptor seems generally appropriate. However, it does not seem to cover the full range of possible dialogues from the GE curriculum, which this testlet is intended to be able cover. This descriptor seems generally appropriate. However, it is questionable whether a spoken can do statement can be assessed with a reading, multiple choice task.			
BERT4	Same as BET Specs 2015/16	Same as BET Specs 2015/16	**CEFR-J A1.3 Spoken Interaction** Be able to ask and answer simple questions about familiar topics such as hobbies, university life, weekend activities provided people speak clearly. **Modified CEFR-J A1.3 Spoken Interaction** Be able to carry out simple classroom activities (checking answers,	This can do statement refers to spoken clarity, so it does not seem appropriate for a reading task. This can to statements' target level is below the lowest new course goal level of A2. This can to statements' target level is also below the lowest new course goal level of A2.	**CEFR A2 Understanding a Native Speaker Interlocutor** Can understand enough to manage simple, routine exchanges without undue effort. **CEFR A2 Information Exchange** Can understand enough to manage simple, routine exchanges without undue effort.	These can do statements are at the appropriate level and they seem to reflect the nature of the task of choosing appropriate responses in a dialogue better than CEFR reading can do statements.	None

			brainstorming, borrowing items). **Modified CEFR-J A1.3 Spoken Interaction** Be able to exchange basic information and simple opinions about familiar topics (e.g. travel destinations, music, relationships and advice), using simple words a limited range of expressions. **CEFR-J A2.2 Spoken Interaction** Be able to exchange opinions and feelings, express agreement and disagreement, and compare things and people using simple English. **CEFR-J A2.2 Spoken Interaction** *Be able to interact in predictable everyday situations (e.g. a restaurant, a shop), using a wide range of words and expressions.*	This descriptor seems generally appropriate. However, it does not seem to cover the full range of possible dialogues from the GE curriculum, which this testlet is intended to be able cover. This descriptor seems generally appropriate. However, it is questionable whether a spoken can do statement can be assessed with a reading, multiple choice task.			
BERT5	BETs 2015 Grammar Section **Modified CEFR-J A1.3 Reading** Be able to understand texts of general interests (e.g. articles about sports, music, travel, etc.) written with simple words supported by illustrations and pictures. **CEFR-J A2.2 Reading** Be able to understand the main points of texts dealing with everyday		**Modified CEFR-J A1.3 Reading** *Be able to understand texts of general interests (e.g. articles about sports, music, travel, etc.) written with simple words supported by illustrations and pictures.* **CEFR-J A1.3 Reading** *Be able to understand short narratives with illustrations and pictures written in simple words.* **CEFR-J A2.2 Reading** *Be able to find the information they need, from practical, concrete, predictable texts (e.g.*	These two CEFR-J statements are off-level for the course goals. There are also no supporting illustrations for this task. These two can dos are at the appropriate level, and grammar knowledge tested by this task may underlie this can do statement. However, a can do statement dealing directly with grammar knowledge would seem more appropriate for this gap-fill task	**CEFR Grammatical Accuracy** **A2** Uses some simple structures correctly, but still systematically makes basic mistakes - for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what he/she is trying to say. **B1** Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	These can do statements are at the appropriate level, and seem to reflect the grammar focus of this task.	None

	topics (e.g. personal information, goals and wishes, hobbies) and obtain the information they need.		*restaurant reviews, recipes, travel itineraries), provided they are written in simple English.* **CEFR-J A2.2 Reading** *Be able to understand the main points of texts dealing with everyday topics (e.g. personal information, goals and wishes, hobbies) and obtain the information they need.*	which focuses on selecting the correct grammar.			
BERT6	BETs 2015 Reading Part 4 Same as BET Specs 2015/16		**Modified CEFR-J A1.3 Reading** *Be able to understand texts of general interests (e.g. articles about sports, music, travel, etc.) written with simple words supported by illustrations and pictures.* **CEFR-J A1.3 Reading** *Be able to understand short narratives with illustrations and pictures written in simple words.* **CEFR-J A2.2 Reading** *Be able to find the information they need, from practical, concrete, predictable texts (e.g. restaurant reviews, recipes, travel itineraries), provided they are written in simple English.* **CEFR-J A2.2 Reading** *Be able to understand the main points of texts dealing with everyday topics (e.g. personal information, goals and wishes, hobbies) and obtain the information they need.*	These first two CEFR-J statements are off-level for the course goals of the new GE curriculum. There are also no supporting illustrations for this task. This can do statement does not seem to match the testlet text genres specified as "a short text adapted from authentic newspaper,magazine articles, or classroom materials." This can do statement seems generally appropriate.	**CEFR A2 Reading for Information and Argument** Can identify specific information in simpler written material he/she encounters such as letters, brochures and short newspaper articles describing events. **CEFR A2 Overall Reading Comprehension** Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items. **A2+** Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.	These can do statements seem appropriate to the nature of the task and the level.	None
BERT7	BETs 2015 Reading Part 5 *Not Stated*		*Not Stated*	Appropriate can do statements needed to be added for this testlet.	**CEFR Reading for Information and Argument**	These can do statements seem generally appropriate	Add a modified version of the following can do statement.

					B1 Can identify the main conclusions in clearly signalled argumentative texts. **B1** Can recognise the line of argument in the treatment of the issue presented, though not necessarily in detail.	for this task. As three of the items in this testlet target specific information from the text it might also be appropriate to add a modified version of the following can do statement **CEFR B1 Reading for Information and Argument** *Can recognise significant points in straightforward newspaper articles on familiar subjects.*	**CEFR B1 Reading for Information and Argument** *Can recognise significant points in straightforward newspaper articles on familiar subjects.*
BERT8	N/A	N/A	**CEFR B1 Overall Reading Comprehension** Can read straightforward factual texts of subjects related to her field and interest with a satisfactory level of comprehension. **EAQUALS Read for Orientation B1** I can find and understand the information Ineed in brochures, leaflets and other short texts related to my interests. **EAQUALS Read for Orientation B1+** I can look quickly through simple factual texts in magazines, brochures or on a website, and identify information that might be of practical use to me.	These can do statements seem generally appropriate for the task.	**CEFR B1 Overall Reading Comprehension** Can read straightforward factual texts of subjects related to her field and interest with a satisfactory level of comprehension. **EAQUALS Read for Orientation B1** I can find and understand the information I need in brochures, leaflets and other short texts related to my interests. **EAQUALS Read for Orientation B1+** I can look quickly through simple factual texts in magazines, brochures or on a website, and identify information that might be of practical use to me.	These can do statements seem generally appropriate for the task.	None
BELT1	**CEFR-J A1.3 Listening** I can understand phrases and expressions related to matters of immediate relevance to myself or				**CEFR Overall Listening Comprehension B1+** I can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details,	These can do statements seem appropriate to the task.	Suggest adding a modified version of this can do statement to better reflect the task focus of understanding key points in short dialogues.

