VALIDATING A SET OF JAPANESE EFL PROFICIENCY TESTS: DEMONSTRATING LOCALLY DESIGNED TESTS MEET INTERNATIONAL STANDARDS

Jamie Dunlea

VALIDATING A SET OF JAPANESE EFL PROFICIENCY TESTS: DEMONSTRATING LOCALLY DESIGNED TESTS MEET INTERNATIONAL STANDARDS

JAMIE DUNLEA

A thesis submitted to the University of Bedfordshire in fulfillment of the requirements for the degree of Doctor of Philosophy

University of Bedfordshire
Centre for Research in English Language Learning and Assessment (CRELLA)

December 2015

VALIDATING A SET OF JAPANESE EFL PROFICIENCY TESTS:
DEMONSTRATING LOCALLY DESIGNED TESTS MEET INTERNATIONAL STANDARDS

JAMIE DUNLEA

ABSTRACT

This study applied the latest developments in language testing validation theory to derive a core body of evidence that can contribute to the validation of a large-scale, high-stakes English as a Foreign Language (EFL) testing program in Japan. The testing program consists of a set of seven level-specific tests targeting different levels of proficiency. This core aspect of the program was selected as the main focus of this study. The socio-cognitive model of language test development and validation provided a coherent framework for the collection, analysis and interpretation of evidence. Three research questions targeted core elements of a validity argument identified in the literature on the socio-cognitive model. RQ 1 investigated the criterial contextual and cognitive features of tasks at different levels of proficiency, Expert judgment and automated analysis tools were used to analyze a large bank of items administered in operational tests across multiple years. RQ 2 addressed empirical item difficulty across the seven levels of proficiency. An innovative approach to vertical scaling was used to place previously administered items from all levels onto a single Rasch-based difficulty scale. RQ 3 used multiple standard-setting methods to investigate whether the seven levels could be meaningfully related to an external proficiency framework. In addition, the study identified three subsidiary goals: firstly, to evaluate the efficacy of applying international standards of best practice to a local context: secondly, to critically evaluate the model of validation; and thirdly, to generate insights directly applicable to operational quality assurance. The study provides evidence across all three research questions to support the claim that the seven levels in the program are distinct. At the same time, the results provide insights into how to strengthen explicit task specification to improve consistency across levels. This study is the largest application of the socio-cognitive model in terms of the amount of operational data analyzed, and thus makes a significant contribution to the ongoing study of validity theory in the context of language testing. While the study demonstrates the efficacy of the socio-cognitive model selected to drive the research design, it also provides recommendations for further refining the model, with implications for the theory and practice of language testing validation,

Acknowledgements

My deepest gratitude and respect goes to all of my former colleagues who work on the EIKEN tests. Their dedication to all of the details, large and small, is at the heart of the program, and ultimately is what made this study possible. Much of their hard work often goes unheralded behind the senses. I hope this study has contributed to shining a light on the professionalism and high standards they possess.

To my supervisors, Professor Cyril Weir, Dr Fumiyo Nakatsuhara, and Dr Chihiro Inoue, I will always be grateful for the opportunity they gave me to start this journey, their professional and academic advice on the way, and their friendship, without any of which the journey may never have been completed. I owe a special debt of appreciation to Professor Barry O'Sullivan, whose support and encouragement has helped me keep sight of the goal and stay the course.

And to my wife and my daughter, who have travelled with me, thank you. Your patience has been tested the most, your perseverance has been the longest, but your faith also the strongest, and your support the most important.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Aims of the Thesis

The primary goal of this study is to apply the latest developments in language testing validation theory to derive a body of evidence which can contribute to the validation of a large-scale, high-stakes English as a Foreign Language (EFL) testing program in Japan. A theoretical model of validation has informed the creation of three research questions, introduced in Section 1.3, which are designed to generate sufficient evidence for the most high-priority, core aspects of the testing program in the study. In addition to this primary goal there are three subsidiary goals which are integrally related to the research design and data collection procedures, and these are discussed below.

The primary goal of this study requires the application of theories currently accepted as best practice at the international level to a preexisting local testing program with a body of assumptions both explicit and implicit underpinning an already accepted range of local uses and interpretations. The focus of this study, the EIKEN testing program, was founded in 1963 (Eiken, n.d.-a), and has primarily developed through close interaction with the needs of the educational and societal context in which it is intimately embedded and used. Internationally, the theory of validation in general educational measurement, and the application of that theory to language testing in particular, has seen a great deal of development and consensus built around theoretical models over the last three decades. These developments have provided a core set of concepts and terminology with a degree of common usage and agreement, in addition to an array of quantitative and qualitative research techniques useful for building a body of evidence and a coherent argument to justify the uses and interpretations of a test. The validity argument to which this study aims to contribute, then, must be accessible, plausible, and convincing for educators, language testing experts, and primary users both within the local context of use and at the level of the wider language testing research community internationally. The first subsidiary goal is thus the evaluation of the effectiveness of applying what might be described as international standards of best practice in language testing validation research to

an assessment designed for and used in a particular local context.

The model of validation which underpins the data collection and evaluative framework for this study is the socio-cognitive model of language test development and validation proposed by O'Sullivan and Weir (2011) and Weir (2005a) (for a full description of the model see the literature review in Chapter 2). O'Sullivan and Weir (2011, p. 27) suggest that the model was informed by practical experience in test development and "allows the test developer to define focal language objectives and to collect evidence in a more comprehensive and satisfactory way than earlier models." The second important subsidiary goal of the study will be to evaluate the usefulness of these claims regarding the model itself and its contribution to meeting the primary goal of deriving evidence to contribute to a comprehensive validation argument for the EIKEN testing program.

This study was designed around a large-scale testing program and thus was cognizant from the outset of the importance of marrying the theoretical to the operational. To justify the resources required to undertake the comprehensive and large-scale data collection and analysis entailed by the research questions in this study, the results needed to be applicable and useable beyond providing evidence for a static validation argument for a testing program at one point in time. In practice, this meant that the data collection and analysis procedures used to address the core research questions needed to be ultimately transferable to and useful for the operational procedures utilized in ongoing test development and operational quality control. At all levels, across all three research questions, the key data collection and analysis procedures were evaluated, selected, and adapted from the perspective of the third subsidiary goal: to derive a comprehensive operational framework of language test validation based on a clear theoretical model which is amendable to being integrated into the ongoing operational test development process.

Validation requires an evaluation of both the technical properties of the test in conjunction with a comprehensive understanding of the context of use. The context in which the EIKEN testing program has developed and is used is described in section 1.2.

1.2 Rationale for a Validation Study of the EIKEN Testing Program

1.2.1 Overview of the EIKEN Testing program

The EIKEN testing program consists of a set of seven, stand-alone tests, each targeting a different level of EFL proficiency. The seven levels are referred to as *grades*, with Grade 1 being the most advanced level and Grade 5 the most elementary level (the seven levels also include two bridging levels, Grade Pre-1 and Grade Pre-2.) The tests are shown in Table 1.1. The EIKEN website (Eiken, n.d.-b) states that, "The grades are designed to provide well-defined steps that can act as both motivational goals and concrete measures of English ability as learners move through the spectrum of commonly recognized ability levels." Each grade, then, in this system is posited by the test developer to target a clear and definable level of proficiency, with each grade building on the ones below, and targeting aspects of performance considered important for progression to higher levels within the system. The levels-based grade system is thus conceived as representing a number of steps through a coherent and consistent framework of EFL proficiency.

The seven grades which comprise the system are shown in Table 1.1 (from Eiken, n.d.-b). In addition to the name of each grade, the table gives an indication of level in terms of what the test developer, the Eiken Foundation, considers to be a relevant level of the Common European Framework of Reference for Languages (CEFR), along with local uses, including recommended uses as benchmark levels by the Ministry of Sports, Culture, Science and Technology (MEXT). The table also includes an overview of the number of test takers for each grade in the academic year from April 2008 to March 2009 (Eiken, n.d.-c), which gives some indication of the large scale of the program.

Table 3.1 Overview of the EIKEN levels

EIKEN Grade	CEFR Comparison	Example of recognition/use	Examinees in 2008
1	C1	International admissions to graduate and undergraduate programs; MEXT benchmark for English instructors (Pre-1)	22,055
Pre-1	B2		71,533
2	B1	MEXT benchmarks for high school graduates	312,034
Pre-2	A2		503,638
3	A1	MEXT benchmark for junior high school graduates	661,798
4			464,819
5			306,745

1.2.2 The structure of the EIKEN tests

The EIKEN tests are administered in two stages, as shown in Figure 1.1. The First Stage tests consist of the grammar and vocabulary, reading comprehension, writing, and listening comprehension components. Examinees who pass the First Stage tests go on to take the Second Stage speaking test interviews (Eiken, n.d.-a). Only examinees who pass both stages are certified as having demonstrated a standard of performance consistent with that demanded by a specific grade.

FIRST STAGE					SECOND STAGE
Grade	**Vocab & Grammar**	**Reading**	**Writing**	**Listening**	**Speaking**
1	Vocab	Yes	Direct	Yes	10 minute interview
Pre-1	Vocab	Yes	Direct	Yes	8 minute interview
2	Yes	Yes	Indirect	Yes	7 minute interview
Pre-2	Yes	Yes	Indirect	Yes	6 minute interview
3	Yes	Yes	Indirect	Yes	5 minute interview
4	Yes	Yes	Indirect	Yes	No
5	Yes	Yes	Indirect	Yes	No

Figure 1.1 Overview of the First and Second Stages

The tests thus have a strong criterion referenced element to their design focus, with each test targeting a specific level of proficiency, and the pass/fail decision being premised on the assumption that examinees have demonstrated sufficient proficiency for certification at a specific grade (Eiken, n.d.-c). For the First Stage tests, cutscores for pass/fail decisions are set at 70 percent for both Grades 1 and Pre-1, and 60 percent for the remaining grades, while the pass mark for the Second Stage speaking tests is set at 60 percent for all grades (Eiken, n.d.-d).

As Figure 1.1 shows, the particular demands of each level have impacted on the design and administration of the tests. This is most apparent in the approach to the testing of productive skills, and is an important factor in the decision to employ a two-stage testing process. The two-stage testing process was the result of practical considerations arising from the desire to maintain a direct test of speaking ability in a face-to-face interview format for the majority of grades, while taking into account the practical limitations of maintaining a large-enough pool of trained examiners to deal with the required number of examinees (M. Fouts, Eiken Research Section, personal communication, 19 July

2015). The same approach to balancing the sometimes competing aspects of practicality, validity, and positive impact were involved in the decision to include different approaches to the testing of writing across the grades.

A detailed breakdown of the structure of the First Stage tests for each grade is included in Appendix A. The format of the test structure is standardized for each grade. Test forms from the previous three live administrations in public test centres are made freely available on the EIKEN website, including the listening files for the listening test components[1]. The standardized test formats shown in Appendix A have been in place since the most recent revision process for each grade, with the date of the introduction of the current format for each grade shown in Table 1.2.

Table 1.2 Date of introduction of the current format for EIKEN grades

Grade 1 and Pre-1	Grade 2 & Pre-2	Grades 3, 4, 5
2004	2003	2002

1.2.3 Changes in the context of use for the EIKEN testing program

1.2.3.1 Historical context

Sasaki's (2008) 150-year overview of the history of English language assessment in Japan provides a useful backdrop against which to chart the development of the EIKEN testing program. While Sasaki's overview explicitly focuses on one particular aspect of that history, namely the interaction between English assessment and formal educational contexts, this perspective is particularly relevant for EIKEN given the close association of the program with Japanese government policies and goals. Sasaki (2008, p. 65) divides the period from the second half of the 19th century through to the early part of the 21st century into the following four periods based on differences in the "intended goals and the degree of popularization of school-based English education in Japan."

1. Period 1 (1860-1945): English for the elite
2. Period 2 (1945-1970): English for everyone

[1] http://www.eiken.or.jp/eiken/en/downloads/

3. Period 3 (1970-1990): English for practical purposes in the era of rapid globalization
4. Period 4 (1990-2012): Introduction of innovative policies.
5.

1.2.3.2 Founding principles: design and development in Period 2

The tests were implemented with three of the current seven grades in 1963, which falls towards the end of Period 2. This period was marked by a rapid increase in the number of students studying English as the first nine years of schooling were made compulsory in post-war Japan. As is still the case today, English was not in fact a compulsory subject in the Japanese school system, but as Sasaki notes (2008, p. 67), it "virtually became a required subject" given its place in high school and university entrance exams, and remains for all intents and purposes a de-facto compulsory subject as the only foreign language taught in the majority of junior and senior high schools.

Two features marking this period influenced the design and implementation of the EIKEN tests. First, as Sasaki (p. 67) notes, "postwar Japan suffered from a serious shortage of both school buildings and teachers," and this shortage was magnified in the case of English, with teachers from other subject areas drafted in to cover the shortfall. Second, despite the stated goals of the ministry of education, high school and university teachers who produced the influential entrance exams for their institutions tended to maintain a focus on traditional grammar-translation approaches, reflecting their own training and experience (Sasaki, 2008, pp. 67-68).

Two significant features of the EIKEN program which continue today can partly be seen as originating in response to these pressures. Firstly, in response to the severe lack of EFL resources and trained teachers, the decision was made to make all of the test materials public following live administrations, with the intention of making high-quality English texts and testing materials available to teachers and learners for incorporation into both the classroom and self-study (Eiken, n.d.-c; Fouts, 2013). This practice has become ubiquitous in high-stakes entrance examinations in Japan and has become deeply ingrained as evidence of fairness and transparency by the Japanese public Yoshida (1996). As Yoshida

(1996) notes however, this presents very serious obstacles to the implementation of modern approaches to quality control in assessment, such as pretesting. The second significant aspect of the EIKEN program is the inclusion of face-to-face speaking tests from the outset of the EIKEN testing program, which can be considered partly a response to the entrenched influence of the grammar-translation approach noted by Sasaki. While the inclusion of a speaking component may not seem innovative by current standards in EFL teaching and assessment, it should be noted that the ministry of education continues to struggle to encourage the introduction of productive skills components into university entrance exams. Indeed, even with receptive skills, listening was only introduced into the influential National Centre Test for University Admissions in 2006, 26 years after that test's implementation (Watanabe, 2013). In this respect, the EIKEN program can be seen as innovative in the local context in having incorporated a focus on all four skills from its inception. Both of these responses are consistent with the explicitly stated intention of the EIKEN program to facilitate positive impact at both the classroom teaching level and at a wider societal level, something which has been a part of the EIKEN program's approach and philosophy from its inception (Eiken,, n.d.-c).

Finally, it is worth noting the close interaction between the test developers and the ministry of education which is evident at the program's introduction in Period 2. As noted on the Eiken website (Eiken, n.d.-e), the precursor of the present Eiken Foundation of Japan, the public-interest corporation which develops and administers the tests, was established explicitly with the purpose of "popularizing and improving practical English in Japan" in response to a government strategy initiative to promote the development of certificated proficiency exams. This relationship has continued to play an important role through the history of the program.

1.2.3.3 Rapid increase in acceptance and use in Period 3

At the wider societal level, Sasaki (2008) notes that this period was marked by the rapid economic development of Japan, and an accompanying rapid expansion of the number of Japanese travelling overseas and being exposed to the use of

English for real communicative purposes. The EIKEN testing program too, saw rapid uptake and acceptance during this time, with a further three grades (Grades, 5, 4, and Pre-1) added to the levels already in place. A significant development was the official recognition by the education ministry in 1968 of the EIKEN tests as authorized proficiency examinations for certification purposes, and by 1987 the number of examinees had grown to 2 million a year.

The rapid uptake of the exams can be seen as partly related to the official authorization. However, this does not wholly account for the wide acceptance of the tests, particularly within the formal educational contexts of junior and senior high school. As Sasaki (2008) notes, this period saw growing criticism of the entrance examination system as progressively larger numbers of students continued on from the compulsory schooling system to high school and higher education. Government responses included encouraging more diverse approaches to high school entrance procedures and introducing a precursor to the current National Centre Test for University Admissions in 1979. An important use of the EIKEN tests which continues today is the use of EIKEN grade certificates in high school entrance applications. Various high schools take different approaches, from taking certificates into consideration in the application process to waiving the need to take the English exam component of the school entrance exams for holders of the relevant grade certificate. This usage contributed to the program's growing popularity.

A further important aspect of the testing program which has contributed to its widespread use is the commitment to accessibility. The EIKEN program from its inception has established public test sites in all areas of Japan, including rural districts and has maintained a relatively low cost, particularly for the lower grades (Eiken, n.d.-f). What is perhaps equally important is the integration with the local educational community of English teachers. Given the public release of all test materials, the test developer has had a strong interest in ensuring the relevance and utility of the materials to the primary stakeholders, amongst which junior and senior high school teachers and learners feature prominently. The public release of the materials allowed for an ongoing process of open validation in that all interested parties had access to the materials used in live examinations.

At the same time, the Eiken Foundation “placed most of its validation resources into the content validation area, investing in an extensive network of expert committees and outside reviewers, made up of testing experts and practicing teachers, to constantly review and revise test materials during all stages of item and test development.” (Eiken, n.d.-c). This network was supplemented by a growing local network of educators who set up and supervised public test administration sites as the number of sites expanded rapidly to cope with the growing demand, including in rural and remote areas.

1.2.3.4 Responding to changing needs in Period 4

Period 4 saw the addition of Grade Pre-2, completing the seven-level testing program. The latter part of this period also saw a series of revisions to test structure and content, including the most recent revisions noted in Table 1.2, leading to the current standardized formats for the First Stage tests shown in Appendix A. As the demand for access to the tests rose, the Eiken Foundation instituted a system of group test sites to complement the public test sites at which the exams had been administered from their outset. Group test sites are institution-specific test sites catering solely to the students at a particular institution who register to take one of the EIKEN grades (Eiken, n.d.-f). Application and administration procedures are handled centrally through a teacher designated as the administrator for a site. This system greatly expanded the number of sites at which the First Stage tests were able to be administered. The commitment to accessibility of the tests is demonstrated by the number of test sites currently available shown in Table 1.3 (adapted from Eiken, n.d.-f). Tests are currently administered three times during the Japanese academic year which runs from April to March. Tests at public test sites are offered on Sundays only. Tests at group test sites are offered over a three-day period from Friday through to Sunday, which has also increased the number of test forms required for each grade.

Table 1.3 Overview of current test administration sites

	Public	Group (institution-specific)
Number of sites	400	18,000
Grades offered	All grades	Grades 2, Pre-2, 3, 4, 5
Stages offered	First and Second Stages	First Stage Only

Sasaki (2008) notes that Period 4 has been dominated by a series of policy initiatives by the Ministry of Education, Culture, Science and Technology (MEXT) aimed at the way English is taught and assessed in schools, and importantly at the way English is assessed for high-stakes, high school and university entrance purposes. Several of these initiatives have had direct effects on the development of the EIKEN testing program during this period**.**

In 2003, MEXT published the Action Plan to Cultivate Japanese with English Abilities. A significant development in the Action Plan (MEXT, 2003) was the use of EIKEN grades as recommended benchmark levels of proficiency for graduates in junior and senior high school, as well as for professional contexts including proof of English proficiency for teacher certification. These recommendations have been reiterated in later policy documents such as MEXT (2011), and are shown in Table 1.1. What is significant about this development is the transition from a largely implicit understanding of the relevance of the various grades to specific levels of attainment in the formal educational system to the explicit use of the EIKEN grades as external proficiency standards. The relationship prior to the publication of the Action Plan had been somewhat the reverse, with preparation materials for EIKEN grades including very general level descriptions pointing to the formal education level that the test was considered most relevant for, with, for example, Grade 3 described as being targeted at a level appropriate for graduates of junior high school. This implicit understanding of

what each grade represented had been developed through the close interaction between the test developer and stakeholders in the educational community and reinforced through the public's ongoing access to the test materials. As noted by Sasaki (2008), given the standardized nature of the Japanese education system, guided as it has been throughout the post-war period by the national Courses of Study curriculum plans produced by MEXT and implemented nationally, reference to these educational levels was in fact not as vague a level-description as it may first seem. The content of material to be studied at the school level was fixed and well understood by the relevant stakeholder groups. In Period 4, however, the rapidly changing national and international educational contexts had led to rising dissatisfaction by various elements of society with the results of the English education and assessment systems. This dissatisfaction was manifested in the policy initiatives of MEXT to set explicit proficiency standards which could be used as both teaching and learning goals and as a means of evaluating the degree to which the educational reforms had contributed to meeting those goals and improving outcomes for the English education system. The move towards using the EIKEN grades as these external benchmarks of proficiency by MEXT in turn highlighted the need to develop more detailed explicit test specifications, as reliance on the implicit understanding of content specified by the Courses of Study was no longer sufficient for the growing expectations of the prime stakeholders for a clearer expression of the proficiency standards represented by each grade.

. The Action Plan (MEXT, 2003) also included recommendations for universities to consider using external proficiency exams such as EIKEN and also international exam systems as standardized measures of English proficiency in place of the English sections of entrance exams which were generally produced in-house by each institution. As Sasaki (2008) notes, the entrance exams have been subject to a great deal of controversy. While some authors (e.g. Mulvey, 2001; Watanabe, 1996, 2003) have pointed out that the relationship between the exams and the impact on teaching and learning is not necessarily as clear-cut as is sometimes assumed, the variable quality, lack of clear specifications, lack of standardization in terms of content, and inappropriate level of difficulty in relation

to the targeted population have been consistently highlighted as problematic (e.g. Brown and Yamashita, 1995; Kikuchi, 2006). The recommended use of the EIKEN exams by MEXT, along with other internationally recognized large scale proficiency exams, as a means of improving the high-stakes university entrance exams through the use of standardized proficiency tests presented new opportunities for the testing program, but also once again increased the demands and expectations of stakeholders in terms of specification and validation evidence. One response to these demands was the EIKEN Can-do Project (Eiken, n.d.-c, Dunlea, 2010), a large-scale survey of 20,000 test takers to develop a set of comprehensive level descriptors describing what test takers at each grade level were able to accomplish in real-world language use situations.

The changes in the context of use for the EIKEN testing program described above have been related to the needs and expectations of stakeholders within the largely local Japanese context of use for which the tests were originally designed and developed. The 2003 MEXT Action Plan, however, precipitated a further—and in some ways more dramatic—change to the range of uses and interpretations and typical stakeholders for which the testing program had been designed, developed and used. The Action Plan (MEXT, 2003) included a number of initiatives to specifically increase the number of Japanese high school and university students studying abroad. In response to this MEXT goal, the EIKEN testing program began investigating the possibility of receiving recognition for EIKEN certificates as proof of English proficiency for admissions purposes by foreign universities (Fouts, 2013). Fouts (2013) suggests that a significant hurdle faced by many students hoping to study abroad, particularly in rural areas of Japan, was the high cost and general lack of accessibility to the international proficiency exams most widely used for entrance purposes in English-medium universities at the time. The cost of taking the upper grades of the EIKEN exam was less than half of the cost of these exams, which were also only available in major population centres. The Eiken Foundation thus hoped to provide students who already possessed an EIKEN grade certificate with the chance to use those certificates for entrance purposes, thus negating the need to undertake costly extra exams for that purpose (M. Fouts, personal communication, 13 July, 2015).

The EIKEN testing program was now presented with possible uses and interpretations which were not explicitly part of the original design of the testing program. In addition, these new potential uses presented a range of new stakeholders who would not be familiar with the common reference points such as the MEXT Courses of Study curriculum guidelines or expected levels of attainment in the formal educational system, which had in the past facilitated communication and common interpretation between local educators and the test developers. The test developers now needed to be able to communicate clearly the design, structure, and content of the tests to international educators and assessment experts in a way that would be familiar to those stakeholders, and present validation evidence collected according to the principles and procedures currently accepted as best practice in the international context. These new demands on the testing program led to a number of initiatives, for example predictive and criterion-referenced validity studies to investigate the appropriacy of using the tests for these purposes (Brown et al., 2012, Hill, 2010).

By the end of Sasaki's (2008) Period 4, then, the EIKEN grades had thus become firmly embedded in the educational and societal context of Japan, as well as developing new uses beyond that local context and involving stakeholders outside the groups of local educators and learners with whom the testing program had previously developed a close interaction. These various uses are described in more detail on the Eiken website (Eiken, n.d.-g), which notes that the number of universities, high schools, and junior high schools recognizing EIKEN for admissions and credit now exceeds 2500 in Japan, and that over 350 English-medium universities outside Japan also recognize one or more of the EIKEN grades for admissions purposes.

1.3 Research Questions

As the overview of the EIKEN testing program in section 1.2 illustrates, it is a large-scale program with two stages of testing covering all four skills targeting seven distinct proficiency levels, with a range of stakeholders, uses and interpretations, and contextual features and constraints which can differ according to the level. In addition, the social context in which the test is embedded has seen

significant changes in recent years, and the testing program itself has seen changes in uses which have extended the range of stakeholders with which the program now needs to communicate effectively.

All three of the primary research questions focus on a key aspect which has defined the program's approach to test design and the key claims it makes to its core stakeholders: the claim to target differing levels of proficiency with level-specific tests that present a structured progression through a definable proficiency framework, acting as both measures of that proficiency and explicit goals for teachers and learners charting a course through those proficiency levels. Focusing on this core aspect, the research questions have been designed through reference to the model of validation selected for use in the study. The levels-based system at the core of the program was further framed in reference to the core aspects of validation identified by the socio-cognitive model, leading to the following three research questions:

> **RQ1:** To what extent and in what ways are the reading tests in the seven levels of the EIKEN testing program qualitatively different in terms of key contextual and cognitive parameters?
> **RQ2**: To what extent and in what ways are the reading tests in the seven levels of the EIKEN testing program quantitatively different in terms of test difficulty?
> **RQ3:** To what extent and in what ways is there a relationship between the reading tests in the seven levels of the EIKEN testing framework and the levels described in an external criterion of EFL proficiency?

The decision to focus on the reading component of the First Stage tests is a pragmatic one taken to balance the demands of generating evidence for the core aspects of validation identified by the socio-cognitive model across the seven levels of the testing program. It would be beyond the scope of this study to attempt to develop a comprehensive validity argument covering all aspects of such an extensive, large-scale program, particularly in relation to adequately covering the description of content for all four skills across all levels while still managing

to adequately focus on the other core aspects of validation called for in this study. This study, as noted in section 1.1, focuses on collecting a robust body of evidence, targeting core features of the program which can contribute to the construction of such a coherent, comprehensive and plausible argument to justify the uses and interpretations of the program.

This issue is discussed further in Section 2.2.4, in the explanation of the socio-cognitive model which underpins the data collection design and analysis, and in Section 6.3, in the discussion of the limitations associated with the study. Explicit choices were made in the identification of these research questions. In addition to the decision to focus on one skill, in the interests of pragmatism, the three research questions provide evidence directly related to core aspects of the test system highlighted by the socio-cognitive model, but do not attempt to address *all* aspects of validity identified by the model. The aspects addressed by these three research questions focus on establishing the outlines of a validity argument for the program by investigating key aspects of how the tests in the program function as measurement instruments within a coherent test system. The research questions identify the criterial features which define those instruments, the empirical distinctiveness of the instruments within the system, and the degree to which those features have relevance and meaning beyond the internal system itself, and importantly beyond the local context in which the tests were designed.

Section 1.2.3 contextualized the program within the history of the wider developments in Japanese education and society, highlighting the widespread use of the tests within that society, and the implicit understanding of the proficiency levels of the program built through an ongoing, consensus-building interaction with the educational community and key stakeholders. This study takes as its starting point the importance of investigating those implicit assumptions regarding the working of the test system itself, and has targeted the research questions with the aim of creating an explicit body of evidence on which to evaluate whether those assumptions hold up in practice. Establishing whether the tests first work in relation to these core respects is a crucial first step, and perhaps more importantly, establishing whether the tests work in these respects not just *individually* at each

grade, but as a coherent system across the grades that constitute the program. If these key assumptions of the testing program did not hold up to scrutiny, then it would not justify the investment of more resources to pursue the other aspects of validation required to flesh out in detail all aspects of a comprehensive validity argument which the socio-cognitive model requires.

Equally, however, as highlighted in Section 6.3, this study is only a first step, and once the outlines of that argument are established, further work will be required to add depth to the argument by targeting the other aspects of validity identified by the model, as well as adding depth by looking at other skills and within individual grades in more detail. In particular, the aspect of consequences and positive impact has been an important part of the design and development of the EIKEN tests from their inception, as described in Section 1.2.3. The goal of this study to address the core aspects of the testing program as a test system was taken as a crucial first step to creating an explicit, evidential basis for evaluating the implicit assumptions of the program, and to form the foundation on which a comprehensive validity argument can be built. It thus should not be taken as an indication that addressing the issue of consequences is not equally relevant. The socio-cognitive model, as described in Section 2.2.4 includes this aspect, and thus provides a quality control mechanism for ensuring that a claim to a comprehensive, coherent, and plausible validity argument justifying the uses and interpretations of the testing program will not be complete without all aspects of the model being addressed. As already noted, this study, for pragmatic reasons including the scale of the data collection and analysis involved has prioritized the investigation of the core aspects of the program identified by the three research questions as a crucial first step.

1.4 Methodology

This section provides a brief overview only of the methodology employed in the investigation of each research question. Each question covers an aspect of the core evidence required for a comprehensive validation study. These aspects contribute to a unified, evaluative validation argument, but each is nonetheless distinct in its approach to quantitative and qualitative data collection and analysis, and the

methodology is thus dealt with in more detail under the relevant chapter dealing with each research question.

The first question draws on the socio-cognitive model to identify contextual and cognitive features of the test tasks which are relevant for describing the stages of EFL proficiency at the core of the testing program. This analysis thus extends beyond the traditional analysis of content validity to include the cognitive dimension of the tests and the cognitive processing demands made on test takers engaging in these test tasks. The methodology for this will involve identifying the features which can be applied to tagging test tasks through both expert human judgment and automated analysis tools. As already noted, there are important subsidiary goals for this study, and one involves the ability to apply the data collection procedures to ongoing test development and quality assurance procedures. The features identified for tagging thus need to be amenable to application on a large scale in a coherent and consistent way within the ongoing operational procedures of the testing program. To facilitate this, in addition to selecting features for tagging test tasks in accordance with the key principles of the socio-cognitive model, four further important criteria were used for the selection of parameters for use in this study: a) relevance b) transparency c) interpretability d) comparability.

Relevance relates to the relevance in establishing the validity of the uses and interpretations of the EIKEN tests for the purposes for which the tests have been developed. In this respect, relevance means helping to define the criterial features of the reading test items and tasks to better enable a principled comparison of these features within and across grades. The latter aspect is particularly important as the taxonomies created for classification needed to be able to capture both the criterial features *within* each grade, but equally importantly be useful for identifying distinctions across all seven grades in the program. Maintaining relevance in this context meant finding a balance between comprehensiveness and practicality.

Transparency refers to the existence of documentation to clearly define what a particular parameter measures and how it is to be applied in practice. This is particularly important for using the findings of this validation study to inform

item writing and test specifications. Such transparency is of course crucial for ensuring consistency in expert human judgment, but even in cases where automated software can return complex measures that human judgment would not be capable of, it is essential that the factors impacting on those measures are transparent and clear. If this is not the case, then there would be no facility for item writers, for example, to act upon the finding that a particular text had higher or lower measures than another text. Similarly there would be no opportunity for teachers and learners to use information derived from such measures to identify relevant learning goals or manipulate learning materials to focus on the factors which might impact on the calculation of a particular measure.

Interpretability, while closely related to transparency, refers specifically to the accessibility of the definition of a particular measure given that explicit documentation of its definition exists. The substantive meaning of parameters for item writers, teachers and learners must be clear in terms of how different aspects of an item or input text can be manipulated to either increase or decrease the resulting measure as appropriate. Interpretability is also thus closely related to relevance in that it must be clear and easily communicated to relevant stakeholders as to why a particular parameter which will be used to define test specifications is relevant to the construct of language proficiency which the test claims to measure.

Finally, comparability refers to the degree to which a parameter has been applied in practice both in terms of the analysis of tasks and texts relevant to the target language use (TLU) domain for each grade of the EIKEN program, but also to learning and teaching materials and other proficiency tests posited to be relevant to similar levels of proficiency for similar purposes. This criterion thus relates to how much research is publicly available on the use of a particular parameter. This is in order to allow for the establishment of principled benchmark levels of measures against which results for the same parameter, when applied to particular items or texts from particular grades of the EIKEN program, can be compared.

The second research question addresses the contribution of the technical measurement properties of the test through the aspect of scoring validity. This

question will be addressed through the application of vertical scaling methodology. Vertical scaling provides the quantitative tools to empirically verify the claim to differential levels of proficiency. Through vertical scaling it will be possible to place test items which are administered at different levels of the EIKEN program onto a common ("vertical") scale of item difficulty.

The third question addresses the need to determine criterion referenced meaning for the locally developed and integrated benchmarks of performance around which the EIKEN testing program has been used through the investigation of the relevance of the EIKEN program to an outside, international benchmark of proficiency. The Common European Framework of Reference (CEFR) produced by the Council of Europe (2001) has been selected as the basis for this investigation. The background to the widespread adoption of the CEFR in a range of international contexts is dealt with in more detail in Chapter 2, including the caveats, criticisms, and limitations associated with the framework. While taking note of these caveats, the widespread usage of the CEFR makes it useful for this study in relation to the four criteria noted above. The approach to linking the EIKEN tests to the CEFR involves a comprehensive application of multiple standard-setting approaches.

1.5 Overview of the Thesis

Chapter two provides a review of the literature relevant to the study, including developments in validation theory to provide a rationale for the adoption of the socio-cognitive model as a basis for the study, and background information on the methodology for the three research questions. Chapters 3, 4, and 5 each focus on one of the three research questions of the study. Each chapter provides an overview of the methodology, data collection, analysis, and results specific for that research question.

Chapter 6 will present the limitations, conclusions, and implications for further research which arise from the study. This concluding chapter will synthesize the results across the three research questions, providing an evaluation of the degree to which the three research questions have contributed to the primary goal of generating a core body of evidence towards the construction of a validation argument for the EIKEN testing program. Chapter 6 will also assess

the results of the study in reference to the subsidiary goals identified in section 1.1.

Chapter 2 Literature Review

2.1 Introduction

The literature review will focus on providing an overview of developments in validation theory in order to provide a rationale for the adoption of the socio-cognitive model, which underpins the data collection and analysis for the study. Section 2.2 thus covers developments in validation theory up to and including the socio-cognitive model. The literature review will then provide an overview of the background literature relevant for the methodology of each of the three research questions. Section 2.3 addresses the literature relevant to answering RQ1, Section 2.4 the literature for RQ2, and Section 2.5 the literature for RQ3.

2.2. Validation Theory and Practice: Selecting a Model for the Study

2.2.1 The acceptance of validity as a unified concept

This review takes the publication of Samuel Messick's seminal 1989 paper as a focal point. Messick's paper articulated a growing consensus around three core concepts which have come to underpin the approach to validation in the three decades since its publication (Bachman, 2005; Fulcher & Davidson, 2007; McNamara, 2006; O'Sullivan & Weir, 2011). While the key concepts put forward by Messick continue to receive a high degree of general consensus, measurement specialists have nonetheless continued to struggle to define a consistent and coherent approach to answering the more specific questions of *how much of what kind of evidence* will constitute sufficient justification for the uses and interpretations of particular assessments (Mislevy et al, 2003; O'Sullivan & Weir, 2011; O'Sullivan, 2011). The models which will be examined have developed as different approaches in the quest to answer these questions.

Messick's 1989 paper strongly advocated a unified conceptualization of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). This succinct description encapsulates two of the three important concepts which the paper helped crystalize into core principles underpinning modern approaches to validation. Firstly, it provided the final push to overturn the categorization of validity evidence into "three separate and substitutable types—namely content, criterion and construct validities" (Messick, 1995, p. 741). This tripartite categorization was replaced with construct validity as a superordinate, unifying concept to which the collection and evaluation of multiple forms of evidence, including content-related and criterion-related evidence, would contribute (Fulcher & Davidson, 2007; Messick, 1989, 1995, 1996; Shepard, 1993). Secondly, the paper shifted the focus of validation away from a static focus on the test as the object of validation, to focus instead on justifying the interpretations and uses made of test scores (Fulcher & Davidson, 2007; Shepard, 1993; Weir et al (2013). Validity was no longer to be seen as "a property of the test or assessment as such, but rather of the meaning of the test scores" (Messick, 1995, p. 741). An important subscript to this reconceptualization was the evaluative nature of the process of validation: it was not to be considered a straightforward yes-or-no decision but would require an evaluation of both empirical evidence and logical argument to determine the *degree* to which test uses and interpretations were justified.

The unified model proposed by Messick went beyond simply subsuming already existing approaches to collecting validity evidence beneath the umbrella of construct validity. The third core principle, and perhaps the most profound paradigm shift which Messick's work facilitated, was to explicitly position the consequences of test use as an important focus of enquiry within the same unified, integrated approach to the investigation of construct validity (McNamara, 2006; Shepard, 1993; Weir et al, 2013). As Weir et al (2013, p. 98) note, "Messick authorized the growing interest in test use as a critical feature, and opened the way to the ethical concerns that now dominate much language testing debate." Messick

presented the facets of a unified validity framework incorporating value implications and social consequences in what he referred to as a "progressive matrix" (Messick, 1986, 1989, 1995, 1996), shown in Figure 2.1. He suggested that "each cell represents construct validity, with different features highlighted on the basis of the justification and function of the testing. From another perspective, the entire progressive matrix represents construct validity, . . .validity and values are one imperative, not two, and test validation implicates both the science and the ethics of assessment. (Messick, 1995, p. 749).

	Test interpretation	**Test use**
Evidential	Construct validity (CV)	CV + Relevance / Utility (RU)
Consequential	CV + Value Implications (VI)	CV + RU + VI + Social Consequences

Figure 2.1 Facets of validity (from Messick, 1986, 1989, 1995).

2.2.2 Building on a growing consensus

While Messick's 1989 paper remains the touchstone for the current acceptance of the unified concept of validity, it is important to note that the paper was not proposing radically new ideas but was in fact building on a growing consensus around these core principles (Shepard, 1993). Prior to the 1990's, the categorization of validity evidence into three essentially stand-alone types had indeed achieved such an entrenched status as to lead some authors to refer it as a "holy trinity" (Shepard, 1993, p 409). Yet by the publication of the 1985 *Standards for Educational and Psychological Testing*, the clean distinctions of this orthodoxy had already been critically challenged in favor of a more unitary interpretation of validity evidence, with the Standards noting that "although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the score. . . . An ideal validation spans all three of the traditional categories" (AERA et al, p. 9, quoted in Fulcher and Davidson, 2007, p. 15).

The key concepts which would become part of the dominant paradigm of validity following Messick's 1989 paper were thus already prominent within the

field prior to its publication. Messick (1986, p. 4) nonetheless worried that the 1985 *Standards* offered a potential opt-out clause to test developers by suggesting that "whether one or more kinds of validity evidence are appropriate is a function of the particular question being asked and of the context and extent of previous evidence." Messick (1986 p. 5) was concerned that this caveat provided a justification for the continued reliance on separate forms of validity evidence at the discretion of the investigator, which "when it occurs, is tantamount to reliance on one kind of validity as the whole of validity, regardless of how discredited such overgeneralization may have become and of how much lip service is paid to validity as a unitary concept."

While much of the discussion in the 1980s and 1990s focused on replacing the "traditional" distinctions with the newly emerging consensus around a unitary conception of validity, the tripartite distinction itself is better seen as a transitory phase in the development of validity theory. Although the concepts underlying criterion validity and content validity models had been developed in the first half of the 20th century (Kane, 2011, 2013; Shepard, 1993), the first explicit categorization of validity evidence to include *construct validity* was presented by the American Psychological Association in 1954 (Cronbach & Meehl, 1955; Fulcher & Davidson, 2007; Kane, 2011, 2013). At the time, however, the taxonomy was presented as a four-way distinction: *predictive validity, concurrent validity*, *content validity* and *construct validity*. In their detailed follow-up explanation of the newly-established concept of construct validity, Cronbach and Meehl (1955, pp. 281-282) suggested that predictive and concurrent approaches could be subsumed under the umbrella of criterion validity evidence, and this tripartite distinction was to be reflected in the 1966 and 1974 editions of the *Standards* (Davidson & Fulcher, 2007; Shepard, 1993). Mounting criticism of the tripartite categorization in the 1970s and 1980s highlighted concerns that the distinction encouraged an opportunistic approach to validation in which "the choice of model was often arbitrary, depending mainly on the availability of data" (Kane, 2011, p. 7).

Cronbach and Meehl's seminal 1955 paper in fact laid out many of the tenets which would eventually come to be associated with the predominant view

of validity now generally recognizable to researchers (Fulcher & Davidson, 2007; Kane, 2011, 2013; O'Sullivan & Weir, 2011). It clearly focused on validation as an evaluative summary of both empirical evidence and logical argumentation as the justification for the interpretations associated with a test, rather than on the test itself: "The proper goals in reporting construct validation are to make clear (a) what interpretation is proposed, (b) how adequately the writer believes this interpretation is substantiated, and (c) what evidence and reasoning lead him to this belief" (Cronbach & Meehl, 1955, p. 297). The paper also presaged a focus on construct validity as a unifying, superordinate concept, noting that "construct validation is important at times for every sort of psychological test: aptitude, achievement, interests, and so on" (Cronbach & Meehl, 1955, p. 283). Despite these important contributions, the paper did not propose a unifying framework to integrate the different forms of evidence under construct validation, and instead left intact the separation of different forms of evidence into separable *validities* (Kane, 2001a, 2011, 2013; Shepard, 1993). The foundations of construct validity as a central, organizing principle were, however, laid. Indeed, only three years after the introduction of construct validity in the *Technical Recommendations*, Loevinger (1957) would suggest that "'since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (cited in Kane, 2011, p. 7).

2.2.3 Problems with the Messickian legacy

McNamara (2006) credits Bachman with introducing and popularizing the principles of Messick's approach within the field of language testing with his 1990 work *Fundamental Considerations in Language Testing*, and language testing has been at the forefront of applications of further developments in validity theory in the decades since Messick's 1989 paper. As with the broader field of educational measurement, despite the clear points of consensus that permeate the writing on validity within language testing, three broad areas of concern can be identified that have driven much of the post-Messick discussion: the role of consequences, developing a model for operationalizing Messick's approach in the practice of validation, and the definition of construct.

2.2.3.1 Consequences

Messick was aware that the consequential basis for test justification in his progressive matrix was considered problematic by some, noting that "some measurement specialists argue that adding value implications and social consequences to the validity framework unduly burdens the concept" (1995, p. 748). He strongly rejected these arguments, suggesting that the four-fold matrix "makes explicit what has been latent all along, namely, that validity judgments are value judgments" (Messick, 1995, p. 748). Nonetheless, it is fair to say that the theoretical acceptance of consequences into a unified concept of validity has perhaps been observed more in the breach than through operationalization in the practice of validation. The lack of concrete applications of consequences in validity argumentation in language testing has been noted by Brown (2008). In the wider field of measurement, Cizek et al (2008, p. 411), after reviewing the sources of validity evidence reported in 283 test reviews, suggest that "in actual test validation practice, the operationalization of so-called consequential validity . . . is so great a burden that it is simply ignored." A follow-up study by Cizek et al (2010) found a similar picture. Based on the dearth of studies implementing consequences in validation practice, the authors (p. 741) endorsed the recommendations of Cizek et al (2008) to essentially de-couple consequences from validity.

Messick (1995, p. 746) in fact indicates some limits to the scope of consequences, noting that "the primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance." Within the post-1989 discussion of validity and consequences in the field of language testing, there has certainly been a strong emphasis on consequences, with a particular focus on the impact of tests on teaching and learning. Bailey (1999, p. 9) in a review of the literature concluded that "language testing washback (1) has often been discussed; (2) is widely held to exist; (3) that there are differing points of view about what the construct may encompass; and (4) that positive washback is viewed as an important criterion in

the development and evaluation of language tests." Messick (1996, p. 243) in fact suggests some limits to the inclusion of washback-related evidence in validation by noting that "technically speaking, evidence of teaching and learning effects should be interpreted as washback—either in general or in particular as contributing to the consequential aspect of construct validity—only if that evidence can be linked to the introduction and use of the test."

While Messick's work firmly positioned the concept of consequences within the current validity and validation paradigm, the difficulty of defining and measuring consequences, and further integrating that information into an evaluative summary of validity evidence, has presented problems to test developers and researchers.

2.2.3.2 Implications for a practical model of validation

As early as 1993, Shepard (p. 407) was already lamenting that despite the new consensus on validity as a unified concept championed by Messick, "examples of the gap between validity theory and measurement practice are numerous." Shepard postulated several reasons for this gap, including the lack of clear examples of how the principle of a unified approach to validity could be operationalized in practice. In addition, she suggests that the presentation of Messick's progressive matrix itself is a potential source of confusion for several reasons:

> (a) The faceted presentation allows the impression that values are distinct from a scientific evaluation of test score meaning. (b) By locating construct validity in the first cell and then reinvoking it in subsequent cells, it is not clear whether the term labels the whole or the part. … (c) The complexity of Messick's analysis does not help to identify which validity questions are essential to support a test use (1993, p. 427)

Similar concerns in relation to the lack of guidelines for the practice of validation were offered by Kane (1992, 2011, 2013) and Kane et al (1999) as reasons for the adoption of an argument-based approach to structuring the presentation of validity evidence. Chapelle et al (2008, 2010) echo these concerns

when discussing their approach to the development of a validity argument for the revised TOEFL, noting that neither Messick's 1989 paper nor the 1999 *Standards* (AERA et al), which had been revised to reflect the new consensus, offered a coherent structure to guide their validation project. They describe Messick's 1989 paper as "an extensive scholarly treatment" which does not offer the practical guidelines they required (Chapelle et al, 2010, p. 3). Bachman (2005) takes a similar approach, but further emphasizes the lack of a clear methodology for incorporating consequences and test use in the process of validation. Kane (2011, p. 7) suggests that the "unified model of construct validity was conceptually elegant, but not very practical." He suggests that it failed to offer guidelines for how to evaluate how much of what kind of evidence would be enough: "In the absence of strong theories, construct validity tends to be very open-ended, and it is not clear where to begin or how to gauge progress.' Shepard (1993, p. 429) also cites "the sense that the task is insurmountable" as an impediment to researchers fully engaging with the model, noting that "current standards do little to prioritize validity questions. . . . Therefore they do not help to answer the question 'How much evidence is enough.'"

It is, however, interesting to note that much of the criticism centers on the commonly-referenced four-fold progressive matrix and the more abstract aspects of the unified approach, notably the inclusion of values and consequences. Surprisingly, these critical examinations do not take up the discussion of the six aspects of validity which Messick (1989, 1994, 1995, 1996) suggested did indeed offer clear guidelines for generating concrete evidence to justify the uses and interpretations of a test.

Messick (1995, p. 744) highlighted six "distinguishable aspects of construct validity", noting that the distinctions do not detract from the unified nature of validity but rather "provide a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences." These six aspects are shown in Table 2.1 (Messick, 1995, 1996). Messick (1989, 1995, 1996) suggested that addressing these six aspects was crucial for validating score-based inferences and test use within a unified approach. Perhaps more

importantly, he also suggested that centering validation around these six aspects of validity evidence would in fact be sufficient for doing so, noting the six aspects were applicable "to all educational and psychological measurement," and provided "a way of addressing the multiple and inter-related validity questions that need to be answered in justifying score interpretation and use" (Messick, 1995, p. 746). These six aspects of evidence collection and appraisal would thus ensure that "the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases" (Messick, 1995, p. 747).

At the same time, Messick stressed that these distinctions did not allow a return to the selective use of one or more types of evidence as convenient to the researcher: "Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications …which is what is meant by validity as a unified concept" (Messick, 1995, p. 747). Importantly, he also emphasized the practical constraints which impact on the practice of validation, including the resources available for data collection, by noting that the inability to collect some forms of evidence does not automatically invalidate a test interpretation or use. However, what differentiates the approach to the old opportunistic nature of evidence collection was the need to touch all of the relevant bases by explicitly including a reference to all six aspects in an integrated validity argument, and if "the bases are not covered, an argument that such omissions are defensible must be provided" (Messick, 1995, p. 747).

The six aspects of construct validity proposed by Messick thus provide concrete lines of evidence collection, with more or less well established methodologies for their analysis and evaluation (with the newer consequential aspect remaining the least clearly specified in terms of established research and precedent). At the same time, they go some way to offering a way out of the seemingly endless nature of validation criticized by Shepard, offering in contrast a potential way of deciding when it is reasonable to stop by evaluating that all of the bases have indeed been sufficiently covered, and the evidence integrated into a plausible and defensible rationale for the proposed uses and interpretations of the assessment being validated. Nonetheless, these six aspects were not been taken up as the explicit focal point of the subsequent argument-based approaches to validity

theory and validation practice prominent in the 1990s and 2000s.

Table 2.1 Six aspects of construct validity (from Messick 1995, 1996)

Aspect	Description
Content aspect	The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989b);
Substantive aspect	The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks
Structural	The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957; Messick 1989b);
Generalizability aspect	The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test criterion relationships (Hunter, Schmidt, & Jackson, 1982);
External aspect	The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965);
Consequential aspect	The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989b

2.2.3.3 Defining the language proficiency construct

The importance of defining the construct of interest for a test has become a well-established part of the general tenets of the unified approach to validity. The understanding in the field of what that means in practice, however, has changed considerably from the early presentations of the concept of construct validity. Cronbach and Meehl (1955, p. 283) defined a construct as "some postulated attribute of people, assumed to be reflected in test performance." Reflecting the philosophy of science which framed their definition of construct, they laid out a

further series of more stringent criteria which associated constructs with scientifically verifiable laws embedded within an explicit theory:

> Scientifically speaking, to "make clear what something is" means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a nomological network. … A necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least some of whose laws involve observables (Cronbach and Meehl, 1955, p. 290).

Shepard (1993, p. 417) refers to this early presentation of construct validity as "overly ambitious to the extent that it hoped to prove a hard-wired system of regularities." The logical positivist philosophy of science underpinning the original position in Cronbach and Meehl has since been largely repudiated in the social sciences (Fulcher and Davidson, 2007; Shepard, 1993). Even Cronbach would later suggest "it was pretentious to dress up our immature science in positivist language" (Cronbach, 1989, p. 159, cited in Shepard, 1993, p. 417). Nonetheless, he still favored the "strong program of hypothesis dominated research" proposed in Cronbach and Meehl over the alternative which he characterized as "a weak program of dragnet empiricism" (Cronbach, 1989, p. 162, cited in Kane, 2013, p. 7).

In fact, Cronbach and Meehl (1955) recognized that the state of knowledge regarding the constructs underlying most psychological tests was far from the ideal implicit in the approach they had staked out, noting that rather than empirically supported, well defined theories, "psychology works with crude, half-explicit formulations" (p. 294). The tenuous nature of the theories underlying psychological and educational tests was in fact a contributing factor to their presentation of construct validation as an ongoing, perhaps never-ending process, in which they stressed that, "the construct is at best adopted, never demonstrated to be correct" (Cronbach and Meehl, 1955, p. 294). While ongoing investigation may add to confidence in the plausibility of the theory-based networks into which the constructs fit, "it is always possible tomorrow's investigation will render the theory obsolete" (Cronbach & Meehl, 1955, p. 298). The strong program of construct validity thus proved untenable in the face of the reality of much

educational and psychological testing. Indeed, regarding the nature of constructs, almost half a century after Cronbach and Meehl's paper, Kane noted that the strong program of construct validity remained unfeasible "given the dearth of highly developed formal theories in education and the social sciences" (Kane, 2001a, p. 326).

The field of language testing and assessment has been faced with the same issues regarding construct definition. While a number of models of second language proficiency have been proposed, there remains no consensus model with universal support, and no one model has established a sufficiently large body of empirical backing to suggest the primacy of its claims regarding the putative components, or interlocking constructs, which might make up that proficiency (Chalhoub-Deville, 1997; Fulcher & Davidson, 2007; McNamara, 1996).

Fulcher and Davidson (2007, pp. 38-51), suggest that the "first and most influential model" of communicative competence was that presented in Canale and Swain's seminal 1980 paper. Canale and Swain's model of communicative competence comprised three elements: *grammatical competence*, *sociolinguistic competence* and *strategic competence* (p. 27). Canale and Swain (1980, pp. 6-7) use the term competence to "refer to underlying knowledge in a given sphere," and distinguished the components that comprise communicative competence from performance, which was characterized as "the realization of these competences and their interaction in the actual production and comprehension of utterances." Canale later revised the model to separate out discourse competence, comprising coherence and cohesion, from sociolinguistic competence, making it instead a stand-alone fourth component of the model (Chalhoub-Deville, 1997; Fulcher & Davidson, 2007).

In describing the implications of their model for second language testing, Canale and Swain (1980, p. 34) point to a possible program of research by suggesting that "it is important to empirically study the extent to which competence-oriented tests are valid indicators of learners' success in handling actual performance." They do not, however, go on to discuss in detail the role their model may play in providing the kind of theory-based network which would facilitate construct validation, nor indeed what particular approach to validity

might inform the evaluation of whether such tests were "valid." Chalhoub-Deville (1997) notes that the model has been criticized for being primarily descriptive and for lacking an empirical basis for its development. She further cites several unsuccessful attempts to empirically validate the distinctiveness of the competence components, but cautions that "potential shortcomings in terms of how the variables are operationalized and in terms of methodology used may have contributed to the researchers' inability to validate the communicative competence model" (p. 6).

Bachman, as noted earlier, presented an influential application of the unified approach to construct validity in the field of language testing in his 1990 book. Bachman's definition of construct reflected that of Cronbach and Meehl, emphasizing the importance of theory-based networks which would predict empirically verifiable components of language ability, "Thus constructs can be viewed as definitions of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behavior" (p 254). Bachman (1990) attempted to provide an explicit model of proficiency, called Communicative Language Ability (CLA), which would enable a program of research in line with the strong program of construct validation encouraged by Cronbach. The CLA model drew on Canale and Swain as well as the work of other prominent applied linguists (Bachman, 1990; Fulcher and Davidson, 2007). It attempted to go further, however, by providing a comprehensive model which would specify "the processes by which the various components interact with each other and the context in which language use occurs" (Bachman, 1990, p. 81). Bachman's CLA model comprises three components, *language competence*, *strategic competence*, and *psychophysiological mechanisms*. Language competence was further broken down into gradually finer levels of detail. At the next level lay organizational competence and pragmatic competence. The former was broken down further into grammatical and textual competence, and the later into illocutionary competence and sociolinguistic competence (Bachman, 1990, p. 87). Below these lay again finer detail, for example grammatical competence contains vocabulary, morphology, syntax, and phonology/graphology, while

textual competence contains cohesion and rhetorical organization. Bachman (1990, p. 111) describes this complex model as "a means for characterizing the traits, or constructs, that constitute the 'what' of language testing." Importantly for a model of construct validation for use in language testing, he also provided an explicit framework of test method characteristics, or facets, which he claimed would "constitute the 'how' of language testing" (Bachman, 1990, p 111).

The model has been praised for its depth and breadth, and for attempting to provide a focus for a principled program of empirical research (Chalhoub-Deville, 2003, McNamara, 2006; Weir et al, 2013). At the same time, Weir et al (2013, p. 77) note that "the model has contributed less than might be hoped to empirical test validation." Paradoxically, the very comprehensiveness of the model is cited as one impediment to its accessibility in language testing development and validation research (McNamara, 2003; O'Sullivan and Weir, 2011; Weir et, 2013). O'Sullivan and Weir (2011) and Weir et al (2013) also note that the model does not provide a clear enough description of the cognitive processing entailed by the skills components to adequately differentiate levels of proficiency during test development.

A different line of criticism of CLA has questioned the very notion of a stable underlying language proficiency construct, or what Chalhoub-Deville (2003, p. 373) refers to as an "ability – in language user" approach to construct definition. McNamara (2006, p. 468), concurs, suggesting the CLA model is "essentially psychological, seeing communicative language ability as a mental ability while the context of use is increasingly understood theoretically as a social arena." Bachman (1990, p. 111) does emphasize the importance of defining both the language ability construct and context in which that ability is used, suggesting that "the characteristics of the test method can be seen as analogous to the features that characterize the context of situation, or speech event, as this has been described by linguists." However, Chalhoub-Deville (2003, p. 372) characterizes this as "an 'ability – in language user' based on 'language user – in context' construct representation" which still fails to capture the dynamic relationship between context and underlying ability, with neither being fixed but indeed impacting on and influencing the other. While recognizing the tension between such a dynamic

model and the need for tests to generalize across contexts of use, Chalhoub-Deville (2003, p. 380) calls on language testing researchers to "develop local theories that detail the L2 'ability – in language user – in context' interactions." Weir et al (2013, pp. 99-100) suggests that the full impact of this approach has yet to be explored in language testing, noting that "testing researchers in the future will need to explore these interrelationships further and determine more closely if and how individual ability and contextual factors interact, and whether and how the ability changes as a result of that interaction."

The field of language testing thus continues to lack a clearly defined, empirically supported *consensus* model of language proficiency. Faced with this dilemma, language testing researchers have nonetheless continued to espouse the unified approach to construct validity as an underlying core principal. At the same time, they have accepted a looser interpretation of the notion of construct, one which encompasses both descriptions of the underlying abilities relevant to language use for particular purposes but also clear descriptions of the contextual features of tasks relevant to the target language use domain which is the target of testing. Indeed over a decade after his 1990 introduction of CLA, Bachman (2004, p. 15) defined construct as "an attribute that has been defined in a specific way for the purpose of a particular measurement situation. Bachman and Palmer (1996) point out that, "construct definitions are generally based on either a theory of language ability or proficiency, *or* [emphasis added] on the content of an instructional syllabus." This broader definition of construct allows for more concrete descriptions of the content of a language syllabus or course to be admitted as the object, or construct, of validation, even in the absence of a comprehensive model of language proficiency. In fact, Messick (1994) also described a similar distinction in relation to performance assessments, referring to the different approaches as *task-centered* and *construct-centered* performance assessment, but recommending the latter as the driver of construct validation.

Chappelle et al (2008, 2010) noted the importance of *both* perspectives for the purposes of language testing validation, and turned to Kane's (1992) interpretation of the argument-based approach to validation, which they claimed would allow them to include "both the competency and the task based

perspectives as grounds for score interpretation and use." Indeed, (Kane, 1992, p. 534) emphasized the ability of the argument-based approach to accommodate the lack of hard theory-based construct definitions in educational measurement as one of its main advantages:

> The argument-based approach to validity is similar to what Cronbach (1989) called the strong program of construct validation. … The term argument-based approach to validity has been used here instead of construct validity or the strong program of construct validity to emphasize the generality of the argument-based approach, applying as it does to theoretical constructs as well as to attributes defined in terms of specific content or performance domains.

The Evidence Centered Design (ECD) approach promoted by Mislevy et al (2003, p. 7), emphasizes the integration of descriptions of the underlying conceptualizations of proficiency, descriptions of the kinds of evidence which can be interpreted as demonstrating that proficiency, and descriptions of the features of tasks relevant to situations in which test takers will be asked to demonstrate that proficiency, noting that "good assessment comes not from 'choosing the right one' but by synthesizing them."

The working definition and practical exemplification of what is meant by construct has thus changed considerably from the way the concept was presented in Cronbach and Meehl (1955). The practical realities of carrying out validation for language tests in the absence of strong theories or models of language proficiency has led perhaps to a more realistic approach to defining constructs in terms of the particular contexts of use to which we want to generalize the results of language tests. At the same time, certain concepts in relation to construct have clearly become well established and continue to influence the development of validity theory and validation practice. The scientific approach to validation, with the concept of proposing and challenging hypotheses through the collection of empirical evidence and by proposing and investigating rival hypotheses has become a strong part of the unified concept of validity (Fulcher and Davidson, 2007; Kane, 2011, 2013).

It is also true that while there is no universally accepted model of language proficiency, the importance of understanding and measuring the underlying cognitive processes associated with test tasks and their relationship to TLU domain tasks has been firmly established (Chappelle, 2008, 2010; Fulcher and Davidson, 2007; Messick 1992; Mislevy et al 2003; O'Sullivan and Weir, 2011; Weir, 2005a). It is the description of the target language use domain—and the kinds of TLU tasks to which tests are intended to generalize—which has become the focus of construct definition and validation, particularly in the field of language testing. In one sense, the working definitions of construct may be more general, not relying on or attempting to elucidate an all-embracing theory-based network of law-like behavior. In another sense, however, it is also arguably more comprehensive thanks to the integration of aspects of both proficiency and context. Indeed, the effort to accommodate the practical realities dictated by the limitations in the various models may actually have taken us closer to being able to explicitly conceptualize the dynamic nature of language proficiency and the interaction of that proficiency with features of the context of use precisely.

2.2.4 Building on Messick: recent trends in the theory of validation

In the years since Messick's 1989 paper, one of the most influential approaches in validity theory has been the promotion of the argument-based approach to validation suggested by Kane (1992, 2001a, 2002, 2011, 2013) and Kane et al (1999). In order to overcome the lack of clearly defined theories of ability for many kinds of educational assessments, and to provide a set of procedures to structure the process of validation, Kane (1992, 2001a, 2011, 2013) proposed the use of *interpretive arguments* to structure the collection and evaluation of evidence to support the interpretations of test scores:

The argument-based approach to validation adopts the interpretive argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. One (a) decides on the statements and decisions to be based on the test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c)

identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments. (Kane, 1992, p. 527)

Kane et al (1999, p. 9) used the metaphor of a series of bridges, shown in Figure 2.2, to illustrate the chain of inferences which link the steps from performance on a test, to the scoring of that performance, and eventually to interpretations of those scores as representing the test taker's ability to perform in the target domain outside the test itself. The observed score for a performance on a test or assessment is premised on some kind of *evaluation* of that performance, with the assumption that the scoring rubrics and methodology are clear and applied appropriately. The *generalization* inference reflects the fact that in most educational assessments, "the observations are treated as if they have been sampled from some universe of observations, involving different occasions, locations, and observers that could have served equally well" (Kane, 1992, p. 529). The *extrapolation* inference reflects the way that test scores are used to predict performance on a wider range of possible tasks likely to be encountered in the target domain. For language testing this is the TLU domain.

While the three-bridge chain of inferences from Figure 2.2 has often been reproduced in discussion of Kane's approach (for example, in Bachman, 2004 and Chapelle et al, 2008), Kane in fact allowed for further inferences. Although, as noted above, Kane stressed that the argument-based approach did not *require* theory-based construct interpretations, importantly he allowed for the chain of inferences to include such a link. Equally importantly, he also described the link leading from interpretations to decisions, stressing that "if the test scores were not relevant to any decision, it is not clear why the test would be given" (Kane, 1992, p. 530).

In the field of language testing, Chapelle et al (2008, 2010) have adapted Kane's approach in their development of a validity argument for the revised TOEFL test. Chapelle et al's (2008, 2010) adaptation includes six inferences. In addition to the extra inferences of *explanation* (for theory-based interpretations of score meaning) and *utilization* (for decisions), they include as their first bridge *domain analysis*, which "links performance in the target domain to observations

of performance in the test domain" (2008, p. 14). Their interpretation of domain analysis draws on Evidence Centered Design (Mislevy et al, 2003).

The argument based approach as advocated by Kane (1992, 2001, 2011, 2013) and Kane et al (1999) has also influenced the Assessment Use Argument (AUA) approach developed by Bachman (2005) and Bachman and Palmer (2010). Bachman (2005) suggests that the validity argument approaches applied by Kane and Chapelle et al "provide a logic and set of procedures for investigating and supporting claims about score-based inferences but do not address issues of test use and the consequences of test use." The AUA formulation attempts to fill this perceived gap by structuring "an assessment utilization argument, linking an interpretation to a decision, and an assessment validity argument, which links assessment performance to an interpretation" (Bachman, 2005, p. 1). Kane has in turn recognized the contribution of the AUA approach in focusing attention on the importance of specifying intended test uses for validation, and has adjusted his terminology accordingly, referring to the interpretative argument as "an 'interpretation/use argument' (or 'IUA') where the IUA includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use)" (Kane, 2013, p. 2).

The argument-based formulations described above and the Evidence Centered Design approach promoted by Mislevy et al (2003) have all drawn heavily on the approach to informal or practical argumentation proposed by Toulmin (Bachman, 2005; Bachman and Palmer, 2010; Chapelle et al, 2008, 2010; Kane, 1992). In this structure, inferences proceed from *grounds*, where grounds are the actual performances or evidence collected from the test taker. A *claim* is made based on these grounds. For example, a claim, or interpretation, of a test taker's ability to perform adequately on a similar TLU task may be made based on the grounds of an observed performance on a test task. The claim will be premised on a *warrant*, which is "a law, generally held principle, rule of thumb, or established procedure" (Chapelle et al, 2008,, p. 6). Warrants require *backing* in the form of evidence to support them. All claims are also subject to potential *rebuttal*. If a plausible counter claim can be made regarding the inference from the observed performance or grounds, then the original claim may be undermined.

Chapelle et al (2008) adopt the approach suggested by Mislevy et al (2003) in which an interpretative argument to support score based interpretations requires two kinds of grounds: a description of the test taker's performance as well as a description of features of the test task used to elicit that performance (Chapelle et al, 2008; Mislevy et al, 2003). This approach thus allows them to include both task-based and competency-based approaches to validation in their chain of interferences, thus making both part of the interpretative argument and the focus of evidence collection and justification.

Kane's approach (2001, 2002) actually specified two layers or stages to an argument-based approach to validation. The interpretive argument lays out the claims and chain of inferences leading to score-based interpretations and uses (Kane, 1992, 2001, 2002, 2011, 2013), and provides a framework for determining the kinds of evidence that will need to be collected to support the interpretive argument. The validity argument describes the explicit examination of the interpretive argument and "evaluates the plausibility of the interpretive argument by examining whether the conclusions follow from the assumptions and whether the assumptions are reasonable, a priori, or are supported by adequate evidence" (Kane, 2002, p. 32). This two-stage process implies a temporal distinction, with the "development of the interpretive argument as a part of the test design process, and the development of the validity argument as an activity that occurs later, primarily after the test is operational" (Chapelle et al, 2008, p. 23). In later writing, however, Kane (2013) has emphasized the iterative nature of the validation process, describing *a development stage* and an *appraisal stage*. Evidence is collected and the interpretative argument is evaluated and if needed adjusted during both stages. What differentiates the two stages is the perspective or approach to evaluating the interpretative argument. During the development stage, the approach is naturally more confirmatory in nature; as evidence is collected, the proposed interpretative argument is tested, and if flaws are found, the test or interpretative argument is adjusted prior to operational use" (Kane, 2013, p. 17). Once the test is developed, however, "a more critical and arm's length evaluation of the proposed interpretation and use can be adopted" Kane (2013, p. 17),

Consistent with the retreat from the "overly ambitious" approach to

construct validation in Cronbach and Meehl (1955), modern applications of validity theory do not expect to develop incontrovertible empirical proof to support detailed construct networks or models of proficiency. As Kane (1992, p. 527) notes, "it is not possible to verify [the] interpretive argument in any absolute sense. The best that can be done is to show that the interpretive argument is highly plausible, given all available evidence." The argument-based approach does however offer guidelines for evaluating the arguments generated. Kane (1992) offers three criteria suggested by Toulmin; 1) the clarity of the argument; 2) the coherence of the argument; and 3) the plausibility of the argument. These criteria can be supplemented by the more general suggestions offered by Cronbach for selecting and evaluating the contribution of evidence to support a validity argument:

1. Prior uncertainty: Is the issue genuinely in doubt?
2. Information yield: How much uncertainty will remain at the end of a feasible study?
3. Cost: How expensive is the investigation in time and dollars?
4. Leverage: How critical is the information for achieving consensus in the relevant audience? (Cronbach, 1989 cited in Kane, 2013, p. 165; Cronbach, 1988 cited in Shepard, 1993).

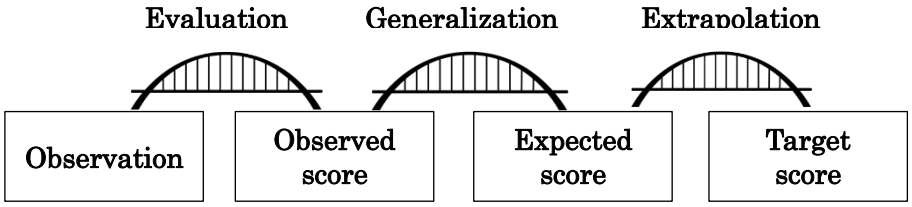


Figure 2.2 Chain of inferences shown as bridges linking test performance to score interpretation, from Kane et al (1999) and Chapelle et al 2008)

The unified approach to validity proposed by Messick (1989), including the six aspects of validity evidence which he claimed would be sufficient for touching all the necessary bases, as well as the subsequent argument-based approaches to validation prevalent over the last three decades, remain at a very

broad level of generality and are not specific to language testing. For Messick and Kane, this is understandable as they were writing to propose solutions to problems in validity theory and validation practice relevant to a very broad spectrum of psychological and educational tests. Kane (1992, p. 534) indeed stresses this generality as one of the main advantages of the argument based approach:

> It can be applied to any type of test interpretation or use: It is highly tolerant. It does not preclude the development of any kind of interpretation or the use of any data collection technique. It does not identify any kind of validity evidence as being generally preferable to any other kind of validity evidence.

The Evidence Centered Design approach offered by Mislevy et al (2003) and Almond et al (2015) does attempt to provide finer levels of detail that move from general design considerations to a model for specifying the structure of assessments derived from that design process and the model for actually delivering and scoring those assessments. Almond et al (2015, p. 23) describe the stages of ECD as involving *domain analysis*, *domain modeling*, and the *conceptual assessment framework* or CAF. The CAF contains a design blueprint for turning the design phases of domain analysis and domain modeling into concrete assessments through the use of *student models*, *evidence models*, *assembly model*, *task models*, and *a presentation model.* Nonetheless, ECD is also, by design, intended to operate at "a level of generality that supports a broad range of assessment types" (Almond, et al, 2015, p. 21). The parts of the ECD approach, including the models in the CAF, remaining essentially spaces to be filled by the specifics relevant to a particular assessment purpose and paradigm: what those specifics would be for language testing at even the most general level remain a blank page on which each particular test development and validation project would write.

Indeed, the practical argumentation approach promoted by Kane was developed without the intention of specifying the *what* of a particular validation project. As Kane notes (2013, p. 9), "the argument-based approach to validation was developed mainly as a way of facilitating *the process* [emphasis added] of

validation." In other words, it remains focused on facilitating the *how* of conceptualizing a validation argument at a very broad level of generality.

Interestingly, Chapelle et al (2008, 2010) who have adapted the argument-based approach to use specifically for language testing also deliberately shun the attempt to develop their model in a way which would point to concrete kinds of evidence necessary for language testing development and validation. Chapelle et al (2011, p. 12) note that "discussions about how much and what kind of validity evidence is needed to support the inferences and uses of test scores has been ongoing since [Cronbach and Meehl]." Nonetheless their adaptation of the argument-based approach does not provide clear indications of the what, and explicitly prefers to avoid providing lists or taxonomies of evidence, stating that "a taxonomy is not an argument, and in working with a taxonomy one is not prompted to look at the strength of the evidence or to organize it in a way that presents a validity argument" (Chapelle et al, 2011, p. 9). Crucially, perhaps, for the success of their approach, Chapelle et al (2008, 2011) state that in fact their application of an argument based approach was applied largely retrospectively to collate and evaluate the large body of evidence that had already been collected during the TOEFL revision project over almost two decades; the evidence, then, was largely in hand when the approach was applied to creating a validity argument for the test. They suggest that *the Standards* (AERA et al, 1999) already identifies various sources of evidence relevant to validation, and that "the TOEFL program maintains a taxonomy consisting of types of research that have been conducted" (2011, p. 9).

It is important to note however, that the *Standards* aims to identify quantitative and qualitative data collection and analysis procedures relevant to validation at the level of generality to be expected for the field of educational measurement generally, and does not provide information specific to language testing. While the validity argument approach espoused by Chapelle et al does provide a clear, accessible and transparent methodology for *structuring* a coherent description of the justification for the uses and interpretations of any test or assessment, not just for language tests, it is squarely focused on the *how*, and provides no clear guidelines for the selection of the *what* as it would relate to

language testing. While it may provide useful tools for conceptualizing an investigation of validity, it was not used to generate the design and development research agenda from the beginning of the process of test development (having been applied retrospectively, as noted above).

By leaving the *what* open to local interpretation for each application of their model, Chapelle et al may indeed have maintained a high degree of flexibility in their model, ensuring its applicability to a broad range of assessment contexts. The drawback of this approach in terms of the implications for the field of language testing generally, however, is that even if widely applied, the result will be a series of validation claims and argumentation, with supporting evidence presented to justify those claims, which will remain essentially local. In their efforts not to constrain or limit the vision of researchers in the search for evidence to justify score-based interpretations and uses of language tests, their model risks creating, by design, a multitude of sources of evidence across different validation projects. This situation will militate against the comparability of validation claims related to language testing.

2.2.5 The socio-cognitive model for language test development and validation

The socio-cognitive model for language test development and validation was first fully elaborated, with validation frameworks describing criterial features across each of the four skills, in Weir (2005a), and has been elaborated and developed further in O'Sullivan (2011, 2012, 2015a) and O'Sullivan and Weir (2011). Figure 2.3, taken from O'Sullivan and Weir (2011), shows how the model identifies five major areas of evidence relevant to the evaluation of language test development and use: *context validity*, *cognitive validity*, *scoring validity*, *consequential validity*, and *criterion-related validity*, with an understanding of the test taker being central to defining both context and cognitive validity. Importantly, as O'Sullivan and Weir (2011, p. 6) note, the model "starts to address an area essentially ignored by earlier theorists, that of the interaction between the different types of validity evidence."

Arguably, however, the most important contribution of the model is that skill-specific variants have been elaborated identifying concrete types of evidence

relevant to each of the four major skills commonly used in language testing: reading, listening, speaking, and writing (Weir, 2005a). The application of the model specifically to reading will be discussed in detail in the literature review for RQ1 in Section 2.3. Here a brief overview is provided of the application of the model across skills to give an indication of its scope and potential for overcoming some of the issues with the other main approaches to validation noted above. The model has been applied to comprehensive construct validation projects for the large-scale, high-stakes examination programs of Cambridge English, with listening the focus of Geranpayeh and Taylor (2013), reading the focus of Khalifa and Weir (2009), writing the focus of Shaw and Weir (2007), and speaking the focus of Taylor (2012). The model has also been applied to validation in the context of linking to the Common European Framework of Reference (CEFR) in O'Sullivan (2008, 2010) and to investigating criterial features of a reading test in Taiwan, the GEPT, in Wu (2012). The applications of the socio-cognitive model applied in these projects bears some similarity to the application of the argument-based approach of Chapelle et al (2008, 2010), in that they were retrospective applications of the model to the collation, evaluation and presentation of evidence to critically review the existing uses and interpretations of a series of operational examinations Importantly, however, in light of the caveats of the Chapelle et al (2008, 2010) approach noted above, the model has also been applied to drive the design and development research agenda of *new* assessments, for example in the development of the Aptis testing system by the British Council (O'Sullivan, 2015a, O'Sullivan and Dunlea, 2015), and in the production of a comprehensive design, development and validation research agenda for the Test of English for Academic Purposes (TEAP) in Japan (see, for example, Nakatsuhara, 2014; Taylor, 2014; Weir, 2014).

As O'Sullivan (2015a) notes, "the real strength of this model of validation is that it comprehensively defines each of its elements with sufficient detail as to make the model operational." The components in Figure 2.3 provide only the superordinate category labels, which on their own, would provide no more specificity than many of the other approaches to validation reviewed above, requiring individual applications to fill in the blanks. Fortunately for the purposes

of this study, the model for each skill provides great detail within each of these components, elaborating criterial features associated with each aspect of validity. The model for reading is discussed in detail under Section 2.3 in the literature review for RQ1. The model provides a detailed taxonomy of the linguistic demands associated with reading tasks. These lists or taxonomies of criterial features, rather than constraining or limiting the validation focus as feared by Chapelle et al (2010), have provided a clear focus for the use of the model to design relevant data collection procedures across a range of projects in various contexts, as noted above. The experience of each of those projects has added to this list of relevant criterial features, and led to adaptations and clarifications. At the same time, the incremental accruement of data using a core of related criterial features and a coherent body of methods and procedures for measuring and evaluating those features has facilitated a growing body of data to flesh out what those criterial features mean in empirical terms. The use of a consistent body of features, operationalized through a body of comparable measures and analysis techniques has thus facilitated the comparison of how these aspects of the different components of validity are realized in practice across different levels of proficiency, across different contexts of use, and for different purposes.

As noted earlier, O'Sullivan and Weir (2011) and Weir et al (2013) suggested one of the shortcomings of Bachman's CLA model was the lack of specificity to the cognitive processing element of the ability components. The socio-cognitive model in Figure 2.3 prioritizes the positioning of cognitive aspects of validity for the specification of test tasks[2]. Weir (2005a, p. 85) stresses that "approximation to the construct in a measurement instrument is essentially the result of the interactions between its context and [cognitive]-based elements. . . . Establishing the nature of these interactions is what will take forward our understanding of language testing and the constructs it attempts to measure."

This comment underscores the point made at the end of Section 2.2.3.3 regarding the nature of present knowledge of language proficiency constructs and

[2] In Weir (2005a), the cognitive aspects of processing were labelled as *theory-based validity*. Subsequent elaborations of the model have re-labelled this *cognitive validity* (e.g. O'Sullivan and Weir, 2011). For the sake of consistency and clarity, cognitive validity is used here for all references to this aspect of validity evidence in the model

their relationship to the context of use. A program of focused research that attempts to identify a range of criterial features relevant to language testing that can be operationalized and in turn contrasted and compared across a range of contexts of use in different testing programs will help us to build a greater understanding of the criterial features of interest of the cognitive processing underlying language proficiency, the contexts in which language is used, and the interactions between them.

As Weir (2005a) and O'Sullivan and Weir (2011) note, the separation into various components is to a large extent an artificial distinction to facilitate description. The components of the model are likely to interact and overlap in many dynamic ways. However, as Weir (2005a) and O'Sullivan and Weir (2011) note, for ease and clarity of description and in order to tease out the impact on each of these components at the point of interaction between test taker, test task, and the context of use in the testing situation, distinguishing these aspects is useful. Perhaps crucially from the perspective of test task design and development, the model has promoted the identification of cognitive processing models relevant to each skill which can be used to identify criterial features of test tasks relevant to different levels of proficiency, and those features are then amendable to empirical validation. The cognitive processing model for reading is discussed in detail in Section 2.3.

Another aspect of the model highlighted by Weir (2005a) and O'Sullivan and Weir (2011) is the temporal aspect to the relationship of the different components. Weir (2005a, p. 43) describes this aspect in the following way;

> The timeline runs from top to bottom: before the test is finalized, then when it is administered, and finally what happens after the test event… Thus as well as a priori (before the test event) validation components of context and [cognitive]-based validity, we also include a posteriori (after the test event) components of scoring, consequences and consequential validity.

The prioritization of contextual and cognitive aspects of validity as aspects of the test design and development that need to be taken into consideration at *the very beginning*, or a priori stage, of test development has similarities to the position of

domain analysis and modeling at the beginning of the development process in Chapelle et al (2008, 2011), Mislevy et al (2003) and Almond et al (2015). In these approaches the grounds of any claims regarding score-based interpretations begin with descriptions of both the performance and the task which elicited that performance. The step forward that the socio-cognitive model makes is to make explicit the role of cognitive processing, and requiring it to be addressed directly, whereas in Chapelle et al (2008), Mislevy et al (2003), and Almond et al (2015), the implications remain implicit with the potential for elaboration of the cognitive aspects of tasks, but equally the potential for these to be overlooked and overwhelmed with descriptions of the contextual features of both performance and the task which elicited that performance.

Weir (2005a) prefers the term *framework* in reference to the socio-cognitive model, and describes each of the skill-specific iterations of the model as validation frameworks for speaking, writing, reading, and listening respectively. O'Sullivan and Weir (2011), however, in places use model and framework seemingly interchangeably, and the use of *socio-cognitive model* is also used in O'Sullivan (2011, 2015a) and O'Sullivan and Dunlea (2015). Fulcher and Davidson (2007) and Chalhoub-Deville (1997), make a principled distinction between the choice of *model* and *framework*. Fulcher and Davidson (2007), following Chalhoub-Deville (1997), differentiate between the terms in the following way:

> We take 'models' to be over-arching and relatively abstract theoretical descriptions of what it means to be able to communicate in a second language, and we reserve 'frameworks' to be a selection of skills and abilities from a model that are relevant to a specific assessment context.

For Fulcher and Davidson (2007, p. 36), then, a "framework document mediates between a model, which is a high-level abstract document, and test specifications which are generative blue prints for a specific test." While this distinction is potentially useful, there is in fact little consistency in the way the terms are often used in the wider literature (Fulcher and Davidson, 2007). Nonetheless, in this dissertation a general distinction will be maintained between

the terms, and the approach to validation proposed by Weir (2005a) will be referred to as the socio-cognitive *model* for language test development and validation rather than a framework. This term more aptly captures the superordinate level of abstraction which the model provides as a guiding frame of reference for making explicit the necessary aspects of language proficiency and test task design necessary for us to build comprehensive, coherent, and plausible validity arguments. At the same time, it is suggested that the socio-cognitive model has the facility to be iteratively applied in ever finer detail. Exemplification of the specific aspects of contextual, cognitive, scoring, and other aspects of validity most relevant to a specific testing context, or perhaps to a specific area of testing, could be fleshed out as a framework for that context. And as demonstrated in O'Sullivan and Dunlea (2015) frameworks derived from the socio-cognitive model can then be used to generate blueprints for test task specification at the most detailed level.

Some aspects of the most recent elaborations of the model are worth mentioning here and will be taken up in slightly more detail in Chapter 6. Firstly, O'Sullivan and Weir (2011) have suggested that criterion-related validity could be usefully subsumed within scoring validity. For O'Sullivan and Weir (2011) this then focuses attention on the core of the validation model for construct validation, combining contextual, cognitive, and scoring aspects of validity, which they see as "essentially inward looking, in that they are focused on aspects of the test itself" (p. 24). In relation to consequences, O'Sullivan and Weir (2011) concur with the issues related to the inclusion of consequences in validity theory generally that were discussed in Section 2.2.3.1. In relation to the original presentation of the socio-cognitive model, they suggest that rather than the temporal positioning of consequences as an a posteriori concern as suggested by Figure 2.3, that it is important to foster an "awareness of consequence as impacting on all elements of validity and forming a guiding principle for the development process from the very beginning of the cycle" (O'Sullivan and Weir, 2011, p. 23). Indeed, O'Sullivan (2011, 2015a) has further questioned the presentation of the temporal relationship between all components of the model, not just in relation to consequences, and has highlighted the implications for test development. While

consequences do not form part of the three research questions driving this study, the temporal aspect of the original presentation of the model will be taken up in more detail in the discussion in Chapter 6.

The socio-cognitive model has been selected for the purposes of driving the research agenda for this study for the reasons noted above. It offers a comprehensive approach to validation that has built on and incorporated the developments in the field of validity theory over the last three decades. It is grounded in practical test development and validation experience, and a growing body of literature on its application make it particularly suitable from the perspective of the four important subsidiary criteria we noted in Chapter 1: a) relevance b) transparency c) interpretability d) comparability. As also noted in Chapter 1, this study does not aim to develop a comprehensive validity argument for the EIKEN tests, but rather aims to derive a core body of evidence crucial for the construction of such an argument. As noted above, O'Sullivan and Weir (2011) have identified the aspects of contextual, content, and scoring validity as being at the core of validation of the uses and interpretations of language tests through the socio-cognitive model—with criterion-related validity aspects posited by them as being usefully included within the remit of scoring issues.

These core aspects of validity align with the three research questions identified in Chapter 1 and have driven the focus on these aspects as being the most relevant for the purposes of this study. At the same time, it is essential to stress that, just as with Messick's presentation of the six aspects of validity described earlier, in the final synthesis of a comprehensive, coherent, and plausible validity argument, *all* components of the model will need to be addressed. This issue has already been highlighted in Section 1.3, and is taken up again in Section 6.3 under the discussion of limitations. It is, however, important to reiterate that the aspects of the socio-cognitive model which underpin this study are those identified by O'Sullivan and Weir (2011) as constituting core aspects for evaluating the evidential basis for the uses and interpretations of the test system as a measurement instrument. These aspects address the questions of how well the test or tests measure what they claim to be testing. At the same time O'Sullivan and Weir (2011), and all iterations of the socio-cognitive model described in this

literature review, emphasise the importance of also addressing the issue of consequences and impact in the development of a plausible validity argument. While important for all tests, it is particularly important for the EIKEN testing program, given the emphasis on positive impact on EFL education and assessment in Japan which has been an explicit part of the program's aims from its inception, as noted in Section 1.2.3.

Despite the difficulties encountered by all post-Messick attempts at operationalizing the inclusion of consequences in approaches to validation (see Section 2.2.3.1), it is equally true that all approaches to validity described here, including the socio-cognitive model, emphasise the importance of addressing consequences. The model is useful in this respect, then, as the inclusion of consequences in Figure 2.3 provides a mechanism for forcing the test developer to consider this aspect in the collection and evaluation of evidence for a comprehensive validity argument, and equally provides a quality assurance mechanism for test users to be able to question validity arguments that do not address this aspect. Addressing consequences for the EIKEN testing program will require a large-scale effort as it will require the investigation of very different uses and interpretations, and quite different typical test takers and stakeholders, across the grades. It is beyond the scope of this study, but must not be beyond the scope of an ongoing, comprehensive validation agenda for the test developer. This study has focused on particular aspects of the socio-cognitive model to derive three research questions which will contribute a core body of evidence towards the construction of a validity argument. The model also provides the safety check of ensuring that the other aspects not included in this study due to practical constraints cannot be ignored by the test developer in the development of that argument.

While a discussion of consequences and aspects of the model not directly addressed by the three research questions may at first impression seem superfluous, examining the model in its entirety—including its relationship to the major trends in validity theory—is crucial for two reasons. Firstly one of the important subsidiary goals of this study is to evaluate the model itself. This is taken up in Section 6.2.5.2, and the usefulness of the model, and the approach

taken in this dissertation to focusing on a subset of core aspects, can only be properly evaluated against an understanding of the model in its entirety. Secondly, To avoid any potential misinterpretation or potential misuse of the application of the model in this study, it is important to stress to future users of the model and readers of this dissertation the importance of considering all aspects of the model for the construction of a comprehensive validity argument in order to touch all of Messick's bases. This is stressed further in Section 6.3, under the limitations.

While McNamara (2006) has referred to Bachman's work, particularly his 1990 book, as the bearer of the Messick legacy, at the present time, the socio-cognitive model is best placed to be the standard bearer of the key concepts promoted by Messick's unified model of validity, particularly in relation to language testing. The clear relationship that can be drawn between the components of the socio-cognitive model and Messick's six aspects of validity, which it was suggested above he intended as the concrete methodology for operationalizing a unified approach to construct validation in practice, support this view. The socio-cognitive model provides us now with the conceptual model for an explicit description of the key concepts which will be sufficient for touching all the bases in relation to collecting evidence to justify the uses and interpretations of test scores for language tests. Further, this model has the level of specificity appropriate for language testing to enable it to be iteratively applied at ever greater levels of detail, allowing for its use to derive frameworks for specific contexts of use and further to drive the generative blueprints which underpin detailed task specification. The model thus provides the means to begin addressing, in relation to language testing, the answers to the question: "*How much* of *what* do we need to collect to construct a coherent, comprehensive, and plausible validity argument for language tests?"

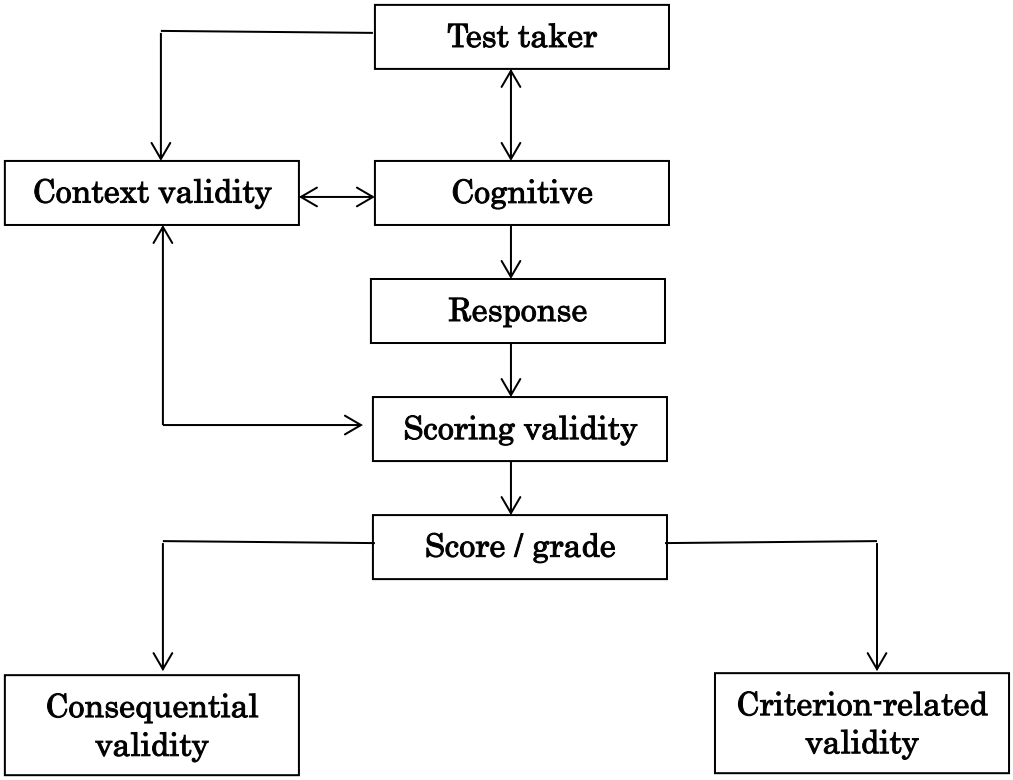


Figure 2.3 Socio-cognitive model for test development and validation (From O'Sullivan and Weir, 2011)

2.3 Research Question 1: Criterial Features of Reading Test Tasks

Section 2.2 provided the rationale and justification for the selection of the socio-cognitive model to drive the data collection and analysis agenda for this study. The socio-cognitive model, as noted above, provides the level of specificity required to examine the criterial features of language tests in sufficient detail to enable a concrete research agenda to facilitate validation. Importantly, given the subsidiary criteria for this study, the components of the model, and the level of specificity of the criterial features within them, also provide the means of actual task specification (see O'Sullivan and Dunlea, 2015, and Taylor, 2014, for examples). Research Question 1, then, takes the model as it has been applied specifically to reading as the overarching framework within which criterial features have been identified and investigated.

Weir (2005) provided skill-specific frameworks with lists of criterial features relevant to each of the major components of the model shown in Figure 2.3 above. The components of that model most relevant for RQ1 are the

contextual and cognitive aspects of validity. The framework for reading was developed further in Khalifa and Weir (2009), particularly in relation to the cognitive model used to differentiate levels of reading considered appropriate for different levels of proficiency. The elements of contextual and cognitive validity described in Khalifa and Weir (2009) are shown in Figure 2.4.