	my family, school, neighborhood etc, provided they are delivered slowly and clearly. **CEFR-J A1.2 Listening** I can catch concrete information (e.g. numbers, times, dates, days of the week, prices), provided they are delivered slowly and clearly. **Modified CEFR-J A1.3 Listening** I can understand instructions and explanations necessary for simple transactions (e.g. teacher classroom instructions), provided they are delivered slowly and clearly				provided speech is clearly articulated in a generally familiar accent. **CEFR Overall Listening Comprehension B1** I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives. **CEFR Listening to Radio Audio & Recordings. B1** I can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly		**CEFR B1 Understanding Interaction Between Native Speakers** Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect.
BELT2	Same as BET Specs 2015/16		Same as BET Specs 2016/17		**CEFR-J A1.3 Listening** I can understand phrases and expressions related to matters of immediate relevance to myself or my family, school, neighborhood etc, provided they are delivered slowly and clearly. **CEFR-J A1.2 Listening** I can catch concrete information (e.g. numbers, times, dates, days of the week, prices), provided they are delivered slowly and clearly. **CEFR-J A1.3 Listening** I can understand instructions and explanations necessary for simple transactions (e.g. teacher classroom instructions), provided they are delivered slowly and clearly	These three CEFR-J statements are off-level for the course goals of the new GE curriculum.	Suggest the following two can do statements to better reflect the nature of the task of understanding specific information from a dialogue. **CEFR A2 Listening to Radio and Audio Recordings** Can understand and extract the essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly**.** **EAQUALS A2 Listen in Discussion**

							I can understand short conversations about family, hobbies and daily life, provided that people speak slowly and clearly
BELT3	Same as BET Specs 2015/16		Same as BET Specs 2016/17		**CEFR-J A1.3 Listening** I can understand phrases and expressions related to matters of immediate relevance to myself or my family, school, neighbourhood etc, provided they are delivered slowly and clearly. **CEFR-J A1.3 Listening** I can catch concrete information (e.g. numbers, times, dates, days of the week, prices), provided they are delivered slowly and clearly. **CEFR-J A1.3 Listening** I can understand instructions and explanations necessary for simple transactions (e.g. teacher classroom instructions), provided they are delivered slowly and clearly.	These three CEFR-J statements are off-level for the course goals of the new GE curriculum.	Suggest the following two can do statements to better reflect the nature of the task of understanding specific information from a dialogue. **CEFR A2 Listening to Radio and Audio Recordings** Can understand and extract the essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly**.** **EAQUALS A2 Listen in Discussion** I can understand short conversations about family, hobbies and daily life, provided that people speak slowly and clearly.
BELT4	Same as BET Specs 2015/16		Same as BET Specs 2016/17		**CEFR-J A1.3 Listening** I can understand phrases and expressions related to matters of immediate relevance to myself or my family, school, neighborhood etc, provided they are delivered slowly and clearly. **CEFR-J A1.3 Listening** I can catch concrete information (e.g. numbers, times, dates, days of the week, prices), provided they are delivered slowly and clearly. **CEFR-J A1.3 Listening**	These three CEFR-J statements are off-level for the course goals of the new GE curriculum.	Suggest the following two can do statements to better reflect the nature of the task of understanding specific information from an informational monologue. Suggest the following two can do statements to better reflect the nature of the task of understanding specific

					I can understand instructions and explanations necessary for simple transactions (e.g. teacher classroom instructions), provided they are delivered slowly and clearly.		information from a dialogue. **EAQUALS A2 Listening to Announcements and Instructions** I can understand short, clear and simple messages at the airport, railway station etc. For example: “The train to London leaves at 4:30”. I can understand the main information in announcements if people talk very clearly. For example: weather reports, etc
BELT5	*Not Stated*	N/A	*Not Stated*	*N/A*	*Not Stated*	Appropriate can do statements need to be added to the BET specifications for this testlet.	Suggest the following two can do statements to reflect the nature of the task of understanding specific information from “extracts of radio or TV shows or recorded messages.” **CEFR B1 Listening to Radio and Audio Recordings** Can understand the information content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech. Can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly.

							CEFR B1 Listening as a Member of a Live Audience Can follow in outline straightforward short talks on familiar topics provided these are delivered in clearly articulated standard speech.
BELT6	N/A		Same as BET Specs 2016/17	See	**EAQUALS B1 Listen in Discussion** I can understand the main points of discussion on familiar topics in everyday situations when people speak clearly, but I sometimes need help in understanding details. **EAQUALS B1+ Overall Listening** I can understand straightforward information about everyday, study- or work-related topics, identifying both general messages and specific details, provided people speak clearly in a familiar accent.	These can do statements seem appropriate to the task.	None

Appendix H
Analysis of CEFR Reading and Listening Subscales Covered by the BETs

CEFR Level	CEFR Can Do Statement	Can Do Statement from BET Specifications Analysis	Covered by BET task types
READING CORRESPONDENCE			
B1	Can understand the description of events, feelings and wishes in personal letters well enough to correspond regularly with a pen friend.	N/A	NO
A2	Can understand basic types of standard routine letters and faxes (enquiries, orders, letters of confirmation etc.) on familiar topics.	N/A	NO
A2	Can understand short simple personal letters	N/A	NO
READING FOR ORIENTATION			
B1	Can scan longer texts in order to locate desired information, and gather information from different parts of a text, or from different texts in order to fulfil a specific task.	**EAQUALS Read for Orientation B1+** I can look quickly through simple factual texts in magazines, brochures or on a website, and identify information that might be of practical use to me.	YES
B1	Can find and understand relevant information in everyday material, such as letters, brochures and short official documents.	**EAQUALS Read for Orientation B1** I can find and understand the information I need in brochures, leaflets and other short texts related to my interests.	YES
A2	Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables.	N/A	NO
A2	Can locate specific information in lists and isolate the information required (e.g. use the "Yellow Pages" to find a service or tradesman).	N/A	NO
A2	Can understand everyday signs and notices: in public places, such as streets, restaurants, railway stations; in workplaces, such as directions, instructions, hazard warnings	BERT1 — **CEFR A2 Reading for Orientation** Can understand everyday signs and notices: in public places, such as streets, restaurants, railway stations; in workplaces, such as directions, instructions, hazard warnings.	YES
READING FOR INFORMATION & ARGUMENT			
B1	Can identify the main conclusions in clearly signalled argumentative texts	BERT7 — **CEFR Reading for Information and Argument** **B1** Can identify the main conclusions in clearly signalled argumentative texts.	YES
B1	Can recognise the line of argument in the treatment of the issue presented, though not necessarily in detail	**B1** Can recognise the line of argument in the treatment of the issue presented, though not necessarily in detail.	YES

B1	Can recognise significant points in straightforward newspaper articles on familiar subjects.	**CEFR B1 Reading for Information and Argument** *Can recognise significant points in straightforward newspaper articles on familiar subjects.*	YES
A2	Can identify specific information in simpler written material he/she encounters such as letters, brochures and short newspaper articles describing events.	BERT6 — **CEFR A2 Reading for Information and Argument** Can identify specific information in simpler written material he/she encounters such as letters, brochures and short newspaper articles describing events.	YES
READING INSTRUCTIONS			
B1	Can understand clearly written, straightforward instructions for a piece of equipment	N/A	NO
A2	Can understand regulations, for example safety, when expressed in simple language.	N/A	NO
A2	Can understand simple instructions on equipment encountered in everyday life — such as a public telephone.	N/A	NO
UNDERSTANDING INTERACTION BETWEEN NATIVE SPEAKERS			
B1	Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect.	BELT1 **— CEFR B1 Understanding Interaction Between Native Speakers** Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect. **EAQUALS Listen in Discussion B1** I can understand the main points of discussion on familiar topics in everyday situations when people speak clearly, but I sometimes need help in understanding details.	YES
A2	Can generally identify the topic of discussion around her that is conducted slowly and clearly.	**EAQUALS A2 Listen in Discussion** I can understand short conversations about family, hobbies and daily life, provided that people speak slowly and clearly.	YES
LISTENING AS A MEMBER OF A LIVE AUDIENCE			
B1	Can follow a lecture or talk within his/her own field, provided the subject matter is familiar and the presentation straightforward and clearly structured.	N/A	NO
B1	Can follow in outline straightforward short talks on familiar topics provided these are delivered in clearly articulated standard speech.	**CEFR B1 Listening as a Member of a Live Audience** Can follow in outline straightforward short talks on familiar topics provided these are delivered in clearlyarticulated standard speech.	YES
LISTENING TO ANNOUNCEMENTS & INSTRUCTIONS			
B1	Can understand simple technical information, such as operating instructions for everyday equipment.	N/A	NO
B1	Can follow detailed directions.	N/A	NO

A2	Can catch the main point in short, clear, simple messages and announcements.	**EAQUALS A2 Listening to Announcements and Instructions** I can understand short, clear and simple messages at the airport, railway station etc. For example: "The train to London leaves at 4:30 I can understand the main information in announcements if people talk very clearly. For example: weather reports, etc	YES
A2	Can understand simple directions relating to how to get from X to Y, by foot or public transport.	N/A	NO
LISTENING TO RADIO AUDIO & RECORDINGS			
B1	Can understand the information content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech.	**CEFR B1 Listening to Radio and Audio Recordings** Can understand the information content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech.	YES
B1	Can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly.	**CEFR Listening to Radio Audio & Recordings. B1** I can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly	YES
A2	Can understand and extract the essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly.	**CEFR A2 Listening to Radio and Audio Recordings** Can understand and extract the essential information from short recorded passages dealing with predictable everyday matters that are delivered slowly and clearly**.**	YES
WATCHING TV AND FILM			
B1	Can understand a large part of many TV programmes on topics of personal interest such as interviews, short lectures, and news reports when the delivery is relatively slow and clear.	N/A	NO
B1	Can follow many films in which visuals and action carry much of the storyline, and which are delivered clearly in straightforward language.	N/A	NO
B1	Can catch the main points in TV programmes on familiar topics when the delivery is relatively slow and clear.	N/A	NO
A2	Can identify the main point of TV news items reporting events, accidents etc. where the visual supports the commentary.	N/A	NO
A2	Can follow changes of topic of factual TV news items, and form an idea of the main content.	N/A	NO

Appendix I
Representation of BET task types in the GE curriculum

Count of Lesson Tasks and Assessments in the GE 2015 First Year Curriculum which were Similar to BERT Testlets

BERT Part	FE Semester 1	FE Semester 2	FE BET Review Lesson	FE Semester 1 Assessments	FE Semester 2 Assessments	Total
1	0	0	1	0	0	1
2	0	0	0	3	3	6
3	0	0	0	0	0	0
4	0	0	1	0	0	1
5	0	2	0	0	0	2
6	8	5	2	2	1	18
7	0	1	0	0	0	1
8	0	0	0	0	0	0

Count of Lesson Tasks and Assessments in the GE 2015 First Year Curriculum which were Similar to BELT Testlets

BELT Part	FE Semester 1	FE Semester 2	FE BET Review Lesson	FE Semester 1 Assessments	FE Semester 2 Assessments	Total
1	0	0	1	0	0	1
2	1	1	0	0	0	2
3	0	3	0	2	0	3
4	0	2	0	0	0	2
5	0	0	0	0	0	0
6	0	0	1	0	0	1

Count of Lesson Tasks and Assessments in the GE 2016 Curriculum which were Similar to BERT Testlets

BERT Part	FE Sem 1	FE Sem 2	FE Sem 1 Assessments	FE Sem 2 Assessments	FE BET Review Lesson	FE Total	SE Sem 1	SE Sem 2	SE Sem 1 Assessments	SE Sem 2 Assessments	SE BET Review Lesson	SE Total	GE Total
1	0	0	0	0	1	1	0	1	0	0	1	2	3
2	0	0	3	3	0	6	0	0	3	3	0	6	12
3	0	0	0	0	1	1	0	0	0	0	1	1	2
4	0	0	0	0	1	1	2	1	0	1	1	5	6
5	0	2	0	0	0	2	1	0	0	1	0	2	4
6	8	5	1	1	0	15	2	3	0	1	0	6	21
7	0	1	0	0	1	2	0	0	0	0	1	1	3
8	0	0	0	0	1	1	0	0	0	0	1	1	2

Count of Lesson Tasks and Assessments in the GE 2016 Curriculum which were Similar to BELT Testlets

BELT Part	FE Sem 1	FE Sem 2	FE Sem 1 Assessments	FE Sem 2 Assessments	FE BET Review Lesson	FE Total	SE Sem 1	SE Sem 2	SE Sem 1 Assessments	SE Sem 2 Assessments	SE BET Review Lesson	SE Total	GE Total
1	0	0	0	0	1	1	0	2	0	1	1	4	5
2	1	1	1	0	1	4	0	0	0	0	1	1	5
3	0	1	0	1	0	2	3	1	1	1	1	7	9
4	0	2	0	0	1	3	0	0	0	0	1	1	4
5	0	0	0	0	1	1	0	0	0	0	1	1	2
6	0	0	0	0	0	0	1	3	0	0	1	5	5

Appendix J
Student Attitudes to the BETs and Class Streaming Surveys

Survey Statements	April–May 2015 1st year students	Jan–Feb, 2016 1st year students	April–May 2016 1st year students	Jan 2017 1st year students	Jan 2017 2nd year students
1. BETの日本語説明（アナウンス）は全てわかりやすかった。 BET spoken instructions were easy to understand.	✓	✓	✓	✓	✓
2. BETの質問用紙のインストラクション（説明）はわかりやすかった。 BET test paper instructions were easy to understand.	✓	✓	✓	✓	✓
3. BETの質問用紙のインストラクション（説明）はわかりやすかった。 BET test paper instructions were easy to understand.	✓	✓	✓	✓	✓
4. BETのリスニングセクションのインストラクション（説明）はわかりやすかった BET DVD spoken instructions were easy to understand.	✓	✓	✓	✓	✓
5. BETのリスニングセクションのインストラクション（説明）はわかりやすかった。 BET listening section instructions are easy to understand.	✓	✓	✓	✓	✓
6. BETの語彙・文法セクションのインストラクション（説明）はわかりやすかった。 BET vocabulary and grammar section instructions are easy to understand	✓	✗	✗	✗	✗
7. BET読解セクションのインストラクション（説明）はわかりやすかった。 BET reading section instructions are easy to understand.	✓	✓	✓	✓	✓
8. 自分のBECC英語クラスの学生の英語レベルは自分と同じくらいと思う。 I think that the other students in my BECC English class have similar English language ability to myself.	✓	✓	✗	✓	✓
9. 私のクラスでのほとんどの学生の英語のスキルは同じくらいと思う。 In my class, most students' English skills are similar to mine.	✓	✓	✗	✓	✓
10. クラスの英語レベルは私の英語レベルとあっていると思う。 I feel that the level of my class is appropriate for my level of English.	✓	✓	✗	✓	✓
11. 授業で配布されるプリントのレベルは私の英語のレベルにあっていると思う。 The level of the classroom handouts is appropriate for my English level.	✓	✓	✗	✗	✗
12. 授業中にダウンロードする教材は私の英語のレベルにあっていると思う。 The level of the classroom materials I download is appropriate for my English level.				✓	✓
13. 私はクラスメートの話す英語をはっきりと理解している。 I can clearly understand what my classmates are saying in English.	✓	✓	✗	✓	✓