The arrow between the context and cognitive reflects the discussion in section 2.2.3.3 above, that these aspects are likely to interact in dynamic ways. The distinction into separate aspects is maintained here, as it is useful both for conceptual reasons and to facilitate the collection and analysis of data. Nonetheless, it is important to reiterate that these distinctions are somewhat artificial. Khalifa and Weir (2009) suggest that:

> Undoubtedly a close relationship exists between these elements, for example between context validity and cognitive validity, which together with scoring validity constitute for us what is frequently referred to as construct validity. Decisions taken with regard to parameters in terms of task context will impact on the processing that takes place in task completion. The interactions between, and especially within, these aspects of validity may well eventually offer further insights into a closer definition of different levels of task difficulty. (p. 8)

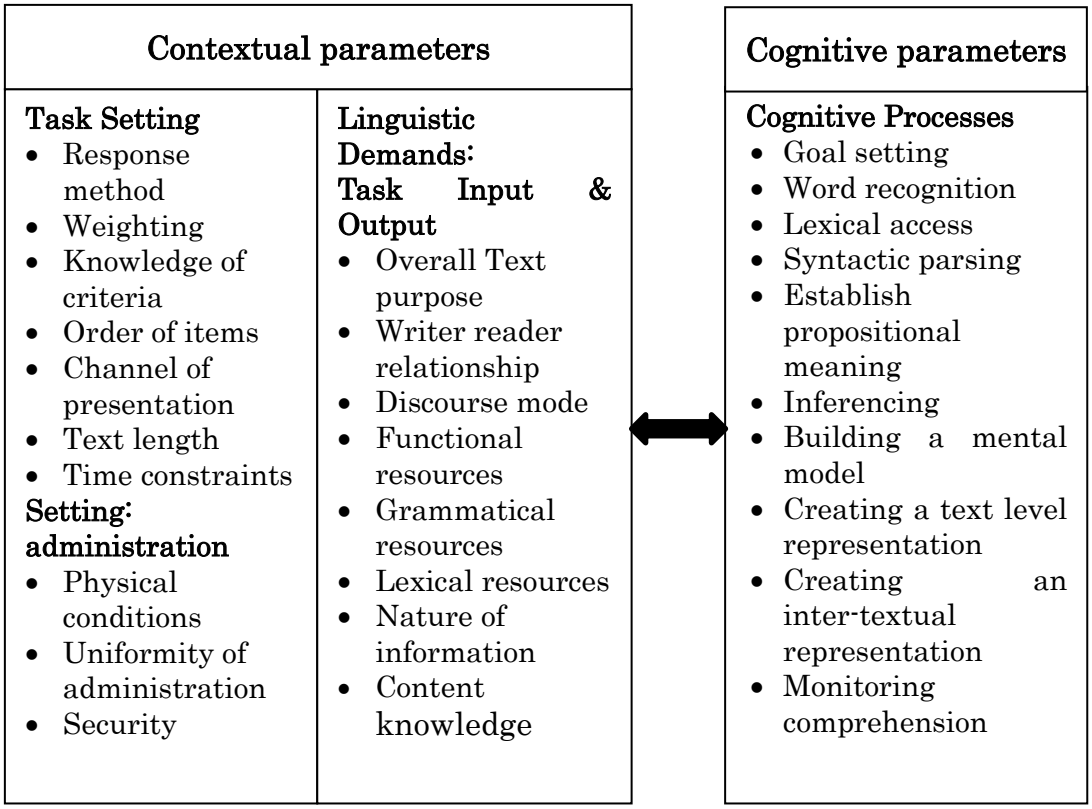


Figure 2.4 Contextual and cognitive parameters for reading (Khalifa and Weir, 2009)

As noted in Section 2.2.4 above, enabling the focused investigation of cognitive processing as a part of the specification of criterial features of test tasks is one of the advantages of the socio-cognitive model. Figure 2.5 shows a diagrammatic representation of the model of cognitive processing for reading provided in Khalifa and Weir (2009, p. 43). The theoretical underpinnings of the model draw on an extensive review and synthesis of the literature on reading "in order to devise a model of the L1 reading process – supported by empirical evidence – which can be treated as the goal towards which the L2 reader aspires" (Khalifa and Weir, 2009, p. 43).

The diagram in Figure 2.5 can be divided into three areas representing the different aspects of the model which are activated in the process of a reader interacting with a text, or "visual input." The far left of the diagram contains metacognitive activities, including the complex act of self-monitoring undertaken by the reader (Bax, 2013; Bax and Weir, 2012; Brunfaut and McCray, 2015; Khalifa and Weir, 2009). Crucial to this section of the model is the "the goal

setter," which is important "because in deciding what type(s) of reading to employ when faced with a text or texts, critical decisions are taken which affect the levels of processing to be activated in the central core" (Khalifa and Weir, 2009, p. 44). The central core of the processing model is divided into an eight-level hierarchy representing increasing levels of cognitive demand as the kind of processing required progresses from the lowest level—recognizing individual words—through to the highest level, which involves integrating information and ideas from multiple texts into an integrated representation of those texts (Khalifa and Weir, 2009; Weir et al, 2013). One important dimension reflected in the levels of processing is the greater levels of integration of information required at each level. The integration is both quantitative and qualitative, in that it reflects processing of larger amounts of text, but also the integration of information across those greater chunks of text into coherent propositions and ideas which are eventually integrated into whole-text representations. This aspect of the levels of processing reflects then the integration required at each level, and this is one aspect that lends itself to operationalization in test task specification and analysis. The right side of the model identifies the various sources of knowledge which the reader will need to draw on in order to activate the different levels of processing, for example lexical knowledge, syntactic knowledge, and knowledge of the wider world outside the text itself (Bax, 2013; Brunfaut and McCray, 2015; Khalifa and Weir, 2009; Weir et al, 2013).

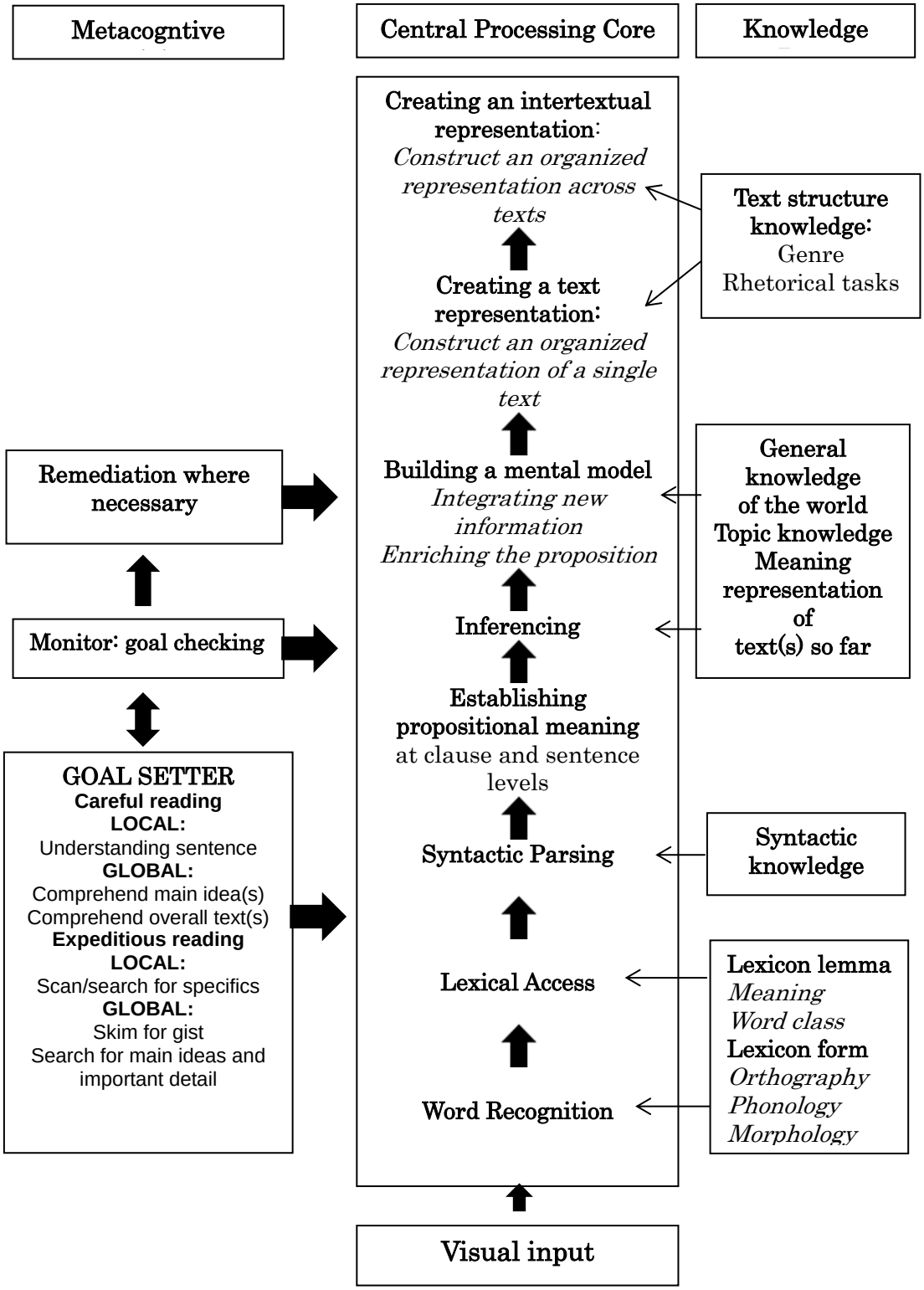


Figure 2.5 Khalifa and Weir (2009) cognitive processing model for reading

This cognitive process model of reading has facilitated the investigation of construct validity very much in line with the original strong program of

validation through scientific hypothesis testing advocated by Cronbach. Khalifa and Weir (2009, p. 42) note that "the cognitive validity of a reading task is a measure of how closely it elicits the cognitive processing involved in contexts beyond the test itself." Not only has the model proved amenable to operationalization and thus to verification in terms of reading processes, it has also provided a means of verifying whether test tasks elicit those processes. Khalifa and Weir (2009) carried out a review of Cambridge Main Suite reading test papers to investigate the degree to which the types of reading and the levels of processing were associated with test tasks targeted at different levels of proficiency. Taylor (2014) also provides an example of how the levels of processing can be applied through expert judgment to the validation of reading test tasks. Taylor (2014) compared the levels of processing judged to be elicited by a set of reading tasks for a test of academic English proficiency in Japan with the explicit intentions of the test developers as exemplified in the test specifications.

While these applications do provide some evidence of the utility of the model for these purposes, as Khalifa and Weir (2009) note in relation to their analysis, "the analysis is based on the opinions of a group of expert judges only and findings will need to be more firmly grounded in the future by having students take the various reading tests and complete verbal protocols of their experience" (p. 63). Subsequent applications of the model have taken up this challenge, including Weir et al (2009) and Wu (2012) which employed recall protocols to collect information from test takers through questionnaires on the levels of processing they employed while completing reading test items. Both studies used a three-way distinction to operationalize the degree of integration of information required by the different levels of processing noted above, asking students if the information they needed to answer the test items could be found within one sentence only, across sentences, or required reading the whole text. Recently, more complex mixed-method studies have employed eye-tracking in conjunction with recall protocols collected through questionnaires (Bax and Weir, 2012). In this study, the questionnaire only contained two responses, finding information within a single sentence or across sentences. Bax (2013) and Brunfaut

and McCray (2015) used a combination of eye tracking with stimulated recall interview protocols to provide a comprehensive picture of the levels of processing employed by test takers, providing empirical backing for the distinctions between the levels and more importantly in the ability to associate test items with particular levels of processing in the model. The Brunfaut and McCray (2015) study is particularly insightful for the subsidiary goals of this study. The study used tasks created from test specifications described in O'Sullivan and Dunlea (2015) which had operationalized the cognitive processing model, and demonstrated how processes targeted by the tasks are amenable to explicit verification through techniques such as eye-tracking and stimulated recalls.

Khalifa and Weir (2009) describe context validity as "the appropriateness of both the linguistic and content demands of the text to be processed, and the features of the task setting that relate to task completion" (Khalifa and Weir, 2009, p. 81). The difficulty of a test task will result from the interaction between test takers, the contextual features of the text to be processed as well the wider context in which the task is carried out, and the level of processing required by the task. In a comprehensive historical review of the contextual and cognitive parameters of Cambridge Main Suite reading tests, Weir et al (2013) focused on a reduced set of criterial features for context, noting that: "full situational authenticity is not normally attainable within the constraints of the testing situation," leading to a choice of contextual parameters for the study that were "most likely to have an impact on reading test performance" (p. 115). The parameters suggested were *response method*, *test length*, and *time constraints* for task setting; and d*iscourse mode (genre/rhetorical task)*, *grammatical resources*, *lexical resources*, *abstractness/concreteness*, and *content knowledge* for *linguistic demands*. Within each of these features, Khalifa and Weir (2009) provide a list of possible options, or in the case of linguistic demands, a number of useful quantitative indices, which they applied to a range of Cambridge Main Suite reading papers.

Alderson (2000, p. 74) in summing up the range of "variables that affect text difficult" provides a very similar list: *topic, syntactic complexity, cohesion, coherence, vocabulary,* and *readability*. In an attempt to create a set of task specification grids which would be useful for specifying criterial features of

reading and listening tasks for the purposes of linking examinations to the CEFR, Alderson et al (2006) derived a set of parameters which were later incorporated into the *Manual for Linking Examinations to the Common European Framework of Reference* (Council of Europe, 2009). The final grid derived by that project provides a range of characteristics split between: a) characteristics of the input text, and b) characteristics of the item. For reading texts, the relevant characteristics are: *topic, text source* (which relates to *genre i*n Khalifa and Weir (2009)), *authenticity, discourse type, domain, topic, nature of content* (degree of *abstractness*), and *text length*. The framework also contains a section for features of the item, which relates more closely to the cognitive processing dimension in Khalifa and Weir (2009). The Alderson et al (2006) item processing dimension addresses the response type (e.g. constructed response or selected response) and the operations targeted. Operations, as shown in Table 2.2, were further defined according to a three-way distinction. Another influential framework in test development has been that proposed for reading by Enright et al (2000) as part of the TOEFL revision program. Enright's taxonomy of characteristics includes *participants, setting, content, purpose, and register. Syntax, vocabulary, discourse organization, rhetorical feature* (rhetorical task in Khalifa and Weir, 2009). Wu (2012) reviewed a number of frameworks designed to specify contextual parameters likely to impact on test task difficulty which included those mentioned above, noting that there is "broad consensus on the features that are likely to impact on reading performance" (p. 55). Green et al (2009, p. 4), provide a list of features derived from Khalifa and Weir (2009) which they suggest is "predicted from theory and previous research to have an impact on reading comprehension." They go on to caution that:

> There is a pragmatic balance to be struck between the comprehensiveness of the text description and the feasibility of application in test development and validation. The chosen features were intended to be either directly measurable or readily judged by teachers or test developers with minimal training.

This reflects the subsidiary goals described in Chapter 1, in which it was suggested that the measures selected for this study should be amenable to

adoption for use within an operational testing program by item writers and developers. The criteria listed by Green et al (2010) are: *text length, grammatical and vocabulary characteristics, cohesion and rhetorical organization, genre and rhetorical task, subject knowledge, cultural knowledge, and abstractness*. Green et al (2010) demonstrate the use of automated text analysis tools to derive estimates of those measures which are open to being "directly measurable" and are in many cases more accurately obtained through automated analysis tools rather than expert judgment.

Table 2.2 Operations in CEFR Grids, from Alderson et al (2006)

Behavior	What is understood	From information that is
Recognize Make inferences Evaluate	Main idea/gist Detail Make inferences Opinion Speaker's/writer's attitude/mood Conclusion Communicative purpose Text structure between parts	explicit implicit

Automated analysis tools lend themselves particularly to the analysis of the vocabulary used in reading texts. Research has consistently shown vocabulary to be an important element of reading proficiency, with knowledge of vocabulary correlating highly with reading comprehension scores (Alderson, 2000; Geranpayeh, 2007: Harsch and Hartig, in press; Khalifa and Weir, 2009; Laufer and Ravenhorst-Kalovski, 2010; Milton, 2010, Shiotsu, 2010; Shiotsu and Weir, 2007). While Shiotsu (2010) and Shiotsu and Weir (2007) found that tests of syntax explained slightly more of the variance in a series of studies designed to investigate the relative importance of these two variables, Shiotsu and Weir (2007, p. 122) caution that, "the results must not be interpreted as indicating that vocabulary is unimportant . . . our attention to the development of lexical competence for improved reading proficiency should continue." It is also important to note that the role of vocabulary and grammatical knowledge has been shown to extend beyond just reading comprehension, with various studies demonstrating the role they play as predictors of receptive language proficiency

skills generally, consistently showing moderate to high correlations with both reading and listening (Geranpayeh, 2007; Harsch & Hartig, in press; Milton, 2010; Stæhr, 2008).

In relation to the role which vocabulary plays in reading comprehension, Van Zeeland and Schmitt (2012, p. 2) note that "of particular interest to the field of second language pedagogy is the percentage of words in written or spoken discourse which enables successful comprehension." They see this aspect, variously referred to as *lexical coverage* or *text coverage*, as "an essential measure, for it allows the calculation of estimates of the vocabulary size necessary for comprehension of written and spoken texts" (Van Zeeland & Schmitt, 2012, p. 1). Van Zeeland and Schmitt (2012, p. 3) refer to a *coverage threshold* to describe the concept of a minimum lexical coverage point required for comprehension. Laufer and Ravenhorst-Kalovski, 2010, p. 15) use the term *lexical threshold* to refer to "the minimal vocabulary that is necessary for 'adequate' reading comprehension." The lexical threshold then associates an estimate of the vocabulary size required to reach the coverage threshold necessary for comprehension. Laufer (1989, cited in Laufer and Ravenhorst-Kalovski, 2010) originally suggested a coverage threshold of 95 percent, meaning readers should be able to recognize and understand 95 percent of running words in a text in order to achieve adequate comprehension. Hu and Nation (2000) and Nation (2006) later suggested a higher threshold of 98 percent, suggesting the higher figure was necessary for the majority of participants in their study to achieve understanding. Schmitt et al (2011) suggested that rather than a single threshold, the coverage threshold may change depending on *the degree* of understanding required. Laufer and Ravenhorst-Kalovski (2010) replicated earlier studies attempting to define a coverage threshold with a much higher number of participants. They (p. 15) recommend two coverage thresholds with associated lexical thresholds defined in terms of the vocabulary size needed to reach them: "an optimal one, which is the knowledge of 8,000 word families yielding the coverage of 98% (including proper nouns) and a minimal one, which is 4,000–5,000 word families resulting in the coverage of 95% (including proper nouns)." The figures for the number of word families required to reach the two different thresholds reflect the more

conservative estimates derived by Nation (2006). Similar conclusions were made by Van Zeeland and Schmitt (2012) in relation to lexical coverage thresholds for listening, recommending 98 percent as an optimal figure, but suggesting 95 percent as sufficient for reasonable understanding in many situations. The difference is important as it relates to the setting of realistic learning and teaching goals. Van Zeeland and Schmitt (2012, p. 7) comment that the "difference between the [different thresholds] is vast and may require quite different teaching approaches and investments of time and effort." Although their study was focused on listening, the principal of defining realistic learning goals is equally relevant to defining lexical thresholds for reading. Nation in fact has recognized that the 98 percent threshold level may indeed be unrealistic in many circumstances and is only reached in many texts after including a small number of words across higher frequency levels (personal communication cited in Taylor, 2014). A number of studies have thus taken the 95 percent criterion as a reasonable and realistic threshold for reading especially in EFL contexts (Chujo and Oghigian, 2009; Taylor, 2014).

Investigating the number of words needed to achieve the coverage threshold in the studies noted above has been facilitated by the development of frequency lists derived from large scale corpora. Those developed by Nation (2006) have been amongst the most widely used for such studies and particularly in the investigation of criterial features of input texts in testing. Nation (2006) created a 14-level frequency list derived from the British National Corpus. Each level contains 1000 word families, with the first level containing the most frequent 1000, the next level the second most frequent 1000, etc. The words were sequenced "largely according to their range and frequency in the 10 million spoken section of the BNC" (Nation, 2006, p. 80). The lists contain two supplementary levels: level 15 contains lists of commonly encountered proper nouns, and level 16 contains exclamations. The lists have been made available by Nation in a version of the Range (Heatley, Nation, and Coxhead, 2002) vocabulary analysis software. The first 14 levels of the lists were also adapted for use in the online analysis tools available in VocabProfile (Cobb, 2015). In VocabProfile the lists are referred as the BNC-20 lists as six additional levels were

added, but Cobb (2015) notes that the first 14 base levels remain the same as Nation (2006) and the lists supplied with the Range software. The BNC lists, either utilizing the Range software or VocabProfile have been applied by Nation (2006) to a large range of authentic reading and listening texts, and Green et al (2010) to a comparative analysis of IELTS reading texts and authentic undergraduate texts,. They have also been used in a number of studies specifically as an important criterial contextual parameter in the analysis of input texts used in EFL tests (Chujo and Oghigian, 2009; Khalifa and Weir, 2009; Taylor, 2015, Dunlea, 2010; Dunlea, 2014; Weir et al, 2013).

An additional useful source of vocabulary profiling is offered by the Academic Word List (AWL) developed by Coxhead (2003). The list contains 570 word families which Coxhead (2003) demonstrated were useful to reading in academic contexts. In academic contexts, Coxhead (2003) found the AWL provided a lexical coverage of approximately 10 percent, and this finding was replicated by Green et al's (2009) study of undergraduate reading texts from a British university. The lists have been applied in a number of the studies of reading texts in language tests mentioned above and have proven to be robust in differentiating between both proficiency levels and texts designed to reflect criterial features of different genres (Dunlea, 2014; Green et al, 2009; Khalifa and Weir, 2009: Taylor, 2014, Weir et al, 2013; Wu, 2012). The AWL is available in the original version of the Range software which combines the General Service List (GSL) and AWL, and is also available using VocabProfile.

Automated tools also allow for the easy analysis of a number of other features of the linguistic parameters defining input texts which are considered to impact on text comprehensibility. As Alderson (2000) notes, sentence length is a "crude" measure of syntactic complexity, as "short sentences tend to be syntactically less complex than longer sentences." Alderson (2000) also notes the usefulness of readability indices as measures of text complexity. Commonly used readability indices combine two proxies for aspects of texts which have proven good predictors of textual complexity: syntactic complexity, operationalized through sentence length measures, and vocabulary frequency, operationalized through measures of word length—longer words tend to be less frequent and less

frequent words will tend to be harder to learn and increase the processing burden on readers who are not familiar with them (Alderson, 2000; Crossley, Greenfield, and McNamara, 2008; Khalifa and Weir, 2009). Traditional readability formulas have been criticized as being inappropriate for EFL contexts because they were generally developed for native speakers of English (Alderson, 2000; Greenfield, 2004: Crossley et al, 2008; Khalifa and Weir, 2009). They have nonetheless proved to be consistently robust predictors of text difficulty and have been used in a number of studies (Alderson, 2000; Greenfield, 2004: Khalifa and Weir, 2009). Indeed, Greenfield (2004, p. 19) compared a number of traditional readability formulas along with formulas developed specifically for the EFL context of Japan and concluded that: "the new formulas have only a narrow, if any, advantage over the time-tested traditional formulas, especially the Flesch and Flesch-Kincaid, and Bormuth formulas. We may therefore use those formulas with some new confidence that they are valid for EFL." Crossley et al (2011) also compared traditional formulas including the Flesch Reading Ease and Flesch-Kincaid Grade level formulas with a new formula they developed for use with second language learners using the Coh-Metrix (Graesser et al, 2004) analysis tool, and which they called the Coh-Metrix L2 Index. Although the Coh-Metrix index correlated most highly with measures of text difficulty derived from the cloze scores of Japanese university students, the Flesch-Kincaid Grade level and Flesch Reading Ease formulas also showed strong correlations which were statistically significant. In a study to investigate the relationship between reading tasks targeting different levels of the CEFR with empirical item difficulty, Dunlea (2014) found that the Flesch-Kincaid Grade level outperformed both the Flesch Reading Ease and Cohmetrics L2 Index. In terms of applications to studies investigating contextual validity parameters as criterial features of reading test tasks, the Flesch-Kincaid Grade Level and Flesch Reading Ease were used by Khalifa and Weir (2009), and in conjunction with the Coh-Metrix L2 index by Green et al (2010), Wu (2012), and Weir et al (2013), while Taylor (2014) utilized the Flesch-Kincaid Grade Level only.

Another aspect of the criterial contextual parameters of reading texts often included in studies of text difficult are measures of lexical diversity. The

simplest measure of lexical diversity is the Type-Token Ratio. The TTR calculates the "ratio of the ratio of types or different words to tokens: the total number of words occurring in the text" to provide a measure of "the number of different words the reader will need to know to understand a passage" (Green et al, 2010). Higher lexical diversity may contribute to the processing demands placed on a reader as it increases the number of new words that need to be processed and incorporated into the ongoing message (Green et al, 2010; Weir, 2013). It is recommended that the TTR should only be used to compare texts of the same length, as the measure is sensitive to the length of a text (Green et al, 2010; Jarvis, 2002; Khalifa and Weir, 2009: McCarthy and Jarvis, 2010); This is because the longer a text is, the more likely it is that individual types will be repeated. In order to overcome this deficiency in the measure, a number of more sophisticated alternatives have been proposed. McCarthy and Jarvis (2010, p. 383) carried out a validation study of such measures, focusing on what they refer to as "the most sophisticated indices of LD that are currently available." Referring to the measures they had investigated, they concluded that studies of lexical diversity should "consider using MTLD, vocd-D (or HD-D), and Maas in their studies, rather than any single index," because "lexical diversity can be assessed in many ways and each approach may be informative as to the construct under investigation." Jarvis (2002) also provided strong support for the use of vocd-D as a robust and accurate measure of lexical diversity, though he was dealing with spoken narrative texts and not reading texts.

As Green et al (2010) report, the online automated analysis tools provided as a part of Coh-Metrix (Graesser et al, 2004) provide the possibility of reporting on a number of indices that extend beyond the measures described above. As well as a range of measures designed to address syntactic complexity, Coh-Metrix is also designed to provide measures of coherence and cohesion across a text. Coh-Metrix has been utilized to supplement other forms of textual analysis in a number of projects including Green et al (2010), Weir et al (2013), and Wu (2012). However, Weir et al (2013) notes the lack of transparency and interpretability of some of the Coh-Metrix indices, which directly impacts on the four subsidiary criteria listed as being relevant to the selection of measures for this

study. The version of Coh-Metrix currently available, Version 3.0, has also subsumed many of the individual indices previously available only separately in Coh-Metrix version 2.0 into a series of super-ordinate "text easability" indices as described in McNamara et al (2011). While these easability indices may be more user-friendly for educators or researchers simply wishing for quick and easily obtained measures of comparability, it provides a further layer of opaqueness for users who need to make decisions about what aspects of a text to adjust based on those measures. A further issue that was encountered in the early stages of the study was that the Coh-Metrix website experienced quite large variations in terms of accessibility and the time taken to process texts. These issues also influenced Taylor (2014) who notes that while Coh-Metrix offers potential for further investigation it was not selected for use in that study primarily for the reasons mentioned above.

An alternative online text analysis tool is offered by Text Inspector (2015). Text Inspector provides measures of the criterial linguistic parameters of texts noted above, including the total number of words, sentence length in words and syllables, readability indices including the Flesch-Kincaid Grade level, the MTLD and vocd lexical diversity measures, as well as other useful measures. In addition, Text Inspector offers a range of 13 indices grouped under the heading of *metadiscourse markers* (Text Inspector, 2015). Metadiscourse, according to Hyland and Tse (2004, p. 156), provides "a motivated way of grouping under one heading the range of devices writers use to explicitly organize their texts, engage readers, and signal their attitudes to both their material and their audience."

Hyland (1999) provided a detailed taxonomy of 10 metadiscourse markers with exemplification, divided under two broad headings: *textual metadiscourse* "used to organize propositional information in ways that will be coherent for a particular audience and appropriate for a given purpose;" and *interpersonal metadiscourse* which "allows writers to express a perspective towards their propositional information and their readers." Hyland (1999) employed the taxonomy in a study of metadiscourse in university textbooks. Camiciottoli (2003) adopted Hyland's taxonomy to investigate the relationship between comprehension by EFL readers in an Italian university and the

metadiscourse features of texts. She (p. 37) concluded that the "results of this study lend further support to the idea that metadiscourse can have a positive influence on comprehension" but also noted that "certain types of metadiscourse may be more facilitating than others during reading." Hyland and Tse (2004) adapted the taxonomy to investigate a four million word corpus of postgraduate dissertations, concluding that "differences in metadiscourse patterns can offer an important means of distinguishing discourse communities" (p. 175).

Bax et al (2013) used Text Inspector to analyze 900 expository essays produced by EFL test takers to investigate patterns of metadiscourse markers. They also carried out a manual analysis of a sub-sample of 200 texts to further refine the examples of metadiscourse markers used in Text Inspector. Bax et al's (2013) analysis modified Hyland's taxonomy, producing a list of 13 metadiscourse markers. The exemplification of each type of metadiscourse marker used by Text Inspector has drawn on the extensive categorization in Hyland (2004), but uses the list of 13 metadiscourse markers refined by Bax et al (2013) as its organizing taxonomy (Bax, personal communication, July, 2015). The metadiscourse features of Text Inspector, then, offer the potential to address aspects of coherence and cohesion as well as a means of investigating criterial features relevant to various genres of texts. This in turn offers the means of not only differentiating between texts targeted at different levels of proficiency, which is the prime goal of this study, but also a means of validating those texts in relation to the criterial features of TLU domain texts for different genres. An attractive feature of the metadiscourse markers used in text inspector, given the subsidiary criteria, particularly transparency, is the detailed lists of metadiscourse markers on which the measures in Text Inspector are based that are available in the literature.

A final list of criterial parameters for use in this study is shown in Chapter 3 as Table 3.1. The list was derived based on the review of the literature in relation to criterial features of contextual and cognitive parameters relevant to reading, and taking into account the four important subsidiary criteria noted in Chapter 1 as important for the selection of measures for use in this study.

2.4 Research Question 2: Vertical Scaling and Scoring Validity

Research Question 2 involves the use of vertical scaling methodology to investigate the degree to which the seven grades of the EIKEN tests can be shown to represent empirically distinct levels of proficiency in terms of item difficulty. This element of the study falls under the scoring validity aspect of the socio-cognitive model. Underlying all three research questions, however, is the assumption that the instruments used in the study are technically adequate in terms of their measurement properties, which is of course also the remit of scoring validity. Chapter 4 is thus divided into two sections: 4.2 which will provide an overall evaluation of the degree to which the tests can be shown to be psychometrically fit for purpose, based on an analysis of operational data, and 4.3, which will deal directly with the methodology, analysis and results of the vertical scaling undertaken for RQ2. The literature review will also follow this division, first providing an overview of the statistics and relevant benchmarks for their interpretation which will be used to evaluate the measurement properties of the test. This will be followed by a review of the literature relevant to the selection of methodology for vertical scaling.

2.4.1 Statistics for evaluating the technical quality of the EIKEN tests

A number of descriptive statistics and indices for evaluating test and item performance will be reviewed in Section 4.2. As most of these indices are commonly used in Classical Testing Theory (CTT) analyses, and are described in detail in texts such as Bachman (2004), this brief overview of the literature will be restricted to reviewing information available on recommended levels for these measures. The aim of this review is to establish benchmarks against which can be used to evaluate the performance of the EIKEN tests.

Firstly, the degree to which score distributions diverge from normality has an impact on the assumptions necessary for Norm Referenced (NR) score interpretations, particularly on indicators of NR reliability (Bachman, 2004; Brown, 1997). The *skewness* and *kurtosis* statistics are indicators of the shape of the score distribution and the extent to which the scores are normally distributed (Bachman, 2004, Field, 2009,). Bachman (2004, p. 74) suggests that “values for

skewness and kurtosis of between -2 and +2 indicate a reasonably normal distribution."

Within the CTT approach to estimating reliability, Bachman (2004, p. 160) notes four sources of measurement error associated with inconsistent measurement:

1. internal inconsistencies among items or tasks within the test
2. inconsistencies over time
3. inconsistencies across different forms of the test
4. inconsistencies within and across raters.

Bachman (1990, p. 184) suggests that internal consistency should be investigated first since "if a test is not reliable in this respect, it is not likely to be equivalent to other forms or stable across time." Weir, (2005a, p. 31) notes that "the use of internal consistency coefficients to estimate the reliability of objectively scored formats is most common and to some extent this is taken as the industry standard." Kuder-Richardson-20 and Cronbach's alpha internal consistency estimates are the most commonly used (Brown, 2002). Cronbach's alpha is more flexible than KR-20, as it is applicable to both dichotomously scored items and polytomously scored items (Brown, 2002; Bachman, 2004, p. 163).

There are of course caveats associated with the interpretation of internal consistency estimates. Cronbach's alpha, as with all internal consistency estimates, provides an estimate of the reliability of the test scores from one administration of a test, and is not an indication of an inherent statistical property of the test itself (Alderson et al, 1995; Brown, 2002; Bachman, 2004; Weir, 2005a). When considering estimates of reliability it is important to be aware of the characteristics of the sample of test takers from which they were derived. Several important criticisms have also been leveled at NR internal consistency estimates. These estimates tend to be depressed when the sample of test takers has a limited range of ability, as with tests such as EIKEN or Cambridge Main Suite tests, which are targeted at specific proficiency levels (Bachman, 2004; Jones, 2001; Saville, 2003; Weir, 2005a). The same can occur when the test consists of a variety of task types targeting different aspects of the same construct which may not correlate highly, but are nonetheless relevant to the comprehensive coverage

of the TLU domain (Alderson et al, 1995; Bachman, 2004; Brown, 2002; Khalifa & Weir, 2009; Jones, 2001; Saville, 2003; Weir, 2005a).

All of the main standards guiding professional practice in language testing make reference to the need to report reliability indices appropriate for the use of the test (for example, AERA et al, 1999; Joint Committee on Testing Practices, 2004; EALTA, 2006; ILTA, 2007), but little specific advice is given on the degree of such indices which should be expected. This is not surprising given that they are generally aimed at users dealing with assessments designed for a wide range of purposes. As Bachman and Palmer (1996, p. 135) caution: "The most important consideration in setting a minimum acceptable level of reliability is the purposes for which the test is intended."

Some very general guidelines have been proposed as broad benchmarks. Khalifa and Weir (2009) suggest that "the minimum KR-20 estimate for a test is normally set at 0.7 though for a multi-item high-stakes test with a normally distributed population acceptable standards are raised to .8 or .9". This range is broadly consistent with Frisbie (1988) who suggests that: "the reliability coefficient should be at least .85 if the scores are the only available useful information." Kaftandjieva (2004, p. 22) offers the recommendations in Table 2.3, based on the number of classification decisions that will need to be made. They are derived from a statistical index of separation which identifies "the number of statistically different performance strata that the test can identify in the sample" (Wright, quoted in Kaftandjieva, 2004, p. 22).

Table 2.3 Reliability recommendations Kaftandjieva (2004)

Number of Levels	2	3	4	5	6
Number of Cut-off Points	1	2	3	4	5
Minimum test reliability	≥ 0.61	≥ 0.80	> 0.88	> 0.92	≥ 0.95

Elsewhere, Weir (2005a, p 29) suggests that "a reliability estimate of .8 is normally considered the minimum acceptable level but we would normally expect something in excess of .9 in tests of importance." Indeed, it is not uncommon to see suggestions that internal consistency for very high-stakes tests needs to be

above .9 (for example Rudner and Shafer, 2001). However a review of actual practice in large scale language tests would suggest that the .8 to .9 range is more reasonable. The technical manual for the Michigan English Language Assessment Battery recommends that "for high-stakes exams such as the MELAB, a reliability figure of 0.80 and above is expected and acceptable." Chapelle et al (2008, p. 283) describe figures of between .87 to .92 for Reading and Listening tests as typical "for tests used to make high-stakes decisions." For Cambridge Main Suite exams, Saville (2003) suggests that internal consistency for a reading test with 40 items is expected to fall within a range of .8 to .85, and for longer tests of more homogeneous grammar and vocabulary item types .85 to .9. Khalifa and Weir (2009, p. 153) report figures for the IELTS Academic Reading test in 2007 ranging from .83 to .86, with an average of .86..

An important measure for interpreting the precision of scores is the Standard Error of Measurement (SEM). The SEM can be used to estimate a score band, or *confidence interval*, around observed scores within which the examinee's "true score" is likely to fall (Bachman, 1990, p 200; Bachman, 2004, pp. 172-174). Bachman (2004, p 173) notes that by convention confidence intervals based on 68%, 95%, and 99% degrees of confidence are the most common. Few guidelines are offered for the interpretation of SEM, as this ultimately depends on the uses and applications of test scores for any particular purpose. Kaftandjieva (2010, p. 20) suggests that "the standard error of the measurement of a good test with average difficulty and consisting of 50 items is usually about three points on the measurement scale." It is important to remember that estimates of SEM are scaled to the measurement scale used in the assessment (Bachman, 2004, p. 172). Estimates of SEM are thus evaluated in terms of the total number of points on the score scale in assessing the magnitude of the estimated fluctuation around examinees observed scores. In terms of reported practice amongst high-stakes EFL exams, TOEFL iBT reports levels of SEM for the Reading and Listening sections, each of which has 30 total scale score points, of 3.35 and 3.20 scale score points respectively (ETS, 2011). Khalifa and Weir (2009, p. 155) report levels of SEM for the 35-item FCE test ranging from 2.30 to 2.45 across operational versions in 2007, which they consider to be "within an acceptable

range." Estimates for the MELAB (Cambridge Michigan Language Assessment, 2012, p. 5) range from 3.73 to 4.48 across operational forms of the language knowledge and reading test on a score scale ranging from 15-100.

The next measure to be reviewed is a commonly used indicator of item discrimination: the point-biserial correlation. The point-biserial is a variation of the Pearson product-moment correlation coefficient measuring the correlation between each dichotomously scored item and the total scores on the test (Bachman, 2004, Crocker and Algina, 1986). The point-biserial, often represented by the symbolr_{pbis}, is interpreted in the same way as other correlation coefficients. Table 2.4 provides an overview of some recommendation in the literature, including from technical and test manuals.

Table 2.4 Recommendations for interpreting point-biserial in item analysis

Level of r_{pbis}	Source
.30	Milanovic (2002, p. 42)
.25	Khalifa and Weir (2009, p. 145)
.25	Pisa Technical Manual (2006, p. 147)
.20	Pisa Technical Manual, 2012, p. 149)
>.3= good, > .1= fair <.1 = poor	Office of Educational Assessment, University of Washington
.15	California English Language Development Test, 2009-2010, p 29
.15	Technical Documentation for the Maryland High School Assessment and Modified High School Assessment, p. 76

Although it is not uncommon to see benchmarks used as cutoffs below which items are discarded, many recommendations instead suggest that items below a certain level are flagged for further review. For example the technical report for the Maryland High School Assessment notes that "values less than 0.15 were flagged . . .and deserve careful consideration by ETS staff and MSDE before including them on future forms." Many of these individual benchmarks are derived from recommendations for interpreting a range of values originally suggested by Ebel and Frisbie, and shown in Table 2.5 (adapted from PISA Technical Manual, 2012, OECD, 2014, p. 149).

Table 2.5 Recommendations on point-biserial from Ebel & Frisbie (1986)

Value of r_{pbis},	Recommendation
> 0.39	Excellent Retain
0.30 – 0.39	Good Possibilities for improvement
0.20 – 0.29	Mediocre Need to check/review
0.00 – 0.20	Poor Discard or review in depth
< -0.01	Worst Definitely discard

The primary decision of interest for test takers taking the EIKEN test is the pass/fail decision leading to certification at the level of proficiency targeted by the grade of the test they have taken. It is necessary, then, to consider the reliability from both norm-referenced (NR) testing and criterion-referenced (CR) testing perspectives. As Brown and Hudson (2002, p 168-169) note: "Whereas NRT results are concerned with *relative decisions* (i.e decisions based on their standing among examinees relative to each other, such as admission or placement decisions), a CRT approach is primarily concerned with the consistency of *absolute* decisions." Brown and Hudson (2002, p. 169) prefer the use of the term *dependability* in relation to CR approaches to estimating reliability. Bachman (2004, p. 193), however, uses the term reliability for both kinds of indices, and the latter approach is followed here.

The first two indices considered are the agreement coefficient (p_o) and the kappa coefficient (κ). Both of these indices are also referred to as *threshold loss agreement indices* (Bachman, 2004, pp. 199-202; Brown & Hudson, 2002, p. 169). They are designed to measure the consistency of classification decisions by estimating the number of test takers classified in the same way in two administrations (Bachman, 2004, pp. 199-202; Brown & Hudson, 2002, pp. 169-170; Cizek & Bunch, 2007, pp. 308-309; Subkoviak, 1988, p. 47). The formula for p_o is

Formula 2.1 $p_o = \frac{a+b}{N}$

a is the number of test takers classified as masters/passing

b is the number of test takers classified as nonmasters/failing

N is the total number of test takers

The kappa coefficient adjusts this index to take account of the fact that some consistent classifications occur by chance and uses the following formula:

Formula 2.2 $\kappa = \frac{p_o - p_c}{N}$

p_o is the agreement coefficient

p_c is the proportion of classification agreement due to chance

One disadvantage of these two approaches is that they require two administrations of the same, or parallel, measures, which is often logistically difficult or simply not possible. However, Subkoviak (1988) provides a method for estimating approximate values of p_o and κ from tables which require only information readily available from a standard statistical analysis for a single administration. This method, along with Subkoviak's tables, is reproduced in Brown and Hudson (2002, pp. 171-174) and Cizek and Bunch (2007, pp. 309-312). In order to derive p_o and κ from the tables, a reliability estimate is needed, and the statistic Z needs to be calculated. Subkoviak (1988, p. 48) describes Z as "the cutoff score of the test described as a standardized score." The formula for Z is:

Formula 2.3 $Z = \frac{C-M-0.5}{S}$.

C is the cutscore of the test

M is the mean of the test scores

S is the standard deviation of the test scores

Both indices are interpreted as "the proportion of correct mastery/non-mastery decisions that would be made at a particular cutpoint," Bachman (2004, p. 202), As Bachman notes, however, "p_o estimates the proportion of agreement that happens for whatever reason… κ, on the other hand

considers only the measure itself as a source of agreement." In interpreting these values, it needs to kept in mind that κ will always return lower values than p_o. The value of κ in terms of real-world impact on the consistency of classification decisions across administrations is a more abstract concept. The value of p_o is more interpretable as it quantifies the proportion of consistent decisions in absolute terms, and can be interpreted as the proportion of classification decisions which would be made consistently across repeated administrations. Subkoviak (1988, p. 52) suggests that for p_o "tests used to make serious decisions should be sufficiently long to guarantee an agreement coeffieient exceeding .85." For kappa, he suggets the range for similarly high-stakes test should be in the range .6 to .7 (Subkoviak, 1988, p. 53).

The threshold loss agreement indices treat all incidences of misclassification as equally important, regardless of how far a misclassified test taker's score is from the cutscore (Bachman, 2004, pp. 199-202; Brown & Hudson, 2002, p. 170). Another approach to estimating the consistency of classification decisions made by a test treats misclassification decisions further from the cutscore as more serious and results in indices referred to as *squared lost agreement indices* (Bachman, 2004, p. 203; Brown & Hudson, 2002, p. 193). Both Bachman (2004) and Brown and Hudson (2002, p. 195) present two squared loss agreement indices—phi lambda and kappa squared—but Brown and Hudson (2002, p. 195) recommend using the former over the later as phi lambda (ϕ_λ) "is much more firmly linked with criterion referenced test theory." Brown and Hudson (2002, p. 195) show how to derive this statistic by first calculating the relevant variance components required. In cases such where only descriptive statistics from the analysis of scores are available, a short-cut method of estimating ϕ_λ is described by both Bachman (2004, p. 203) and Brown and Hudson (2002, p. 196). The formula for deriving ϕ_λ in this way is presented below:

Formula 2.3 $\phi_\lambda = 1 - \frac{1}{K-1}\left(\frac{M_p(1-M_p)-S^2{}_p}{(M_p-C_p)^2+S^2{}_p}\right)$

K= number of items

M_p = mean proportion correct score

$S^2{}_p$ = Standard deviation of the proportion-correct scores

2.4.2 Vertical scaling

As discussed in Chapter 1, the research questions are designed to focus on validation for three claims which are central to the uses and interpretations of the EIKEN testing program as a set of level-specific tests targeting a common construct of general English language proficiency. RQ2 is designed to elucidate the empirical level of difficulty underlying the common frame of reference within which all of the level-specific tests are assumed to belong. The methodology used to investigate RQ2, vertical scaling, is defined by Kolen and Brennan (2004, p. 372) as a procedure in which "tests that differ in difficulty, but are intended to measure similar constructs are placed on the same scale." This definition parallels that by Harris (2007, p. 233) as "the process of linking different levels of an assessment, which measure the same construct, onto a common score scale." Kolen and Brennan (2004, pp. 3-4) note that this process is also referred to as "vertical equating" but they prefer to distinguish between vertical scaling and equating, reserving the latter for the process of adjusting "scores on test forms that are built to be as similar as possible in content and statistical characteristics." Vertical scaling in the context of language testing, and in particular in relation to linking exams to the CEFR, is discussed by North and Jones (2009, p. 2), who emphasize the benefits of creating a bank of items calibrated to a common scale through the use of IRT in order to provide "a single measurement scale upon which we can locate items by their difficulty and learners by their ability, as well as criterion levels of performance."

A number of methodological issues need to be resolved when planning vertical scaling. Kolen and Brennan (2004) and Tong and Kolen (2010) discuss three data collection designs: a *common item design* in which students take a test

with items appropriate to their level but which also includes items from adjacent levels; the *scaling test design* in which students across all levels take a separate composite test containing items from across all levels to be scaled; and the *equivalent groups design* in which examinees at a particular level are randomly assigned to take either the test for their level or a test for an adjacent level.

In terms of scaling methods, options exist within both Classical Testing Theory (CTT) and Item Response Theory (Kolen and Brennan, 2004; Harris, 2007; Young, 2006). However, Probabilistic IRT models provide a methodology to overcome the inability of CTT analysis to distinguish between item and person effects when considering changes in test difficulty across different test forms and test populations (Bachman, 2004). The IRT approach thus provides potential advantages in test development and item analysis in general. This includes allowing for what Kolen and Brennan (2004) describe as a pre-equating approach, in which items are calibrated to the common scale in pretesting using anchor items, and then stored in an item bank. When new test forms are constructed, they can be built to pre-set difficulty levels from already calibrated items in the item bank, avoiding the need for further equating or vertical scaling of new test forms (North and Jones, 2009; Kolen and Brennan, 2004). In the context of state-based testing programs in the United States, Reckase (2010, p. 4) notes that "a review of the literature identified no state testing programs that used other than IRT approaches for forming vertical scales." Looking at the same context, Briggs and Weeks (2009, p. 4) also note that IRT-based approaches to vertical scaling are the "predominant method employed by commercial test developers."

Within an IRT-based approach, the choice of a 3-parameter model, 2 parameter model, or a Rasch-based model also needs to be addressed (Harris, 2007: Kolen and Brennan, 2004; Young, 2006). Bachman (2004, p. 141) describes the main distinction between these models in the following way:

> [The models] vary in terms of the number of parameters they include: a 1-parameter IRT model, often referred to as the Rasch model, includes only a difficulty parameter…a 2-parameter IRT model includes a difficulty parameter and a discrimination parameter…while a 3 parameter IRT model includes, in addition to parameters for difficulty and

discrimination, a pseudo-chance, or guessing parameter.

In practical terms, the choice of model has serious implications, as the 2- and 3-parameter models require much large sample sizes to derive stable estimates than the Rasch model (Harris, 1989; Kolen and Brennan, 2004). Considerable debate has ensued around the appropriacy of the three models, both for vertical scaling and in educational measurement generally. McNamara and Knoch (2012) provide an overview of how this debate developed over several decades from the 1980s in the field of language testing, noting that the Rasch model is now generally accepted and widely used in the field. In a comprehensive overview of the state of the art in language testing at the turn of the century, Bachman (2000, p. 5) also noted that "the Rasch model, in its various forms, is still the most widely used in language testing." In relation to vertical scaling, Harris (2007) notes that previous research has provided conflicting results regarding the superiority of particular models. State-based testing programs in the United States use either 3-parameter or Rasch based approaches to vertical scaling (Briggs and Weeks, 2009; Reckase, 2010). Briggs and Weeks (2009) compared these two approaches, noting that "in practice many states" used Rasch based approaches. They concluded that there was nothing to recommend one model over the other, but cautioned that interpretation of growth will be different depending on the model used. McNamara and Knoch (2012, p. 9) in fact suggest that one of the original sources of criticism of the Rasch model, "the deliberate simplification of the assumption of equal discrimination of items" is actually one of the model's strengths as it "permits exploitation of the property of specific objectivity in Rasch models, which means that the relationship of ability and difficulty remains the same for any part of the ability or difficulty continuum," a property that is "essential for … tests to be vertically equated."

If an IRT-based scaling method is selected, estimation procedures must also be selected. Young (2006) describes three possible estimation procedures: *concurrent calibration*, *fixed estimation*, and *separate estimation*. Pomplun, Omar, & Custer (2004) note a number of studies which have arrived at the conclusion that the choice between concurrent and separate estimation makes little substantial

difference to parameter estimation, a conclusion supported by the results of a simulation study by Paek, Young, and Yi (2008).

The widespread use of Rasch-based scales is language testing is also reflected in examples of vertical scaling for language assessment. The common measurement scale linking the five-level set of tests in the Cambridge Main Suite exams utilizes a Rasch-based vertical scale (Cambridge English, n.d.). The English Access for ELLs tests produced by the WIDA Consortium, which are designed to measure academic language proficiency across the spectrum of grade levels in United States schools (K-12), also uses a Rasch-based vertical scale (Kenyon et al, 2011). The study by Kenyon et al (2011) gives a detailed account of the use of the Rasch model to create vertical scales for a series of tests at different proficiency levels in listening, reading, writing, and speaking, and is thus particularly instructive for this study. The Language Training and Testing Centre (LTTC) in Taiwan has undertaken vertical scaling studies using Rasch to link the set of level-specific EFL proficiency tests referred to as GEPT (Wu, 2012). Of particular relevance to RQ2 is the study by Brown et al (2012) which utilized the Rasch model to explore the feasibility of linking the upper grades of the EIKEN tests in order to facilitate comparisons with a single external proficiency test spanning multiple levels in a criterion--related validity study. The Brown et al (2012) study provided an important reference point for the selection of methodology for the current study. However, there are also important differences, which will be discussed under the methodology section of Chapter 4.

The Common European Framework of Reference (CEFR) represents a special case of vertical scaling in language testing. The CEFR proficiency framework is not itself a test, but provides a set of proficiency descriptors calibrated to a vertical scale to which examinations can be linked. The CEFR does not prescribe any specific set of tests or indeed testing procedure (Council of Europe, 2001, 2009). The proficiency descriptors used to illustrate the increasing levels of proficiency within the framework were calibrated using vertical scaling methodology that employed the Rasch model and a common-item non-equivalent groups design to link sets of can-do descriptors administered to teachers in sets of questionnaires of increasing difficulty. (Council of Europe, 2001; North, 2000;

North and Schneider, 1998) This enabled the collection of empirical data on a large bank of descriptors calibrated to a common vertical scale.

Once a vertical scale has been created, various statistics are available for evaluating the effectiveness of the scaling procedures selected. Kolen and Brennan (2004), Tong and Kolen (2007, 2008), and Young (2006) provide useful overviews of the statistics available for this purpose.

An important follow-up issue is the development of an appropriate strategy for the maintenance of those scales (Kolen & Brennan, 2004; Harris, 2007; Tong & Kolen, 2008). Tong and Kolen (2008) investigated strategies for maintaining vertical scales developed with the Rasch model by utilizing real data from a large-scale state testing program for English Language Arts and Mathematics. They note two principal strategies. The first, horizontal equating, involves new tests at each grade level in subsequent years being linked *horizontally* back to the original vertical scale through the use of common linking items which were used in the original scaling. The linking items are only inserted into the appropriate grade level in subsequent years, meaning the vertical links in the scale are only created once; hence the description of equating in subsequent years as horizontal. The alternative is to construct a new vertical scale in each subsequent year and to equate the new vertical scale to the original, baseline scale. Tong and Kolen (2008, p. 14) concluded that, for the data in their study:

> The horizontal equating approach is the more straightforward and is easier to apply in practice. The multiple vertical scales approach is more complex; it also demands vertical linking items be administered in multiple years. Furthermore, decisions need to be made on how horizontal equating can be conducted to link the vertical scales. It appears that in the present context of linking scales from two adjacent years, the horizontal equating approach might be preferable because it produces results similar to those for multiple vertical scales but is easier to implement.”

The overview of literature relevant to the selection of procedures for the application of vertical scaling to investigate RQ2 for the EIKEN tests highlights

the lack of a consensus model in terms of scaling methods, linking designs, and other aspects including the choice of IRT models and estimation procedures (Harris, 2007; Kolen and Brennan, 2004; Young, 2006). Indeed Harris, (2007, p. 241) notes that "no single combination of methodology, data collection design, and sample has been found to be superior to others to a generalizable extent, and most designs seem to work well in at least some of the settings." As with many aspects of language testing, the selection of methodology does not involve the search for "the right" way, but rather requires a principled approach to considering the purpose of scaling within the relevant features of the context involved, and arriving at a balance of the sometimes competing demands of those features. The methodology selected for carrying out the vertical scaling for RQ2 will be discussed further in Chapter 4.

2.5 Research Question 3: Linking Examinations to the CEFR

The third research question falls under the paradigm of criterion-referenced validity in the socio-cognitive model shown in Figure 2.3. As noted in Section 2.2.4, more recent interpretations of the model (O'Sullivan, 2011; O'Sullivan, 2012; O'Sullivan, 2015a, O'Sullivan and Weir, 2011) have suggested placing criterion-related validity within the context of scoring validity, one of the aspects posited by them as constituting the core of test validity. Establishing how the seven grades of the EIKEN tests relate to a clear, accessible description of proficiency is directly relevant to the overall aim of this study, which is to derive evidence that can contribute to the creation of a comprehensive, clear, and coherent validity argument. As noted in Chapter 1, the changing context of uses and interpretations for the EIKEN grades make RQ3 a vital part of that validity argument. The proficiency levels targeted by the EIKEN tests should have meaning derived from both the comprehensive description of criterial features identified in RQ1, and in terms of how those features are relevant to interpretations of the EFL proficiency construct beyond the implicit assumptions on which the original test development was based. For these reasons, and taking into account the subsidiary criteria for selection of instruments in the study identified in Chapter 1—relevance, transparency, interpretability,

comparability—the Common European Framework of Reference for Language: Learning, teaching, assessment (CEFR) was selected as the basis for investigating RQ3. While the extensive body of literature documenting the experience of testing programs in varied contexts, including Asia, of linking their tests to the CEFR underpins its choice as the focus of RQ3, this choice brings with it potential criticism. While the literature on linking to the CEFR and potential caveats regarding its use in any context are discussed in more detail below, it is first necessary to address the issue of whether using a proficiency framework putatively developed for one context, Europe, is justified in another very different context, Japan.

In answering this question, it is important to emphasize that the CEFR was chosen specifically because it is an external criterion—both to the set of tests which is the focus of the study and also to the local context in which the tests were developed. As noted above, it is specifically the changing context of use for the test which has driven the focus of RQ3. That changing context, outlined in Section 1.2.3, has placed demands on the test developer to communicate effectively with stakeholders outside the original context of use. At the same time, this changing context means that for local educators too, an important claim requiring validation is that the seven-level set of tests hold wider relevance as explicit benchmarks of proficiency with transparent interpretations outside the implicit assumptions built through close interaction with the educational context in which the tests were first developed. From this perspective, it should be clear that the use of the CEFR in RQ3 is not intended to recommend that the CEFR be imposed upon the context of Japan as a fixed standard to which the EIKEN tests, or any other EFL tests in that context, must be mapped. Indeed, as Section 1.3 detailed, the EIKEN tests have derived much of their utility through the close interaction of the testing program with the local educational community and context, and the establishment of the EIKEN testing program predates the publication of the CEFR by more than three decades. A complete and direct correspondence between such a locally embedded program and the CEFR would not be expected, or indeed desired. Differences that may be discovered in the process of examining the relationship may be as instructive as similarities. What is

useful to remember is that it is the existence of a common framework such as the CEFR, designed to provide a common set of reference points to facilitate such comparison, that can lead to such insights. Indeed, as Milanovic and Weir (2010) emphasize in a foreword to an important volume of case studies on aligning tests to the CEFR, the authors of the CEFR have consistently warned against the use of the CEFR in a prescriptive way to dictate practice in any particular context, something emphasized again by North, Martyniuk and Panthier (2010) in the same volume. Milanovic and Weir (2010, p. x) refer to the CEFR as deliberately "underspecified and incomplete," a feature which they suggest "makes it an appropriate tool for comparison of practices across many different contexts in Europe and beyond." This is a theme also emphasized by Davidson and Fulcher (2007. p. 232) in the context of test development. They suggest that the CEFR is underspecified for any particular local context, but this is in fact its advantage, as it can be used as a "series of guidelines from which tests (and teaching materials) can be built to suit local contextualized needs."

The various cautions and limitations suggested for the CEFR in general are described in more detail below. They are mentioned here first to highlight the fact that the use of the CEFR as a prescriptive tool to impose a particular definition of proficiency on a local context would run counter to the intentions of the original authors regardless of whether such a misuse took place within any of the diverse contexts inside Europe in which the CEFR is being used or indeed outside in cases such as Japan. A full discussion of the limitations and advantages of the CEFR is beyond the scope of this study. The following review of the literature is designed to underscore the usefulness of the framework for the purposes of RQ3 by demonstrating its widespread adoption and the body of literature describing the experience of testing programs inside and outside Europe in linking tests to the CEFR, while framing the caveats and limitations that need to be kept in mind when doing so.

The Common European Framework of Reference (CEFR) was published by the Council of Europe in 2001 following a 10-year development process (Morrow, 1994; North, 2010). However, the foundations on which it rests go further back, with Alderson (2005), Morrow (2004), and Trim (2010) suggesting

that the CEFR represents over 40 years of research and development in language education in Europe. Although the framework is intended to encompass more than just assessment, the failure of many commentators to move beyond the proficiency scales contained within the full document has been criticized (e.g. Morrow, 2004), As North (2007) notes in reference to the full name of the CEFR: "Assessment is in third place; the language testing profession is a service industry to support teaching and learning." Nonetheless, Morrow (2004, p. 8) recognizes the importance of the scales to the framework's descriptive system, leading him to refers to the Common Reference Levels as being "at the heart" of the CEFR.

These levels and the Illustrative Scales which define them were developed in two major projects carried out in Switzerland in 1994 and 1995 (North, 2000; North & Schneider, 1998). North (2000) defines the Common Reference Levels as a user-oriented scale of language proficiency according to the three-way distinction made by Alderson (1991). According to North (2000), what distinguishes the development of the CEFR from previous descriptive scales of proficiency, such as the American Council for the Teaching of Foreign Languages (ACTFL) scale, is the use of Rasch analysis to empirically validate the allocation of descriptors to difficulty levels. The calibrated descriptors are used to define six broad levels of language proficiency across a total of 54 separate scales describing communicative activities, strategies, and communicative language competences. A number of authors have suggested there are significant shortcomings in the CEFR. Weir (2005b) and O'Sullivan and Weir (2011) claim that a lack of comprehensiveness in coverage and the absence of an explicit theory of language mean that it would be inappropriate to apply the scales as they are to language test development. Alderson et al (2006) note there are gaps and inconsistencies in the use of specific terminology across levels and different scales. Davidson and Fulcher (2007, p. 232) suggest that the CEFR "does not detail particular contexts in which it is to be used, and so lacks the necessary detail on which to build test specifications." Morrow (2004, p. 8) also notes that there is significant ambiguity in some of the terminology used within the can-do proficiency descriptors, asking, "What are the 'main' points…Is what is 'clear' in my opinion, 'clear' in yours. . . . And how many is most?" McNamara (2007, p. 7) has also

criticized the CEFR from a critical language testing perspective, noting that "the CEFR gives power to those who mandate that outcomes be reported in terms of this set of claims—governments, ministries, authorities, funding agencies, testing agencies." Fulcher (2010, p. 15) also cautions against the "reification" of the CEFR and its use "as a tool for harmonization of language teaching."

North (2007) has responded that the CEFR is in fact an attempt to describe real-life language use, and is not a constructor-oriented scale for test development, something that Alderson (2005) also notes the developers of the CEFR have emphasized. North, Martyniuk and Pantheir (2010, p. 13) have responded to the criticisms with a list of seven principles as guidelines, noting that the CEFR is intended to be "context neutral" and thus needs to be "applied and interpreted with regard to each specific educational context." They further note that while it "attempts to be comprehensive . . . it cannot claim to be exhaustive," recognizing the need for further development of the framework. Perhaps most relevant to the usefulness of the framework for the EIKEN tests in the context of this study is the principle that "the CEFR offers a common language and point of reference as a basis for stakeholders to reflect upon and critically analyze their existing practice and to allow them to better 'situate their efforts' in relation to one another." Indeed Davidson and Fulcher (2007) have suggested that the lack of specificity is in many ways an advantage for test developers, as it allows them to flesh out the framework with aspects of their local context. They recommend that researchers treat the CEFR as "a series of guidelines from which tests (and teaching materials) can be built to suit local contextualized needs" (p. 232).

Despite some obvious shortcomings, Alderson (2005) suggests that the wide and enthusiastic adoption of the CEFR by language educators points to its potential practicality and usefulness. A number of testing programs have undertaken research to demonstrate compatibility with the CEFR levels (for example, Bechger, Kuiper, & Maris, 2009; Kecker & Eckes, 2010; O'Sullivan, 2008, 2010, 2015b; Papageorgiou, 2007; 2009). Of relevance particularly to this study are examples of applications of the framework to contexts outside Europe, such as Tannenbaum and Wylie (2005, 2008), Wu (2012), and Wu and Wu (2010). In Japan the CEFR has also been used as the basis for test development (see

Nakatsuhara, 2014 and Weir, 2014), and formed the basis for a locally developed adaption, CEFR-J, for use in formal educational contexts (Negishi, Takada, Tono, 2012). Many of these projects reflect the principles recommended by North et al (2010) in that they have approached the process of linking to the CEFR from a critical perspective, each noting recommendations for adapting the framework itself or the recommended processes for linking to it (see for example O'Sullivan, 2008 and Wu and Wu, 2010).

Figureas et al (2005, p. 266) describe the development of a manual in response "to the need for guidance to assist examination providers in relating their examinations to the CEFR." This resulted in a Preliminary Pilot version of a manual being published by the Council of Europe in 2003, and a subsequent revised edition in 2009. All of the studies noted above which have undertaken projects to link an exam to the CEFR have made some reference to one or both versions of the Manual. The Manual (Council of Europe, 2009) lists four steps in the process of building an argument to justify a claim of linkage to the CEFR: *familiarization, specification, standardization, and validation*. The Manual is supported by a series of Reference Supplements dealing with technical issues related to linking examinations to the CEFR, including Reference Supplement B: Standard Setting by Kaftandjieva (2004).

Kaftandjieva (2004, p. 1) describes standard setting as being "at the core of the linking process." Standard setting is described by Cizek (1993, p. 100) as: "The proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance." In the case of linking to the CEFR, such a number is the point on a test score scale at which a test taker can be considered to have demonstrated a level of proficiency described in one of the CEFR levels.

Among the many standard-setting methods available to practitioners, the Basket Method has been widely employed across Europe in relation to linking to the CEFR (Kaftandjieva, 2009, 2010). However, in the wider international context, particularly in the United States, the Angoff method, or modifications collectively referred to as the Modified Angoff method, is often cited as the most frequently used method (Cizek & Bunch, 2007; Cohen, Kane, & Crooks, 1999). It is also one

of the most widely researched, with papers comparing it to other methods (Bowers & Shindoll, 1989; Livingston & Zieky, 1989) or investigating modifications of the Angoff method (Clauser et al, 2009; Hurtz & Auerbach, 2003; Norcini et al, 1987). Although the Angoff method has been criticized as placing too great a cognitive burden on participants (Cizek & Bunch, 2007), studies have shown it to be robust (Plake, Impara, & Irwin, 2000) and less prone to statistical bias than other methods (Reckase, 2006). Zieky (2001) has also refuted claims that the judgment task is too cognitively demanding for standard-setting panelists. The Angoff method has not only remained one of the most widely used methods, but has been employed in a number of studies linking exams to the CEFR (e.g. Tannebaum & Wiley, 2005, 2008; O'Sullivan, 2008; O'Sullivan, 2015b).

Standard setting has a long history of use in educational measurement in the United States (Kaftandjieva, 2010; Papageorgiou, 2010). However in relation to the situation in Bulgaria, Kaftandjieva (2010, p. 23) describes the most commonly used methods for setting cutscores and passing standards on exams as: "tradition, authority and the Goldilocks method," with the latter referring to an arbitrary process of setting a cutscore such as 80 percent simply because "70 percent is too little and 90 percent is too much." However, it would not be unreasonable to suggest that a similar lack of familiarity with standard-setting methodology was also common in other European countries prior to the introduction of the CEFR. It was indeed the perceived lack of familiarity with procedures for setting cutscores and linking examinations which led to the production of the Manual (2003, 2009) and Reference Supplements (2004). A similar process of growing familiarization with the principles of standard setting in conjunction with exposure to the CEFR can also be noted for Japan (see Chapter 5 for details of the familiarity of participants in standard-setting panels in this study).

Kaftjandjieva (2010, p. 29) gives a comprehensive account of standard-setting methods, describing 62 documented methods, but cautioning that even this list "is not complete." Cizek and Bunch (2007), Hambleton et al (2000), Kane (1998), and Livingston and Zieky (1982) also provide useful

overviews and descriptions of the most prominent methods. These methods are often separated into one of two categories—*student centred* or *test centred*—based on a classification system originally suggested by Jaeger in 1989 (Kane, 1998; Kaftandjieva, 2010). Kane (1998, p. 131) describes the two approaches in the following way:

> In the test-centered methods, the judges review the tasks or items in the test and decide on the level of performance on these tasks that would indicate attainment of the performance standard. . . . In the examinee-centered methods, performances of real examinees are evaluated relative to the performance standard, and the test scores of these examinees are used to set the cutscore. For example, in the borderline-group method, the judges identify examinees who just meet the performance standard and the cutscore is set equal to the median score for these examinees.

It is important to note that all forms of standard setting involve some form of human judgment (Cizek and Bunch, 2007; Kane, 2001b), and so the selection of judges is a crucial part of the process. In terms of the criteria for selecting judges, or raters, to participate in standard-setting panels, Jaeger (1991, p. 4) suggested that "expert judges should be well experienced in the domains of expertise we demand of them." In terms of the number of judges, Raymond and Reid (2001) note a wide range in the literature, ranging from "admissions" of 5 to recommendations of 15 to 20. Hurts & Hertz (1999, p. 885) applied generalizability theory to eight studies using the Angoff method and concluded: "10 to 15 raters is an optimal target range."

A number of criteria have been suggested for evaluating the results of standard setting (Cizek et al, 2004; Cizek & Bunch, 2007; Hambleton, 2001; Kaftandjieva, 2010). Cizek and Bunch (2007, pp. 59-63) describe three main categories of evidence. *Procedural validity evidence* involves a description of the processes employed, including the training for participants, the degree of correspondence of the procedures to the requirements of the methods used, and also includes feedback from participants. *Internal validity evidence* looks at the

accuracy and consistency of the results of the standard-setting methods used, including the degree to which participants converge toward a common standard over the course of standard setting rounds. *External validity evidence* includes comparison of results obtained from other standard-setting methods and other sources of information. In terms of strengthening the plausibility of results obtained from standard setting, Kane (2001b, p. 75) recommends replicating standard setting with different methods, suggesting that using different methods and participants "would provide an especially demanding empirical check on the appropriateness of the cutscore." Cizek and Bunch (2007), take the opposite view, warning that there is no consensus methodology for reconciling the different cutscores likely to be generated by different methods. The use of multiple methods was in fact taken up as an important approach to external validation in this study, and is described in more detail in Chapter 5. The approach taken to ameliorating the concerns of Cizek and Bunch (2007) regarding this approach is discussed in Section 5.2. Cizek and Bunch (2007, p. 63) also describe the *reasonableness* of the decisions made as an important criterion for evaluating the decisions made through standard setting, and the interpretation of reasonable decisions in the context of this study is also discussed further in Section 5.2

While adopting a clearly documented, principled approach to collecting and analysing data can inform cutscore decisions, Cizek and Bunch (2007) caution that the results "are seldom, if ever, purely statistical, psychometric, impartial, apolitical, or ideologically neutral activities." However, Cizek and Bunch (2007) also emphasize that decisions taken within the context of educational measurement always involve, to some degree, evaluative judgments by those tasked with making those decisions. Standard setting does not remove that burden or the difficulties inherent in carrying out those responsibilities. There will be no magic statistical procedure, technique or software application which will remove the need for principled decisions to be taken in relation to setting cutscores. This situation reflects the discussion of validity and validation at the beginning of this chapter, in which it was stressed that validity is not an absolute decision, but a matter of degree established through a thorough evaluative argument. Standard setting should certainly be viewed in this light, and decisions

should be made within a clear framework of reference and an understanding of the goals and contextual constraints under which the process is carried out. Linking to the CEFR does provide a principled approach to investigating the relationship between an examination and a descriptive proficiency framework that can act as an external criterion to the test itself. Gathering evidence to support claims of relevance to this framework would add meaning to the interpretation of the seven grades of the EIKEN testing program. In addition to providing a means of investigating RQ3, establishing such a link would add to the understanding of local test users of what certification at a particular grade of this widely used local measure of EFL ability means in relation to internationally used standards, and provide a common frame of reference to facilitate communication about the EIKEN tests with language teachers and learners outside Japan. Nonetheless, in evaluating the standard setting studies undertaken as a part of investigating RQ3 in Chapter 5, it will be prudent to bear in mind Kaftandjieva's (2004, p. 31) caution that:

> There is no gold standard, there is no true cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence. In three words, nothing is perfect.

Chapter 3 RQ1: Criterial Features of Test Tasks at Each Grade

3.1 Introduction

This chapter addresses Research Question 1: To what extent and in what ways are the seven levels of the EIKEN testing framework qualitatively different in terms of key contextual and cognitive parameters? Establishing these core aspects of contextual and cognitive aspects of validity are also central to the approach of the socio-cognitive model which has been taken to guide the focus and structure the analysis in the process of developing a coherent, comprehensive validity argument for the uses and interpretations of the EIKEN tests.

3.2 Methodology

3.2.1 Overview

Investigating RQ1 involves looking at aspects of what the socio-cognitive framework of language testing development and validation refers to as contextual and cognitive aspects of validity, with the particular categories of criterial features referred to as parameters. Contextual validity parameters deal with many aspects traditionally addressed under the issue of content validity, and it is here that the study looks at core aspects of the actual content of the tests. As noted in Chapter 2, while the distinction between contextual and cognitive is to some extent an artificial device, with these features in fact interacting closely, the focus on identifying and investigating the cognitive processes engaged by language test tasks is a defining aspect of the socio-cognitive model.

As already noted in Chapter 1, the ability to develop explicit criteria for the inclusion of contextual and cognitive parameters in test specifications was an important concern and impacted on the selection and application of parameters for

inclusion in the study. The features selected were required to be operationally accessible to item writers and content developers in the ongoing process of maintaining criterial features and distinctions both within and across grades.

The measures chosen for investigation, based on the review of the literature in Chapter 2, Section 2.3—and taking into account the four subsidiary criteria of *transparency, relevance, interpretability* and *comparability*—are shown in Table 3.1. As already noted in Chapter 1, this study focuses on investigating reading as a practical consideration, given the scale of retrospectively investigating multiple test forms from seven to nine years of administration across all seven grades of the EIKEN program.

Table 3.1 categorizes the measures together into various groups. The first important distinction is between those measures which were developed and applied to items and texts through expert human judgment and those which were evaluated through the use of automated analysis software. These two groups will be discussed separately below. Within each of these two main groups, there are then two other distinctions. Firstly whether the parameter applies to the test item or input text on which an item is based. An item defines the action undertaken by the test taker in completing an aspect of a test task and which is scored. For the reading section of the First Stage tests, all items are dichotomously scored, multiple choice response formats. In the case of the short grammar and vocabulary MC gap-fill items in Section 1 of the test, although items and input texts overlap closely, the distinction is still maintained. The item parameters relate to the selection of the correct option from the response alternatives to fill a gap, whereas the text parameters relate to the whole sentence or sentences used as the contextual stem for the item. The distinction is clearer in MCQ-type tasks, such as the long reading passages in Section 4, which may have several items attached to one input text.

A further distinction is then made in Table 3.1 in terms of whether the parameter relates primarily to contextual or cognitive aspects of validity. As already noted, these aspects interact closely. However, it is clearly important to identify explicit cognitive parameters which can distinguish between levels of cognitive demand in the completion of test items and tasks. The three parameters

selected for this purpose in the evaluation and classification of criterial features of EIKEN reading test tasks are thus associated with the items, while the remaining contextual parameters are associated with the input texts (though it worth repeating the caution that while useful for analysis purposes, this classification scheme is not intended to imply an absolute distinction in practice).

This study has made an explicit choice to focus on measures drawn from the literature which lend themselves to the use of both human expert judgment and automated analysis tools to analyze *actual test content.* As outlined in the literature review in Section 2.3, research on the contextual and cognitive features of reading tests undertaken from a socio-cognitive perspective has utilized a number of approaches in addition to the methodology employed in this study. Additional approaches not employed in this study include questionnaires, think aloud protocols and eye-tracking technology to elicit data directly from test takers to identify the processes they engage in while completing tasks. This study took the decision to focus only on the analysis of test content directly, and not to employ procedures for eliciting data directly from test takers. The decision, as with all choices informing the study design—including the decisions to focus only on reading and concentrate on a subset of core aspects of the socio-cognitive model--reflects the pragmatic realities of balancing the scale and scope of this study across the three research questions. As with the other explicit decisions for which a rationale has already been provided, it is not intended to suggest that research methodologies targeting test takers directly are not important. Indeed as the literature review emphasized, it is the potential for operationalization of the socio-cognitive model through a varied range of methodologies, particularly the cognitive processing model of reading, which imbue it with such potential for empirical validation. The decision to select measures and methodology focusing on the analysis of test content—the use of which is also extensively described in the literature review—is premised on two key drivers for the study.

Firstly, as noted in Sections 1.1 and 1.4, an important element of this study from the beginning has been the subsidiary goals, which are intimately connected with the main research questions in terms of determining the study design. An important subsidiary goal from the beginning has been the ability to

relate the findings and methodology involved directly to operational quality assurance and ongoing procedures for the test development teams. As described in Section 1.4, this rationale has underpinned the selection of parameters for tagging test content for both contextual and cognitive parameters. The process, described in more detail in Section 3.2.2 below, involved the selection and refinement of parameters through iterative interaction with focus groups consisting of test production team members, the development of a manual to make the definition of tags transparent to teams, and the refinement of the process through the tagging of large numbers of previously administered test items. The undertaking was extremely resource intensive, but was also necessary to secure funding and support, as explained in Section 1.1, as the results of the study had to be demonstrably applicable to improving operational processes.

A second imperative, scale, also drove the decision to focus on expert judgment and automated analysis of test content. This is also linked to the subsidiary goals of the study of evaluating the usefulness of the model itself. As noted in the literature review, various approaches have been taken to eliciting data regarding the contextual and cognitive features of tasks, many of them including not only analysis of test content but also investigation of processes elicited in test takers by the test tasks. However, these studies have been largely restricted to small-scale investigations. The use of think-aloud protocols and other procedures with test takers directly, by their nature, involve small numbers and are usually restricted to non-operational settings. While the studies described in the literature review have demonstrated the usefulness of the various approaches and the socio-cognitive model, including the various parameters adopted for this study, their generalizability is limited because of the limited scale. This study took the deliberate decision to focus on operationalizing the model with large amounts of operational data to make an important contribution to evaluating the use of the socio-cognitive model on a scale which can provide direct insights and generalize to the operational testing environment and the large amounts of data generated directly from that operational context.

Given these imperatives underpinning the design of methodology for RQ1, and bearing in mind that an important aspect of the study is the integration

of RQ1, RQ2 and RQ3, all of which pose significant demands on resources, it was decided to focus on the methodology described in more detail below to create a comprehensive picture of the contextual and cognitive parameters of a large bank of operational test items. The methodology focuses on expert judgment and automated analysis of that content directly, using procedures which would be amenable to ongoing application to actual test production and quality assurance in an operational setting. As with the caveats noted from the outset in Section 1.1, and reiterated in the rationale for study design decisions throughout, this study is intended to provide a core body of evidence which can contribute to the construction of a validity argument for the tests. It is not intended to be the validity argument. Neither is it intended that the focus taken in this study or the results, however robust, should be used to preclude further data collection to add both breadth and depth to that validity argument. From this perspective, future data collection for an ongoing, dynamic validation agenda could usefully focus on test taker processes directly to provide an extra dimension to compliment the data collected for this study.

Table 3.1 List of measures used to investigate RQ1

Parameter	Expert	Auto-mated	Item	Input text	Con-textual	Cog-nitive
Operation	✔		✔			✔
Key information	✔		✔			✔
Explicitness	✔		✔			✔
abstractness	✔			✔	✔	
Discourse type	✔			✔	✔	
Genre R	✔			✔	✔	
Topic	✔			✔	✔	
Percentage of AWL words		✔		✔	✔	
BNC level for 95% threshold		✔		✔	✔	
Number of words				✔	✔	
Average Sentence Length		✔		✔	✔	
Average syllables per sentence		✔		✔	✔	
Average syllables per word		✔		✔	✔	
Syllables per 100 words		✔		✔	✔	
Flesch-Kincaid Grade		✔		✔	✔	
Lexical diversity (MTLD)		✔		✔	✔	
Lexical diversity (VOCD)		✔		✔	✔	
Announce Goals tokens count		✔		✔	✔	
Attitude marker token count		✔		✔	✔	
Code gloss token count		✔		✔	✔	
Emphatic token count		✔		✔	✔	
Endophoric token count		✔		✔	✔	
Evidential token count		✔		✔	✔	
Hedge token count		✔		✔	✔	
Label stage token count		✔		✔	✔	
Logical connective token count		✔		✔	✔	
Person marker token count		✔		✔	✔	
Relational marker token count		✔		✔	✔	
Sequencing token count		✔		✔	✔	
Topic shift token count		✔		✔	✔	

3.2.2 Parameters tagged through expert judgment

For the purposes of investigating reading texts and items, there are a total of eight parameters in Table 3.1 applied through expert judgment. The text genre is used to

describe features of text types from the TLU domain which EIKEN reading texts can be considered to best reflect. In many cases, particularly at the lower levels, specific genres were not explicitly specified for item writers developing reading texts. This study was used as an opportunity to identify genres that best capture the intended nature of texts as they have been developed, and at the same time, consistent with the goal to be forward looking, to in turn be useful for more explicit future specification of the genres most relevant to each grade.

An initial list of draft parameter tags was created from the review of the literature as described in Chapter 2. Definitions of the parameters and of the choices for tagging items and texts within each parameter were prepared, and these definitions were reviewed by a focus group of content specialists working with each EIKEN grade. As one of the subsidiary goals of this study was to develop item and text specification categories which could be used operationally, it was imperative from the outset to have the input of content specialists and to ensure that the categories chosen indeed met the criteria for selection, particularly relevance, from the perspective of hands-on item developers. In refining the list of possible options for use in each parameter, and in particularly in relation to topic and genre, reference was made to the existing item specifications for the EIKEN grades to identify areas of overlap with the documentation in the literature of the parameters chosen, for example in Alderson et al (2006), Khalifa and Weir (2009), and Wu (2012), and to identify areas specific to the EIKEN grade specifications which may not have been included in those previous studies. A series of meetings with the content specialist focus groups was undertaken to refine these categories to derive a list of options within each parameter which would meet the four criteria mentioned above,

Once a revised version of the parameters, options within parameters and their definitions had been prepared through iterative review and feedback from the focus group, a draft tagging manual was prepared. The focus group employed the manual to tag items and texts from three operational test forms for each grade (one test form from each of the first, second, and third test administrations in one yearly test cycle). The focus group comprised representatives from the content teams for each grade. All members of the focus group tagged all of the reference

forms across all grades, and the tags were reviewed in a series of meetings which entailed discussing discrepancies and resolving differences of interpretation through an iterative process of discussion until a consensus had been achieved. The reference test form sets were then used to derive examples for illustrating the definitions of each parameter and the option tags in the tagging manual, and were also used as training sets for the expert judges used to tag the large number of items and texts involved in the study. Appendix B contains the table of contents of the final manual to give an idea of the scope of information provided to guide the tagging process. As can be seen from the table of contents, each parameter description covers a minimum of three pages of information which included the definition and examples of how the tags were applied in practice on actual EIKEN items from the training sets used by the focus groups.

All items and texts for all operational tests administered for each grade since the latest revision for that grade up to and including the final administration in the 2010 academic year were to be tagged using the manual to provide a large body of data for the investigation of RQ1. Three expert judges who were experienced item writers and content reviewers for the EIKEN tests were trained in the use of the manual and carried out a standardization exercise in which they tagged the reference set of tests which had been previously tagged through consensus by the focus group of content specialists. Any discrepancies were resolved through discussion, and revisions made to the manual to reflect feedback and insights from the external judges. One of the judges was designated as a quality assurance coordinator who would liaise regularly with the author to resolve issues raised by the other judges in ongoing tagging. Each judge was then allocated to a grade reflecting the item content with which they had the most experience. A sample of tags applied by each judge was cross-checked by the quality assurance coordinator and the author, and any discrepancies became the focus of periodic review and discussion sessions between the judges, the coordinator and the author.

3.2.2.2 Analysis

The parameters used for tagging items through expert judgment constitute

nominal categories in terms of scale definition, thus precluding the use of statistical tests appropriate for the ordinal and interval level scale data obtained through the automatic software analysis described in the next section. In addition, as already noted in chapter one, the number of items and texts in a test form for each grade differs, and the number of tests administered in a test cycle across one academic year differs between the advanced Grades 1 and Pre-1 and the other grades from Grade 2 to Grade 5. The number of years of test administrations included in this study also differs across grades because of the staggered nature of the most recent test revision process, with the revised versions of Grade 1 and Pre-1 being first used in 2004, Grades 2 and Pre-2 in 2003, and Grades 3, 4, and 5 in 2002. To facilitate comparison, tallies of tags observed within each parameter for a grade will be converted to a percentage of total occurrences within that grade. The percentage of occurrence of tags within grades will be used to contrast and compare trends across grades through the collation of tables and graphically where appropriate. For *operation*, *key information*, *explicitness*, and *abstractness* Chi-square tests will also be used to test the strength of association between these nominal variable parameters and EIKEN grade level.

3.2.3 Parameters tagged through automated text analysis

3.2.3.1 Overview

The parameters derived through automatic textual analysis software are split into three groups: 1) measures useful for establishing a lexical profile of texts and identifying lexical thresholds for vocabulary size; 2) a range of quantitative indices useful for capturing features of the text which are useful indicators of textual and syntactic complexity, which as noted in Chapter 2 may impact on the cognitive demand associated with processing a text, (e.g. readability measures, average sentence length, word length, lexical diversity measures); and 3) a range of metadiscourse markers useful for investigating aspects of coherence and cohesion.

To derive the parameters, two automated analysis tools were used. The first group of parameters were derived with the software program Range (Heatley, Nation, and Coxhead, 2002), and the second and third group of parameters were

both derived with the Text Inspector (2015) online analysis tool. A sub corpus of the texts used for tagging through expert judgment, consisting of 126 reading texts, was created for the analysis. A sub corpus of texts was created for several reasons, including the aspect of practical efficiency, given the time required to prepare separate text files for each text and to run these files through the analysis tools. Balancing the number of texts in each grade was also important to meet the assumptions of both parametric and non-parametric tests of statistical significance used to investigate differences in each measure across grades. Perhaps most importantly, some text types are only used in certain grades, and by design target very different aspects in terms of topic, genre, length, abstractness, etc. It was thus considered that a balanced set of texts covering the same period of time and from the same long reading passages used in the long reading comprehension task section would provide the best form of comparison of how textual features associated with the core construct of reading were treated across the grades. The long reading comprehension task section is included from Grade 4 onwards. Grade 5, the most elementary level of the tests in the EIKEN program, does not contain a long reading comprehension task, and so texts from Grade 5 were not included in the sub corpus for analysis.

A total of 21 texts were selected from each grade from Grade 4 to Grade 1. To derive a balanced set of texts for each grade, one text from the long reading comprehension task section (W4) was selected from each test form administered at public test sites on Sunday administrations, and only tests from administrations across the same period covering the years 2004 to 2010 were used. The long reading comprehension task type is consistent in terms of format across Grades 4 to 1, with a long text followed by a number of multiple choice questions targeting comprehension of the text. The final long reading task is considered the most demanding in terms of reading comprehension within each grade, and generally utilizes lexical and syntactic complexity at the upper end of what is considered appropriate for each grade. Within this section, there are several texts, classified in the order in which they appear as simply A, B, or C. The final long reading passage, text C, was used from Grades Pre-1, 2, 3, and 4. For Grades 3 and 4, texts A and B are notices and emails or letters respectively, whereas C is a long

passage. For Grade 2, text A is an email, and text B and C are both long passages which are intended to be generally comparable, but C is by design written to reflect the upper end of the range of complexity considered appropriate for the grade. For Grade Pre-2, which only has two texts in this section, text A is an email text, and so the long passage in text B was used. Grade 1 contains three texts, of which text C is designed to incorporate several features which distinguish it from the other two long reading passages, texts A and B. Text C for Grade 1 is longer in length than the other two texts and is formatted to fit across two facing pages to more closely resemble an authentic magazine layout and journalistic style, with the questions placed below the section of the text to which they relate. Texts A and B are both designed to be long expository or argumentative texts of a generally similar format to the final long reading passage used across the other grades. Although texts A and B are written to be generally comparable, text B is by design written to reflect the upper end of the range of complexity considered appropriate for the grade.

As described above the use of automated analysis tools to derive quantitative measures of the linguistic features was directed primarily at the input texts, excluding the actual question stems and response options from the analyses. For the majority of measures—for example meta-discourse markers targeting coherence and cohesions or readability measures—the rationale for this is clear from the literature review in that these parameters are measures of the features of extended, connected text, and indeed would not provide stable or interpretable results for the short, unconnected pieces of text constituting the questions and options. Some automated measures, such as the average number of words in the question stems and options and the vocabulary profile may indeed provide useful insights if applied to these features. For the purposes of this study, however, the questions and response options in the test content were not analyzed using the few quantitative indices which may have been applicable, as the existing item writing specifications already provide very clear and stringent constraints for constructing the questions and responses. Instructions limit the number of words to a very narrow range, with clearly delineated differences between the permissible ranges across the different test levels. Admissible vocabulary and grammar are

constrained, particularly at the lower levels to be at a level lower than those used in the school text books and Courses of Study curriculum guidelines appropriate for that grade, with once again clear distinctions built into differentiate between the grades. Examining these criterial features with the few measures that would be applicable would tend to produce results indicating differences which had already been explicitly built into the system of item construction, and this may in fact provide a misleading representation of the actual criterial differences between grades.

For the purposes of this study, then, the decision not to include question stems and response options as separate targets of analysis for linguistic features was thus considered the most reasonable approach given the nature of the measures used and the heavily constrained nature of the questions and options as described above. As with other aspects of this study, however, including the investigation of cognitive parameters described earlier, this does not mean that further investigation of the impact of different features of the questions and items would not be instructive. As continually noted, this study cannot aim to collect and analyze all relevant information for a comprehensive validity argument, and RQ1 focuses on addressing criterial features which may distinguish key differences between the seven levels of the testing program. The ongoing validation agenda may indeed benefit from future studies which might address the possible interaction between features of the questions and options and empirical item difficulty. To do so however, may require an experimental approach in which items targeted at the same level were manipulated to contain questions and options with distinct criterial features, such as different length, in order to investigate the impact of such features on item difficulty. Such avenues of research, while potentially important, as noted earlier, are beyond the scope of this study.

3.2.3.2 Lexical threshold and coverage by the AWL

For the lexical profile of the input texts, two measures were chosen for the focus of analysis: the percentage of coverage of running words, or tokens, in a text which is covered by the Academic Wordlist (AWL), and the lexical threshold of

the texts in relation to the 14-level frequency lists developed by Paul Nation (2006). A full description of both measures, and the rationale for their selection as principal indicators of lexical resource demands, is contained in the literature review in Chapter 2.

For the BNC coverage, an estimate of the vocabulary resources needed to ensure that a learner would be familiar with 95 percent of the running words of a text is derived. To derive this lexical level estimate, a text is analyzed and the cumulative percentage of coverage across the 14 1000-word levels of the list is examined. Following Nation (2006), the percentage of proper nouns is added to the cumulative coverage by the 14 1000 word levels. The level at which 95 percent is reached is then fixed as the vocabulary level required by a learner in order to ensure the accessibility of the text from the perspective of lexical resources.

Given the importance of defining clear and explicit vocabulary benchmarks in test specifications, both from the perspective of creating parallel test forms and from the perspective of providing clear learning and teaching goals, a second corpus of texts was prepared for analysis with the Range software. All non-listening sections of the First Stage tests, using the same test forms from which each long reading passage had been selected for the sub corpus described above, was also prepared for analysis of AWL coverage and the BNC level. These texts thus included the entire grammar, vocabulary and reading sections of each test form analyzed. The text files contained questions and response options. The writing prompt for the constructed response writing tasks for Grade 1 and Pre-1 was not included, though the reordering tasks used as indirect tests of writing for Grades 2, Pre-2, 3, and 4 were included. Text files for Grade 5 were included in this analysis.

3.2.3.3 Measures of textual and syntactic complexity

The second group of parameters was measured using the Text Inspector online analysis tool. The range of measures provided allows for a principled comparison across a set of features which have been shown to be good indicators of both textual complexity, and thus of the cognitive demand placed on readers by the

linguistic features of the text. Many of these measures, such as sentence length and readability measures are considered to be good indicators of syntactic complexity. Lexical diversity measures are also included here, as they are considered to be indicators of one aspect of textual complexity that also impacts on the cognitive demand placed on readers by the text. Six measures were included in the analysis, the total number of words (or tokens), the sentence length in words, the number of syllables per word, lexical diversity (MTLD, lexical diversity (VOCD), and the Flesch-Kincaid Grade level measure of readability.

3.2.3.4 Coherence and cohesion through metadiscourse markers

As described in Bax et al (2013), the metadiscourse markers used in Text Inspector have a very clear and explicit framework, based on Hyland (2004). As the texts analyzed across grades differ in total length, it would be misleading to simply rely on the total counts of tokens for metadiscourse markers provided in the output from Text Inspector. To enable comparison the total number of tokens for a particular metadiscourse measure were divided by the total number of tokens in the text, and this was converted to a percentage. The percentages derived in this way are used for the comparison of metadiscourse measures across grades. As Bax et al (2013) clearly demonstrate, quantitative differences alone can be misleading, and it is the variety or range of usage of the options within a particular metadiscourse marker category which can be most instructive. For example, Bax et al (2013) found that the quantity of logical connectors decreased as expected as the proficiency level of writers increased. At the same time, the higher proficiency writers were shown to use a wider range of logical connectors, indicating a greater access to these resources and a greater sophistication in terms of their writing proficiency, rather than relying on a narrow range of explicit logical connectors likely to be taught at lower and intermediate levels of proficiency. For that reason, the range of metadiscourse options within each category used in the reading texts across grades may be more useful as a possible indicator of important criterial features distinguishing texts targeted at different levels of proficiency.

3.2.3.5 Analysis

Where the parameters are shown to meet the assumptions of parametric tests of significance, one-way Analysis of Variance (ANOVA) tests will be used, with grade as the independent variable and the parameter in question as the dependent variable. Where the assumptions of parametric tests are not met, for example when the distributions are not normally distributed or in the case of the BNC-level, the level of measurement is ordinal, the Kruskal-Wallis test, a non-parametric counterpart to a one-way ANOVA will be used with the Man-Whitney test, a non-parametric equivalent of independent t-tests, used for post-hoc tests of significance between groups (Field, 2009). The analysis of statistical differences across all parameters uses SPSS Statistics 21.

3.3 Results

3.3.1 Expert judgment tags

3.3.1.1 Topic

The results for topic are collated and displayed in Appendix C and Appendix D. In Appendix C, Figures C1 to C7 present the results for topic usage across all non-listening sections of the First Stage tests graphically as pie charts. Figures C8 to C13 display the topics used in the long passage-based reading comprehension tasks only—labeled as W2 for the gap-fill tasks and W4 for the MCQ reading comprehension tasks according to the task coding system used by the EIKEN tests (see Appendix A). Figures C8 to C13 show results for Grade 4 to Grade 1, as Grade 5 has no long passaged-based reading section. In the pie charts, topics with a total representation of less than five percent are collapsed within an overall *other* category in order to highlight the most salient topics in each grade.

In Appendix D, the tables give the topic usage observed for *all* of the topics included in the list of contextual parameters used for tagging without collapsing categories. Table D1 presents usage across all seven grades for the non-listening sections in the First Stage tests. Table D2 gives the results for the long, passage-based reading sections only. The total number of texts tagged for topic in the non-listening sections was 10798. The number of texts tagged within

the passage-based reading comprehension sections only totaled 1176.

Examining the trends in topic usage illustrated in Figures C1 to C7, there appears to be a very clear distinction between the two advanced-level tests of Grade 1 and Grade Pre-1 and the remaining grades. For Grade 1, six topics cover almost 55 percent of texts used, while for Grade Pre-1 five of these same topics account for 40 percent of the texts, with *science and technology* falling below the five percent threshold. The topics common to these two grades generally cover more abstract areas related to issues associated with current affairs, social trends and issues, and topics likely to be observed in newspaper or magazine articles or introductory university texts, genres which are identified as being relevant to the TLU domain of these grades. Interestingly, *daily life*, the most prominent topic for all other grades, continues to account for just over five percent in both Grade 1 and Grade Pre-1.