Survey Statements	**April–May 2015 1st year students**	**Jan–Feb, 2016 1st year students**	**April–May 2016 1st year students**	**Jan 2017 1st year students**	**Jan 2017 2nd year students**
14. BETスコアは私が英語でどんなことができるのかを知るのを助けてくれると思う。 I think my BET results help me to know what I can do in English.	✓	✓	✓	✓	✓
15. BETのスコアは自分の読解力、リスニング力、そして語彙・文法の英語能力レベルを知るのを助けてくれると思う。 I think my BET scores help me to know my English proficiency level in reading, listening and grammar.	✓	✗	✗	✗	✗
16. BETのスコアは自分の読解力、リスニング力レベルを知るのを助けてくれると思う I think my BET scores help me to know my English proficiency level in reading, listening.	✗	✓	✓	✓	✓
17. BETスコアは私の英語の苦手な部分を知ることを助けてくれると思う。 I think my BET scores help me to know my weak points in English.	✓	✓	✓	✓	✓
18. BETスコアは私の英語の得意な部分を知ることを助けてくれると思う。 I think my BET scores help me to know my strong points in English.	✓	✓	✓	✓	✓
19. BETスコアは英語学習のプランを立てるのに役立つと思う。 My BET scores are useful to help me to plan my English study	✓	✓	✓	✓	✓
20. 次のBETに向けて、BETのスコアを良くしたい。 I want to improve my BET score on my next BET	✓	✓	✓	✓	✗
21. 私はBETスコアを上げるために熱心に勉強をする意欲がある。 I'm motivated to study harder to improve my BET score.	✓	✓	✓	✓	✗
22. BETのリスニング問題はBECC英語クラスのリスニング問題と似ている。 BET listening tasks are similar to listening tasks in my BECC English classes.	✗	✗	✗	✓	✓
23. BETのリーディング問題はBECC英語クラスのリーディング問題と似ている。 BET reading tasks are similar to reading tasks in my BECC English classes.	✗	✗	✗	✓	✓
24. BETは私がBECC英語クラスで勉強した広範囲の内容が含まれていた。 The BET included a wide range of content from what I studied in my BECC English classes.	✗	✗	✗	✓	✓

Appendix K
Teacher Attitudes to the BETs Surveys

Survey Statements	June–July, 2015	July, 2016	Jan–Feb, 2017
The BETs do a good job of streaming students into classes by English language ability.	✓	✓	✓
Students in the higher-streamed classes I teach clearly have higher overall English language proficiency than students in the lower-streamed classes I teach.	✓	✓	✓
I think that GE students have the same opportunity to study the material covered by the BETs in all GE classes.	✓	✓	✓
I think the content of GE classes is generally the same across all of the GE Freshman English and Sophomore English classes taught by different teachers.	✓	✓	✓
I think the BETs are an effective way to measure GE students' English reading proficiency.	✓	✓	✓
I think the BETs are an effective way to measure GE students' English listening proficiency.	✓	✓	✓
I refer to the BET specifications when designing lessons for the new GE curriculum.	✓	✓	✓
I referred to the BET specifications when I made or revised lessons for the GE curriculum in semester 2.	✘	✘	✓
I think about how my teaching will affect my students' BET scores when preparing my GE classes.	✓	✓	✘
I thought about how my teaching would affect my students' BET scores when preparing my GE classes.	✘	✘	✓
I refer to the guidelines in the GE Curriculum Overview document for making listening and reading tasks, when designing lessons for the new GE curriculum.	✓	✓	✘
I referred to the guidelines in the GE Curriculum Overview document for making listening and reading tasks, when I made or revised lessons for the GE curriculum in semester 2.	✘	✘	✓
BET content is representative of GE curriculum content.	✘	✘	✓
BET reading tasks are similar to reading tasks in the GE curriculum.	✘	✘	✓
BET listening tasks are similar to listening tasks in the GE curriculum.	✘	✘	✓

Appendix L
CEFR self-assessment survey can do statements for reading and listening

Reading
R A1 I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues. 例えば、掲示やポスター、カタログの中のよく知って いる名前、単語、単純な文 を理解できる。
R A2-1 I can read very short, simple texts. ごく短い簡単な文章なら理解できる。
R A2-2 I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters. 広告や内容紹介のパンフレット、メニュー、予定表のようなものの中か ら日常の単純な具体的に予測がつく情報を取り出せる。
R A2-2-3 I can understand short simple personal letters. 簡単で短い個人的な手紙は理解できる。
R B1-1 I can understand texts that consist mainly of high frequency everyday or job-related language. 非常によく使われる日常言語や、自分の仕事関連の言葉で書かれたテ クストなら理解できる。
R B1-2 I can understand the description of events, feelings and wishes in personal letters. 起こったこと、感情、希望が表現されている個人的な手紙を理解できる。
Listening
L A1 I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly. はっきりとゆっくりと話してもらえれば、自分、家族、 すぐ周りの具体的なものに関する聞き慣れた語やごく 基本的な 表現 を 聞き 取れる
L A2-1 I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal and family information, shopping, local area, employment). (ごく基本的な個人や家族の情報、買い物、近所、仕事などの） 直接自分に関連した領域で最も頻 繁に使われる語彙や表現を理解することができる。
L A2-2 I can catch the main point in short, clear, simple messages and announcements. 短い、はっきりとした簡単なメッ セージやアナウンスの要点を聞き取れる。
L B1-1 I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. 仕事、学校、娯楽で普段出会うような身近な話題について、明瞭で標準 的な話し方の会話なら要点を理解することができる。
L B1-2 I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear. 話し方が比較的ゆっくり、はっきり としているなら、時事問題や、個人的もしくは仕事上の話題について も、ラジオやテレビ番組の要点を理解することができる。

Appendix M
Correlations of BERT, BELT Scores and CEFR Self-assessment survey results