Figures C1 and C2 clearly show that a wide range of topics is being covered in these two grades, with the *other* category (i.e. the combined coverage for all topics not reaching the five percent threshold) accounting for just over 45 percent for Grade 1 and 60 percent for Grade Pre-1. Examining Table D1 allows us to see how the remaining topics are distributed and there is clearly a wide range of topics spread across the texts used in both grades. Only two topics on the list are completely unused for Grade 1, and only one of these is completely unused at Grade Pre-1. Many of the topics, while registering some use have figures below one percent. Some, such as *plants and animals* or *sports*, appear at fairly consistent levels across all of the grades, while some, such as *leisure and entertainment*, although observed in the upper grades have a much greater usage at the lower grades, typically exceeding the five percent threshold used in the pie charts. Other topics which fit much more closely with the kind of abstract subject areas appropriate for the more advanced levels, such as *arts and literature* and *history and archeology*, while not exceeding the five percent threshold have noticeably higher representation at the advanced levels.

All of the grades from the intermediate level Grade 2 to the elementary level Grade 5 show a clear trend toward dominance by a set of topics related to more concrete, everyday, routine aspects of interaction and language use often

associated with more functional aspects of communication. *Daily life* clearly is the largest single topic across all of these grades, with *leisure and entertainment* and *education—school life* constituting a large percentage of all of these grades. Other topics consistent with this general field of everyday activity are also represented to a greater or lesser degree across these grades. There is some variability, however, with individual topics dropping below the five percent threshold for some grades while being prominent in other grades. *Shopping and obtaining services,* for example, constitutes between five and eight percent for Grades 2, Pre-2, and 3 but drops below the five percent threshold for Grades 4 and 5. The lower percentage of topics subsumed within the *other* category for the pie charts for Grades 3, 4, and 5 highlights the reduced range of topics and the greater salience of a smaller range of overlapping topic areas in these grades.

Grades 2 and Pre-2 show some interesting features in topic use which highlight their role as bridging levels designed to build from the lower elementary level grades in order to take learners through intermediate stages towards the more advanced usage required for Grades Pre-1 and 1. This is demonstrated by the greater range of topics covered at these levels, with the *other* category accounting for 55.5 percent for Grade 2 and 43.8 percent for Grade Pre-2. This role as a bridging level, with overlapping features of levels both below and above is particularly prominent for Grade 2, where *work and job related* also appears as one of the more prominent topics, crossing the five percent threshold. Examining Table D1 also reveals that many of the more abstract topics prominent at the upper grades are also being used at Grade 2, though to a lesser extent, such as *business, finance, industry* and *science and technology.*

The above trends become much more pronounced when the results for the longer passage-based reading comprehension sections are examined more closely. The most prominent group of topics, each accounting for five percent or more of topic usage, now account for more than 60 percent of all topics used for the upper three grades (1, Pre-1, 2), while the *other* category now accounts for less than 40 percent in these grades, emphasizing a greater concentration of topic usage in certain areas. For Grade 1, four topics now account for the majority of topic usage, and three of these also feature in the most prominent seven topics

used at Grade Pre-1. The longer passage-based reading comprehension sections also reveal a greater overlap between Grade 2 and the advanced levels. Four of the most prominent seven topics at Grade 2 now overlap with the most prominent topics at Grade Pre-1 or both Grade Pre-1 and Grade 1. Interestingly, *work and job related* no longer features in the most prominent topics for the two most advanced levels. This is probably due to the nature of the text types used in the passage-based reading comprehension sections for these levels, which feature expository and argumentative texts intended to represent genres such as journalistic magazine articles and introductory-level university texts. The *work and job* related topics observed for these grades were situated within the very short sentence-completion tasks used in the vocabulary section, which cover a wider range of situations likely to occur in the relevant TLU domains for these grades to contextualize the vocabulary targets. At Grade 2, on the other hand, the longer passage-based reading comprehension section for W4 also includes an e-mail exchange which is situated within either the public or employment domains of activity, and this accounts for the prominence of this topic for Grade 2 in the long reading sections.

The difference in the nature of the texts used in the vocabulary sections as opposed to the long reading sections for Grades 1 and Pre-1 also accounts for some other differences in the observed topic coverage between Tables D1 and D2. Topics such as *history and archeology*, while still logging higher usage than the lower grades did not make the five percent threshold when topic coverage in all non-listening sections was examined in Table D1. However, when only the long expository and argumentative texts used in the gap-fill and MCQ passage-based reading sections is examined, this topic joins the most prominent topics exceeding five percent usage for both grades. For Grade Pre-1, *social trends* also moves into the most prominent group, and for Grade 1, *science and technology* increases dramatically from 6 percent to 29.5 percent.

In terms of topic usage, the passage-based reading comprehension sections for Grade 2 show greater similarity to the upper grades, whereas the role of pivoting or transitioning between levels seems to be more appropriately located at Grade Pre-2, where the *other* category now accounts for 57.3% of topic usage,

with the fewest number of categories logging zero usage (five) also now seen at this grade. A wide range of topics both from the more abstract topics common to the more advanced level reading passages as well as the more concrete, everyday topics common to the lower grades is observed in the texts used in Grade Pre-2, albeit many sparingly. Partly this is due to the mixed nature of the texts used in these sections for this grade, covering both narrative story-telling type gap-fill passages, expository passages on more objective, socially-relevant topics used in both the gap-fill and MCQ sections, and transactional e-mails used in the MCQ section.

Interestingly a distinction is also seen occurring between Grades 3 and 4, with topics such as *history and archeology* and *biographies* now featuring in the most prominent topics for Grade 3. Both Grade 3 and Grade 4 include a range of fixed text types in the MCQ passage-based reading comprehension section, which always includes one notice or sign, one letter or e-mail, and one longer text. For Grade 3, this longer text is always an expository text, which while constrained for difficulty on other features, is an explicit attempt to move away from the more personal, narrative-type passages used at Grade 4 for the same task type.

3.3.1.2 Genre

Genre is used to describe the intended TLU domain text types which the texts used in the test are intended to reflect. All of the texts are original, created by item writers for use in the test. Item writers must clearly indicate a variety of source texts for the information used in the texts, and the specifications for each grade give varying degrees of explicit instruction to item writers in relation to suitable genre types. The genres used here are intended to capture the criterial features of the text type and to focus attention on the relationship to TLU domain texts.

Table 3.2 Genres actually used in non-listening sections

Genre	G1	GP1	G2	GP2	G3	G4	G5
Advertising material					1.7%	1.7%	
Broadcast and recorded spoken text			24.8%	62.2%	76.9%	76.5%	98.8%
Business letters / e-mail			8.9%	.4%			
Greeting cards, postcards, invitation cards						.3%	
Job advertisements					.2%		
Magazines	100%	100%	40.0%	13.3%	7.4%		
Messages and short memos						1.0%	
Not clear			25.2%	11.2%		.5%	1.2%
Notices and regulations					5.8%	5.1%	
Personal letters / e-mail			1.1%	6.2%	7.7%	7.7%	
Textbooks and readers for language learning				6.7%	.3%	7.1%	
Tickets and timetables						.1%	

Table 3.2 provides an overview of the genres used across all grades in the non-listening sections of the First Stage tests. Because of the short nature of the texts used in the grammar and vocabulary section (W1) of each grade, these texts were not tagged for genre. Only sections with texts that included sufficient context to identify criterial features relevant to a potential TLU domain text were tagged. The total number of texts tagged for genre was 4208. Table 3.2 presents results for only genres that were actually used in at least one grade. Cells shaded in grey indicate grades at which that genre was not observed. One text type, *broadcast or recorded spoken text* applies only to texts in task types used in Grades 2 through 5. The short conversation gap fill sections (W3 and W10) were all tagged with this genre as these texts reflect transcribed excerpts of spoken interactions and dialogues. For the reordering tasks used as indirect tests of writing (W6, W7), the texts fell into either the *broadcast or recorded spoken text* category—when the section of the text to be reordered was part of a short dialogue interaction—or were classed as *not clear* when they were clearly a short written text, but, as with the tasks in the Grammar and Vocabulary sections, did not provide enough context

for determining a relevant genre in the TLU domain.

The results for the long passage-based reading sections (W2 and W4), in which a greater variety of texts types are used, are shown graphically in Figure 3.1 below. As with *topics*, the results for W2 and W4 cover Grade 1 through Grade 4, but do not include Grade 5. The total number of texts tagged for genre in these sections is the same as for topic, 1176. There is a very clear distinction between the upper three grades, Grades 2, Pre-1, and 1, which cover CEFR levels B1, B2, and C1 respectively, and the genre usage reflected for Grades 4, 3, and Pre-2, , targeted at CEFR levels A1 and A2. The upper grades use exclusively *magazines* as the genre types for the long reading passages. A distinction was made in consultation with the focus group of content specialists during the construction of the tagging lists between *magazines* and *newspapers*. *Newspaper* was taken to refer to shorter, factual descriptions of news-worthy events. The genre of *magazines* was applied to longer feature-style articles intended to present a topic in more depth in an expository or argumentative style of writing.

The distinct dominance of the longer feature-style magazine article as a genre type for reading comprehension tasks at the upper grades accords with the criterial features for the grades across other contextual parameters. The topics used for these grades, as noted above, tend to be more abstract themes related to social issues and current affairs, etc. There is also a clear interaction with discourse type, discussed below, as the upper grades are dominated by the expository, and to a lesser extent, argumentative discourse types which would complement the genre type utilized most often at these levels. For Grade 2, there is also a sizeable representation for *business letters/emails* which reflects the test specifications for this grade, in which one of the long reading texts in the MCQ reading comprehension section is always a business-related e-mail.

The lower grades, also reflecting the test specifications for these grades, cover a wider range of text genres. All three grades from Grade Pre-2 to Grade 4 also include a letter or e-mail as a fixed text type in the passage-based MCQ reading comprehension section. The difference between these grades and grade 2 is that the written interaction is a *personal* letter or e-mail in which there is a familiarity between the sender and intended reader. The topics are also

constrained to be more relevant to the level of the typical test takers targeted by these lower grades. The use of different types of greeting cards, memos and notices at Grades 3 and 4 also reflects the specifications for these grades. The MCQ reading task section for these grades always includes one text type from these source genres in which learners are encouraged to use more expeditious reading skills to find relevant information rather than focus on close reading of a full text. At Grades Pre-2 and 3, we also see a noticeable portion of genre taken by *magazines*, reflecting the attempt at these grades to introduce experience with reading more expository, factual texts as learners progress through the levels-based system, working towards the intermediate B1-level Grade and higher, in which the texts become gradually more demanding and reflective of real texts likely to be encountered in the TLU domain for these grades. As noted above, however, texts at the lower grades in particular, are modified across several parameters to control for difficulty. The magazine genre captures the attempt of the more difficult reading comprehension tasks at Grade 3 to move away from the more narrative, personal, story-telling type texts seen at Grade 4.

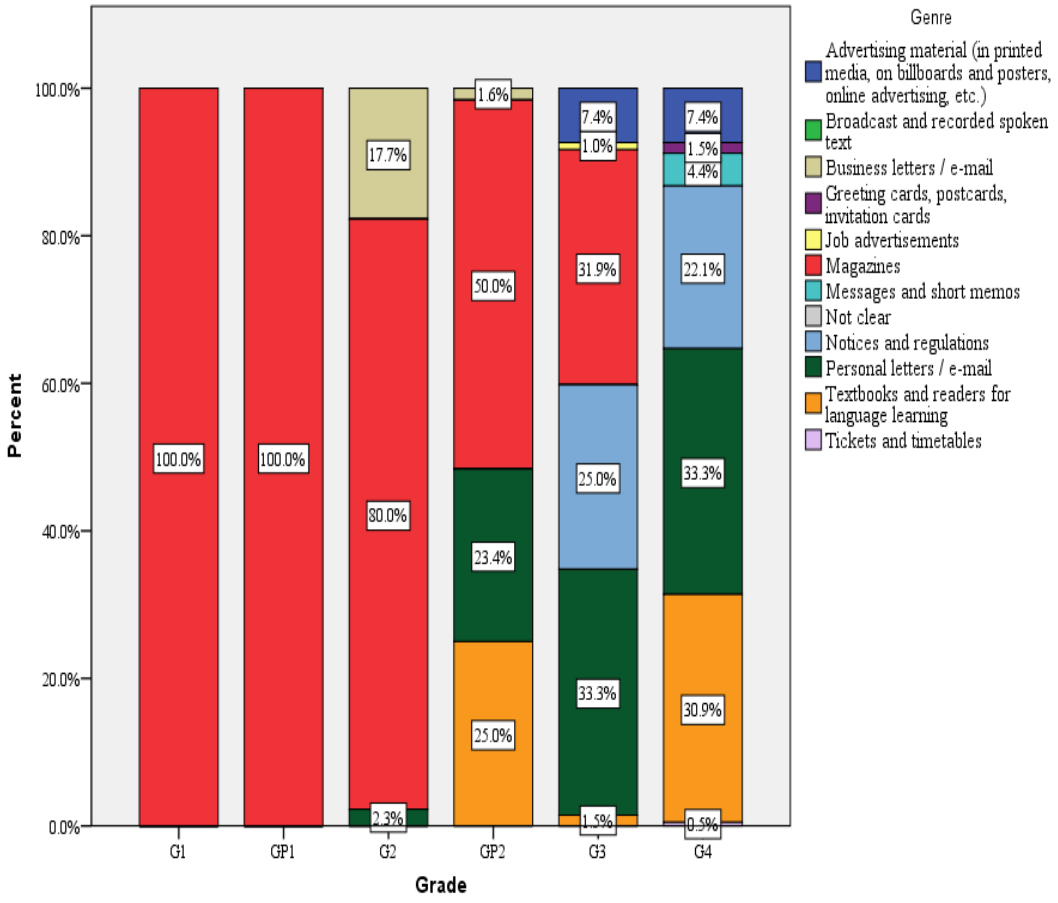


Figure 3.1 Genres reflected in long reading comprehension sections (W2, W4)

3.3.1.3 Discourse type

Discourse type has been applied to only the long passage-based reading sections, and so is only relevant for Grade 1 through Grade 4. Further, within the passage-based MCQ reading task section (W4), discourse type was not applied to letters and e-mails. The focus of these interactional written communication texts is, as is often the case in real-world communication, not focused on developing any one particular idea or theme, and often touch on several different functions including making requests, providing personal information, or updating others on recent events etc, in the same text. It was thus felt these texts did not lend themselves to an easy classification according to the limited discourse types available, and were better described by other contextual parameters. The total number of texts to which this parameter was applied then is 916.

Looking to the results in Figure 3.2, it can be seen that expository texts are the most prominent discourse type noted from Grade 3 upwards to the higher grades. At Grade 2, expository texts completely dominate, and at the advanced levels, the emergence of the use of *argumentative* texts is seen for Grades 1 and Pre-1.

Texts with an *instructive* discourse type constitute almost half of the texts used at Grades 3 and 4, reflecting the use of notices, memos, regulations, and advertising material such as posters etc. in long MCQ reading tasks for these grades. Texts with a narrative discourse type constitute a sizeable proportion of texts at Grade 4 and Grade Pre-2, but are not noted in Grade 3. The long reading passages at grade 3 and 4 include three texts: one notice (or similar text type), one letter/email, and one long passage-based reading task—and as already noted, the letters and e-mails were not tagged for discourse type. The long passage at Grade 4 is structured around a fictionalized account of a central character involved in a simple narrative account of a special event, such as visiting a famous city or taking part in a festival, etc. In terms of genre, these texts were generally tagged as *textbooks and readers for language learning* as this captured the core TLU domain situation in which one would be likely to encounter such modified texts. At Grade 3, the final, long reading task is, as noted above, designed to be a factual, expository account explicitly moving away from the personal, familiar narrative reading texts in Grade 4. Grade Pre-2, on the other hand, does not include any notices in the MCQ reading section, explaining the very low use of instructional texts, and only one e-mail (not tagged for discourse type) and one long passage, The long passages have all been tagged as expository, or for a small number, descriptive. However, Grade Pre-2 includes a separate passage-based gap-fill reading comprehension section which is not used in Grades 3 and 4. The test specification for Grade Pre-2 for this section designates two types of texts, one being a narrative story with a central character taking part in relatively familiar events or activities, and the other being a factual expository type text, and it is the first of these that has accounted for the large percentage of texts with a narrative discourse type at this grade.

The lack of narrative texts at the upper grades, but also in the long

passage-based MCQ comprehension task for Grade 3 reflects to a certain extent a deliberate decision to restrict the definition of this discourse type. An excerpt from the extra information section of the tagging manual for this parameter describes this restricted definition:

> Biographical accounts of famous people that appear in Grade 3 reading passages (stories) will be Expository, not Narrative, as they are a more objective form of writing intended to present information and would be seen in TLU texts such as short articles, textbooks or readers that introduce important people, or in expository essays in which students are to research and introduce famous writers, etc

This definition differs from the classification taxonomy suggested by Enright et al (2000) and utilized by Wu (2012), in which *historical biographical/autobiographical narratives* are defined as a separate category from expository and argumentative texts. The association of historical biography-type articles with the narrative discourse type is not made in Alderson et al (2006), however, and it was felt that separating biographical type articles and accounts from primarily narrative text types was important for two reasons. Firstly, it was felt that the kind of biographical texts used at the higher grades describe not only the life of an important or famous individual, but do so in a way which attempts to position that person's contribution within a wider sphere of activity or evaluate their contribution to society more broadly. A salient feature of these texts seems to be less the relating of events in sequence and more that of creating a clear explanation of the importance or relevance of those events to wider social concerns. Secondly, the focus group input from the content specialists made clear that there was an explicit attempt to make the texts used at the lower grades accessible through making them more personalized and situated within everyday, familiar spheres of activity for the expected test takers, particularly at the A1-level grades.

The emergence of argumentative texts at the advanced levels, while still relatively limited at Grade Pre-1, identifies another important criterial difference. The distinctions seen in practice once again reflect the intentions of the test specifications, as at the intermediate and lower grades, item writers are instructed

to present clear, factual texts that are balanced and objective, appropriate for the level of the test takers being targeted. At the advanced levels, a distinct attempt is made to reflect more demanding text types representative of real-world texts likely to be encountered in the TLU domain. Argumentative texts of course lend themselves to more demanding question types, targeting author intentions and the need to make appropriate inferences regarding the way pieces of the text are integrated to support the author's stance. All of these features underscore important criterial features characterizing the advanced levels. At the same time, the relatively low level of argumentative texts at Grade Pre-1 raises questions about the consistency of the criterial features of texts across test forms. Identifying areas in which test specifications could be tightened to improve the consistent and comprehensive inclusion of the features considered criterial for a grade is of course one of the potential benefits of implementing a comprehensive approach to tagging as has been undertaken in this study. .

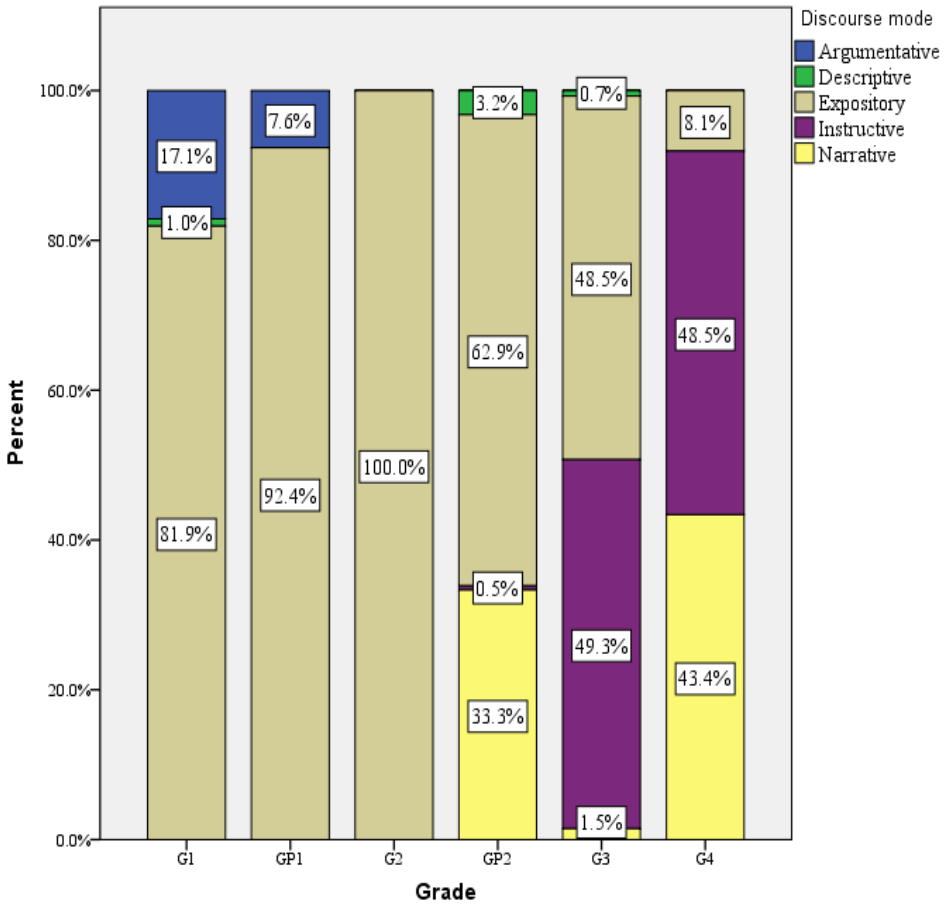


Figure 3.2 Discourse types in long reading comprehension sections (W2, W4)

3.3.1.4 Abstractness

Figures 3.3 and 3.4 show the abstractness level of texts used across all non-listening sections of the First Stage tests and texts used only in the longer, passage-based reading comprehension sections respectively. The total number of texts tagged for this parameter is the same as for topic, 10798 for all sections and 1176 for texts in the long reading comprehension sections only. For Grades 5 through 2, the major pattern of the level of abstractness of the texts does not change noticeably when comparing all non-listening sections to the long passage-based reading comprehensions sections. For Grades Pre-1 and 1, however, a noticeable increase is seen in the percentage of *fairly abstract texts*, and a corresponding decrease in *mostly concrete* texts. This difference is most likely due to the nature of the short sentence-completion texts used to contextualize target words in the vocabulary sections, which contain a wider range of topics than the long reading comprehension sections for these grades, where the focus is more clearly on more abstract topics related to broader social issues and trends.

In terms of trends across grades in the long reading comprehension sections, there are clearly three distinct groups which might be called elementary/beginner for the A1-level grades comprising Grades 4 and 3, lower intermediate/intermediate for the A2-B1 level grades of Grades Pre-2 and 2, and advanced for Grades Pre-1 and 1. This is admittedly a very broad distinction, but importantly also reflects the similar pattern of results for topic usage. At the advanced level, the strong representation of abstract texts is seen, while there is a distinction between the B2-level Grade Pre-1 and C1-level Grade 1 in the percentage of *fairly abstract* and *mainly abstract* texts used. For Grade 1, mainly abstract texts constitute the second largest category, but still account for slightly less than ten percent.

Grade 2 had considerable overlap in topic usage with the advanced grades, and the difference in the level of abstractness may indicate that the texts at this grade, while making a deliberate attempt to focus attention on topics which are pointing away from the familiar, everyday topics of the lower grades and towards texts focused on a broader sphere of activity in terms of social issues etc.,

are still being made accessible to the intended test takers in line with the stated intentions for this grade to act as an important transition level towards advanced proficiency.

Pearson's chi-square test was used to test if there is a statistically significant association between the text abstractness in the long reading comprehension passages and the EIKEN grades. There was a significant association, $\chi^2(15) = 1060.57$, $p < .001$. The strength of association was further tested through Cramer's V, which Field recommends as the most useful test for this purpose when there are more than two levels or categories for each variable, resulting in contingency tables greater than 2X2, as is this case with this data. The value for this test statistic was .548, which was highly significant ($p < .001$). Cramer's V can be interpreted in a similar way to correlation coefficients in terms of effect size (Field, 2009, p. 699), and so this result indicates a large effect size for the strength of association. It should be noted however, that 25 percent of cells had expected frequencies of less than five, although no cell had expected values less than 1. For contingency tables greater than 2X2, Field (2009, p. 692) suggests that no more than 20 percent of cells should have expected counts less than 5, and no cells should have counts less than 1. Violating this assumption, however, is likely to increases the probability of *false negative* decisions, meaning that a real effect might be missed (Field, 2009, p. 692). Given the highly significant results for the association between grade and text abstractness, and the large size of the test statistic for the strength of that association, it is reasonable to suggest that the results of the chi-square confirm that the trends clearly evident in the visual examination of the data in Figure 3.4 have not occurred merely by chance.

.

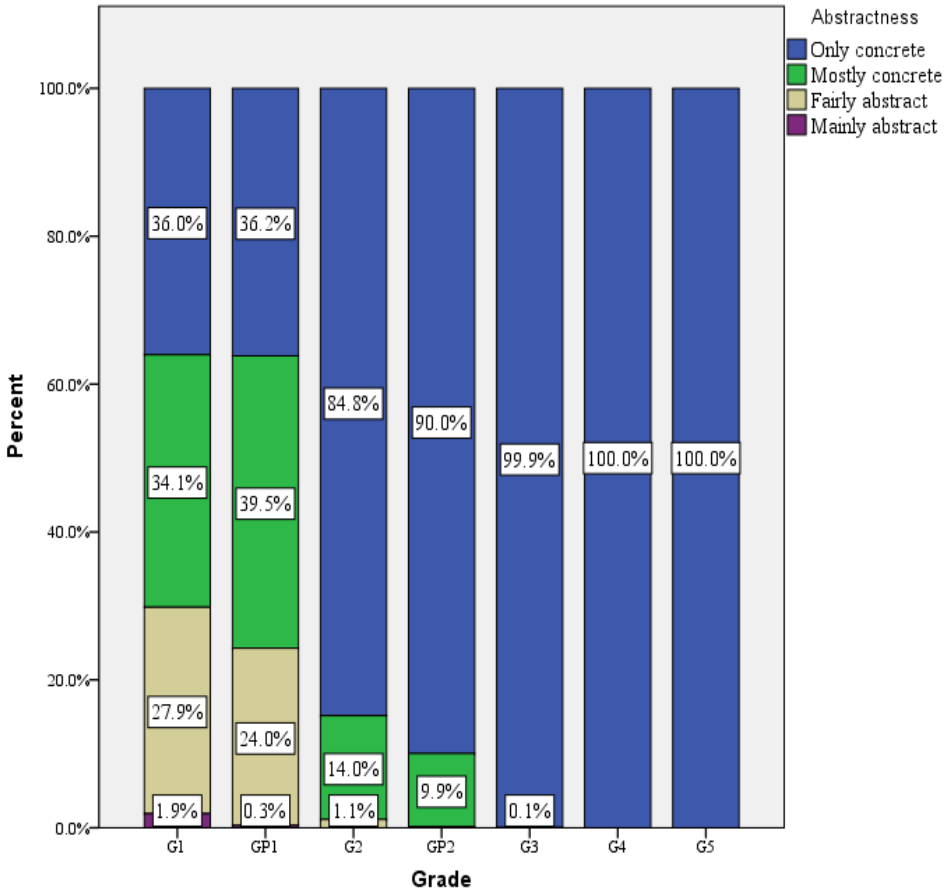


Figure 3.3 Abstractness of texts across all non-listening sections

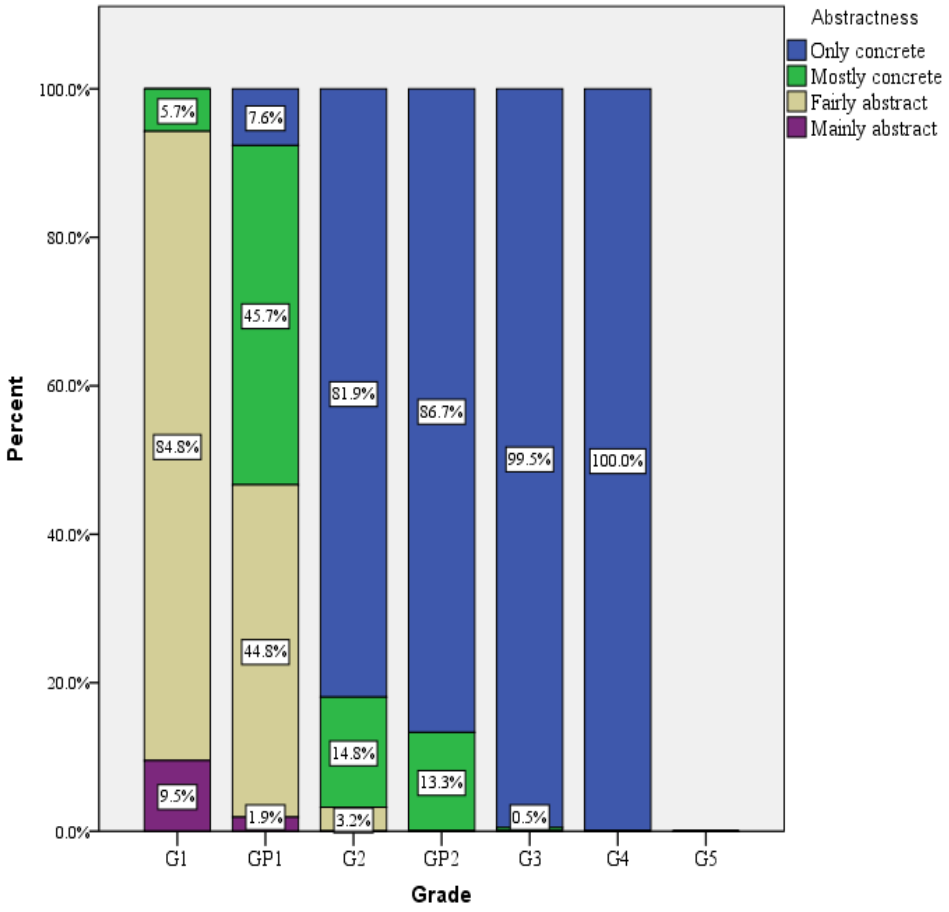


Figure 3.4 Abstractness of texts in long reading comprehension sections

3.3.1.5 Explicitness

An extract from the definition of Explicitness in the Tagging Manual is presented below to help interpret the results for this parameter:

> Implicitness refers to information targeted by an item and which a test taker needs to understand in order to successfully complete that item, and the two categories, explicit or implicit refer to the degree to which that information is explicitly expressed in the text. . . . For the purposes of the EIKEN item analysis, implicit will refer to information that would require pragmatic inferences. Pragmatic inferences require the reader/listener to access background information or cultural knowledge not explicitly contained within the text in order to understand the information targeted. Implicit items then will be determined by the amount of pragmatic inference involved – in other words, reading

between the lines. Implicit is not simply a matter of having to synthesize information from several parts of text. It refers to unsaid elements of a speaker's or writer's message.

As noted already, this parameter is item-based. It is closely related to the degree of inferencing in which a test taker must engage in order to identify, and where necessary integrate, the information required to successfully answer questions or complete other test tasks. Importantly, however, the definition of implicit employed in this taxonomy has restricted the kind of inferencing associated with items classified as implicit to the kind of *pragmatic* inference described by Khalifa and Weir (2009, p. 51). This is opposed to textual inference requiring test takers to integrate and interpret information provided within the input text which the item relates to. As Khalifa and Weir (2009, p. 51) note, pragmatic inference, due to the range of possible interpretations which are likely to be highly idiosyncratic and related to the personalized experience of each test taker, is difficult to introduce appropriately into a testing situation. Khalifa and Weir (2009, p. 51) further note that "text-based inference…may be more amenable to inclusion within tests." The explicit/implicit distinction in this study, then, was from the start intended more as a tool for quality control. When consistency of test content and difficulty—in order to ensure comparability of interpretations across test forms--is taken into consideration, the kind of pragmatic inference which would be entailed by items being tagged as *implicit* (according to this definition) would indicate potentially problematic items. From that perspective, this dimension is from the outset quite different from the other parameters, in that it was predicted, or expected, that one of the categories—the implicit category—would essentially not be observed in "good" items, regardless of the grade.

Table 3.3 shows the results of all items included in the data set used for this analysis. Across all non-listening sections in the First Stage tests, this totaled 13,762 items. The results clearly show the overwhelming majority of items conform to the expectations of the item specifications in terms of being explicit according to the definition employed here. Across all seven grades, a total of

only .2% of items were tagged as implicit. The majority of these are located in Grades 1 and Pre-1, in which the boundaries between pragmatic and textual inferencing may indeed be harder to define when dealing with abstract topics and cognitively demanding reading comprehension tasks appropriate for the TLU domain for these grades. It is important to stress that the implicit/explicit distinction employed for this analysis does not imply that cognitively demanding levels of *text-based* inference are not required. This aspect of cognitive processing was deliberately separated from the explicitness dimension and is evaluated though the *key information* parameter reviewed below.

Table 3.3 Explicitness dimension of items across non-listening sections

Grade		Explicit	Implicit	Total
G1	Items	855	6	861
	%	99.3%	.7%	100%
GP1	Count	847	14	861
	%	98.4%	1.6%	100%
G2	Count	2788	2	2790
	%	99.9%	.1%	100%
GP2	Count	2789	1	2790
	%	100.0%	.0%	100%
G3	Count	2380	0	2380
	%	100.0%	.0%	100%
G4	Count	2380	0	2380
	%	100.0%	.0%	100%
G5	Count	1700	0	1700
	%	100.0%	.0%	100%
Total	Count	13739	23	13762
	%	99.8%	.2%	100%

3.3.1.6 Operation

The definition of operation in the Tagging Manual restricts this parameter specifically to "the type of information targeted by a question in Q&A-type items." There are two possible categories, *main idea* and *specific detail*, and this parameter was thus applied only to the MCQ long, passage-based reading comprehension section (W4), and thus applies to Grade 4 through Grade 1.

A total of 2958 items were tagged for this parameter, with Figure 3.5 providing a breakdown of the results. A very small number of items, 0.5 percent in Grade Pre-2 (a count of 2) were tagged as *other,* indicating the judges were confident of identifying the operation required by the item in the overwhelming majority of cases. A chi-square test indicated the association between this parameter and grade was statistically significant ($\chi^2(5)$ =326.84, $p < .001$.), and Cramer's V indicated a moderate strength, or effect size, for this association with a test statistic of .333. No cells had counts of lower than five, indicating that the cross-tabulation met the assumptions for the chi-square test.

As expected, the lower, A1-level grades (Grade 3 and 4) have the smallest proportions of items targeting *main idea.* Items at these levels are clearly aimed at targeting more local understanding of explicit, factual information for texts that are concrete dealing with everyday, topics familiar to the intended test takers. There is a noticeable section of items at Grade 3, just over 10 percent, which target *main idea*. The definition of *main idea* in the Tagging Manual in fact subsumes two types of operation in terms of the nature of the information that test takers need to process and identify in order to correctly answer the question: Those items which require careful global reading to synthesize propositions across the text in order to understand the main idea or central message of the writer; and those items targeting expeditious global reading to skim the text for gist to obtain an understanding of the overall theme or topic area at a general level. For Grade 3, the items tagged as *main idea* focus on the latter. Indeed item writing guidelines for this grade explicitly call for, where possible, the last of the four items attached to the reading text in this section to ask about the general theme or topic areas and for item writers and reviewers to require that test takers need to look for relevant key words and themes across the text in order to correctly answer these items. In such cases, however, it is generally limited to identifying the theme at a very general level, for example *the passage is about the history of a famous festival*.

Interestingly, the highest proportion of *main idea* items is actually seen at Grade 2 rather than in a smooth progression increasing as the grades increase in level. One possible explanation for this may lie in the different task specifications used for the grades. While the grades have been integrated within a common

conceptual framework from the inception of the EIKEN testing system, actual test and item development has been carried out by teams of content specialists specializing in specific grades, resulting in grade-specific approaches. For Grade 2, the number of items described as targeting "overall" ideas is explicitly built into the guidelines for the creation of the MCQ reading tasks. For the upper grades, while overall, or what have been categorized as *main idea* items with a careful global reading perspective, are encouraged, there is no quantitative quota for the number of these items to be produced for each reading task, and this may account for the difference in the proportion of these items across these grades.

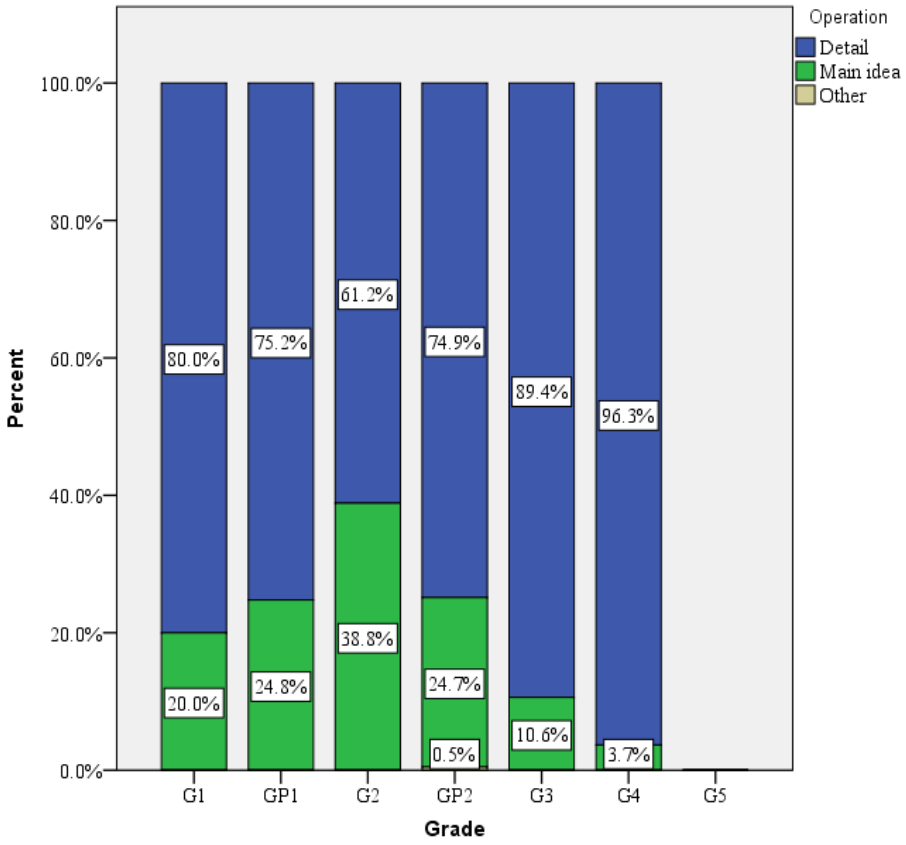


Figure 3.5 Operations targeted by items in long reading comprehension section

3.3.1.7 Key information

Key information is a cognitive parameter for items which is designed to identify the degree to which the item requires test takers to identify relevant information

and integrate that information across the input text used for a test task. As noted above in the *operation* section, this is linked to text-based inferencing, and is specifically designed to focus on the kind of inference which is considered by Khalifa and Weir (2009) to be suitable for operationalization in test tasks. While the distinctions are quite broad in the interests of keeping the evaluation practical, accessible and efficient for expert judges and ultimately for use in test specifications for ongoing item-writing development and review, the distinction is explicitly intended to provide a window on cognitive processing and the level of cognitive demand. It is clearly a very useful way of connecting contextual text characteristics with the model of reading described by Khalifa and Weir (2009) in an integrated approach to specifying the level, or difficulty, of a task. Within the non-listening sections of the First Stage tests, this parameter has been applied to all sections except the grammar and vocabulary sections (W1) and sentence reordering (W6, W7) sections of each grade. The total number of items tagged for this parameter is 5532, and the number of items tagged only within the long reading comprehension sections is 4016.

Figures 3.6 and 3.7 show the results for all items tagged for this parameter and for only items in the long reading comprehension sections respectively. Visually in both graphs a general trend is seen across the grades, with the largest number of *within sentence* items being at the very low A1-level grades, and the largest number of *across paragraph* items being at the advanced B2-level and C1-level Grade Pre-1 and Grade 1 tests. The association between this parameter and grade is statistically significant. Looking first to Figure 3.6, the chi-square test for all items tagged is $\chi^2(12) = 1275.55$, $p < .001$. The results for only the long reading comprehension section items were also highly significant ($\chi^2(10) = 330.33$, $p < .001$). The results for the test of the strength of the association produced a moderate effect size in the form of a Cramer's V of .340 for all items tagged, and a small effect size of .203 for the long reading comprehension sections.

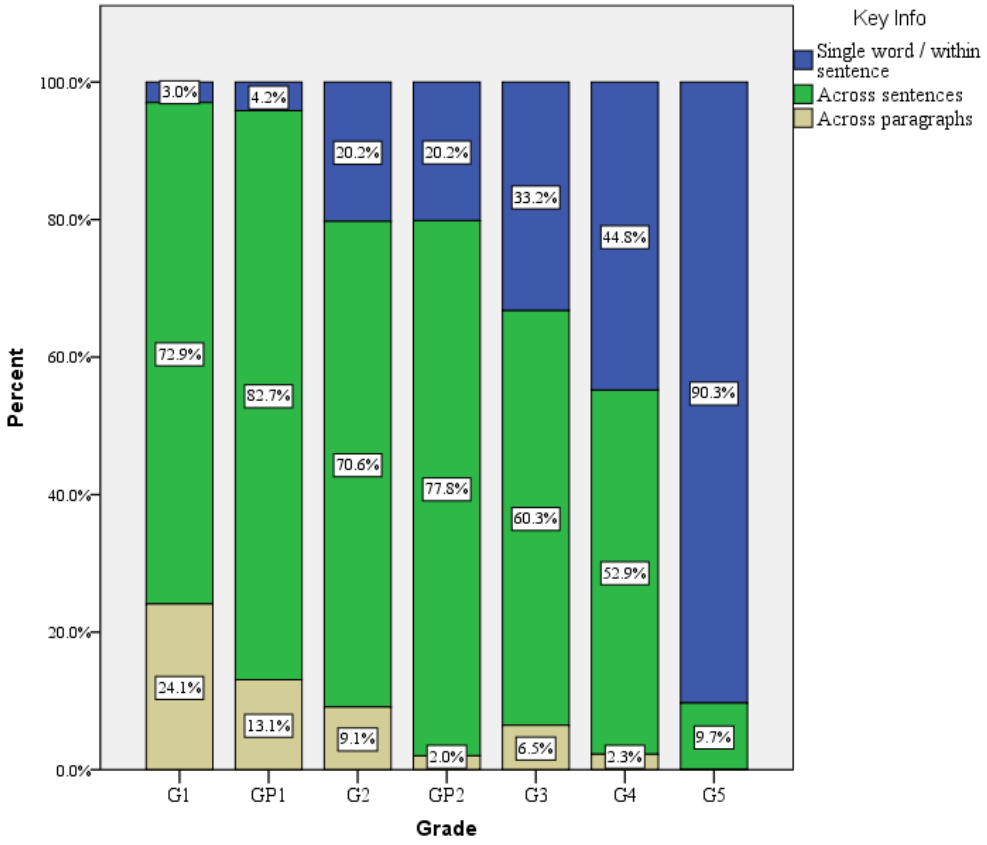


Figure 3.6 Key information for gap-fill dialogues and long reading comprehension sections

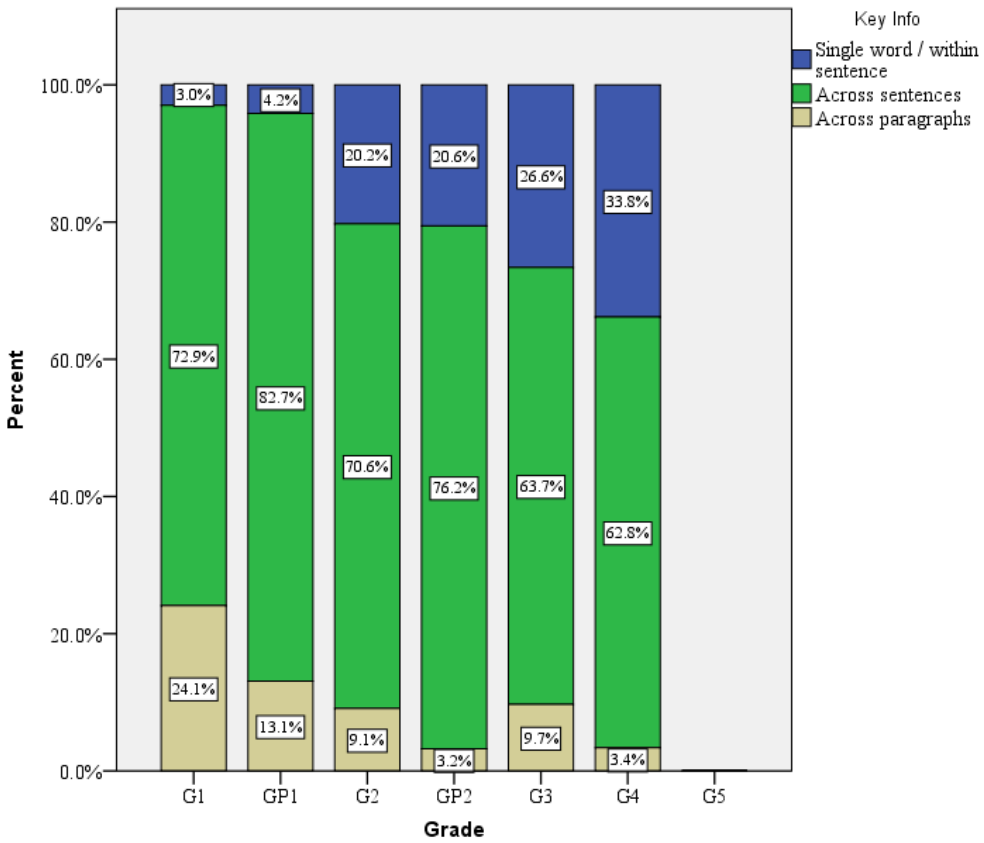


Figure 3.7 Key information targeted in long reading comprehension sections

3.3.2. Automated analysis

3.3.2.1 Vocabulary (AWL and BNC)

Table 3.4 provides the descriptive statistics for the percentage of AWL words and the lexical level at which 95 percent coverage of running words is achieved for the sub-corpus of long reading comprehension texts. Table 3.5 provides the results for the same measures using all non-listening sections of the First Stage Tests.