Test	n size	Survey Reliability Cronbach's Alpha	Test Reliability K-R20	Correlation significance (2 tailed)		Correlation Strength		Disattenuated Correlations	
				Spearman	Pearson	Spearman	Pearson	Spearman	Pearson
BERT1 2015	226	.90	.80	<.001	<.001	.44	.43	.51	.52
BELT1 2015	226	.84	.59	<.001	<.001	.37	.36	.52	.51
BERT1 2016	204	.90	.79	<.001	<.001	.54	.51	.64	.60
BELT1 2016	204	.89	.61	<.001	<.001	.33	.29	.44	.39
BERT2 2016	180	.89	.79	<.001	<.001	.53	.54	.63	.65
BELT2 2016	180	.86	.62	<.001	<.001	.34	.36	.47	.49
BERT2 2017	147	.90	.80	<.001	<.001	.33	.36	.38	.42
BELT2 2017	147	.83	.61	.011	.003	.21	.24	.30	.34
BERT3 2017	146	.90	.79	<.001	<.001	.43	.43	.51	.51
BELT3 2017	146	.88	.68	<.001	<.001	.49	.51	.64	.65

Appendix N
Focus Group Interview Questions

2015 Questions	2016 Questions
1. Tell us your name and how long you've been working at the BECC.	1. Tell us your name and how long you've been working at the BECC
2. What is the first thing that you think of when you think of the BETs?	2. What is the first thing that you think of when you think of the BETs?
3. How do you try to prepare your GE students for the BETs?	3. How do you try to prepare your GE students for the BETs?
4. How much class time do you spend on preparing your students for the BETs?	4. How many of the GE lessons do you usually cover for each unit?
5. To what extent do you think that the BETs are administered in the same way for all classes?	5. To what extent do you think BET listening and reading tasks are representative of the listening and reading tasks in the GE curriculum?
6. To what extent do you refer to the BET specifications and/or the GE Curriculum Overview when designing new lessons for the curriculum?	6. How do you use the BET specifications when designing new lessons for the curriculum?
7. How effective is the use of BET scores for streaming students for GE classes?	7. How effective is the use of BET scores for streaming students for GE classes?
8. What are the effects of current GE steaming policy for teacher classroom management, and class preparation?	8. What are the effects of current GE steaming policy for teacher classroom management, and class preparation?
9. How beneficial is the current GE streaming policy for students?	9. How beneficial is the current GE streaming policy for students?
10. To what extent do you think BETs affect students' motivation for learning English language?	10. To what extent do you think BETs affect students' motivation for learning English?
11. Should BECC teachers be more actively involved in writing BET tasks?	11. Should BECC teachers be more actively involved in writing BET tasks?
12. What is your overall opinion of the BETs?	12. What is your overall opinion of the BETs?
13. What suggestions do you have for improving the BETs, or for improving how information about the BETs is communicated?	13. What suggestions do you have for improving the BETs, or for improving how information about the BETs is communicated?

Appendix O
Management Semi-Structured Interview Questions

Interview 1 Questions	Interviews 2–5 Questions
1. What do you know about the General English curriculum? GE（一般教養の BECC 英語）のカリキュラムについて何を知っていますか?	1. In general, what do you think about the role of English language courses, and the Bunkyo English Communication Center, at Hiroshima Bunkyo Women’s University? 一般的に広島文教女子大学の英語の授業の役割と BECC の役割についてどうおもいますか？
2. What do you know about the Bunkyo English Tests? BET (文教英語テスト)について何を知っていますか?	2. Could you please tell me what you have heard about the BECC General English curriculum, or anything you know about it? BECC 一般教養の GE（General　English） について何か聞いたことはありますか？何か知っていることはありますか？
3. What do you think about streaming English classes for GE courses? GE コースのレベル分け（授業）についてどう思われますか?	3. Have you heard of the Bunkyo English Tests of reading and listening? Could you please explain what you have heard about these tests, or what you know about them? BET（文教英語テスト）のリスニングとリーディングについて何か聞いたことはありますか？これらのテストについてどんなことを聞いたことがありますか、またどんあことを知いますか？
4. What do you think about setting English proficiency performance standards for GE graduates? 一般教養英語の修了時に“英語でここまでできるようになる等”の設定を行うことについてどう思われますか?	4. BECC General English classes are streamed depending on students’ results from the BETs. What do you think about streaming English classes for GE courses? BECC の GE のカリキュラムは can do statements という形で、学生の熟達度レベルの目標を設定します。例えば BECC の can do statements の目標は、「家族や趣味、仕事、旅行、 最近の出来事など、日常生活に直接 関係のあることや個人的な関心事 について、準備なしで会話に入ることができる。」Can do statement－どのようなことを英語でできる能力があるのかを示したリストが書かれた証明書を、一般教養の BECC 英語の修了時に、学生に授与することについてどう思われますか?
5. What do you think about issuing proficiency certificates listing English ability in terms of can do statements to GE graduates? Can do statement－どのようなことを英語でできる能力があるのかを示したリストを一般教養の BECC 英語の修了時、学生に授与することについてどう思われますか?	5. The BECC GE curriculum sets proficiency goals for students in the form of can do statements. An example of a BECC can do statement goal is: I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events). What do you think about issuing proficiency certificates listing English ability in terms of can do statements to GE graduates? BECC の GE のカリキュラムは can do statements という形で、学生の熟達度レベルの目標を設定します。例えば BECC の can do statements の目標は、「家族や趣味、仕事、旅行、 最近の出来事など、日常生活に直接 関係のあることや個人的な関心事 について、準備なしで会話に入ることができる。」Can do statement－どのようなことを英語でできる能力があるのかを示したリストが書かれた証明書を、一般教養の BECC 英語の修了時に、学生に授与することについてどう思われますか?

6. How important is it to give students numerical scores to show their English ability in reading, listening and speaking? 読解、リスニング、スピーキングの英語力を数字で表したスコアを学生に示してあげることはどのくらい重要なことと思われますか?	6. How important do you think it is to give students numerical scores to show their English ability in reading, listening and speaking? リーディング、リスニング、スピーキングの英語力を数字で表したスコアを学生に示すことはどのくらい重要なことと思われますか?
7. How are BET scores used outside of the Bunkyo English Communication Center in other university departments? BET（文教イングリッシュテスト）の成績結果を他学科で使用することがありますか?　ある場合はどのように使用されていますか?	7. How useful will it be for Bunkyo PR to show student improvement in students' English reading and listening ability over two years of study in the GE curriculum based on BET scores? 2年間のGEカリキュラムで学ぶことによって英語のリーディング、リスニングの能力が伸びたことを、BETのスコアで表すことは、文教のPRにとってどれくらい役に立つと思われますか？
8. How useful are BET scores for these purposes in other departments? BETの成績結果は他学科でどれくらい役にたっていますか?	