In Table 3.4, the mean percentage of AWL words shows a consistent trend in line with expectations, with the lowest percentage for Grade 4 and the highest for Grade 1. The tests targeted at CEFR A1 and A2, Grades, 3 and Pre-2, have a noticeably low percentage of AWL words. The intermediate B1 increases substantially to 4.12 percent, followed by another dramatic rise at the B2-level Grade Pre-1 and C1-level Grade 1. The differences between these two

advanced-level grades is smaller, though the trend of increasing across grades is maintained. A one-way ANOVA was used to test the statistical significance of the difference in mean AWL coverage across the grades. Although skewness and kurtosis results for Grade 4 and Grade 3 indicated that the distributions for these two grades diverged from a normal distribution, Feld (2009, p. 360) suggests that "when group sizes are equal the F-statistics can be quite robust to violations of normality." As the sample sizes were equal, the remaining grades all appeared to be normally distributed, and the total sample also showed acceptable results (N126, kurtosis =-0.463747737, skewness = 0.768604297), the one-way ANOVA was chosen due to the loss of statistical power associated with non-parametric alternatives. As Levene's test indicated that the assumption of homogeneity of variance was not met, the Welch F-ratio is reported[3]. There was a highly significant effect for grade on the percentage of AWL words, (Welch: F(5, 55.049) = 65.573, p <. 001). When the assumption of homogeneity of variance is not met, Field (2009, p. 375) recommends the Games-Howell procedure for post-hoc pairwise comparisons. The results are presented in Table 3.6. The differences in means are statistically significant for most comparisons (p<.05).The results for Grade 3 with the adjacent grades, Grade 4 below and Grade Pre-2 above were not significant, and the results for the differences between the two highest levels, Grade Pre-1 and Grade 1 were also not significant at the .05 level.

Table 3.4 also reports on the mean BNC level of the texts, according to the 95 percent coverage criterion adopted for this study. Once again, the trend generally conforms to expectations, with highest grades requiring the largest vocabulary level to cover 95 percent of the texts. At the same time, the three lowest grades are grouped together, and the mean for Grade 3 is slightly lower than that for Grade 4. The results for these three grades, when rounded, would result in a mean level of K3, or the first 3000 words on Nation's (2006) 14-level list, to cover 95 percent of the words in the texts used. Examining the mode, or most frequently occurring level, there does appear to be a distinction between the two A1-level grades, Grade 3 and 4, and the A2-level Grade Pre-2, with the most

[3] SPSS produces two "robust tests of equality of means", Welch and Brown-Forsythe, for when the assumption of homogeneity of variance is not met. As the results were significant for both, only the Welch test results are reported.

frequently occurring level being K2, or the first 2000 words, for the lower grades.

As the BNC-level represents an ordinal scale, the Kruskal-Wallis non-parametric equivalent of ANOVA was used to test for significance between the distributions of BNC levels required for each grade. The results showed a significant effect for grade, $H(5) = 44.962$, $p < .05$. As it is expected that the median BNC level will increase as the grade increases, an additional test, the Jonckheere-Terpstra test was also run. The Jonkheer-Terpstra test tests "for an ordered pattern to the medians of the groups" (Field, 2009, p. 568). The results of this test were also highly significant ($J = 4{,}403.00$, $z = 6.565$, $p < .001$). Post-hoc tests were carried out on all pairwise combinations[4]. The output for the post-hoc comparisons produced by SPSS is provided in Figure 3.8. It is important to refer to the adjusted significance values shown in Figure 3.8. To control Type I errors when making multiple comparisons, i.e. obtaining a significant result when no effect is present, a Bonferroni correction is often made. The Bonferroni correction usually involves dividing the alpha criterion level by the number of comparisons to derive a stricter criterion to determine whether the differences observed are likely to have occurred by chance alone (Field, 2009, p. 782). In SPSS 21, this correction is applied directly to the significance level calculated for each post-hoc comparison, while the alpha level criterion is maintained (in this case, .05). The statistically significant comparisons indicate two distinct groups, with Grade 4, 3, Pre-2, and 2 all showing statistically significant differences to the upper-level Grade 1 and Pre-1 tests.

The AWL and BNC levels were also calculated for the entire non-listening sections of the First Stage tests, treating each entire test form as a single text. The results in Table 3.5 generally demonstrate a similar pattern, but with slightly lower results for both AWL coverage and the mean BNC level in each grade (note that Grade 5 is included in this analysis). This can be interpreted

[4] SPSS 21 produces pairwise post-hoc comparisons when the Kruskal-Wallis test statistic is significant. The test used is the Dunn-Bonferroni test (Dunn, 1964). It differs from another commonly used approach to non-parametric post-hoc comparisons, i.e. running separate Mann-Whitney tests, a non-parametric equivalent of the T-test, for each pair. The Dunn-Bonferroni approach "compares pairs of groups based on rankings created using data from all groups, as opposed to just the two groups being compared"
(http://www-01.ibm.com/support/docview.wss?uid=swg21477370)

as resulting from the wider range of task types used across the whole non-listening sections of the First Stage tests, including personal letters and emails, notices, signs, and conversational dialogues, in which a lower percentage of words associated with more academic genres would be expected. A one-way ANOVA carried out on the AWL coverage for the entire Grammar/Vocabulary/Reading (GVR) sections resulted in highly significant differences: $F(6, 60.648) = 464.358$, $p <. 001$ (as the assumption of homogeneity of variances was not met, the Welch F statistic is reported): Post-hoc comparisons (Games-Howell) are presented in Table 3.7. All results except for the Grade 4/3 comparison are statistically significant ($p < .05$).

For the BNC level, the results of both the Kruskal-Wallis and Jonckheere-Terpstra tests were highly significant ($H(6) = 115.307$, $p < .001$; $J = 7{,}670.500$, $z = 10.891$, $p < .001$). The post-hoc pairwise comparisons provided by SPSS are shown in Figure 3.22. The pattern of statistically significant comparison is generally the same as the results for the long reading passages. While the differences for Grade 2 and Pre-1 are no longer statistically significant, the significance level for both Grade Pre-2/2 and Grade 2/Pre-1 comparisons, even with the very conservative Bonferroni correction, are quite close to significance.

Table 3.4 AWL % and BNC level for long reading texts (W4)

		G4	G3	GP2	G2	GP1	G1
AWL	**N**	21	21	21	21	21	21
	Mean	0.49%	0.86%	1.63%	4.12%	7.13%	8.23%
	Median	0.00%	0.78%	1.37%	3.93%	7.19%	8.41%
	Max	3.23%	5.19%	4.43%	6.67%	12.69%	13.47%
	Min	0.00%	0.00%	0.00%	1.88%	2.60%	2.74%
	SD	0.89%	1.12%	1.24%	1.38%	2.80%	2.41%
	Kurtosis	3.726	11.598	-0.347	-0.871	-0.928	0.603
	Skewness	2.022	3.071	0.621	0.387	0.161	-0.183
BNC Level	**N**	21	21	21	21	21	21
	Mean	3.4	3.2	3.4	3.7	5.4	6.9
	Median	3	3	3	4	5	6
	Mode	2	2	3	4	4	6
	Max	11	8	10	7	10	13
	Min	1	1	2	2	3	3
	SD	2.5	1.6	1.9	1.4	2.0	2.3
	Kurtosis	5.169	1.002	6.845	1.500	-0.137	1.079
	Skewness	2.055	3.071	2.439	1.071	0.868	0.839

Table 3.5 AWL % and BNC level for First Stage non-listening sections

		G5	G4	G3	GP2	G2	GP1	G1
AWL GVR	**N**	21	21	21	21	21	21	21
	Mean	0.15%	0.42%	0.50%	1.22%	4.02%	6.77%	7.82%
	Median	0.00%	0.40%	0.50%	1.30%	4.20%	6.80%	7.90%
	Max	0.60%	1.10%	1.40%	1.80%	4.90%	8.40%	9.50%
	Min	0.00%	0.00%	0.20%	0.60%	2.80%	5.00%	6.30%
	SD	0.19%	0.33%	0.30%	0.34%	0.62%	1.08%	0.86%
	Kurtosis	-0.428	-0.280	3.018	-0.841	-0.470	-1.137	-0.441
	Skewness	0.846	0.852	1.695	-0.446	-0.514	-0.036	-0.190
	N	21.000	21.000	21.000	21.000	21.000	21.000	21.000
BNC GVR	**Mean**	2.0	2.5	2.4	2.2	3.0	4.9	6.6
	Median	2	2	2	2	3	5	6
	Mode	2	2	2	2	3	5	6
	Maximum	2	4	4	3	4	7	8
	Minimum	2	2	2	2	2	4	5
	SD	0.0	.7	.7	.4	.5	.9	.9
	Kurtosis		-0.102	1.276	-0.276	1.864	0.476	-0.728
	Skewness		0.962	1.680	1.327	0.130	0.994	0.363

Table 3. 6 Post-hoc comparisons for AWL coverage in reading texts (W4)

Games-Howell		Mean Difference	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Grade 4	G3	-.37117%	.31251%	.840	-1.3087%	.5664%
	GP2	-1.14501%*	.33217%	.017	-2.1439%	-.1461%
	G2	-3.63289%*	.35859%	.000	-4.7150%	-2.5507%
	GP1	-6.63921%*	.64147%	.000	-8.6228%	-4.6556%
	G1	-7.74550%*	.56139%	.000	-9.4739%	-6.0171%
Grade 3	G4	.37117%	.31251%	.840	-.5664%	1.3087%
	GP2	-.77384%	.36466%	.297	-1.8655%	.3178%
	G2	-3.26172%*	.38888%	.000	-4.4277%	-2.0957%
	GP1	-6.26804%*	.65888%	.000	-8.2908%	-4.2452%
	G1	-7.37433%*	.58120%	.000	-9.1492%	-5.5995%
Grade Pre-2	G4	1.14501%*	.33217%	.017	.1461%	2.1439%
	G3	.77384%	.36466%	.297	-.3178%	1.8655%
	G2	-2.48788%*	.40485%	.000	-3.7000%	-1.2757%
	GP1	-5.49420%*	.66843%	.000	-7.5394%	-3.4490%
	G1	-6.60049%*	.59201%	.000	-8.4019%	-4.7991%
Grade 2	G4	3.63289%*	.35859%	.000	2.5507%	4.7150%
	G3	3.26172%*	.38888%	.000	2.0957%	4.4277%
	GP2	2.48788%*	.40485%	.000	1.2757%	3.7000%
	GP1	-3.00632%*	.68195%	.002	-5.0843%	-.9284%
	G1	-4.11261%*	.60722%	.000	-5.9525%	-2.2727%
Grade Pre-1	G4	6.63921%*	.64147%	.000	4.6556%	8.6228%
	G3	6.26804%*	.65888%	.000	4.2452%	8.2908%
	GP2	5.49420%*	.66843%	.000	3.4490%	7.5394%
	G2	3.00632%*	.68195%	.002	.9284%	5.0843%
	G1	-1.10629%	.80723%	.744	-3.5243%	1.3117%
Grade 1	G4	7.74550%*	.56139%	.000	6.0171%	9.4739%
	G3	7.37433%*	.58120%	.000	5.5995%	9.1492%
	GP2	6.60049%*	.59201%	.000	4.7991%	8.4019%
	G2	4.11261%*	.60722%	.000	2.2727%	5.9525%
	GP1	1.10629%	.80723%	.744	-1.3117%	3.5243%

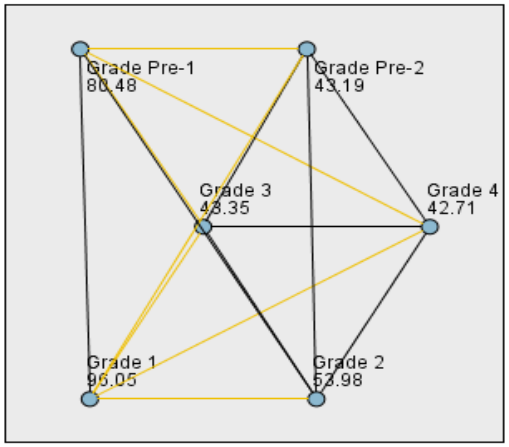
Table 3. 7 Post-hoc comparisons for AWL (all non-listening sections)

		Mean Difference	Std. Error	Sig.	Lower Bound	Upper Bound
Grade 5	Grade 4	-0.2667%*	.08402%	.046	-0.5306%	-.0027%
	Grade 3	-0.3476%*	.07887%	.002	-0.5946%	-.1006%
	Grade Pre-2	-1.0667%*	.08570%	.000	-1.3362%	-0.7971%
	Grade 2	-3.8714%*	.14218%	.000	-4.3282%	-3.4146%
	Grade Pre-1	-6.6190%*	.23902%	.000	-7.3950%	-5.8431%
	Grade 1	-7.6714%*	.19155%	.000	-8.2912%	-7.0516%
Grade 4	Grade 5	0.2667%*	.08402%	.046	0.0027%	.5306%
	Grade 3	-0.00081	.09847%	.981	-0.3866%	.2247%
	Grade Pre-2	-0.8000%*	.10402%	.000	-1.1228%	-0.4772%
	Grade 2	-3.6048%*	.15392%	.000	-4.0900%	-3.1195%
	Grade Pre-1	-6.3524%*	.24619%	.000	-7.1436%	-5.5612%
	Grade 1	-7.4048%*	.20042%	.000	-8.0443%	-6.7652%
Grade 3	Grade 5	0.3476%*	.07887%	.002	0.1006%	.5946%
	Grade 4	0.0008095	.09847%	.981	-0.2247%	.3866%
	Grade Pre-2	-0.7190%*	.09991%	.000	-1.0293%	-0.4088%
	Grade 2	-3.5238%*	.15117%	.000	-4.0020%	-3.0456%
	Grade Pre-1	-6.2714%*	.24448%	.000	-7.0589%	-5.4840%
	Grade 1	-7.3238%*	.19831%	.000	-7.9585%	-6.6892%
Grade Pre-2	Grade 5	1.0667%*	.08570%	.000	0.7971%	1.3362%
	Grade 4	0.8000%*	.10402%	.000	0.4772%	1.1228%
	Grade 3	0.7190%*	.09991%	.000	0.4088%	1.0293%
	Grade 2	-2.8048%*	.15485%	.000	-3.2924%	-2.3171%
	Grade Pre-1	-5.5524%*	.24677%	.000	-6.3449%	-4.7599%
	Grade 1	-6.6048%*	.20113%	.000	-7.2460%	-5.9635%
Grade 2	Grade 5	3.8714%*	.14218%	.000	3.4146%	4.3282%
	Grade 4	3.6048%*	.15392%	.000	3.1195%	4.0900%
	Grade 3	3.5238%*	.15117%	.000	3.0456%	4.0020%
	Grade Pre-2	2.8048%*	.15485%	.000	2.3171%	3.2924%
	Grade Pre-1	-2.7476%*	.27160%	.000	-3.6013%	-1.8939%
	Grade 1	-3.8000%*	.23092%	.000	-4.5201%	-3.0799%
Grade	Grade 5	6.6190%*	.23902%	.000	5.8431%	7.3950%

Table 3. 7 Post-hoc comparisons for AWL (all non-listening sections)

Pre-1	Grade 4	6.3524%*	.24619%	.000	5.5612%	7.1436%
	Grade 3	6.2714%*	.24448%	.000	5.4840%	7.0589%
	Grade Pre-2	5.5524%*	.24677%	.000	4.7599%	6.3449%
	Grade 2	2.7476%*	.27160%	.000	1.8939%	3.6013%
	Grade 1	-1.0524%*	.30040%	.019	-1.9870%	-.1177%
Grade 1	Grade 5	7.6714%*	.19155%	.000	7.0516%	8.2912%
	Grade 4	7.4048%*	.20042%	.000	6.7652%	8.0443%
	Grade 3	7.3238%*	.19831%	.000	6.6892%	7.9585%
	Grade Pre-2	6.6048%*	.20113%	.000	5.9635%	7.2460%
	Grade 2	3.8000%*	.23092%	.000	3.0799%	4.5201%
	Grade Pre-1	1.0524%*	.30040%	.019	0.1177%	1.9870%

Pairwise Comparisons of Grade



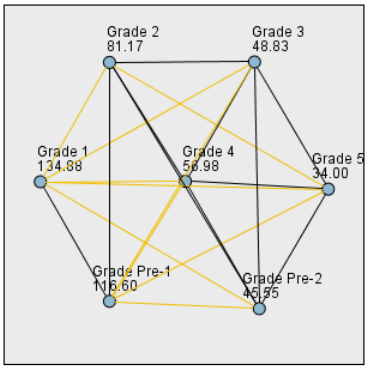
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 3-Grade 1	-52.698	10.724	-4.914	.000	.000
Grade 4-Grade 1	-53.342	11.198	-4.764	.000	.000
Grade Pre-2-Grade 1	-52.857	10.592	-4.990	.000	.000
Grade 2-Grade 1	-42.071	10.592	-3.972	.000	.001
Grade Pre-2-Grade Pre-1	-37.285	10.724	-3.477	.001	.008
Grade 3-Grade Pre-1	-37.125	10.854	-3.420	.001	.009
Grade 4-Grade Pre-1	-37.769	11.322	-3.336	.001	.013
Grade 2-Grade Pre-1	-26.499	10.724	-2.471	.013	.202
Grade 3-Grade 2	-10.626	10.724	-.991	.322	1.000
Grade 4-Grade 2	-11.270	11.198	-1.006	.314	1.000
Grade 4-Grade 3	-.644	11.322	-.057	.955	1.000
Grade Pre-1-Grade 1	-15.573	10.724	-1.452	.146	1.000
Grade Pre-2-Grade 2	-10.786	10.592	-1.018	.309	1.000
Grade Pre-2-Grade 3	.160	10.724	.015	.988	1.000
Grade 4-Grade Pre-2	-.485	11.198	-.043	.965	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 3.8 Post-hoc comparisons for BNC Level in long reading texts

Pairwise Comparisons of Grade



Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 2-Grade 1	-53.714	12.428	-4.322	.000	.000
Grade 3-Grade 1	-86.048	12.428	-6.924	.000	.000
Grade 4-Grade 1	-77.905	12.428	-6.268	.000	.000
Grade 5-Grade 1	-100.881	12.428	-8.117	.000	.000
Grade Pre-2-Grade Pre-1	-71.048	12.428	-5.717	.000	.000
Grade Pre-2-Grade 1	-89.333	12.428	-7.188	.000	.000
Grade 3-Grade Pre-1	-67.762	12.428	-5.452	.000	.000
Grade 4-Grade Pre-1	-59.619	12.428	-4.797	.000	.000
Grade 5-Grade Pre-1	-82.595	12.428	-6.646	.000	.000
Grade 5-Grade 2	-47.167	12.428	-3.795	.000	.003
Grade Pre-2-Grade 2	-35.619	12.428	-2.866	.004	.087
Grade 2-Grade Pre-1	-35.429	12.428	-2.851	.004	.092
Grade 3-Grade 2	-32.333	12.428	-2.602	.009	.195
Grade 4-Grade 2	-24.190	12.428	-1.946	.052	1.000
Grade 3-Grade 4	8.143	12.428	.655	.512	1.000
Grade 5-Grade 3	-14.833	12.428	-1.194	.233	1.000
Grade 5-Grade 4	-22.976	12.428	-1.849	.064	1.000
Grade Pre-1-Grade 1	-18.286	12.428	-1.471	.141	1.000
Grade Pre-2-Grade 3	3.286	12.428	.264	.791	1.000
Grade Pre-2-Grade 4	11.429	12.428	.920	.358	1.000
Grade 5-Grade Pre-2	-11.548	12.428	-.929	.353	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 3.9 Post-hoc comparisons for BNC level in all non-listening sections

3.3.2.2 Linguistic features of text

Table 3.8 presents the descriptive statistics for the six variables analyzed in order to build a profile of linguistic features of the long reading texts through automated textual analysis. All variables show an increasing trend across the grades, all in the direction predicted by the increasing levels of difficulty targeted by the EIKEN grades. One of the lexical diversity measures, VOCD, showed a slightly higher result for Grade 4 than for Grade 3. However, the difference of 0.08 indicated very little difference in this measure for these two grades. Skewness and Kurtosis results for all variables for all grades were within +/- 2 of 0, meeting the generally recommended parameters for indicating normally distributed data, and so one way ANOVAs were conducted to test for statistical significance in the means across the grades for each variable. The total number of words was not included in the ANOVAs as this distinction is strictly maintained through the test specification and quality assurance procedures, and is clearly different across the grades.

The results for Average Sentence Length (ASL), MTLD, VOCD, and the Flesch-Kincaid Grade Level (FKG) did not meet the assumption of homogeneity of variance, and for these the Welch version of the F-statistic is reported. The results of the ANOVA are shown in Table 3.9, with the results for five variables being highly significant ($p < .001$). The post-hoc tests using the Games-Howell procedure[5] are included for each variable in Appendix E. The results are generally statistically significant across all pairs, except for Grade 1 and Pre-1. For these two grades, although the trend in means were in the predicted direction for all five variables, none of the post-hoc tests were statistically significant. Similarly for Grades Pre-2 and 2, although the difference in means was in the predicted direction, with Grade 2 higher for all measures, the results were only statistically significant for the Flesch-Kincaid Grade Level. Differences between Grades 3 and 4 were statistically significant for all measures except the two lexical diversity

[5] Tukey and Bonferroni tests were also carried for all variables and could be reported for the Average Syllables per Word, as this met the equal variances assumption. As the results were identical however, the Games-Howell results are presented for all five variables.

measures. As noted above, the VOCD measure was also slightly higher for Grade 4 than for Grade 3. Grade 3 and Grade Pre-2 demonstrated differences in means in the expected direction for all variables and which were also statistically significant, except for Average Syllables per Word.

Table 3.8 Descriptive Statistics for Linguistic Features

		Words	ASL	ASW	MTLD	VOCD	FKG
Grade 4	**Mean**	159	10.69	1.36	56.03	65.99	4.59
	Median	159	10.53	1.37	56.07	68.36	4.47
	Max	173	14.80	1.46	76.25	91.69	6.59
	Min	146	8.21	1.21	43.57	46.88	2.82
	Range	27	6.59	.25	32.68	44.81	3.77
	SD	7	1.79	.07	8.13	11.52	.99
	Total N	21	21	21	21	21	21
Grade 3	**Mean**	260	13.46	1.43	63.31	65.19	6.50
	Median	260	13.55	1.44	59.70	62.54	6.62
	Max	289	16.47	1.54	97.89	93.62	8.46
	Min	238	10.88	1.29	41.86	45.55	4.09
	Range	51	5.59	.25	56.03	48.07	4.37
	SD	10	1.36	.07	13.21	10.97	1.15
	Total N	21	21	21	21	21	21
Grade Pre-2	**Mean**	307	17.28	1.47	79.79	84.96	8.46
	Median	307	17.29	1.46	82.10	83.33	8.03
	Max	324	20.53	1.60	91.43	112.63	11.04
	Min	294	15.00	1.35	48.87	60.93	6.54
	Range	30	5.53	.25	42.56	51.70	4.50
	SD	8	1.59	.07	9.91	12.57	1.17
	Total N	21	21	21	21	21	21
Grade 2	**Mean**	364	18.80	1.51	86.71	88.94	9.55
	Median	361	18.05	1.51	82.63	86.20	9.44
	Max	388	22.38	1.61	139.81	136.92	11.11
	Min	343	16.65	1.42	60.72	64.60	7.91

Table 3.8 Descriptive Statistics for Linguistic Features

		Words	ASL	ASW	MTLD	VOCD	FKG
	Range	45	5.73	.19	79.09	72.32	3.20
	SD	13	1.77	.06	19.26	18.03	.99
	Total N	21	21	21	21	21	21
Grade Pre-1	**Mean**	508	23.40	1.66	123.16	115.49	13.17
	Median	512	23.73	1.67	127.02	113.03	13.45
	Max	549	29.06	1.88	151.09	138.26	17.39
	Min	465	19.19	1.46	91.92	92.74	9.07
	Range	84	9.87	.42	59.17	45.52	8.32
	SD	20	2.87	.11	16.85	13.01	1.93
	Total N	21	21	21	21	21	21
Grade 1	**Mean**	515	25.24	1.72	135.13	122.27	14.53
	Median	514	25.00	1.71	128.48	126.09	14.62
	Max	546	36.14	1.88	179.99	170.35	18.82
	Min	473	18.89	1.63	77.70	83.68	11.07
	Range	73	17.25	.25	102.29	86.67	7.75
	SD	16	3.84	.08	30.62	21.93	1.85
	Total N	21	21	21	21	21	21

Table 3.9 Overview of ANOVA results for linguistic variables

Equal variances assumption met	**F Statistic**	**Between Groups df**	**Within groups df**	**Sig.**
Average Syllables per Word (ASW)	111.871	5	55.337	.000
Equal variances assumption not met	**Welch F Statistic**	**df1**	**df2**	**Sig.**
Average Sentence Length (ASL)	111.871	5	55.337	.000
Lexical diversity MTLD	75.349	5	54.734	.000
Lexical diversity VOCD	57.600	5	55.632	.000
Flesch-Kincaid Grade (FKG)	143.788	5	55.552	.000

3.3.2.3 Metadiscourse markers

Measures for the thirteen metadiscourse markers provided by the Text Inspector online analysis tool were obtained for the same sub corpus of 126 reading comprehension texts used for the lexical resources and linguistic features analyses reported above. Table 3.10 provides an overview of the total use of all 13 features for each grade. Following Bax et al (2013), a standardized measure per 100 words is also provided. The results are similar to those reported by Bax et al (2013) in that the total use increases as the level of the grade increases, but this appears to be more a function of the longer texts used in the higher grades. When the measures are standardized to a count per 100 words, a tendency for a slightly greater quantity of usage at the lower levels is seen, which also reflects the findings of Bax et al (2013). This trend is not uniform across all grades, however, and the differences between the levels are not as large as those reported in that study.

Table 3.10 Metadiscourse marker in reading passages.

	MM Total	MM Total Std
Grade 4	15.86	9.91
Grade 3	21.90	8.42
Grade Pre-2	27.14	8.84
Grade 2	29.48	8.12
Grade Pre-1	40.57	8.00
Grade 1	43.43	8.43

In order to make meaningful comparisons regarding the use of each metadiscourse marker across the grades, the total count for each measure in each text was divided by the total number of words (tokens) for that text to derive the percentage of words (tokens) in a text accounted for by each metadiscourse marker. Appendix F presents the descriptive statistics for the percentage of words

covered by each metadiscourse marker in Table F1. Only 11 of the markers are reported in Appendix F, as both *Announce Goals* and *Label Stage* markers were almost totally unused across all grades, with the former observed once in one Grade Pre-2 text, and the latter a total of six times, three times each in Grade 1 and Grade Pre-1 texts only. Both of these categories are in fact subcategories of a superordinate category of *frame markers* according to Hyland's categorization. *Announce goals* includes examples such as *I would like to, I will focus on*, and *label stage* includes textual organization markers such as *in conclusion, to sum up.* As noted in the literature review, the study of metadiscourse markers in EFL/ESL contexts has often been associated with writing (as is the Bax et al, 2013, study), and in particular writing in academic contexts, which was principally Hyland's concern. As the examples from these two categories show, they would appear to be more relevant to specific genres of academic writing, such as essays and reports, which would explain their absence from the mainly journalistic genre utilizing expository and argumentative discourse types used at the higher grades.

Table F2 in Appendix F shows the kurtosis and skewness results for the remaining 11 metadiscourse markers. All of the markers except for *logical connectives* showed a marked tendency to non-normality across at least one grade distribution and for most markers across more than one grade. Given the consistency of non-normality across the markers, which replicates the findings of the Bax et al (2013) study, it was decided to follow the approach adopted by those authors and to use the Kruskal-Wallis non-parametric counterpart of a one-way ANOVA to test for statistically significant differences across the grade distributions for these measures. Due to the number of tests, the output from SPSS, including box plots to visually identify patterns, have been grouped together in Appendix G Appendix G further contains the follow-up, post-hoc pairwise comparisons for each metadiscourse marker for which the initial Kruskal-Wallis test was significant. Below the results for individual categories are discussed in relation to the two broad categories of metadiscourse proposed by Hyland (1999), interpersonal and textual.

Hyland (1999, p. 7) defines textual metadiscourse as being "used to organize propositional information in ways that will be coherent for a particular

audience and appropriate for a given purpose." Interpersonal discourse, on the other hand "allows writers to express a perspective towards their propositional information and their readers." This distinction was also maintained by Bax et al (2013), who investigated the differential use of markers in these two categories in writing of test takers at CEFR B2, C1, and C2 levels. They concluded that in their data, when counts for individual markers were grouped together into these two broad categories, "significantly fewer interpersonal markers were used as proficiency levels increased, while the use of textual markers was relatively consistent." Interpersonal markers include *attitude markers*, *emphatics*, *hedges*, *person markers*, and *relational markers*, Textual markers include the four subcategories of Hyland's frame markers, *announce goals*, *label stages*, *sequencing*, and *topic shifts*, as well as *code glosses*, *endophoric markers*, and *logical connectives*.

Turning to interpersonal markers first, the box plots for *attitude markers* show a marked difference between Grade 4 and all other grades, with considerable overlap for grades Pre-2 to 1. Although the median for Grade 3 is the same as for Grade 4, there is considerable overlap for the top 50 percent of usage figures for Grade 3 with the distributions for other grades. Not surprisingly, the post-hoc tests show statistically significant results only for Grade 4 with all other grades. The Kruskal-Wallis test for the use of *emphatics* across grades was not statistically significant. Uses of the *emphatics* marker were observed fairly consistently in texts across all grades. Interestingly, these usage figures are broadly consistent with the standardized counts per 1000 words for a corpus of university texts and research articles analyzed by Hyland (1999, p. 10). Hedges show a marked jump between Grade 4 and Grade 3, with usage then plateauing from Grade 3 on. This is reflected in the post-hoc comparisons which show statistically significant results for comparisons between Grade 4 and all other grades except Grade 3. The comparison for Grade 4 and 3 is, however, very close to significance. The remaining grades, as the descriptive statistics demonstrate, show little difference in the level of usage. As a superordinate category, *hedges* appears to subsume a number of commonly used exponents that are being used in a broadly similar fashion across the generally expository texts from Grade 3 to the higher grades.

For *person markers* and *relation markers*, Grade 4 shows the highest usage, which then drops dramatically, with fairly low usage of both these markers across all subsequent grades. This pattern is reflected in the post-hoc comparisons, which only show statistically significant differences for comparisons between Grade 4 and other grades (Grade 4 and Grades 3, Pre-2, and 2 *for person markers*, and Grade 4 and Grades 3, Pre-2, 2, and Pre-1 for *relation markers*). Both of these markers contain a number of personal pronouns, and the particular genre and discourse type of Grade 4, in which a narrative story is related about a central fictional character, readily accounts for the high usage noted for these grades. For the more expository and argumentative discourse types and journalistic genre noted for the higher grades, there is certainly much less scope for the use of *person markers*, and it appears a less direct appeal to establish a position in relation to the reader through the use of *relation markers*.

Turning next to markers of textual metadiscourse, the Kruskal-Wallis result for *endophorics* was not statistically significant, and although the test for *topic shift* was initially significant, post-hoc tests revealed no significant pairwise comparisons between grades. The actual usage of both *endophoric* and *topic shift* markers was very low across all grades, and may reflect the same point made by Bax et al (2013) regarding the low use of *endophoric markers* and the four sub-categories of Hyland's frame markers seen in their data. They interpreted this as an indication that the length of the texts in their corpus may not be adequate to require such explicit organization for the purposes of establishing coherence and cohesion. It may indeed be only in longer texts, for example full journal articles or longer chapters from textbooks, that such marking would become more prominent. The usage figures are lower than the standardized usage (per 1000 words of text) noted in the university textbooks and research articles analyzed by Hyland (1999), which once again may be due to length. It may, however, also point to genre as a determining factor, with potentially important differences between the kinds of discourse structured in academic texts to the more journalistic style of feature articles employed in the upper grades of EIKEN.

Code glosses show an interesting pattern in which usage spikes from Grade 3 to Grade Pre-1, with the greatest usage in the mid-level Grade Pre-2 and

Grade 2. Post-hoc tests showed significant results for all comparisons between Grade 4 and higher grades, except for Grade 1, which had the second lowest mean and median results for this measure. The more explicit use of code glosses, such as *in other words, for example, such as,* etc, may indeed be a feature of the intermediate expository texts in Grades Pre-2 and Grade 2, in which the reading texts make use of methods of highlighting and structuring a text likely to be taught at these levels.

The lower usage figures at both ends of the grade distribution can be explained by the features associated with the text types appropriate for these levels. Grade 4 texts are short narratives telling a story about a central, fictional character relevant to the lower-secondary school students who are typical test takers for that grade, and will not require the structuring of a complex argument or message. In terms of genre, the texts reflect language-learning materials, whereas the remaining grades focus more on expository texts (and argumentative ones for the higher grades), with factual information attempting to replicate a more objective, and authentic, journalistic genre. At the higher Grade Pre-1 and Grade 1 levels, which deal with more abstract texts and complex topics, there may be a more sophisticated, less explicit approach to structuring text to achieve the same results as the explicit use of code glosses.

For *evidentials,*, as with the interpersonal *person markers* and *relation markers* described above, Grade 4 shows the highest usage, which is markedly higher than the nearest other grades. As with the discussion of person and relation markers, the noticeably high usage for Grade 4 is likely due to high usage of specific exponents, which impacts on the overall results. Although *evidentials,* includes more sophisticated examples such as *according to, argue,* and *claim,* which are not likely to be included in Grade 4, it also includes *said* and *show,* which are likely to be highly represented in Grade 4 texts, but not for usages which would exemplify the categorization of evidentials as a metadiscourse marker. In particular, Grade 4 reading texts contain many instances of direct speech attributed to the main characters in the short narratives with variations of the verb *to say*. At the same time, looking to the remaining grades, a general pattern is seen in which use of these markers increases noticeably from Grades

Pre-2 on, leading to their highest usage in the two highest grades. Statistically significant comparisons are noted between Grade 4 and all other grades, but also between Grade 3 and Grades 1 and Pre-1.

Sequencing, which also represents a subcategory of frame markers in Hyland's original categorization, includes examples such as *finally, first, firstly, last, lastly, next,* etc. There is a distinct pattern of decreasing usage across levels, with the two lowest grades having the highest usage, which then tends to level out across the higher grades. The post-hoc tests reflect this with statistically significant results between Grade 4 and Grade Pre-2, 2, Pre-1 and 1, and between Grade 3 and Grades Pre-2, Pre-1, and 1, though not between Grade 3 and Grade 2.

Logical connectives are dealt with last due to the prominence of this category in the discussion of metadiscourse and its centrality to the concept of textual cohesion. *Logical connectives* show the highest usage at the lower Grade 3 and Pre-2 levels, with usage falling noticeably in Grade 2, and although rising slightly for the higher grades, remaining lower than Grades 3 and Pre-2. Grade 4 shows the lowest usage. Post-hoc tests show the differences, however, are only statistically significant between Grade 4 and Grades 3 and Pre-2, and between Grade Pre-2 and 2. The trend however, is nonetheless interesting. For the Bax et al (2013) data of texts written by L2 learners, no significant difference was found for texts at three levels judged to be CEFR B2, C1, and C2. At the same time, those authors looked at the more detailed usage of individual exponents to reveal a pattern in which the connectives which they classed as being conceptually easier, e.g. *and, but,* etc, decreased as the level increased, while conceptually more difficult connectives tended to increase.. Bax et al (2013) cite studies with similar results which suggest that the prominence of such cohesive devices in the writing of L2 learners may be a result of the prominence of those same devices in teaching and learning materials (e.g. Carlsen, 2010, and Hawkey & Barker, 2004). While not using the same taxonomy of metadiscourse markers, Crossley et al (2011) in relation specifically to the intuitive simplification of reading material for L2 learners, note texts simplified for beginner and intermediate level learners show a greater use of explicit cohesive devices. Crossley et al (2012) used the Coh-Metrix (Graesser et al., 2004) tool to analyze their corpus. While the category

they classed as *connectives* did not show significant differences, the category they refer to as causal particles, and which includes the exponents *because, so*, and *since* classified as logical connectives in the scheme used by Text Inspector, were associated with statistically significantly greater usage in the lower level texts. Crossley et al (2012, p. 106) suggest that intuitive modification of texts in the creation of L2 reading materials "leads to texts that are predominantly more cohesive and less sophisticated as the text level decreases," suggesting that "such linguistic modifications should likely produce texts that are more comprehensible for beginning level learners." This offers the intriguing possibility that the data here also offers some corroboration for this assumption, specifically that the higher incidence of logical connectives in the Grade 3 and Pre-2 texts is associated with a deliberate modification of cohesion in these texts. The results may suggest that, at the higher grades, a more sophisticated and less explicit approach to creating these cohesive links is being employed.

3.4 Conclusions Regarding RQ1

The results demonstrate that across a range of dimensions, it is possible to distinguish distinct differences in the features associated with reading texts at the different EIKEN grades. In terms of RQ1, then, the accumulated evidence across a range of both expert judgment and automated analysis criterial features supports the conclusion that the EIKEN grades do indeed show important distinctions in terms of both contextual and cognitive parameters. These distinctions were clearest, and statistically significant for most grade pairs, for the five linguistic features derived through automated textual analysis. This result supports the selection of these features as useful indices for identifying criterial differences for the purposes of this study. Importantly for the subsidiary goals of this project, they will also be useful for developing clearer explicit specifications for ongoing item development and quality control.

The range of features adopted for expert judgment have importantly demonstrated their usefulness in determining criterial differences, but also their interpretability and practical application in the large-scale analysis of texts and items by trained judges. Of crucial importance is the adoption of item features

which allow for the evaluation and classification of features of cognitive processing in the form of the *operation* and *key information* parameters. Clearly identifying cognitive processing elements of the test task open to manipulation by trained judges provides the opportunity to truly operationalize a cognitive processing model such as that provided by Khalifa and Weir (2009) for reading, incorporating these parameters into an explicit and useable test specification model. This in turn may offer a way to operationalize the interaction of criterial contextual and cognitive parameters in the specifications for test tasks designed to target different levels of proficiency. While the results across parameters such as topic, abstractness and discourse type have generally shown trends that confirm expectations for the grades, they have also demonstrated some variability within and across grades. Identifying the optimal profile for these measures, by making reference to key features of the TLU domain tasks relevant to each grade, and incorporating these profiles explicitly into task specifications would provide useful tools for quality assurance.

At the same time, it can been seen that for many features the distinctions are not statistically significant at adjacent levels, but do define consistent features for broader proficiency bands. This is particularly true for many of the expert judgment features, in which distinctions were often noted for beginner, intermediate and advanced levels which subsumed adjacent grades. This finding replicates the findings of Alderson et al (2006) in relation to the evaluation of texts and test tasks at different levels of the CEFR. They noted that although many of the features they had proposed did not distinguish between the six CEFR levels, they did provide the potential to distinguish across three broader levels of A1 + A2, B1 + B2, and C1 + C2.

For the lexical resources associated with each grade of the EIKEN testing system, the AWL proved a clear and consistent predictor of criterial differences across levels. The BNC level at which 95 percent coverage of a text is reached also demonstrated clear distinctions, particularly between the grades targeting B1, B2, and C1, with a clear distinction between the vocabulary required to access each of these upper grades and that needed for the tests targeting A2 and A1. For these lower-level tests, the BNC criterion did not always distinguish clearly

between the more finely grained lower levels, particularly for the levels subsumed within A1, Grades 3, 4 and 5. This result however, concords with the caveats noted by Nation (2006) in relation to the use of the highest category in Bauer and Nation's (1993) six-level taxonomy of word family, which was employed in the development of the BNC lists. Nation (2006) notes that for lower-level learners, a lower level of the six-level categorization of word family would likely be more appropriate. This is particularly likely to be true for EFL contexts such as Japan where there is a greater linguistic distance between the L1 and the target language, including a completely different writing script. For Grade 5 and Grade 4 in particular, which are clearly associated with a specific learning context in junior high school in Japan, it is clear that a more finely tuned vocabulary learning goal would be more appropriate, and consequently also for the test specification required to ensure consistent and comparable item development for these grades.

The investigation of cohesion and coherence measures in the form of metadiscourse markers has also demonstrated that an easily accessible, transparent online analysis tool can provide additional useful measures for investigating criterial features of texts intended for different levels. At the same time, the relatively narrower range of research generated for these measures means that defining appropriate benchmark criteria is a work in progress. Identifying relevant forms of textual cohesion is a potentially important development for both test task specification and for the potential positive washback this would have on teaching and learning. In this light, the findings in relation to logical connectives are particularly interesting. It will, however, be necessary to explore much more thoroughly the relationship between these measures and authentic TLU domain texts to provide clearer guidelines for the interpretation of these potentially useful indicators of coherence and cohesion.

Chapter 4 RQ2: The Empirical Difficulty of EIKEN Grades

4.1 Introduction

This chapter deals with scoring validity and is divided into two parts. The three research questions addressed by this study deal with contextual and cognitive aspects (RQ1), criterion aspects of validity (RQ3), and aspects of scoring validity in terms of the extent to which the EIKEN tests can be said to measure empirically distinct difficulty levels (RQ2). The interpretation of any results obtained in the course of investigating these three questions, however, is premised on the assumption that the tests are psychometrically sound measurement instruments. Accordingly, the first part of this chapter, Section 4.2, presents a brief overview of the technical scoring properties of the tests, focusing on the test forms used as a basis for the standard-setting carried out to answer RQ3 in Chapter 5 The second part of this chapter, Section 4.3, describes the methodology, data collection, analysis and results of the vertical scaling study carried out to investigate RQ2.

4.2 Scoring validity of the EIKEN Tests

Table 4.1 contains the descriptive statistics, with indicators of test and item performance, for the test forms of each grade used in the CEFR linking study described in Chapter 5 as a part of investigating RQ3. The table also includes Grades 5 and 4, which are not included in the CEFR linking study (see Chapter 5), but still form part of the focus here on the technical performance characteristics of the EIKEN testing program overall. Table 4.1 gives an overview of the results of Classical Testing Theory analyses conducted by the Eiken Foundation on operational data from a live administration of one First Stage test form from each grade of the Eiken tests, including estimates of reliability, Standard Error of Measurement (SEM) and mean point-biserial correlations. Each of the forms was

administered at all public test centres within Japan and internationally on one of the three official test administrations conducted within each academic year (see Chapter 1 for details of the EIKEN tests). The statistics shown here represent the routine, post-hoc CTT analyses conducted on all test forms administered operationally. Note that the number of test takers in the analyses is less than the total number of test takers for typical official test administrations presented in Chapter 1. Post-hoc analysis of all operational test forms is carried out prior to final confirmation of pass/fail decisions and the release of test results for each grade. Due to the very tight schedule for returning results to test takers, a large set of test taker data is sampled to maximize the speed of the analysis software and procedures while ensuring that the analysis results accurately represent the technical performance characteristics of each form for that administration.

While the main focus in this section is on establishing the suitability of the measurement properties of the test forms used as the basis for standard setting for RQ3, it also allows us to consider the measurement properties and analysis procedures used for the EIKEN tests generally. As noted in Chapter 2, NR reliability estimates require an understanding of the sample from which they are derived for interpretation. They should not be generalized to a sample which may be significantly different in key characteristics from the one used for the analysis. In that respect, the sample sizes used as the basis for calculating the statistics in Table 4.1 are large, with 20,000 test takers for each grade from Grades 5 to 2, and just under 20,000 for Grade Pre-1, and over 6000 for Grade 1. The smaller numbers for Grade 1 reflect the position of this grade in operational use as a high-level certification test with a smaller number of total test takers relative to the other grades. The numbers for Grade 1 actually represent almost the entire body of test takers for this grade in one administration (see Table 1.1 in Chapter 1 for an overview of the number of test takers for a full operational year), The samples used here then are sufficiently large to give confidence that they are representative of the general test taker population for typical EIKEN administrations. The First Stage of the EIKEN tests utilize what Cizek and Bunch (2007) describe as a compensatory scoring model, in which performance on one section of the test can compensate for performance on another section as it is the

total score across the whole test which is used as the primary score scale for decisions. While other possibilities for combining scores exist, the compensatory approach is generally the most common (Cizek & Bunch, 2007, p. 20). As Kaftandjieva (2010, p. 15) notes, "as the summarized result usually has a higher reliability than the separate components, in the absence of other considerations this strategy is to be recommended." The statistics in Table 4.1 reflect this approach and were calculated using the responses to all dichotomously scored items contained in the First Stage tests. For Grade 5 to Grade 2, all items are weighted equally and scored as dichotomous multiple-choice items, meaning the total number of items equals the total possible raw score. The structure of the Grade 1 and Grade Pre-1 First Stage tests differs slightly in that some items in the Reading and Listening sections are weighted differently, and the Writing component of both grades is a constructed-response performance task rated by human raters using a holistic rating scale. The statistics reported in Table 4.1 for Grades 1 and Pre-1 are calculated only on the unweighted raw scores for the Vocabulary, Reading and Listening components, which contain only dichotomously scored multiple-choice items.

As the pass-fail decisions for the First Stage test are based on the overall performance on the whole test, the operational test performance statistics in Table 4.1, which are routinely calculated for quality assurance, are thus based on the total scores. It is important to bear in mind that these results thus pertain to sections other than the reading sections which forms the main focus of the three research questions.

Table 4.1 Test statistics for test forms used in CEFR linking study

Statistic	G1	GP1	G2	GP2	G3	G4	G5
Items	68	70	75	75	65	65	50
Test takers	6276	19525	20000	20000	20000	20000	20000
Mean	39.9	41.1	40.8	43.5	44.2	42.7	38.6
SD	8.6	9.9	10.6	10.2	9.3	10.8	7.0
Min	1	1	1	6	0	0	0
Max	68	70	75	75	65	65	50
Median	40	41	40	43	44	43	40
Skewness	-0.174	-0.028	0.386	0.353	-0.082	-0.025	-0.566
Kurtosis	-0.035	-0.837	-0.394	-0.323	-0.769	-0.918	-0.794
mean % correct	0.59	0.59	0.54	0.58	0.68	0.66	0.77
Reliability (alpha)	0.82	0.86	0.86	0.87	0.87	0.90	0.86
SEM	3.6	3.8	3.9	3.7	3.3	3.4	2.6
pbis	0.28	0.30	0.30	0.31	0.33	0.37	0.36

Table 4.2 Passing scores Grade 1 and Pre-1

	G1	GP1
Total raw score VRL (unweighted)	68	70
Total VRL + W (weighted)	113	99
Cutscore as percentage	70%	70%
Cutscore (unweighted)	48	49
Cutscore (weighted)	79	69

Table 4.3 Overview of the passing scores for Grades 2, Pre-2, 3, 4, & 5

	G2	GP2	G3	G4	G5
Total raw score	75	75	65	65	50
Cutscore as percentage	60%	60%	60%	60%	60%
Cutscore	45	45	39	39	30

In this section, we are primarily interested in the indicators of test and item performance in the table—reliability, SEM, and mean point-biserial correlation—in order to evaluate the technical quality of the EIKEN test forms used in Chapter 5 from the perspective of scoring validity. The reliability and SEM estimates in table 4.1 are derived using internal consistency measures appropriate for NR-type inferences (see Chapter 2 for a discussion of the various indices of test and item performance used in this section). The types of feedback provided by the EIKEN tests allow for both NR and CR interpretation. In addition to considering the data contained in Table 4.1, we will make use of several CR-based reliability estimates as well to more fully evaluate the technical adequacy of the tests. The CR reliability estimates were generated specifically for this study by using formulas described in Chapter 2 which allow estimates of these indices to be made using only the results of a typical statistical analysis such as that contained in Table 4.1.