Appendix P
KR-20 Statistics for GE Vocabulary Quizzes

FE 2015 Quizes (n=177)	KR-20	FE 2016 Quizes (n=179)	KR-20	SE 2016 Quizes (N=203)	KR-20
Unit 1	.84	Unit 1	.82	Unit 7	.80
Unit 2	.73	Unit 2	.78	Unit 8	.69
Unit 3	.83	Unit 3	.81	Unit 9	.80
Unit 4	.79	Unit 4	.77	Unit 10	.76
Unit 5	.78	Unit 5	.79	Unit 11	.73
Unit 6	.80	Unit 6	.81	Unit 12	.66

Appendix Q
Summary of OOPT Administration Data

Time administered	Students from Higher A2–B1 Stream Class	Students from Lower A2–B1 Stream Class	Students from Higher A1–A2 Stream Class	Students from Lower A1–A2 Stream Class	Students from Mixed A1–A2 Stream Class	GCD (A2–B1)	Total
End of April 2015 for BET1 2015	20	23	24	20	NA	NA	88
End of April 2016 for BET1 2016	26	NA	24	NA	20	16	86
End of April 2016 for BET2 2016	22	20	NA	NA	16	NA	58
End of January 2017 for BET2	NA	NA	NA	NA	NA	15	15
Late December/Early January 2017 for BET3	37	43	NA	NA	26	NA	106

Appendix R
BERT, KET and PET reading section lexical profiles on the GSL and AWL

Lexical characteristic	BERT 1 2015	BERT1 2016	BERT2 2016	BERT 1 2017	BERT 2 2017	BERT 3 2017	BERT Average	KET (A2)	PET (B1)
Type-token Analysis									
Words in text (tokens)	1274	1804	1905	1,762	1891	1904	1,756.67	1,310	3,962
Different words (types)	446	515	531	538	594	641	544.17	483	1,184
Type-token ratio	.35	.29	.28	.31	.31	.34	0.31	.37	.30
Tokens per type	2.86	3.5	3.59	3.28	3.18	2.97	3.23	2.71	3.35
Frequency Analysis (VP) (Frequency coverage in %)									
Lexical characteristic	**BERT 1 2015**	**BERT1 2016**	**BERT2 2016**	**BERT 1 2017**	**BERT 2 2017**	**BERT 3 2017**	**BERT Average**	**KET (A2)**	**PET (B1)**
K1 Words (1–1000)	80.93	83.70	81.84	79.51	79.27	80.93	81.03	86.95	81.22
K2 Words (1,001–2,000)	5.10	6.10	5.72	5.90	6.77	6.83	6.07	5.04	8.81
AWL Words (academic)	1.41	1.55	1.26	1.87	1.27	1.68	1.51	.61	2.45
Off-list words	12.56	8.65	11.18	12.71	12.69	10.56	11.39	7.40	7.52

Appendix S
BERT and KET/PET BNC frequency levels (%)

Lexical characteristic	BERT1 2015	BERT1 & BERT2 2016	2017 BERTs	KET (A2)	PET (B1)
K1	83.52	86.06	83.25	89.30	84.73
K2	4.95	4.48	5.51	5.04	8.63
K3	1.96	1.19	2.30	.69	2.32
K4	1.65	.92	.76	1.22	.83
K5	.47	.27	.58	.69	.43
K6	.16	.16	.16	.08	.08
K7	.08	.03	.05	.15	.05
K8	.39	.75	1.01	0	.20
K9	0	0	.14	.8	.20
K10	.08	.19	.11		.10
K11	.16	.05	.13		.15
K12–20	.44	.10	.29	0	0
Off-list	8.3	5.80	5.85	2.75	2.27
Tokens per family (on list)	3.66	7.13	7.63	3.54	4.66

Appendix T
BELT tapescript and KET/PET BNC frequency levels (%)

Frequency level	BELT1 2015	BELT1 & BELT2 2016	2017 BELTs	KET	PET
K1	81.39	88.38	88.43	93.72	92.38
K2	6.09	4.42	4.47	3.53	4.64
K3	2.83	2.13	1.64	.74	1.16
K4	1.11	.75	.58	.20	.43
K5	1.63	.92	.59	.29	.36
K6–10	.51	.70	.61	.22	.22
K11–15	.35	.24	.27	.03	.04
K16–20	.09	.03	.03	0	0
Off list	4.72	2.43	3.36	1.28	.78
Words	1166	3710	6422	6,890	9,081
Types	442	681	934	902	1.354
Tokens per type	2.64	5.17	6.88	7.64	6.71
AWL	1.46	.94	.81	.13	.90

Appendix U
Teacher and Student Survey Analyses for Domain Inference Backings

Analysis of teacher attitudes to the BETs survey statements for domain definition inference backing (*n*=11)

Survey Statement	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Proportion Agreeing
BET content is representative of GE curriculum content.	0	0	0	1	8	2	5.1/6	100%
BET reading tasks are similar to reading tasks in the GE curriculum.	0	0	1	1	7	2	4.9/6	91%
BET listening tasks are similar to listening tasks in the GE curriculum	0	0	1	1	6	3	5/6	91%

Analysis of student attitudes to the BETs survey statements for domain definition inference backing all FE *(n* = 207)

Survey Statement	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std. Deviation	Proportion Agreeing
The BET included a wide range of content from what I studied in my BECC English classes	0	3	13	74	91	26	4.60/6	.84	92.27%
BET reading tasks are similar to reading tasks in my BECC English classes	0	4	18	84	81	20	4.46/6	.86	89.37%
BET listening tasks are similar to listening tasks in my BECC English classes.	0	5	34	74	77	17	4.32/6	.93	81.16%

Analysis of student attitudes to the BETs survey statements for domain definition inference backing SE All (*n* = 193)

Survey Statement	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std. Deviation	Proportion Agreeing
The BET included a wide range of content from what I studied in my BECC English classes	2	4	12	74	78	23	4.51/6	.93	90.67%
BET reading tasks are similar to reading tasks in my BECC English classes	3	4	14	80	76	16	4.40/6	.93	89.12%
BET listening tasks are similar to listening tasks in my BECC English classes.	2	6	25	83	60	17	4.26/6	.97	82.9%

Appendix V
Student Survey Analyses for Evaluation Inference Backings

BET DVD spoken instructions were easy to understand.
BET の日本語説明（アナウンス）は全てわかりやすかった。

BET Administration	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std Deviation	Proportion Agreeing
BET1 2015 (*n* = 175)	2	2	4	39	84	44	4.90/6	.92	95.43%
BET1 2016(*n* = 197)	0	3	10	39	91	54	4.93/6	.90	93.40%
BET2 2016(*n* = 207)	0	0	4	34	87	81	5.19/6	.78	98.06%
BET2 2017(*n* = 209)	2	1	6	30	111	59	5.03/6	.86	95.69%
BET3 2017(*n* = 193)	0	0	9	33	80	71	5.10/6	.85	95.3%

BET test paper instructions were easy to understand.
BET の質問用紙のインストラクション（説明）はわかりやすかった。

BET Administration	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std Deviation	Proportion Agreeing
BET1 2015 (*n* = 175)	2	2	9	46	79	37	4.77/6	.96	92.57%
BET1 2016(*n* = 197)	1	3	5	48	91	49	4.89/6	.09	95.43%
BET2 2016(*n* = 207)	0	1	0	32	92	81	5.22/6	.74	99.51%
BET2 2017(*n* = 209)	2	1	5	35	113	53	4.99/6	.85	96.17%
BET3 2017(*n* = 193)	0	0	5	35	84	69	5.12/6	.79	97.4%

Appendix W
Teacher Survey Analyses for Utilization Inference Backings

A. Teacher Survey Analysis for Utilization Inference Assumption 1 Backing

Analysis of teacher attitudes to the BETs survey statements for utilization inference backing