Firstly, it is necessary to consider the question of whether the score distributions on which the above statistics are based are normally distributed. Following Bachman's (2004) suggested guidelines noted in Section 2.4.1, it can be seen that all of the score distributions can be considered to be normally distributed and all fall within the guidelines for accepting the assumption of normality.

Reliability for the test forms administered in Table 4.1 is reported in the form of Cronbach's alpha. In Chapter 2, rules of thumb were reviewed from authors writing on internal consistency reliability in the field of language testing and assessment, along with reports of operational best practice in large-scale EFL/ESL testing programs. That review supports an interpretation of a broad range of internal consistency estimates of between .8 and .9 as being generally reasonable figures for a set of EFL level-specific tests such as the EIKEN testing program.. From this perspective, all of the figures reported in Table 4.1 are acceptable. The Grade 1 test falls at the lower end of the spectrum, but it should be noted that this test in particular is perhaps more prone to the effects noted in Chapter 2 by Jones (2001), Saville (2003), and Weir (2005a), as it contains a range of more varied item and task types than other levels, particularly in the

listening section and has a much smaller number and narrower range of test takers given the advanced level of the test. As the pass/fail classification is designed to separate test takers into two distinct levels with one cutpoint, taking into consideration the recommendations of Kaftandjieva (2004) in Section 2.4.1 of Chapter 2, and bearing in mind the figures are generally comparable for those reported for other high-stakes EFL testing programs, it is safe to conclude that the levels of internal consistency are appropriate for the stated uses and interpretations of the EIKEN tests for the forms of the seven grades reported here. The issue of the dependability of the classifications from the perspective of CR-type reliability estimates will be considered after reviewing the precision of the test scores and estimates of item discrimination using SEM and the point-biserial statistic respectively.

As noted in Chapter 2, SEM provides an estimate of the precision of the test scores. To illustrate the use of this statistic, consider the SEM for the Grade 1 test form reported in Table 4.1, which is reported as 3.6. An examinee's score will fall within $\pm$ 1 SEM of his or her observed score approximately two-thirds, or 68% of the time, between $\pm$1.96 (approximately 2 SEMs) 95% of the time, and between $\pm$ 2.58 SEMs 99% of the time (Bachman, 2004, p. 173). If an examinee had an unweighted raw score of 40 on the First Stage of the Grade 1 test reported in Table 4.1, the estimate of SEM can be used to say that we are 68% sure that his or her true score falls within a band of 36 to 44 raw score points (see also Chapter 5, Section 5.3.5.3 for a discussion of the use of SEM and examples of the calculation of confidence intervals in relation to cutscores set during standard setting studies for RQ3). Referring back to the generally accepted benchmarks for SEM noted in Chapter 2, as with reliability estimates, the levels of SEM reported for the forms of the EIKEN tests shown in Table 4.1 are comparable to levels accepted as good practice across a range of EFL tests in the field. The values for item discrimination estimated through the point-biserial statistic (r_{pbis}) in Table 4.1 are the mean values across all items in each form of the test.

The primary decision of interest for test takers taking the EIKEN test is the pass/fail decision leading to certification at the level of proficiency targeted

by the grade of the test they have taken. Thus reliability needs to be considered from both NR and CR perspectives. The indices for CR reliability described in Chapter 2 can be calculated from the statistics already available in Table 4.1. Utilizing the values from Tables 4.2 and 4.3 to calculate Z for each grade, and utilizing the reliability indices for each grade, it is possible to provide estimates of p_o and κ for each grade in Table 4.4 below. Note that in deriving the values for p_o and κ from the table, the recommendation of Subkoviak (1988) is followed to round both the reliability estimate and Z to the nearest appropriate value for use in the table. For example, for Grade 1, the reliability will be rounded to .80 and Z will be rounded to 0.90.

Table 4-4 Estimates of p_o and κ for test forms in Table 4.1

	G1	GP1	G2	GP2	G3	G4	G5
Z	0.94	0.79	0.39	0.14	-0.57	-0.34	-1.24
p_o	0.87	0.9	0.87	0.86	0.88	0.86	0.93
κ	0.55	0.69	0.71	0.71	0.7	0.71	0.66

As noted in Chapter 2, p_o is more interpretable in real-world terms as it quantifies the proportion of consistent decisions in absolute terms (e.g. for the form of Grade 1 in Table 4.1, 87% of classification decisions are consistent). From the guidelines given by Subkoviak, (1988, pp. 52-53) reported in Chapter 2, it can be seen that all of the test forms in Table 4.1 have sufficient values of p_o, and for kappa, all grades except Grade 1 fall within the recommended range, with Grade 1 falling slightly below at .55. The squared loss agreement index phi lambda (ϕ_λ) is also calculated as an estimate of CR reliability. Following the worked example provided in Brown and Hudson (2002, p. 196), the values from Table 4.11 are used to calculate ϕ_λ for Grade 1 below. The same process has been followed to derive ϕ_λ estimates for each grade in Table 4.5. For Grade 1, k=68, $M_p = .59$. To calculate $S^2{}_p$ the standard deviation of the test scores in Table 4.1 is divided by the number of items, 68, to derive S_p, or the standard deviation of the proportion correct scores, and then square this to derive $S^2{}_p$.

Thus, $\frac{8.6}{68}$ = 0.126588235, and 0.126588235^2 = 0.016024581. Putting these values into the formula from Chapter 2, the value for ϕ_λ for Grade 1 is:

$$\text{Formula 4.1}\quad \phi_\lambda = 1 - \frac{1}{68-1}\left(\frac{.59(1-.59) - 0.016024581}{(.59-..70)^2 + 0.016024581}\right) = .883$$

Brown and Hudson (2002, p. 196) describe the value of .83 derived in their worked example as "a moderately high value indicating a fair amount of consistency in classifications." The figures for the phi lambda CR reliability index in Table 4.4 indicate sufficiently high results, with the corollary of an appropriately high amount of consistency in the classifications made by the test forms used in Table 4.1.

Table 4.5 Phi lambda estimates based on Table 4.1 statistics

Grade	G1	GP1	G2	GP2	G3	G4	G5
ϕ_λ	0.883	0.902	0.867	0.840	0.886	0.900	0.935

The primary purpose of the current section is to consider whether the test forms used in the standard setting study can justifiably be said to meet the minimum requirements of technical proficiency from the perspective of scoring validity. At the same time, given the large sample size indicative of typical EIKEN administrations, it is suggested that the results can be taken, with due caution, to provide some insights into the general measurement properties of the tests. Many other aspects of the scoring validity, including the impact of the established practice of releasing all test forms publicly on the operational aspects of ensuring scoring validity and the approach to maintaining comparability across forms will need to be addressed in the creation of a fully comprehensive validity argument. As Chapter 4 is focused mainly on answering the requirements of RQ2 through vertical scaling, however, a more detailed discussion of the general scoring properties of the tests addressing specifically the relationship to the typical uses and interpretations of each grade is beyond the scope of this study. The discussion carried out in this section will be taken as sufficient evidence to justify investigating the three research questions on the premise that the scoring validity

and technical properties of the EIKEN tests, in particular the forms discussed in detail in the investigation of RQ3 in Chapter 5, are indeed sufficient.

4.3 The Use of Vertical Scaling to Investigate RQ3

4.3.1 Rationale for vertical scaling study

The rationale for carrying out a vertical scaling study lies primarily in the ability to validate the claim that each grade in the EIKEN testing program contains items at a level of difficulty which is: 1) empirically distinguishable from other grades; and 2) accords with its putative position within the gradually increasing, stepped set of levels covered by all seven grades in the program. As noted in the literature review in Chapter 2, vertical scaling methodology has been used for the purpose of calibrating tests targeting the same construct but which are constructed to differing specifications because they target different parts of what is assumed to be a larger ability or proficiency scale. RQ2 is designed to elucidate the empirical level of difficulty underlying the common frame of reference within which all of the level-specific tests are assumed to belong

As noted in Chapter 2, Brown et al (2012) have already carried out important exploratory work on using Rasch analysis to facilitate vertical scaling with the EIKEN grades. That work serves as an important pilot study in terms of demonstrating the feasibility of vertical scaling with the Rasch model for the EIKEN tests. At the same time, while the present study builds on many important features of the Brown et al (2012) methodology, there are several important differences. Firstly the Brown et al (2012) study focused on the upper three levels of the EIKEN grades, Grades 2, Pre-1, and 1, as these were the grades of primary interest for the research questions in that study. Given the narrower range of levels covered, that study thus was able to make use of the external anchor test design for linking, which was considered inappropriate in the context of linking all seven grades which span a much wider proficiency range. Secondly, the Brown et al (2012) study was conducted as a part of investigating the appropriacy of using EIKEN tests for proof of English proficiency in the context of university admissions for English-medium universities. The study utilized a sample appropriate to the study aims, and was administered in an English medium

university outside Japan. Thirdly, as the primary purpose of that study was to enable the comparison of performance of a common group of test takers on both the EIKEN tests and an external criterion test of English for academic purposes, the study utilized a small sample which took only one form of each relevant grade. For the purposes of this study, it was essential to utilize test takers representative of the typical test taker population spanning all seven grades. In addition, the methodology needed to facilitate the calibration of large numbers of previously administered test forms.

4.3.2 Methodology for RQ2

4.3.2.1 Selecting a scaling methodology and establishing the baseline scale

In determining a suitable vertical scaling strategy, the methodology had to account for two principal processes: firstly, establishing the vertical scale across the seven grades; and secondly, retrospectively calibrating items previously administered in live test forms to that vertical scale. As already noted in Chapter 1, each EIKEN grade is a level-specific test. The tests are designed to measure distinct levels of a common construct of language proficiency. However, these distinctions and the links between grades in terms of relevance to that common construct of language proficiency have traditionally been defined in terms of content specification. In any one operational test administration, no potential mechanism to facilitate linking in the form of overlapping items exists between forms targeted at different levels, which precludes common-item approaches to post-hoc vertical scaling. It is important to reiterate once again that all test content is made publicly available following administration, which impacts on the ability to facilitate research questions such as RQ2. The first major issue for the vertical scaling methodology then was how to establish the vertical links for the initial baseline scale.

In order to establish the vertical scale initially, it was decided to use a common-item non-equivalent groups design, and recruit participants to take part in a specially designed test session in order to establish a baseline version of a vertical scale linking item difficulty across the seven grades. Recruiting participants for research and pretesting purposes is a notoriously difficult process (as recognized by Alderson et al, 1995, p. 99, in their discussion of the lack of

pretesting among UK EFL examination boards). A common-person design was considered impractical, as the requirement to take more than one version of the test greatly increases the burden on test takers and reduces the total number of items that can be administered for the same sample of participants. Brown et al (2012) used a common-item, non-equivalent groups design, utilizing an external anchor test to facilitate vertical linking of test forms from Grades 2, Pre1, and 1. This is also referred to as the scaling test approach noted in Chapter 2. However, a design in which all participants take a single external anchor test in addition to an on-grade test appropriate to their level would not be appropriate for this study. This study needed to encompass test takers from Grade 5, the most elementary level of proficiency (and generally taken by examinees in their early teens) to the advanced-level Grade 1 test targeted at a C1 level of proficiency (and generally taken by adults). As noted by Kolen and Brennan (2004), in such cases large numbers of items on the anchor test will be either far too difficult or far too easy for many test takers. An overlapping, common-item non-equivalent groups design promised to be the most effective and practical approach. Examples of such designs are included in Kenyon et al (2011) and North (2000).

For the EIKEN tests an existing pretesting program is run to administer blocks of grammar and vocabulary items outside of the operational testing sessions for research, item development, and quality assurance purposes. This process allows some pretested items of known technical properties to be included in each operational test form. These items are scored operationally as part of the operational test form, but also facilitate the post-hoc analysis of operational test form performance. However, test forms used in pretesting sessions at the time of this study only contained items targeted at a specific grade, and thus in the same way as operational test forms, did not contain a vertical linking mechanism to connect forms at different levels. The pretesting program, however, offered the possibility of utilizing an existing system of recruiting test takers broadly representative of the general testing population and administering blocks of test items to them. Test forms would need to be modified to ensure sufficient linking across grades, but this approach would avoid the need to undertake an expensive, separate administration for the purpose only of vertical scaling on top of the

ongoing operational commitments of live testing and pretesting. This promised to provide the means of carrying out the first important step noted above, namely the creation of a baseline, vertical scale spanning all seven grades.

Prior to this study, IRT methodology has been employed internally to carry out ongoing research on the feasibility of constructing a vertical scale by using the pretesting items, utilizing both 3-parameter and 2-parameter models. The results of these ongoing analyses have also been used as an important source of information for reference in quality assurance (personal communication, Eiken Foundation Research and Test Development Section, March, 2014). As noted in Chapter 2, however, these models require very large sample sizes to derive stable parameter estimates. Creating and maintaining a large-scale operational vertical scale through pretesting with these models poses a great many logistical issues for a large-scale testing program. In addition, the vertical links established through that research strand had been developed prior to the most recent revisions of the different grades noted in Chapter 1 (personal communication, Eiken Foundation Research and Test Development Section, March, 2014). Building an operationally viable vertical scale with the potential for direct operational usage would thus require the establishment of vertical links across the grades within the paradigm of the integrated evaluation of the current EIKEN tests addressed by this study—across key contextual, cognitive, scoring and criterion aspects of validity—and this required a practically feasible scaling methodology for establishing and maintaining those links. Sample size was thus an important consideration. Various recommendations for appropriate minimum sample sizes for use with the Rasch model range from 100 (North and Jones, 2009) to 250 for high stakes usage (Linacre, 1994). Thus, in consideration of ensuring the ongoing potential practicality of vertical scaling procedures, and considering the wide-spread use of the Rasch model in vertical scaling generally and language testing specifically, using the Rasch model to establish a base vertical scale and recalibrating previously administered items to this new scale promised to provide the best balance of theoretical robustness, technical adequacy, and ongoing practicality and efficiency.

4.3.2.2 Calibrating items in previously administered test forms

The next important methodological issue was how to move from the baseline scale to the calibration of large numbers of previously administered test forms. This would allow more confident generalizations to be made regarding the differences between EIKEN grades than if only a small-scale, experimental design were used. To address this issue, reference was made to an innovative study by Saida and Hattori (2008), which describes a way of calibrating test data obtained from previously administered test forms. Saida and Hattori (2008) equated test data obtained from eight separate test forms from a prefectural achievement test for first year high school students administered in Akita prefecture in Japan. One test form was originally administered each year from 1995 to 2002, but the forms across years did not contain common items or test takers. In order to facilitate the comparison of ability measures for first year students in different years, Saida and Hattori (2008) needed a way of retrospectively equating the performance of test takers across different years. To overcome the lack of any links between the data sets, they created specially constructed equating tests consisting of subsets of items from the main tests. Each equating test contained some items from the baseline test, the 1999 form, and some items from one of the forms administered in other years, referred to as the target forms. Each equating test was administered to approximately 400 first year high school students, allowing the item parameters from each target form to be equated to the baseline test form measurement scale. The original data sets for the live administrations of the eight forms, comprising response data from approximately 140,000 test takers across eight years, were then analyzed using a 2-parameter IRT model, with the items that were included in the equating forms anchored at the item difficulty parameters estimated during the equating analysis. The ability measures for the original 140,000 test takers were thus estimated on a common measurement scale, enabling the comparison of average group performance across years.

It is important to note that there are significant difference between the Saida and Hattori (2008) study and this one. Firstly, that study was an equating study, designed to link tests created to equivalent specifications and targeting the same level of ability. Secondly, Saida and Hattori were interested in calibrating

items from a relatively small number of test forms, eight, and had access to a large pool of participants for the administration of equating tests through the cooperation of secondary schools. Thirdly, their primary focus was on estimating ability parameters for the 140,000 first year high school students who had taken the tests over an eight-year period, whereas this study is primarily focused on empirical item and test form difficulty. Nonetheless the methodology employed by Saida and Hattori (2008) provides a useful model for adaptation for the purposes of carrying out vertical scaling of previously administered EIKEN test items.

In order to answer RQ2 by calibrating large numbers of previously administered test forms rather than focusing on a small number of test forms as an indicative example of empirical difficulty across grades, and at the same time enable the creation of an item bank containing empirical difficulty measures for all previously administered items on a common scale, the following three steps were devised:

1. Construct the vertical scale by administering a subset of previously administered grammar and vocabulary items to a sample of test takers taking part in the pretest program. One test form is constructed for each grade and administered to test takers at the appropriate level, but each form will be linked both up and down to the forms at adjacent levels through common items. The data is analyzed using the Rasch model to create a common vertical scale.
2. For each grade, multiple forms of previously administered grammar and vocabulary items are constructed and administered to test takers participating in the pretesting program. Each test form will contain anchor items from the test forms administered in Step 1, and these items will have their Rasch difficulty values anchored to the values calibrated in Step 1. Each form is taken by a group of on-target examinees appropriate for the level of the form. Non-anchor items are equated horizontally to the vertical scale through the anchor items.
3. The original data sets obtained during previous administrations of complete test forms are reanalyzed using the Rasch model. Operational test forms for all grades contain grammar and vocabulary sections, and the

majority of items contained in previously administered test forms will have already been calibrated to the vertical scale in Steps 1 and 2. These grammar and vocabulary items are treated as anchor items with their Rasch difficulty values fixed at the values calibrated in Steps 1 and 2. This will allow all other items contained in the form to be calibrated to the vertical scale through these anchor items.

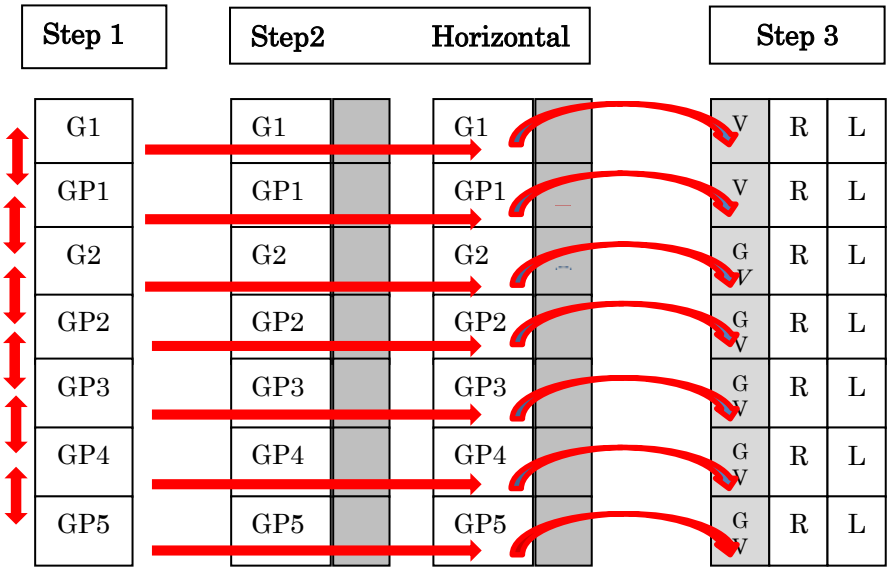


Figure 4.1 Overview of vertical scaling for Steps 1, 2, & 3

Figure 4.1 presents a simplified visual representation of the three steps. Steps 1 and 2 both take place within the pretesting program, utilizing responses from test takers recruited for that purpose. Step 3 utilizes response data previously collected in live administrations. The arrows indicate the direction of linking in each step. In Step 1, the red vertical arrows indicate overlapping items linking each test form to the adjacent forms both above and below. A group of items from each grade form in Step 1 is selected for use as an internal anchor for forms of the same grade in Step 2. Each form contains the anchor items appropriate for that level along with other grammar and vocabulary items previously administered in live tests, and these items are equated horizontally to the vertical scale through the anchor items. A large number of test forms are administered in Step 2 (Figure 4.1 presents two forms for ease of presentation only). In Step 3, the Grammar and

Vocabulary Sections (for G1 and GP1, vocabulary only) act as an internal anchor set. Most of the grammar and vocabulary items in each operational set will have been calibrated in Steps 1 and 2. These items are anchored at the values calibrated in Steps 1 and 2, to allow the calibration of all items in that particular test form. Figure 4.1 presents one test form as a representative example, but the same procedure is followed for multiple test forms across multiple years of past administrations.

The design draws on methodology for scale maintenance from Kolen and Brennan (2004) and Tong and Kolen (2006, 2011) described in Chapter 2. Once a vertical scale has been created, as noted previously, it is important to consider issues of scale maintenance. Creating a vertical scale is a time-consuming and resource intensive project. It may not be practical to implement the kind of data collection designs necessary to link complex assessments spanning different levels in every administration or for every new form. One approach to overcoming this problem is the use of horizontal equating to link subsequent test forms back to the original vertical scale. In the horizontal equating stages, only on-grade anchor items are used within a grade-specific target test form to enable the calibration of the new items in that form back to the vertical scale. The methodology employed for this complex vertical scaling project thus employs a hybrid approach similar to the methodology described in Tong and Kolen (2008), in which vertical scaling data collection designs and analysis methodology is employed in the initial creation of the vertical scale in Step 1, and subsequently horizontal equating methodology is employed to calibrate on-grade items back to the vertical scale in Steps 2 and 3.

4.3.2.2 Scope of vertical scaling

Table 4.6 provides an overview of the test forms for which it was possible to calibrate enough items in Steps 1 and 2. For each grade, the first and last test administration for which it was possible to calibrate items is listed. The number outside of the brackets is the year, and the number inside the brackets refers to one of the three official administrations carried out in each academic year (for example, the forms calibrated for Grade 1 include all those from the first

administration in 2004 to the third administration in 2006). Below the final administration is the total number of items included across all forms of the First Stage tests for that grade in that period, and below that is the equivalent total number of test forms.[6]

Given the large number of test forms across grades in the relevant years, it only proved possible to calibrate forms from part of the full range of tests within the relevant period. A number of practical constraints impacted on the number of test forms recalibrated in practice. These include differences in the number of items that could be included in test forms for different grades in Steps 1 and 2; differences in the minimum requirements for the numbers of anchor items required for Step 3 due to differences in the number of items in operational tests for each grade; differences in the potential pool of test takers and participation rates across grades for the pretesting program; and differences in the number of years to be covered since the most recent revision for that grade.

Table 4.6 Overview of test forms calibrated in Step 3

	G1	GP1	G2	GP2	G3	G4	G5
First	2004 (1)	2004 (1)	2003 (1)	2003 (1)	2002 (1)	2002 (1)	2002 (1)
Last	2007 (3)	2010 (3)	2009 (2)	2009 (2)	2006 (2)	2006 (2)	2005 (3)
Items	828	1491	3750	2250	1885	1885	1200
Forms	12	21	50	30	29	29	24

4.3.2.3 Participants

Tables 4.7 and 4.8 provide an overview of the sample sizes for Steps 1 and 2. As Step 3 utilizes existing response data from operational testing, no recruitment for this stage was necessary. Test forms for analysis in Steps 1 and 2 were administered at the same time as a part of the ongoing pretesting program. The analysis of the data was carried out in sequence, with test forms for Step 1 analyzed first, and then test forms for Step 2 analyzed with anchor values fixed

[6] Note that for Grade 1 and Grade Pre-1, the total number of items differs from Table 4-1. The total number of items for G1 and GP1 here is 69 and 71 respectively, as the extended response writing task was included in the analysis as a single item analyzed with the partial credit model. Winsteps allows for the inclusion of mixed format item types through the specification of models in the control file. This approach was piloted in the Brown et al (2012) study.

from their estimation in Step 1. For the pretesting program, participants are normally recruited for one of two testing sessions. The first session takes place in six public test centres located in different parts of Japan to ensure the sample is spread over a number of regions and cities. This session covers Grades 2, Pre-1, and 1. The second session covers Grades Pre-2, 3, 4, and 5 and is administered at closed sites located at schools around Japan. Participants for pretesting for these grades are recruited directly through schools, as the main testing population for theses grades is drawn from junior and senior high school students. Cooperating with schools to participate in the pretesting sessions for the elementary to lower intermediate EIKEN grades is a practical and efficient way of accessing a comparatively large and representative sample for these grades. However, one trade-off that is entailed by this approach is that a shorter testing time is necessary to reduce the administrative burden on teachers and allow for administration within class schedules. The longer time available for administration at public test sites means that it is possible to include a larger number of items in test forms for those sites.

For the upper three grades, invitations to participate were sent to test takers who had recently taken an EIKEN grade, including both passing and failing candidates. Given the scope of this study, as evidenced by the number of items covered across grades in Table 4.6, and the pool of participants available, it was necessary to send invitations to test takers who had taken an EIKEN grade in one of the previous four official administrations. Test takers were allocated a test set appropriate for the same grade which they had taken in the official administration. For Grades Pre-2 to 5, the schools participating recruited participants and allocated those participants according to the appropriate level of difficulty. Guidelines for this process are given by the Research and Test Development Section, and teachers are instructed to allocate test takers to the grade most appropriate for their level.

Tables 4.7 and 4.8 provide an overview of the numbers of participants in each of Step 1 and Step 2. The analysis in order to create the vertical would create the foundation for all subsequent calibrations, and so a higher minimum number of participants was set as a goal for recruitment for these sets. As Table 4.7

demonstrates a very large number of test takers were recruited for the vertical scaling forms administered as a part of Step 1, with many more than the minimum required thus ensuring the robustness of the analysis for this critical step. The *dropped* and *% dropped* categories refer to the number of test takers dropped from the final estimation of item parameters due to misfit. Table 4.8 presents the planned and actual total number of test takers recruited for the test forms administered for the horizontal equating to be conducted in Step 2. A minimum of 250 test takers per form was set as a goal to meet the recommendation for 250 test takers per form by Linacre (1994). A slightly lower minimum level was initially set for Grade 1 as this advanced level has a much smaller pool of test takers on which to draw. As Table 4.8 demonstrates however, a sufficient number of test takers was recruited to ensure that the average number of test takers per form for all grades exceeded 250.

Table 4.7 Number of test takers participating in Step 1

	G1	GP1	G2	GP2	G3	G4	G5
Goal per form	330	330	330	330	330	330	330
Total	391	435	458	597	1329	456	701
Dropped	14	17	30	50	127	53	89
% dropped	3.6%	3.9%	6.6%	8.4%	9.6%	11.6%	12.7%

Table 4.8 Number of test takers participating in Step 2

	G1	GP1	G2	GP2	G3	G4	G5
Goal per form	220	250	250	250	250	250	250
No. of forms	2	4	8	9	10	10	10
Total N	564	1310	2053	4053	8513	3746	4379
Average per form	282	328	257	450	851	375	438
Dropped	1	4	41	165	211	233	479
% dropped	0.2%	0.3%	2.0%	4.1%	2.5%	6.2%	10.9%

4.3.2.4 Step 1: establishing the vertical scale

Establishing the vertical scale involves a number of important decisions which were identified in the literature review in Chapter 2. Those decision in relation to Step 1 are described below: 1) The data collection design to ensure linking across forms; 2) the calibration method; 3) the content of items to be used in the vertical scaling forms and which will act as anchors for horizontal equating in Step 2; 4) The quality control procedures used in the analysis and estimation of difficulty measures on a common vertical scale.

The common-item linking design is shown in outline in Figure 4.2. A total of seven test forms were constructed. Each form was targeted at a specific grade, and test takers at the appropriate level were allocated to that form so that test takers were on-grade in terms of the core items for that form. All forms were connected to both the grade above and grade below, except for Grade 5, which was connected only to the grade above, and Grade 1, which was connected only to the grade below. All test takers thus took a core block of on-grade items for their level, as well as some items from the grade above and some items from the grade below. For a particular grade-level form, the more difficulty items amongst the core on-grade items would be inserted into the grade-form above, and the easier items would be inserted into the grade form below. In this way, although test takers were taking off-grade items, the distance between these items and the ability of examinees was constrained as far as possible to provide a gradually increasing set of links across all grade forms..

The total percentage of items in each form which are shared with other forms is very high, reaching 100 percent for Grades Pre-1, Pre-2, 3 and 4. As described above in Section 4.2.2.3 on participants, the differences in administration between public and closed test sites impose different time constraints on the grade forms used in the two different types of test centres. The total number of items in Grade Pre-2 was thus fewer than in Grade 2, meaning that more on-grade items were included for that Grade 2. Grade 1 and Grade 5 share links in only one direction. For horizontal equating purposes, a commonly cited rule of thumb for the minimum proportion of common items is 20 percent (Kolen & Brennan, 2004, p. 271). This rule has also been adapted for vertical

scaling (e.g. Reckase, 2010). In relation to vertical scaling when using the concurrent calibration method, North and Jones (2009, p. 4) recommend that "the safest method is that applied by Cito: anchor each test 50% upwards and 50% downwards to its adjacent tests in such a way that every item is an anchor item, except 50% of the items on the highest and on the lowest test forms." That recommendation was followed as far as possible within the constraints of the different permissible test lengths across grades.

Concurrent calibration was selected as the estimation method for creating the vertical scale in Step 1 (see Chapter 2 for a description of the various estimation methods available). Kolen and Brennan (2004) note two principal benefits for concurrent calibration. Firstly, practicality in that only one analysis needs to be conducted with the final parameters estimated already placed on a common vertical scale, as opposed to separate calibration which requires separate analyses for each form and the construction of equating coefficients to link forms. Secondly, concurrent estimation also benefits from a wider frame of reference, utilizing all responses to the common items from across different forms, resulting in more robust estimation. As noted in Chapter 2, recent studies have indicated little difference in results for comparisons of concurrent and separate calibration (e.g. Paek, Young, & Yi, 2008; Pomplun et al, 2004). Brown et al (2012) had also demonstrated the viability of concurrent calibration specifically in the context of vertical scaling for the EIKEN tests, albeit across a smaller range of grades.

The items selected for all forms in Steps 1 and 2, and which would then be used as anchor items in the calibration of previously administered test forms were drawn only from the grammar and vocabulary section which is common to all seven grades. The decision to select anchor items for horizontal equating in Step 3 from the Grammar and Vocabulary items had a particularly pragmatic rationale. The short nature of the items allows for efficient administration of a sufficiently large number of forms in Step 2 to calibrate enough grammar and vocabulary items to cover a large number of the previously administered operational post-revision forms for each grade. At the same time, the consistency of the item format across all grades, and the very common usage of these item types in language testing and teaching contexts in Japan, would reduce the chance

of item/task format effects impacting on test takers at different grades in different ways. It is also easier to control for difficulty in the common items for adjacent levels, ensuring that the off-grade items are not too far from the on-grade test takers.

It is of course necessary to recognize that the use of items targeting a particular aspect of the overall construct being tested is a tradeoff. It would appear that in many situations pragmatic adjustments to the guidelines on common-item sets are made in the course of balancing the complex set of demands usual in vertical linking studies for large-scale assessments. For example, the performance elements posed problems for the creation of a vertical scale for the WIDA Access for ELLs described by Kenyon et al (2011) because of the difficulty of including overlapping items across different levels for the constructed response writing and speaking tests. Kenyon et al (2011) adopted the pragmatic solution of using reading items already calibrated to the vertical reading scale as linking items for the writing, and listening items for the speaking tests. Brown e al (2012) also describe successfully using an anchor test consisting of a reduced set of items taken from vocabulary and multiple-choice gap-fill reading item types common to the upper grades of the EIKEN tests as a scaling test to link the four skill components of the EIKEN grades in their study.

At the same time, as noted in Section 2.3., research has consistently shown grammar and vocabulary to be highly correlated with other aspects of language proficiency and to perform well as predictors of ability on the other aspects of language proficiency covered by the First Stage tests. Thus, the selection of items as anchors was made based partly on reasons of practicality and efficiency, but was also made on the basis of the potential ability of those items to adequately capture the range and variability of performance by examinees across grades and across other aspects of language proficiency in order to make a reasonable vertical scale.

While pragmatic considerations thus underpin the use of grammar and vocabulary items as the key linking items, it can also be said that these pragmatic considerations coincide in important ways with the recommendations for constructing the links between grades for vertical scaling noted above. Although

the rationale also draws on the large body of research into the relationship between grammar and vocabulary with other skills in order to support this choice (see the literature review in Section 2.3 for references), it is clear that it is premised on assumptions which will require further validation. While previous research provides support for the assumption that grammar and vocabulary correlate highly with the other skills generally, this assumption in the context of the EIKEN tests will require empirical validation. What is perhaps more important in the context of vertical scaling, and equating methodology generally, is perhaps less the predictive power of the grammar and vocabulary sections and rather the potential impact on the vertical scale of utilizing anchor items from only one of the content formats in the tests.

Vertical scaling guidelines generally include suggestions for content representativeness among common item, with the recommendation that the common item set should cover the same content areas in the same proportions as the main test (Kolen & Breenan, 2004). The guidelines, as Hardy et al (2011) note, are generally derived from the literature on equating, in which characteristics of the tests to be equated and the samples of test takers will differ in important ways from those for vertical scaling. In particular, the content and level of the tests and ability of the test takers will differ by design. Very few studies have actually explored the impact of varying the design of the common item set for vertical scaling. The few studies addressing this area have focused on varying the number of items (e.g. Fitzpatrick, 2014) or the statistical properties of the common item set (e.g. Sinharay & Holland, 2007), rather than the content coverage.

Sinharay and Holland (2007) indeed recommend relaxing recommendations for strict comparability in terms of the difficulty range covered by the common-item set. In terms of content, Hardy et al (2011) report on a study which looked at varying content in common item sets in mathematics,. They found there were differences in the amount and variability of growth depicted in the scales derived through common item sets with different degrees of content representativeness. However, the direction of trends for different content combinations was not consistent across the different mathematical content strands, making it difficult to draw clear guideless.

As noted above, in practice pragmatic decisions reported in vertical scaling studies, for example Kenyon et al (2011), indicate that principled compromises in the context of vertical scaling are being made. As noted in Chapter 2, many of the aspects of study design for vertical scaling in fact remain unresolved in terms of a clear consensus model favoring one choice over another. This makes it crucial that that the decisions are documented clearly, allowing for further research and evaluation to investigate the impact of the decisions on the scale created. Utilizing different combinations of content coverage for anchoring designs may indeed result in different scales with different interpretations of growth across grades, as has been noted for various other aspects of vertical scaling study designs in Chapter 2. The strong relationship with other skills and predictive power noted in the literature for grammar and vocabulary may prove to ameliorate the lack of content representativeness of the anchors. However, this too will be an area which requires future research and investigation. Not only will the predictive power of these items in the context of the EIKEN tests need to be more clearly established to provide support for the approach taken in this study. Research investigating the impact of varying content coverage in the anchor design—regardless of the predictive power of any one particular format used for anchoring—will also be required. Such research would not only add further important empirical support for the design decisions taken in this study but would provide an important contribution to the literature on vertical scaling. This issue will be discussed further in Chapter 6 under the limitations and implications sections.

The fourth important decision for creating a vertical scale in Step 1, which was listed at the beginning of this section, refers to the quality control measures employed in the construction and analysis of test forms. Firstly, the criteria for item selection in Step 1 are presented, followed by an overview of the quality control criteria and analysis procedures used for estimating difficulty parameters on a vertical scale.

Content teams in charge of item review and production at each grade were asked to select the on-grade items for each form from items that had been previously administered in live tests of the same grade. In selecting items, item performance statistics from the live test administration were required to meet the

following criteria:

a. Proportion correct of between 0.2 and 0.8 to avoid items at the extreme ranges of difficulty for on-grade test takers
b. Point-biseral of greater than 0.2 (See 4.2.2 for an explanation of the point-biserial as a measure of item discrimination)
c. Content consistent with specifications for the grade targeted (items had already passed rigorous content review before use in operational tests but were reviewed by content specialists again before inclusion in Step 1).

As already noted, the Rasch model was used to analyze the data and estimate difficulty parameters for the items. The program used for the analysis was Winsteps (Linacre, 2015). A concurrent estimation analysis was carried out on a data matrix combining responses from all test takers across all seven forms. To format the data for a concurrent analysis, the seven data sets (one for each form) were combined into one large data matrix. Within each examinee response string, all item slots for items on forms not taken by the examinee are first coded as "not reached". The combined data set is then used in the estimation of final parameters on a common, vertical scale (Kolen and Brennan, 2004, p. 388). While a concurrent estimation approach was used, several analysis runs were required in order to iteratively refine the measurement frame through the deletion of examinees and items that did not meet quality control criteria for evaluating how closely the data fir the expectations of the Rasch model. This process is described below:

1. Data is cleaned by removing examinees with no responses and examinees who selected the same response option for all items (for example all 1s, etc.).
2. The first concurrent estimation analysis is run and examinee fit statistics are reviewed. Examinees with infit mean square values of over 1.2 are dropped from further analysis. Person fit statistics are reviewed on the first analysis run only.
3. A second concurrent estimation analysis run is carried out with the remaining examinee responses. Item fit statistics are reviewed. For items, both infit and outfit mean square values are evaluated. Items with infit or outfit mean square

values greater than 1.5 are dropped from further analysis, and the remaining data is re-analyzed. This step is repeated until parameter estimates stabilize with all items having infit and outfit values of 1.5 or less.

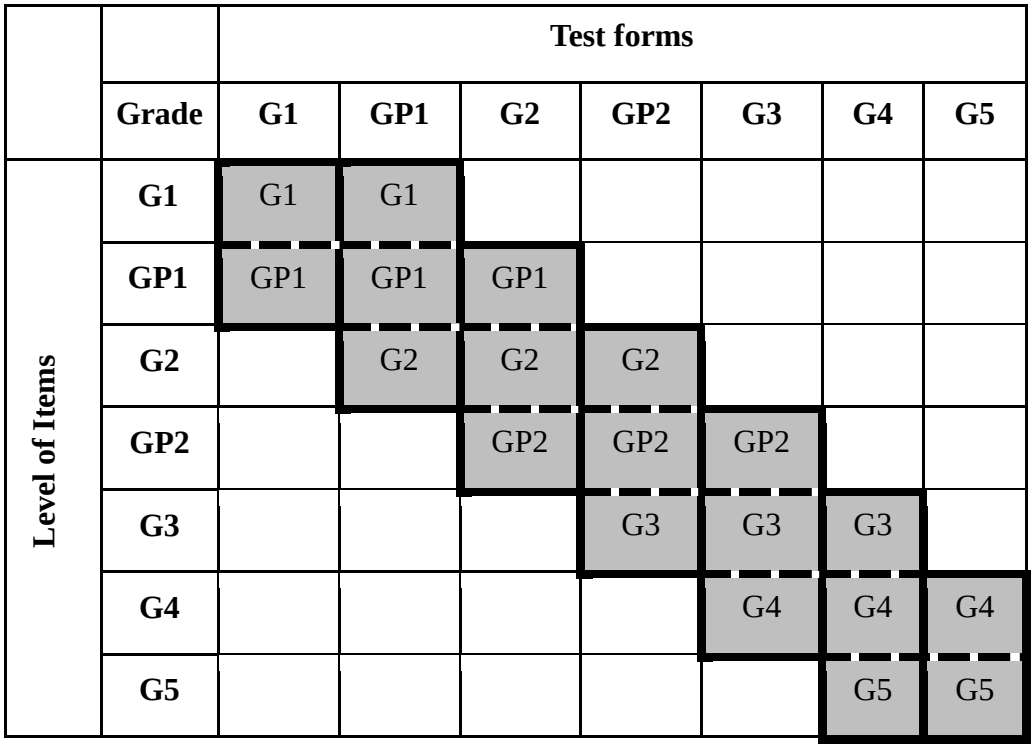


Figure 4-2 Overview of data collection linking design for Step 1

Table 4.9 Breakdown of item distribution across forms in Step 1

	Items shared with grade above		On-grade items not shared	Items shared with grade below		Total items in form	% with grade above	% with grade below	Total % of items shared
	above	On-grade		On-grade	From below				
G1			40	20	20	80	0%	50%	50%
GP1	20	20	0	20	20	80	50%	50%	100%
G2	20	20	15	13	12	80	50%	31%	81%
GP2	13	12	0	13	12	50	50%	50%	100%
G3	13	12	0	13	12	50	50%	50%	100%
G4	13	12	0	13	12	50	50%	50%	100%
G5	13	12	10			35	71%	0%	71%

4.3.2.5 Step 2: Horizontal equating of grammar and vocabulary items

Following the construction of the vertical scale in Step 1, large numbers of previously administered grammar and vocabulary items were calibrated to the vertical scale through horizontal equating. The items calibrated in this step were intended for use as anchors in the horizontal equating of entire previously administered test forms in Step 3. The items used in this step were all drawn from the same range of previously administered operational test forms that would be calibrated in Step 3.

In addition to the items that needed to be calibrated for use as anchors in Step 3, each form contained a block of on-grade anchor items selected from the vertical scaling forms in Step 1. This block of anchor items formed the horizontal equating link to allow all on-grade, non-anchor items in Step 2 to be calibrated to the vertical scale. As already noted above, in reality the administration of test forms in Step 1 and Step 2 took place simultaneously. The order of the steps in fact refers to the order in which the test forms were analyzed and calibrated to the vertical scale. As both vertical linking forms and horizontal equating forms would be administered in the same testing sessions, the items from Step 1 intended for use as anchors in Step 2 had to be identified in advance to enable those items to be included in the horizontal equating forms used for the analysis in Step 2. All forms at the same grade level contained the same block of anchor items. The block of anchor items was preselected to represent between 25 percent and 30 percent of the total number of items on the form. In this way, some leeway was allowed for anchor items not meeting quality control fit criteria in Step 2 to be dropped and still maintain minimum anchor coverage of 20 percent.

The analysis procedures and quality control criteria for Step 2 were essentially the same as for Step 1. In Step 2, the data across forms within the same grade level was combined for a concurrent analysis of all data for the same grade level. Quality control criteria are also introduced for the evaluation of the performance of anchor items. The additional procedures employed are noted below.

1. Prior to the first analysis run, the block of anchor items in each on-grade horizontal equating form are fixed at the difficulty values

estimated in Step 1.

2. Anchor items that showed unacceptable levels of fit in Step 1 (i.e. infit/outfit mean square greater than 1.5) are dropped from the analysis.
3. For each grade level, concurrent data analysis, after the first analysis run, anchor performance is reviewed. The quality control criteria employed for anchor items is the displacement index generated by Winsteps. Anchors with displacement of greater than 0.3 are dropped from the analysis. The analysis procedure is repeated until no anchor items show problematic displacement levels, or until the minimum required percentage of anchor items is reached (20 percent).
4. Person fit statistics are reviewed and persons showing unacceptable levels of misfit are deleted (see description in Step 1 for details)
5. Item fit statistics are reviewed and items showing unacceptable levels of misfit are deleted. This step is repeated until the all remaining items meet the acceptable fit criteria (see description in Step 1 for details).

Table 4.10 shows the total number of on-grade items administered for each grade in Step 2. The table also contains a breakdown of the total number of items selected in advance for each anchor block, the total number of anchors remaining in the analysis following quality control checks, the number of non-anchor items dropped according to the same misfit criteria described for Step 1, and the final number of items in each grade calibrated to the vertical scale.

Table 4.10 Overview of items calibrated in Step 2

	G1	GP1	G2	GP2	G3	G4	G5
Total number of items used in Step 2	210	390	749	420	383	383	259
No of anchor items selected in advance	30	30	29	24	15	16	12
No of anchor items after quality control	24	24	24	16	12	14	7
No. of non-anchor items deleted	8	2	5	7	6	22	17
Total of non-anchor items calibrated	172	358	715	389	362	345	230

4.3.2.6 Step 3: calibrating previously administered test forms

Step 3 utilized operational test data to calibrate items previously used in full test sets during live administrations. The test forms to be calibrated were indicated previously in Table 4.7. The range of test forms to be calibrated in many ways dictated the content of forms used in the horizontal equating stage of Step 2. For each First Stage test form to be horizontally equated to the vertical scale in Step 3, a sufficient number of anchors would be required. The benchmark figure of 20 percent coverage was used to calculate a minimum number of items necessary, based on the total number of dichotomously scored MC items in each First Stage test. As noted above, these items were all drawn from the grammar and vocabulary items in Section 1 of each test. These items were calibrated in advance to the vertical scale in Step 2, as described above. Table 4.11 shows the total number of dichotomously scored MC items in the First Stage tests, the minimum number of anchor items required to reach 20 percent of the total items in the First Stage tests, along with the number of items in Section 1 for each grade. As it was anticipated that some items would fail to meet the quality control criteria in Step 2, more items than the minimum needed to maintain 20 percent anchoring had been incorporated into the horizontal equating forms for calibration in Step 2. This allowed items intended as anchors but which did not meet the quality control criteria to be dropped from the analysis. .

Table 4.11 Number of anchor items in operational test forms for Step 3

	G1	GP1	G2	GP2	G3	G4	G5
Number of items in First Stage test	68	70	75	75	65	65	50
Minimum anchor items (20%)	14	14	15	15	13	13	10
Total number of items in Section 1	25	25	20	20	15	15	15

The quality control criteria for horizontally equating each test form were essentially the same as in Steps 1 and 2. Some important differences need to be noted. Firstly, the most important difference, of course, is the data itself. For Step 3, all response data had been derived from previous live administrations. The sample sizes for forms at each grade were roughly similar to those shown in Table

4.1. Secondly, the response data for each form were analyzed separately (as opposed to a concurrent analysis) in order to calibrate the item parameters onto the vertical scale using the anchor items. The majority of items in Section 1 of each form to be used as anchor items had been calibrated to the scale in Step 2, and the difficulty parameters for anchors were thus fixed at the values derived in Step 2. Some items from section 1 of the operational forms had been used in the vertical scaling in Step 1, and these items were fixed at the item difficulties estimated in Step 1. An overview of the procedures is provided below:

1. Prior to the first analysis run, the grammar and vocabulary anchor items in Section 1 are fixed at the difficulty values estimated in Step 2 (or Step 1 for items used in the vertical scaling sets).
2. Anchor items that showed unacceptable levels of fit in Step 2 are dropped from the analysis.
3. Some items are occasionally used in more than one test form. When an item was used in more than one form, it was first calibrated for one administration and the item difficulty fixed at that value. These items were thus treated as anchor items in the analysis of subsequent forms in which they were used. The inclusion of some anchor items across sections other than Section 1 thus increased the horizontal equating links across test forms in the same grade.
4. After the first analysis run, anchor performance is reviewed. The quality control criteria employed for anchor items is the displacement index generated by Winsteps. Anchors with displacement of greater than 0.3 are dropped from the analysis. The analysis procedure is repeated until no anchor items show problematic displacement levels, or until the minimum required percentage of anchor items is reached (20 percent).
5. Person fit statistics are reviewed and persons showing unacceptable levels of misfit are deleted (see description in Step 1 for details of criteria)
6. Item fit statistics are reviewed and items showing unacceptable

levels of misfit are deleted. This step is repeated until the all remaining items meet the acceptable fit criteria (see description in Step 1 for details of criteria).