Survey Statement	Administration Time	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Mean	SD	Agreeing %
The BETs do a good job of streaming students into classes by English language ability.	June & July 2015 (*n* = 8)	0	0	0	3	5	0	4.63	.52	100%
	July 2016 (*n* = 10)	0	0	0	1	8	1	5.00	.47	100%
	January 2017 (*n*=10)	0	0	0	1	8	2	5.09	.54	100%
Students in the higher-streamed classes I teach clearly have higher overall English language proficiency than students in the lower-streamed classes I teach.	June & July 2015 (*n* = 7)	0	0	0	1	2	4	5.43	.79	100%
	July 2016 (*n* = 9)	0	0	0	1	5	3	5.22	.67	100%
	January 2017 (*n* = 10)	0	0	1	0	3	6	5.4	.97	90%

B. Teacher Survey Analyses for Utilization Inference Assumption 2 Backing

Analysis of teacher attitudes to the BETs survey statements for utilization inference backing

Survey Statement	Survey Dates	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	STD Dev	Proportion Agreeing
I think the BETs are an effective way to measure GE students' English reading proficiency.	June/July 2015	0	0	1	0	7	0	4.75	.71	87.5%
	July 2016	0	0	0	1	9	1	5.00	.45	100%
	Jan/Feb 2017	0	0	0	1	6	4	5.27	.65	100%
I think the BETs are an effective way to measure GE students' English listening proficiency.	June/July 2015	0	0	1	0	7	0	4.75	.71	87.5%
	July 2016	0	0	0	2	7	1	4.90	.57	100%
	Jan/Feb 2017	0	0	1	0	6	4	5.18	.87	90.91%

C. Teacher Survey Analysis for Utilization Inference Assumption 4 Backing

Analysis of teacher attitudes to the BETs survey statements for utilization inference backing

Survey Statement	Administration Time	Never	Rarely	Occasionally	Usually	Always	Mean	SD	Proportion Never/Rarely/ Occasionally	Proportion Usually/Always
I thought about how my teaching would affect my students’ BET scores when preparing my GE classes.	June & July 2015 (*n* = 9)	0	2	3	4	0	3.22	.83	55.55%	44.44%
	July 2016 (*n* = 12)	0	6	1	3	2	3.08	1.24	58.33%	41.66%
	January 2017 (*n* = 11)	0	1	6	3	1	3.36	.81	63.63%	36.36%

Appendix X
Student Survey Analyses for Utilization Inference Backings

A. Student Survey Analyses for Utilization Inference Assumption 1 Backing

Analysis of student attitudes to the BETs survey statements for utilization inference backing

Survey Statement	Administration Dates Student Year	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std. Deviation	Proportion Agreeing
I think that the other students in my BECC English class have similar English language ability to myself.	April/May, 2015 First Years	2 (1.15%)	7 (4.20%)	20 (11.49%)	77 (44.25%)	56 (32.18%)	12 (6.09%)	4.23	.96	83.33%
	Jan/Feb, 2016 First Years	4 (1.94%)	4 (1.94%)	15 (7.28%)	61 (29.61%)	92 (44.66%)	30 (14.56%)	4.57	1.02	88.83%
	Jan, 2017 First Years	1 (.47%)	6 (2.79%)	21 (9.77%)	81 (37.67%)	88 (40.93%)	18 (8.37%)	4.4	.91	87%
	Jan, 2017 Second Years	2 (1.04%)	9 (4.66%)	21 (10.88%)	67 (34.72%)	79 (40.93%)	15 (7.77%)	4.33	1.00	83.42%
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
In my class, most students' English skills are similar to mine.	April/May, 2015 First Years	2 (1.15%)	7 (4.02%)	23 (13.22%)	94 (54.02%)	38 (21.8%)	10 (5.75%)	4.09	.92	81.61%
	Jan/Feb, 2016 First Years	2 (.97%)	4 (1.94%)	22 (10.68%)	77 (37.38%)	81 (39.32%)	20 (9.71%)	4.41	.94	86.41%
	Jan, 2017 First Years	1 (.47%)	9 (4.19%)	29 (13.49%)	83 (38.60%)	81 (37.67%)	12 (5.58%)	4.3	.94	81.90%
	Jan, 2017 Second Years	3 (1.55%)	11 (5.70%)	31 (16.06%)	63 (32.64%)	72 (37.31%)	13 (6.74%)	4.19	1.07	76.68%
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
I feel that the level of my class is appropriate for	April/May, 2015 First Years	3 (1.72%)	3 (1.72%)	21 (12.07%)	73 41.92%	62 (35.63%)	12 (6.90%)	4.29	.95	84.48%
	Jan/Feb, 2016 First Years	2 (.97%)	2 (.97%)	10 (4.85%)	60 (29.13%)	96 (46.60%)	36 (17.48%)	4.72	.91	93.2%

my level of English.	Jan, 2017 First Years	0 (.00%)	7 (3.26%)	19 (8.84%)	75 (34.88%)	92 (42.79%)	22 (10.23%)	4.5	.91	87.9%
	Jan, 2017 Second Years	5 (2.59%)	5 (2.59%)	21 (10.88%)	68 (35.23%)	78 (40.41%)	16 (8.29)	4.33	1.04	83.94
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
The level of the classroom handouts (materials I download) is appropriate for my English level.	April/May, 2015 First Years	2 (1.15%)	4 (2.30%)	14 (8.05%)	72 (41.38%)	71 (40.80%)	11 (6.32%)	4.37	.9	88.51%
	Jan/Feb, 2016 First Years	2 (.97%)	1 (.49%)	13 (6.31%)	61 (29.61%)	95 (46.12%)	34 (16.50%)	4.69	.91	92.23%
	Jan, 2017 First Years	0 (.00%)	3 (1.40%)	11 (5.12%)	80 (37.12%)	100 (46.51%)	21 (9.77%)	4.6	.79	93.5%
	Jan, 2017 Second Years	2 (1.04%)	4 (2.07%)	16 (8.29%)	67 (34.72%)	89 (46.11%)	15 (7.77%)	4.46	.91	88.60%
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
I can clearly understand what my classmates are saying in English.	April/May, 2015 First Years	3 (1.72%)	3 (1.72%)	23 (13.22%)	63 (36.31%)	65 (37.36%)	17 (9.77%)	4.35	1	83.33
	Jan/Feb, 2016 First Years	0 (0.00%)	3 (1.46%)	15 (7.28%)	72 (34.95%)	89 (43.20%)	27 (13.11%)	4.59	.86	91.26%
	Jan, 2017 First Years	0 (.00%)	5 (2.33%)	12 (5.58%)	92 (42.79%)	83 (38.60%)	23 (10.70%)	4.5	.85	92.1%
	Jan, 2017 Second Years	1 (.52%)	5 (2.59%)	15 (7.77%)	66 (34.20%)	90 (46.63%)	16 (8.29%)	4.49	.89	89.12%

B. Student Survey Analyses for Utilization Inference Sub-Assumption 3.2 Backing

Analysis of student attitudes to the BETs survey statements for utilization inference backing

Survey Statement	Administration Dates Student Year	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std. Deviation	Proportion Agreeing
I think my BET results help me to know what I can do in English.	April/May, 2015 First Years	2 (1.16%)	1 (.58%)	11 (6.36%)	68 (39.31%)	69 (39.88%)	22 (12.72%)	4.54	.90	91.91%
	Jan/Feb, 2016 First Years	1 (.49%)	0 (0%)	11 (5.34%)	62 (30.10%)	105 50.97%)	27 (13.11%)	4.70	.80	94.17%
	Jan, 2017 First Years	0 (0%)	2 (.93%)	11 (5.12%)	76 (35.35%)	99 (46.05%)	27 (12.56%)	4.64	.80	93.95%
	Jan, 2017 Second Years	0 (0%)	2 (1.04%)	19 (9.84%)	70 (36.27%)	71 (36.79%)	31 (16.06%)	4.57	.91	89.12%
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
I think my BET scores help me to know my English proficiency level in reading, listening (and grammar).	April/May, 2015 First Years	1 (.58%)	2 (1.16%)	6 (3.47%)	62 (35.84%)	78 (45.09%)	24 (13.87%)	4.65	.85	94.80%
	Jan/Feb, 2016 First Years	1 (.49%)	0 (0%)	8 (3.88%)	61 (29.61%)	103 (50.00%)	33 (16.02%)	4.77	.80	95.63%
	Jan, 2017 First Years	0 (0%)	2 (.93%)	10 (4.65%)	63 (29.30%)	109 (50.70%)	31 (14.42%)	4.73	.80	94.42%
	Jan, 2017 Second Years	0 (0%)	3 (1.55%)	10 (5.18%)	66 (34.20%)	83 (43.01%)	31 (16.06%)	4.67	.86	93.26%
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
I think my BET scores help me to know my weak points in English.	April/May, 2015 First Years	1 (.58%)	2 (1.16%)	8 (4.62%)	58 (33.53%)	77 (44.51%)	27 (15.61%)	4.67	.88	93.64%
	Jan/Feb, 2016 First Years	1 (.49%)	1 (.49%)	13 (6.31%)	61 (29.61%)	97 (47.09%)	33 (16.02%)	4.70	.86	92.72%
	Jan, 2017 First Years	0 (0%)	3 (1.40%)	13 (6.07%)	72 (33.64%)	93 (43.46%)	33 (15.42%)	4.65	.86	92.52%
	Jan, 2017	1	2	15	66	75	34	4.63	.93	90.67%

	Second Years	(.52%)	(1.04%)	(7.77%)	(34.20%)	(38.86%)	(17.62%)		
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
I think my BET scores help me to know my strong points in English	April/May, 2015 First Years	1 (.58%)	2 (1.16%)	7 (4.07%)	69 (40.12%)	71 (41.28%)	22 (12.79%)	4.59	.85	94.19%
	Jan/Feb, 2016 First Years	1 (.49%)	0 (0%)	12 (5.83%)	62 (30.10%)	102 (49.51%)	29 (14.08%)	4.70	.82	93.69%
	Jan, 2017 First Years	0 (0%)	2 (.93%)	17 (7.91%)	69 (32.09%)	97 (45.12%)	30 (13.95%)	4.63	.85	91.16%
	Jan, 2017 Second Years	0 (0%)	3 (1.55%)	12 (6.22%)	64 (33.16%)	81 (41.9%)	33 (17.10%)	4.67	.89	92.23%
Survey Statement	**Administration Dates Student Year**	**Strongly Disagree**	**Disagree**	**Somewhat Disagree**	**Somewhat Agree**	**Agree**	**Strongly Agree**	**Average**	**Std. Deviation**	**Proportion Agreeing**
My BET scores are useful to help me to plan my English study	April/May, 2015 First Years	1 (.58%)	2 (1.16%)	12 (6.98%)	65 (37.79%)	70 40.70%	22 (12.79%)	4.55	.89	91.28%
	Jan/Feb, 2016 First Years	1 (.49%)	2 (.98%)	12 (5.85%)	82 (40.00%)	82 (40.00%)	26 (12.68%)	4.56	.86	92.68%
	Jan, 2017 First Years	0 (0%)	3 (1.4%)	14 (6.51%)	82 (38.14%)	91 (42.33%)	25 (11.63%)	4.56	.83	92.09%
	Jan, 2017 Second Years	1 (.52%)	3 (1.55%)	13 (6.74%)	75 (38.86%)	74 (38.34%)	27 (13.99%)	4.55	.91	91.19%

.

C. Student Survey Analyses for Utilization Inference Sub-Assumption 3.3 Backing

Analysis of student attitudes to the BETs survey statements for utilization inference backing

Survey Statement	Administration Dates Student Year	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std. Deviation	Proportion Agreeing
I want to improve my BET score on my next BET.	April/May, 2015 First Years	1 (.58%)	1 (.58%)	5 (2.89%)	41 (23.70%)	69 (39.88%)	56 (32.37%)	4.99	.91	95.95%
	Jan/Feb, 2016 First Years	0 (0%)	2 (.97%)	4 (1.94%)	51 (24.76%)	75 (36.41%)	74 (35.92%)	5.04	.88	97.09%
	Jan, 2017 First Years	0 (0%)	0 (0%)	6 (2.79%)	52 (24.19%)	91 (42.33%)	66 (30.70%)	5.01	.81	97.21%

Survey Statement	Administration Dates Student Year	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree	Average	Std. Deviation	Proportion Agreeing
I'm motivated to study harder to improve my BET score.	April/May, 2015 First Years	1 (.58%)	1 (.58%)	8 (4.62%)	57 (32.95%)	67 (38.73%)	39 (22.54%)	4.76	.91	94.22%
	Jan/Feb, 2016 First Years	0 (0%)	3 (1.46%)	5 (2.43%)	80 (38.83%)	76 (36.89%)	42 (20.39%)	4.72	.86	96.12%
	Jan, 2017 First Years	0 (0%)	2 (.93%)	9 (4.19%)	79 (36.74%)	83 (38.60%)	42 (19.53%)	4.72	.86	94.88%

Appendix Y
Phi(lambda) Coefficients for Alternate BET Course Placement Cut-Scores

BET	Alternate Cut-score proportion of test score (%)	Alternate Cut score	Alternate Phi(lambda)	Proportion of Students in A1-A2 Stream (%)	Proportion of Students in A2-B1 Stream (%)
BET1 2015	52.94	36	.91	14.35	85.65
	58.82	40	.85	24.78	75.22
	63.24	43	.80	34.35	65.65
	72.06	49	.81	61.74	38.26
	75.00	51	.85	71.74	28.26
	79.41	54	.90	83.48	16.52
BET1 2016	47.67	41	.90	16.52	83.48
	51.16	44	.86	21.30	78.70
	55.81	48	.80	34.35	65.65
	63.95	55	.80	62.17	37.83
	67.44	58	.85	74.78	25.22
	72.09	62	.90	84.35	15.65
BET2 2016	58.14	50	.90	11.26	88.74
	61.63	53	.85	19.37	80.63
	65.11	56	.80	28.83	71.17
	74.42	64	.80	60.81	39.19
	77.97	67	.85	76.58	23.42
	81.40	70	.90	87.84	12.16