7. Once stable item parameters are estimated for all items remaining in the data set for each form (i.e. those items not deleted through the application of quality control criteria), these items are fixed at these stable difficulty values.

4.3.3 Results

Table 4.12 provides an overview of the total number of items administered and calibrated in Step 3. The number of operational test forms calibrated in this step which was noted previously is also included for reference, along with the number of items per form. As noted earlier, following Brown et al (2012), the constructed response writing items for Grades 1 and Pre-1 were also included in the analysis for calibration at this stage, bringing the total number of items per form for Grade 1 and Pre-1 to 69 and 71 respectively. Table 4.12 gives a breakdown of the total number of anchor items surviving quality control review which were kept in the analysis. Note that the total number of anchor items exceeds the number of items calibrated in Step 2, as items which were used in multiple forms were also treated as anchors, as described in the outline of procedures for Step 3 above. This greatly increased the amount of horizontal linking across forms within grades at the lower levels. For items deleted, the total of all items deleted due to quality control procedures, including both anchors deleted due to displacement and non-anchor items deleted due to unacceptable misfit is given first, along with this figure as a percentage of all items in the analysis. Following this, a breakdown of deleted items according to the reason for deletion is also given. The majority of items were deleted due to the application of displacement criteria to anchor items, while very few items were deleted due to misfit. Only Grade 5 experienced more items deleted through misfit than displacement.

Before discussing the results in terms of empirical distinctions in difficulty across grades in detail, the main focus of RQ2, it is worth first addressing the issue of unidimensionality and what light the results shed on this

aspect of the psychometric properties of the items within and across grades.

All of the unidimensional IRT models discussed in the literature review for vertical scaling in Section 2.4.2, including the Rasch model, require the data to meet the assumption of unidimensionality (Henning, 1992: Kolen & Brennan, 2004; McNamarra, 1996; Sick, 2010). McNamara and Knoch (2012), in their historical overview of the use of the Rasch model in language testing research, describe the initial debate surrounding this assumption in respect to language tests, citing for example the concerns raised by Buck (1994) and Hamp-Lyons (1992), amongst others. Henning (1992) and McNamara (1996) in response to these concerns emphasized the difference between *psychological* and *psychometric* unidimensionality, with the latter being a necessary feature of the data for Rasch analysis. The now widespread acceptance and use of Rasch in language testing noted by McNamara & Knoch (2012) suggests a general acceptance of this distinction. As Henning (1992, p.10) notes, in terms of psychological abilities, "it is barely conceivable to imagine a language test that would be unidimensional." However, for the purposes of designing a measurement instrument, it is the assumption of psychometric unidimensionality that needs to be met, which McNamara (1996, p. 271) defines as, "loosely, a single underlying patter of scores in the data matrix." In fact, Kolen and Brennan (2004, p. 157) note that the assumption of unidimensionality "might not hold strictly" in educational measurement contexts generally, but "might hold closely enough for IRT to be used advantageously in many practical situations." Henning et al (1985), Henning (1992), and McNamara (1996) demonstrate how language tests covering a range of language skill areas can in fact meet the assumption of psychometric unidimensionality necessary for Rasch analysis based on this definition.

It is also important to note that the assumption of unidimensionality is not just a feature of IRT, and in fact is equally, but implicitly, assumed in CTT approaches which rely on internal consistency (Henning et al, 1985: Henning, 1992; McNamara, 1996; Sick, 2010). The difference is that IRT, and Rasch analysis, provides explicit methods for testing the assumption. As Sick (2010) notes, in the case of Rasch analysis, tests for the requirement for unidimensionality in the data are actually carried in the process of analyzing the

data, rather than as an a priori statistical test of assumptions, as for example in the case of analysis of variance procedures. One of the principal indicators of unidimensionality in Rasch analysis is provided by the item and person fit statistics produced by Rasch analysis (Bonk & Ockey, 2003; Eckes, 2009; Henning, 1992: McNamara, 1996; Sick, 2010), and which were described above in Section 4.3.2.4 under the measures of quality control used in the development of the vertical scale. McNamara (1996, pp. 275-277) provides a detailed description of how fit statistics can be used to evaluate whether items and persons conform to the assumption that there is a single *measurement* [emphasis in original] dimension of ability and difficulty" underlying the test. McNamara (1996, p. 275) notes that "extreme" values for fit statistics would indicate that "the hypothesis is unlikely to be true for the item or individual concerned." He goes on to describe possible explanations for such misfit, and show how patterns of misfitting items can help identify sections of a test which may in fact be targeting something not originally intended by the test developer. Sick (2010) notes how such use of fit statistics with Rasch allows for the iterative refinement of the data set, by dropping items which demonstrate they are not measuring the same underlying measurement ability, so that a robust unidimensional measurement frame can be constructed with the items which do conform to the assumption.

The results for misfit described below in Table 4.12 show that the items across the grades demonstrated sufficient fit according to the quality assurance criteria described above. Indeed, only a small fraction of items from any grade were dropped from the analysis due to any of the quality control criteria and—as described in more detail below—the majority of those were from displacement criteria and not misfit. Across the grades, Grades 1, Pre-2, and 3 showed no misfit, Grades Pre-1 and 2 lost only 1 item each, and 20 items and 103 items were dropped from Grades 4 and 5 respectively. This thus provides confirmation that the assumption of unidimensionality holds generally for the data set, and, importantly, not just across the different skill areas and item formats contained within grades but across the vertical dimension of the seven-level set of tests. At the same time, the tendency for more items to demonstrate misfit at the very lowest levels would suggest that this is worth further investigation to identify

possible implications for how these levels interact with the other grades and the overall frame of measurement. For the purposes of this study, as any items demonstrating misfit were dropped from the analysis, it can is clear that evidence of misfit was minimal and all items used in the final analysis to generate difficulty estimates for items across the seven grades on a common vertical scale did in fact demonstrate sufficient unidimenionality to meet the requirements of the Rasch model.

A number of other methods for investigating dimensionality are also noted in the literature including principal components analysis (PCA), other factor analytic techniques, and the Bejar method, which consists of running separate analyses for possible sub-domains or items grouped according to possible different measurement dimensions and comparing the difficulty estimates generated by the overall analysis with the separate subdomain analyses (Bonk and Ockey, 2003; Henning et al, 1985; Henning, 1992; McNamara, 1996; North, 2000; Sick, 2010). Henning et al (1985) also suggest that internal consistency estimates such as Cronbach's alpha provide an indication of psychometric unidimensionality. As sufficient support for the assumption of unidimensionality of the data for the purposes of this study can be demonstrated through the use of Rasch fit statistics, as described above, these alternative methodologies were note employed. Nonetheless, the CTT reliability estimates described in Section 4.1 above provided evidence of sufficient internal consistency of test forms within each grade, and this also provides some support for the unidimensionality assumption. The important caveat to this is that it only applies to test forms within each grade, as the CTT statistics of course do not provide information regarding the vertical common frame of measurement, which was the purpose of the vertical scaling in the first place. The Brown et al (2012) study, which as noted was an important pilot study demonstrating the applicability of the use of Rasch for vertical scaling of the EIKEN tests, did carry out Principal Components Analysis of all of the data used in that study. The authors investigated both one-factor and two-factor solutions for the data. In the one-factor solution, all of the EIKEN components across the three grades vertically scaled in that study and the components of the TOEFL iBT loaded on one factor. In the two factor solution, all

of the EIKEN components loaded most heavily on the first factor. These findings, as noted above, were carried out under experimental rather than operational conditions and for a different purpose to this study. They nonetheless also provide some support for the assumption of unidimensionality in the EIKEN First Stage tests, both across skill components within grades and across a vertical dimension.

As Henning et al (1985) note, even when evidence indicating that the assumption of unidimensionality can indeed be met by language tests in general, it is still necessary for researchers investigating specific tests to test this assumption with their data. In the case of this study, although cautious indications in support of unidimensionality could be claimed from the investigation of internal consistency estimates of operational data and the vertical scaling study undertaken by Brown et al (2012), as described above, it is still imperative to demonstrate that the data used for the purposes of vertical scaling to investigate RQ2 demonstrate sufficient unidimensionality to allow reasonable interpretations of the difficulty estimates generated. Use of fit statistics generated as a part of the Rasch analysis output has provided sufficient support for this assumption for the data set generally, with only a small fraction of items demonstrating misfit. As these items were dropped from the final analysis, the final output can be said to have been generated through the use of items which meet the requirements of the Rasch model. It is, however, worth noting Kolen and Brennan's comments regarding dimensionality and vertical scaling in their updated volume of *Test Equating, Scaling, and Linking* (2014, p. 469):

> One of the most challenging aspects of applying IRT to vertical scaling is the assumption that the same unidimensional ability is assessed across grades. It is unlikely that this assumption strictly holds in practice, although this assumption might hold well enough that the unidimensional models can be used to construct reasonable vertical scales…More research on psychometric structure across grades and on the use of multidimensional IRT in vertical scaling is needed.

As Table 4.12 demonstrates, using the fit criteria suggested in the literature, the data has certainly been shown to "hold well enough" for the

purposes of this study. Nonetheless, it will be important for future studies to answer Kolen and Brenann's call for more research in the context of vertical scaling, and to employ alternative methods to investigate this aspect of the operational data provided by the EIKEN tests in more detail.

Table 4.12 Overview of items calibrated in Step 3

	G1	GP1	G2	GP2	G3	G4	G5
Total no of items in Step 3	828	1491	3750	2250	1885	1885	1200
Items per form	69	71	75	75	65	65	50
Forms	12	21	50	30	29	29	24
Items calibrated in Step 3	626	1120	2573	1625	1372	1375	784
Anchor items	178	308	983	517	420	408	238
(% of all items)	*21%*	*21%*	*26%*	*23%*	*22%*	*22%*	*20%*
Total items deleted	24	63	194	108	93	102	178
Deleted (% of all items)	3%	4%	5%	5%	5%	5%	15%
Non-anchor Items (misfit)	*0*	*1*	*0*	*0*	*1*	*20*	*103*
Anchors (displacement)	*24*	*62*	*194*	*108*	*92*	*82*	*75*

The reason for deletion of items has implications for the evaluation of item difficulties across grades. Non-anchor items deleted due to misfit have no difficulty values estimated on the common vertical scale and are thus not included in the following comparison of item difficulties. Anchor items deleted during Step 3 due to the application of misfit criteria in fact have difficulty estimates previously estimated either in Steps 1 or 2, depending on when the item was first calibrated to the vertical scale. These items have already contributed to the construction of the vertical scale during Steps 1 and/or 2, and their difficulty values are thus taken into account and included in the evaluation of average item difficulty across grades. As noted above, some items have been used in more than one operational form. Such items were also treated as anchor items after being calibrated first in one form, and then having their item values fixed in subsequent forms. These items were generally in sections other than Section 1, and such cases only occurred in the lower grades (Grades 2 to 5). Any such common-item anchor

items that were deleted due to displacement have been included in the evaluation of average item difficulties by fixing their difficulty values at the level first estimated for the initial form in which they were analyzed. In this way, the largest possible number of items was included in the evaluation of how item difficulty differs across the different grades in the EIKEN program in order to answer RQ2.

For the investigation of RQ2, difficulty for the First Stage test forms as a whole will be investigated, and this will also be broken down to focus on difficulty across grades for reading items only. Table 4.13 shows the descriptive statistics for all items calibrated to the vertical scale in Steps 1, 2, and 3 following the procedures explained previously, and Table 4.15 shows the descriptive statistics for item difficulty across grades for reading items only.

Table 4.13 Item difficulty across all items

Grade	N	Mean	Std. Deviation
G5	1097	-6.556486	1.4164820
G4	1865	-4.179441	1.0646182
G3	1884	-2.817809	.9909091
GP2	2250	-1.170522	.9861549
G2	3750	.465203	.8428168
GP1	1490	2.208187	.8107653
G1	828	3.614388	.9173213

Table 4.14 Item difficulty for non-listening sections only

Grade	N	Mean	Std. Deviation
G5	119	-5.990950	1.3065993
G4	434	-4.264655	.9540719
G3	435	-2.859566	.8548025
GP2	600	-.983915	1.0739417
G2	1000	.581142	.7607425
GP1	336	2.082401	.7261074
G1	192	3.385127	.9161763

There is a clear difference in mean difficulty evident purely from eyeballing the means, and the same trend is evident for both Table 4.13 and Table 4.14. This trend is further clarified in the boxplots of item difficulties shown below in Figure 4.3 and Figure 4.4 for item difficulty across all items and reading items only. Although items in the extreme ends of the distribution for a grade overlap slightly with the grades above and below, the interquartile range for each grade is clearly distinct with no overlap for the bulk of item difficulties calibrated to the common vertical scale.

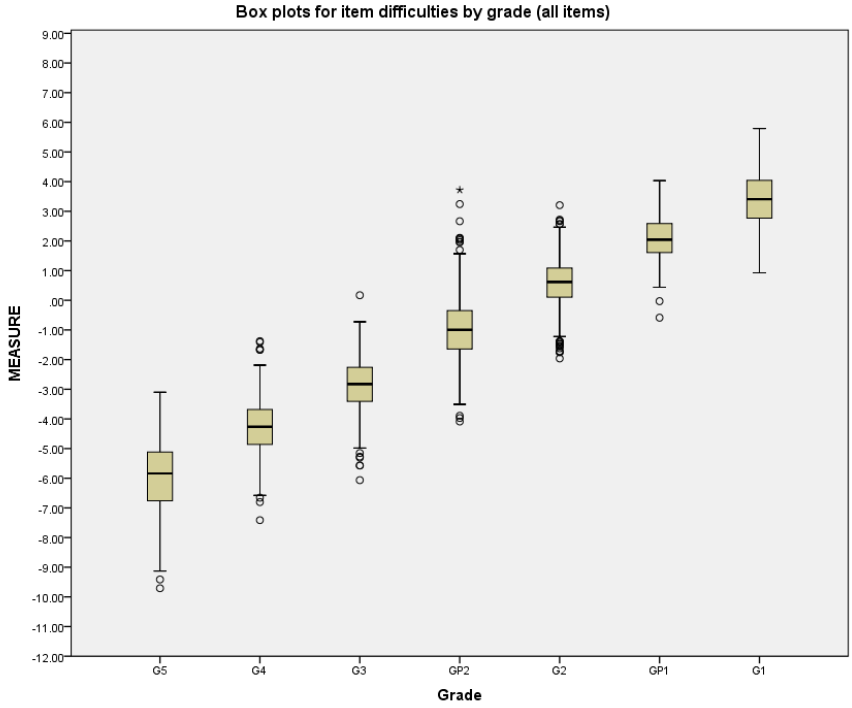


Figure 4-3 Boxplots for item difficulty by grade (all items)

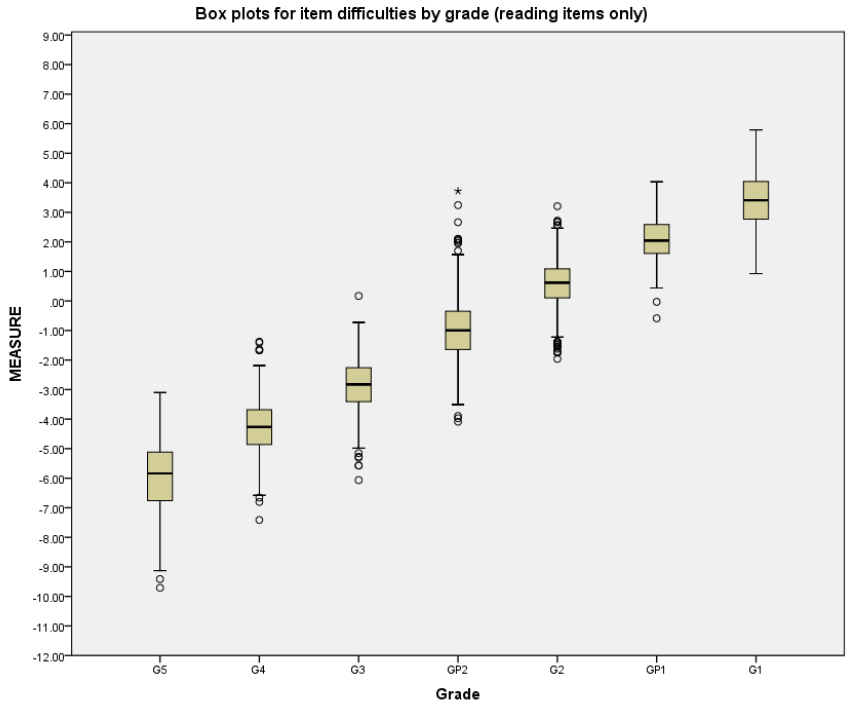


Figure 4-4 Boxplots for item difficulty by grade (reading only)

To further examine these differences, one-way analysis of variance (ANOVA) was carried out with SPSS Statistics version 21 using the item difficulty estimates as the dependent variable and the grade as the independent variable. Results for the one-way ANOVA conducted for all items are summarized first, followed by a second one-way ANOVA conducted on the reading item difficulties only.

Tables 4.15, 4.16, and 4.17 summarize the results from the first one-way ANOVA utilizing difficulty data from all items in the First Stage test forms calibrated during Step 3. The summary statistics for Table 4.16 are shown for reference only, and for use in calculating the effect size. As the assumption of homogeneity of variance was not met, as demonstrated by the statistically significant result for Levene's test in Table 4.15, it would not be appropriate to use the F-ratio in Table 4.16 as evidence of statistically significant differences between the means of the independent variables (Field, 2009). To confirm if there are significant differences between the groups in such cases, SPSS provides adjusted tests of the F-ration in Table 4-17. Both of these tests are significant,

confirming that there are significant differences between the grades. To determine which grades are statistically different in terms of difficulty, the post-hoc tests in Table 4.18 are used. SPSS offers a number of options for conducting follow-up, post-hoc tests which can be broadly divided into two groups, those that require the assumption of homogeneity of variance to be met and those that do not (Field, 2009). When running the analysis, it is thus useful to specify both types of tests in advance. Two tests were selected from the options offered by SPSS, the Bonferroni and Game-Howell tests. The first of these assumes equal variances, while the second can be used when the assumption is not met. Field (2009, p. 374) recommends the Bonnferonni and Tukey tests as offering the greatest control over Type 1 error rates, and notes that Bonnferonni's test "has more power when the number of comparisons is small." Only the results of the Games-Howell test are shown in Table 4.18. Although all grade comparisons using the very tight Type 1 error control of the Bonnferonni test were significant, as noted earlier, this test assumes equal variances. The Games-Howell test is able to test for statistically important differences between the grades when this assumption is not met, and is also robust in the face of unequal sample sizes (Field, 2009), making it appropriate in this case given the different number of items in each grade sample. Referring to Table 4.18, it can be seen that all grade comparisons are statistically different according to the Games-Howell test.

Table 4.15 Test of homogeneity of variance (all items)

Levene Statistic	df1	df2	Sig.
95.157	6	13157	.000

Table 4.16 Main ANOVA summary table (all items)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	99800.437	6	16633.406	17251.589	0.000
Within Groups	12685.540	13157	.964		
Total	112485.977	13163			

Table 4.17 Robust tests of equality of means (all items)

	Statistic[a]	df1	df2	Sig.
Welch	15301.643	6	4391.451	0.000
Brown-Forsythe	15743.416	6	7570.550	0.000

Table 4.18 Games-Howell post-hoc test between grades (all items)

(I) Grade			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Games-Howell	G5	G4	-2.3770450*	.0493633	.000	-2.522747	-2.231343
		G3	-3.7386774*	.0484787	.000	-3.881777	-3.595578
		GP2	-5.3859640*	.0475524	.000	-5.526339	-5.245589
		G2	-7.0216890*	.0449270	.000	-7.154352	-6.889026
		GP1	-8.7646733*	.0476464	.000	-8.905327	-8.624019
		G1	-10.1708740*	.0533412	.000	-10.328312	-10.013436
	G4	G5	2.3770450*	.0493633	.000	2.231343	2.522747
		G3	-1.3616324*	.0335992	.000	-1.460748	-1.262517
		GP2	-3.0089190*	.0322483	.000	-3.104048	-2.913790
		G2	-4.6446440*	.0282339	.000	-4.727942	-4.561346
		GP1	-6.3876283*	.0323867	.000	-6.483173	-6.292084
		G1	-7.7938290*	.0402990	.000	-7.912776	-7.674882
	G3	G5	3.7386774*	.0484787	.000	3.595578	3.881777
		G4	1.3616324*	.0335992	.000	1.262517	1.460748
		GP2	-1.6472866*	.0308772	.000	-1.738369	-1.556204
		G2	-3.2830116*	.0266571	0.000	-3.361654	-3.204369
		GP1	-5.0259959*	.0310217	.000	-5.117513	-4.934478
		G1	-6.4321966*	.0392104	.000	-6.547941	-6.316452
	GP2	G5	5.3859640*	.0475524	.000	5.245589	5.526339
		G4	3.0089190*	.0322483	.000	2.913790	3.104048
		G3	1.6472866*	.0308772	.000	1.556204	1.738369
		G2	-1.6357250*	.0249328	.000	-1.709271	-1.562179
		GP1	-3.3787093*	.0295532	.000	-3.465891	-3.291527
		G1	-4.7849100*	.0380592	.000	-4.897266	-4.672554

(I) Grade			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
	G2	G5	7.0216890*	.0449270	.000	6.889026	7.154352
		G4	4.6446440*	.0282339	.000	4.561346	4.727942
		G3	3.2830116*	.0266571	0.000	3.204369	3.361654
		GP 2	1.6357250*	.0249328	.000	1.562179	1.709271
		GP 1	-1.7429843*	.0251116	.000	-1.817075	-1.668894
		G1	-3.1491850*	.0347232	.000	-3.251741	-3.046629
	GP 1	G5	8.7646733*	.0476464	.000	8.624019	8.905327
		G4	6.3876283*	.0323867	.000	6.292084	6.483173
		G3	5.0259959*	.0310217	.000	4.934478	5.117513
		GP 2	3.3787093*	.0295532	.000	3.291527	3.465891
		G2	1.7429843*	.0251116	.000	1.668894	1.817075
		G1	-1.4062007*	.0381765	.000	-1.518907	-1.293495
	G1	G5	10.1708740*	.0533412	.000	10.013436	10.328312
		G4	7.7938290*	.0402990	.000	7.674882	7.912776
		G3	6.4321966*	.0392104	.000	6.316452	6.547941
		GP 2	4.7849100*	.0380592	.000	4.672554	4.897266
		G2	3.1491850*	.0347232	.000	3.046629	3.251741
		GP 1	1.4062007*	.0381765	.000	1.293495	1.518907

Following the same procedures as described above for the one-way ANOVA looking at the difference in mean item difficulty based on the calibration of all items, a follow-up test was carried out looking at the subset of reading items only. The results of this ANOVA are summarized in Tables 4.19, 4.20, 4.21, and the post-hoc Games-Howell tests in Table 4.22. As with the ANOVA for all items, Levene's test of the homogeneity of variances for the reading items calibrated for all seven grades was significant, indicating the assumption of equal variances was not met by the data. It was thus necessary to refer to results of the Welch and Brown-Forsythe tests in Table 4.21, both of which were significant. The post-hoc

Games-Howell tests once again demonstrated that the differences between all grades was significant at p>0.05.

Table 4.19 Test of homogeneity of variance (reading items only)

Levene Statistic	df1	df2	Sig.
19.936	6	3109	.000

Table 4.20 Main ANOVA summary table (reading items only)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	18343.563	6	3057.260	3773.842	0.000
Within Groups	2518.659	3109	.810		
Total	20862.222	3115			

Table 4.21 Robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	3641.631	6	791.674	0.000
Brown-Forsythe	3254.444	6	1035.237	0.000

Table 4.22 Games-Howell post-hoc tests (reading items only)

(I) Grade			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Games-Howell	G5	G4	-1.7262945*	.1282326	.000	-2.109412	-1.343177
		G3	-3.1313836*	.1265937	.000	-3.509858	-2.752909
		GP2	-5.0070351*	.1275480	.000	-5.388204	-4.625866
		G2	-6.5720916*	.1221678	.000	-6.938093	-6.206090
		GP1	-8.0733511*	.1261562	.000	-8.450593	-7.696109
		G1	-9.3760761*	.1368137	.000	-9.783809	-8.968343
	G4	G5	1.7262945*	.1282326	.000	1.343177	2.109412
		G3	-1.4050891*	.0614581	.000	-1.586720	-1.223458
		GP2	-3.2807406*	.0634004	.000	-3.468051	-3.093430
		G2	-4.8457971*	.0517309	.000	-4.998774	-4.692821
		GP1	-6.3470566*	.0605516	.000	-6.526058	-6.168055
		G1	-7.6497816*	.0804309	.000	-7.888197	-7.411366
	G3	G5	3.1313836*	.1265937	.000	2.752909	3.509858
		G4	1.4050891*	.0614581	.000	1.223458	1.586720
		GP2	-1.8756515*	.0600166	.000	-2.052953	-1.698350
		G2	-3.4407080*	.0475234	.000	-3.581206	-3.300210
		GP1	-4.9419675*	.0569990	.000	-5.110470	-4.773465
		G1	-6.2446925*	.0777914	.000	-6.475414	-6.013971
	GP2	G5	5.0070351*	.1275480	.000	4.625866	5.388204
		G4	3.2807406*	.0634004	.000	3.093430	3.468051
		G3	1.8756515*	.0600166	.000	1.698350	2.052953
		G2	-1.5650565*	.0500098	.000	-1.712815	-1.417298
		GP1	-3.0663160*	.0590880	.000	-3.240921	-2.891710
		G1	-4.3690411*	.0793348	.000	-4.604231	-4.133852
	G2	G5	6.5720916*	.1221678	.000	6.206090	6.938093
		G4	4.8457971*	.0517309	.000	4.692821	4.998774
		G3	3.4407080*	.0475234	.000	3.300210	3.581206
		GP2	1.5650565*	.0500098	.000	1.417298	1.712815
		GP1	-1.5012595*	.0463451	.000	-1.638366	-1.364153
		G1	-2.8039846*	.0703598	.000	-3.013171	-2.594798
	GP1	G5	8.0733511*	.1261562	.000	7.696109	8.450593

(I) Grade			Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
		G4	6.3470566*	.0605516	.000	6.168055	6.526058
		G3	4.9419675*	.0569990	.000	4.773465	5.110470
		GP2	3.0663160*	.0590880	.000	2.891710	3.240921
		G2	1.5012595*	.0463451	.000	1.364153	1.638366
		G1	-1.3027251*	.0770773	.000	-1.531391	-1.074059
	G1	G5	9.3760761*	.1368137	.000	8.968343	9.783809
		G4	7.6497816*	.0804309	.000	7.411366	7.888197
		G3	6.2446925*	.0777914	.000	6.013971	6.475414
		GP2	4.3690411*	.0793348	.000	4.133852	4.604231
		G2	2.8039846*	.0703598	.000	2.594798	3.013171
		GP1	1.3027251*	.0770773	.000	1.074059	1.531391

*. The mean difference is significant at the 0.05 level.

Tests of statistical significance for differences between means, including the post-hoc tests controlling for Type-1 error rates, are not necessarily indicative of important difference, particularly when sample sizes are large, making even small differences likely to be statistically significant (Brown, 2008; Field, 2009). It is now common practice to provide estimates of the effect size to identify when differences are actually meaningful in practice. Two effect size measures are commonly reported for ANOVA analyses, and can be easily calculated from the output from SPSS in tables 4-16 and 4-20 above. The first of these is known as eta-squared, η^2, and is calculated using the following formula:

Formula 4.2 $$\eta^2 = \frac{SS_M}{SS_T}$$

Where SS_M is variance associated with the between-groups effect

SS_T is the total amount of variance in the analysis

Substituting the values for the all-item ANOVA from Table 4-19 and the reading-only ANOVA in Table 4-23 into the formula generates the following results:

$$\eta^2_{all\ items} = \frac{99800.437}{12685.540} = 0.85$$

$$\eta^2_{reading\ only} = \frac{18343.563}{2518.659} = 0.88$$

The results can be interpreted as the proportion of variance accounted for by the variable of grade in determining differences in item difficulty (Brown, 2008). Using this interpretation, it can be said that 85 percent of the variance in the all-item analysis and 88% in the reading-only analysis is accounted for by the difference in grade. As Field (2009) notes, the value of η^2 can also be interpreted according to the widely used benchmark levels suggested by Cohen (1988) for interpreting Pearson's r of small (0.2), medium (0.3) and large (0.5). Clearly, the value of eta squared for both analyses represents a comparatively large effect.

Field (2009) however notes that eta squared is a biased estimator of effect size as it is based on the variance in the sample only, and does not thus generalize to a wider population. He suggests the use of an adjusted effect size indicator, omega squared, ω^2:

Formula 4.3 $\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R}$

Where $df_{M)}$ is the model degrees of freedom, and

MS_R is the within-groups variance, or unsystematic error variance

$$\omega^2_{all\ items} = \frac{99800.437 - (6).964}{112485.977 + .964} = 0.77$$

$$\omega^2_{reading\ only} = \frac{18343.563 - (6).810}{20862.222 + .810} = 0.88$$

Field (2009) suggests slightly different guidelines for recognizing benchmark effects, based on Kirk (1996, reported in Field, 2009) of small .01, medium .06, and large .14. The values for omega squared are also, not surprisingly, well above the guidelines for a large effect.

One final evaluation of the importance of the differences in mean item difficulty will be reviewed. The eta squared and omega squared measures tell us that the overall effect for grade is not only significant but of practical significance, but as Field (2009) notes, this does not address the important question of the

differences between individual pairs of grades. Tong and Kolen (2007) suggest the use of an effect size measure for evaluating the mean differences between person measures across grade levels. The formula for the effect size measure in Tong and Kolen (2007) is:

Formula 4.4 $Effect\ size\ = \frac{M_1 - M_2}{\frac{\sqrt{S_{group\ 1} + S_{group\ 2}}}{2}}$

Where M_1 is the mean of the first group, and M_2 is the mean of the second group

$S_{group\ 1}$ and $S_{group\ 2}$ are the standard deviations of the two groups

Tong and Kolen's (2007) formula is actually the equivalent of a commonly used method for calculating Cohen's d. All of the information necessary for calculating the effect size for the differences between pairs of grades for both the all-items and reading-only analyses is available in Table 4-13 and 4-14. Substituting the appropriate values into the formula generates an effect size estimate based on Cohen's d for each pair of grades. The results are presented below in Table 4-23. Cohen (1988) recommended the following guidelines for interpreting values of d: small, d = .2, medium, d = .5, and large, d = .8. The effect sizes are all clearly very large according to this criterion.

Table 4.23 Cohen's d effects sizes for grade pair comparisons

Grade pairs	All items	Reading only
G5/G4	2.377045	1.726295
G4/G3	1.323999	1.551216589
G3/GP2	1.666392	1.932512426
GP2/G2	1.783215	1.681748469
G2/GP1	2.107736	2.01883502
GP1/G1	1.62438	1.575960456

4.3.4 Conclusions regarding item difficulty across EIKEN grades

In relation to RQ2, vertical scaling has provided the means to make principled comparisons of the difficulty of items used at the different levels of the EIKEN testing program. Most importantly for RQ2, the results clearly demonstrate that the grades are targeted at distinct levels of difficulty in terms of the empirical item difficulty as measured in the Rasch-based logit scale used as a common metric across the grades and created through the vertical scaling methodology.

Chapter 5 RQ3: Linking the EIKEN Grades to the CEFR

5.1 Introduction

Research Question 3 addresses the issue of criterion-related validity evidence for the EIKEN set of tests by examining the relationship to an external, widely used proficiency framework, the Common European Framework of Reference for Languages (CEFR). Section 5.2 explains the methodology used for the investigation of RQ3, and subsequent sections discuss each of the main linking studies undertaken.

5.2 Methodology

The investigation of a relationship to the CEFR is situated within the overall validation of the EIKEN set of tests and to which the three research questions of this study are designed to contribute. The approach described in this chapter conceptualises the linking process presented in the Manual for Linking Examinations to the Common European Framework of Reference (Council of Europe, 2009) as a part of the overall validation process, rather than attempting to place larger validation issues such as scoring validity of the test itself within the framework of linking. The evidence from the investigation of RQ3 thus contributes to the validation of the EIKEN tests, rather than other aspects of validity addressed as a part of RQ1 and RQ2 being subordinated to the process of linking the tests to the CEFR.

Under the heading of Validation in the diagrammatic presentation of procedures for linking exams to the CEFR in the Manual (Council of Europe, 2009, p. 15), several aspects are referred to collectively as *Test Validity:*

- Content validity
- Operational aspects (in pretesting, piloting)

- Psychometric aspects

The above elements of validation are more appropriately dealt with under separate aspects of the socio-cognitive validation model. For this study, issues of content validity are thus dealt with in Chapter 3 in the discussion of contextual and cognitive validity as a part of investigating RQ1. Issues related to scoring validity are the focus of RQ 2, and are discussed in Chapter 4. Chapter 3 also addresses the specification stage outlined on page 15 of the Manual, as RQ1 involves creating detailed profiles of the criterial contextual and cognitive features of each grade of the EIKEN system. Chapter 6 integrates the results for RQ1 and RQ2 with the results of the linking studies described here in discussing the conclusions and implications of the study.

This chapter deals specifically with the aspects of the linking process most directly related to establishing a link to the CEFR: standard setting and the validation of that standard setting. Accordingly, Section 5.2.1 gives a more detailed description of the approach taken to standard setting in addressing RQ3, and Section 5.2.2 describes the approach taken to the validation of that standard setting. Subsequent sections will describe specifically the participants, instruments, procedures, and results of the specific linking studies and validation activities undertaken to address RQ3.

5.2.1 Standardisation

The Manual uses *standardisation* to describe the use of standard-setting procedures to determine cutscore points for a test being linked to the CEFR. An overview of the literature on the application of standard setting in the context of linking exams to the CEFR is contained in Chapter 2. As emphasised in the Manual (2009), standard setting is central to the establishment of a link to the CEFR. The investigation of RQ3 thus focuses on this phase of the linking process.

5.2.1.1 The Standard setting panels

Two standard-setting panels were used to link the reading tests from five of the seven levels of the EIKEN suite of tests to the CEFR. Standard-setting Panel 1

focused on the advanced levels of Grade 1 and Grade Pre-1, while Standard-setting Panel 2 focused on Grades 2, Pre-2, and 3. A further standard-setting study was carried out as a part of external validation, and is described in Section 5.5.

Only the reading tests from Grades 1, Pre-1, 2, Pre-2, and 3 were included in this study (for a description of the EIKEN levels, see Chapter 1). As explained in Chapters 1 and 3, Grades 5, 4 and 3 of the EIKEN tests have a strong achievement-test focus, taking close account of the curriculum used in Japanese junior high schools. Grades 5 and 4 in particular are positioned to be relevant for test takers whose main exposure to English has been in the EFL context of Japan, within the formal education system. The content of these two grades is targeted to be relevant to learners who have completed the first two years of junior high school within this context. Following the recommendations of the CEFR (Council of Europe, 2001, pp. 31-33), these three Grades of the EIKEN test can be represented in relation to the CEFR as three branching levels increasing in terms of difficulty, with Grade 5 being A1.1, Grade 4 being A1.2, and Grade 3 being at the upper end of the A1 level as A1.3. Such a branching approach is recommended in the CEFR for tests closely connected to formal educational contexts in which smaller degrees of progression are defined in terms of specific learning goals tied to a specific curriculum (Council of Europe, 2001, p. pp. 32-33),. While the usefulness of such an approach is emphasised in the CEFR, such finer levels of distinction are not specified in the main document or supporting materials such as the Manual for Linking Exams to the CEFR (Council of Europe, 2009) and Reference Supplements (Council of Europe, 2004). Grade 3 is targeted at test takers who have a level of English proficiency commensurate with the goals of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) for junior high school graduates (MEXT 2003, 2011). After reviewing the descriptions of performance which characterize the A1 level in the illustrative scales of proficiency within the CEFR, it was considered that this would be the minimum level of the EIKEN tests to which it would be appropriate to apply standard-setting procedures.

In order to maximise the aspect of practicability emphasised by Berk

(1986), two separate standard-setting panels were planned, each covering a different range of grades. Grouping the five grades to be investigated into two separate standard-setting events, each with a different panel, promised to provide the best balance in terms of time and resources, and in terms of identifying suitable experts with relevant expertise. The first standard-setting panel thus focused on only the upper, advanced-level Grade Pre-1 and Grade1, while the second panel addressed the intermediate and upper-beginner levels, Grades 2, Pre-2 and 3.

As already noted in Chapter 1, the two upper levels of the seven-level set of EIKEN tests differ in some significant ways from the other levels. Grade Pre-1 is posited to be relevant to an approximately B2 level of proficiency, while Grade 1 is considered to be relevant to the more advanced C1 level. Both of these examinations are advanced-level, general English proficiency examinations and are used for a range of purposes from university admissions to proof of English ability for employment purposes, including in teacher certification programs administered by prefectural boards of education. The panel participants were thus drawn from the higher-education domain (for details of participants, see section 5. 3.2).

The lower grades of the tests, to varying degrees, are required to take more account of the school curriculum, as the formal educational context in which most test takers acquire the language constitutes the TLU domain in the EFL context of Japan. This gives the tests at the lower levels an element of an achievement test focus, and while this is strongest at the levels geared towards junior high school, the intermediate-level Grade 2 and lower-intermediate Grade Pre-2 exams are also listed by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) as recommended desired benchmarks for high school graduates (MEXT 2003, 2011). The panel participants were thus recruited from educators experienced in the junior high school and high school educational context of Japan (for details of participants, see section 5.4.2).

5.2.1.2 The choice of standard-setting methods

Two standard-setting methods were used by both panels: the Basket method and a

modified Angoff method. The Basket method has been described as a modification of the approach used in the Angoff yes/no approach, modified specifically to allow for multiple cutoffs for different CEFR levels to be set with a single round of judgments (Kaftandjieva, 2009). The wording used as the judgment criteria for this method in both of the main standard-setting panels of this study follows the description in the Manual (2009) and Kaftandjieva (2009): "At which CEFR level can a test taker already answer the item correctly?" The method is intuitively easy to grasp and this certainly makes training and the actual judgment process easier. The main advantages of the Basket method relate to the relative simplicity of the judgment task and its practicability, both important criteria for the evaluation of standard setting suggested by Berk (1986). The intuitive ease of application of the Basket method has led to it being widely used for linking tests to the CEFR (Kaftandjieva, 2004; Council of Europe, 2009; Kaftandjieva, 2009; Kaftandjieva, 2010).

The second method used for standard setting was a modified Angoff method. Variations of the Angoff method which usually involve making probability judgments remain the most widely applied methods in standard setting (Cizek & Bunch, 2007; Cohen, Kane, & Crooks, 1999), though in recent years applications of the Bookmark method have begun to overtake it in popularity as the use of Rasch and other IRT applications spread in State-wide testing programmes in the United States. The modified Angoff method relies on a definition of a borderline or minimally competent examinee. To make the probability judgements required by the method less cognitively demanding to judges, the judgment tasks for the panels used in this study asked for an estimate of the number of candidates in a group of 100 minimally competent test takers who would correctly answer a test item. For Grade 1, the judges were asked "For 100 test takers minimally competent at CEFR level C1, how many will correctly answer the item?" For Grade Pre-1, the question was "For 100 test takers minimally competent at CEFR level B2, how many will correctly answer the item?" Similarly for the other grades, the wording was adjusted according to the CEFR level most relevant to that grade.

There were two reasons for the decision to employ two standard-setting

methods in each panel. The primary reason for the selection of the two methods was to compensate for the inherent limitations of each method that have been noted in the literature on standard setting, particularly in the context of linking to the CEFR. As noted in Chapter 2, standard setting does not have a well-established place in either Europe or Japan, and so it was thus thought that the Basket method would prove more accessible to the panels used in this study. At the same time, the Basket method has serious limitations (Kaftandjieva, 2009, 2010). Kaftandjieva (2010, p. 129) recommends against its use for high-stakes testing programmes due to, among other reasons, "considerable distortion of the cut scores in terms of underestimation of the lower and overestimation of the higher cut score." The Angoff method, on the other hand, has been criticized for placing too great a cognitive burden on participants, particularly the variations of this method that utilize estimates of the probability of borderline test takers correctly answering a test item, (Cizek & Bunch, 2007).

The decision was taken therefore to employ the Basket method in the first round of judgments to be made by each panel, as this would provide a more accessible introduction to evaluating the test items under review in relation to the CEFR. This preliminary round of judgments would be followed by a subsequent round of judgments employing the more conceptually demanding modified Angoff method. The Basket method was thus used first as a "primer," to help judges form an initial impression of items in terms of the CEFR before using the more conceptually difficult Angoff.

For each grade, judges first judged items using the Basket method. After that, they then rated the same items using the Modified Angoff procedure. Judges were given feedback in the form of the actual proportion correct for each item when it was administered in a live test, and given the chance to change their judgments for both Basket and Angoff ratings (the use of feedback and discussion differed between the two panels and is explained in more detail under the relevant procedure section of each standard-setting panel below). It was expected that this approach to estimating the probability of test takers correctly answering the items would make the Modified Angoff procedure more accessible to the panel judges and also improve the precision of those judgments. From the outset it was decided

that the second round of judgments employing the more statistically robust Modified Angoff procedures would form the basis of determining the cutscores for the relevant CEFR level on each of the EIKEN Reading tests under investigation.

The second reason for selecting multiple standard-setting methods is to take up Kane's (2001b) call to replicate standard setting with different methods as a powerful source of external validity evidence. The use of multiple standard-setting methods by each panel could thus provide additional validity evidence by allowing for the comparison of the cutscores obtained by the two methods. If the cutscores showed a reasonable degree of consistency, this would add strength to the evidence obtained by the standard-setting panels. It is important to note, however, that features of the way the cutscores are set in the Basket method make it unsuitable for tests targeted at a restricted, narrow range of ability (Council of Europe, 2009; Kaftandjieva, 2010). This meant that using cutscores derived from the Basket method as a source of validity evidence by way of comparison with the Angoff cutscores was only applicable for the more advanced grades. This important limitation of the Basket method is described further in the discussion of the results for Panel 2.

5.2.2 Validation

Validation in this chapter refers to the evaluation of the quality and accuracy of the standard setting carried out for the purposes of RQ3 to investigate the relationship between the reading tests used at different EIKEN levels and the CEFR proficiency levels to which those reading tests were hypothesised to be relevant in Chapter 1. As already noted above, the establishment of a link between an examination and the CEFR is premised on the assumption that the examination is fit for purpose, i.e. the uses and interpretations to which the examination is applied can be supported through a comprehensive and coherent validity argument. The Manual (2009) devotes only a small section of its chapter on validity to wider validity issues in relation to the quality of the test itself, as the purpose of the Manual is clearly focused on procedures for linking through standard setting, and the discussion of validity in it focuses on issues related to validating that standard

setting. In the context of this study, RQ3 contributes criterion-related validity evidence to the wider validity argument for the EIKEN reading tests, and issues not related directly to standard setting, such as content and scoring validity issues, are thus discussed in RQ1 and RQ2.

It is worth considering here what would constitute a reasonable result for the separate standard-setting panels used to investigate RQ3. In Chapter 1, each of the EIKEN grades was posited to be relevant to a particular CEFR level. The results of the standard setting will be discussed in terms of the reasonableness of the claim that a test taker who has passed a particular EIKEN grade can be considered to have demonstrated sufficient proficiency to be classed as having crossed the threshold of the relevant CEFR level from the one below it. For Grade 1, for example, this would require a level of proficiency to pass which would also be sufficient to achieve a classification of competent at the C1-level of proficiency in terms of the CEFR. It needs to be remembered that RQ3 is not designed to change, alter, or amend the pass/fail decisions of the current EIKEN tests in any way. What the study is designed to do is verify whether a test taker who has been classified as passing a particular grade, for example who is a certificate holder for Grade 1, would also be considered to have demonstrated proficiency at the relevant CEFR level, which for Grade 1 would be C1. The final cutscore point required to achieve classification at the relevant CEFR level will be examined in terms of whether each of those cutscores falls below the score needed to pass the EIKEN test using the established pass/fail decisions for each EIKEN grade.

This application of the concept of reasonable decisions in relation to the specific focus of RQ3 has implications for the interpretation of the external validity evidence collected through the application of multiple standard-setting methods. It is now widely accepted that different standard-setting methods will derive different cutscores (Cizek, 2001; Zieky, 2001). Indeed, Kaftandjieva (2010) also notes that different cutscores will be obtained if standard setting is replicated using the *same* method. Cizek and Bunch (2007) take the view that the use of multiple methods should be avoided as there is no consensus on how to synthesize the different results. As RQ3 is designed to investigate whether it is reasonable to

claim relevance between the level of performance required to pass an EIKEN grade and a particular level of the CEFR, and given that differences between results obtained from different standard-setting methods are to be expected, for the current study, it was decided to define an acceptable degree of difference which would constitute reasonable results for the validation of RQ3. In effect, the concern is whether the application of multiple methods replicates the finding that a test taker who would have passed a particular EIKEN grade would also have demonstrated sufficient proficiency for classification at the relevant CEFR level. The purpose is not, then, to attempt to replicate exactly the same cutscores through different methods, which, as already noted, has been recognized as unrealistic. The problem Cizek & Bunch (2007) note when using multiple standard-setting methods—in terms of how to decide which cutscore is actually the "right" cutscore—becomes less of an issue in this context provided that both cutscores fall below or close to the score required by a candidate to pass the EIKEN grade in question.

The evidence for procedural and internal validation will be discussed separately within the results sections for each of the two main standard-setting panels. The use of two standard-setting methods for each of these panels, as already noted, constitutes an aspect of validity evidence, and evidence from this perspective will also be addressed under the results section for each panel. A separate section, Section 5.5, will focus on describing a third standard-setting study which was carried out specifically to collect external validity evidence.

As already noted, replication of standard setting utilizing different methods constitutes one form of an external verification check. In fact Kane (2001b, p. 75) goes further than simply recommending the use of different methods, suggesting that combining different methods with different participants, in effect complete replication as independent standard-setting events, "would provide an especially demanding empirical check on the appropriateness of the cutscore." The standard-setting study described in section 5.5 was devised and carried out after the two main standard-setting panels described in sections 5.3 and 5.4, and was specifically designed as an attempt at external validation. The rationale, procedures, participants and results of this separate study, and the

contribution the study makes to the validation of RQ3 are thus discussed separately to the two main standard-setting panels.

5.3 Standard setting panel 1

5.3.1 Introduction

The first standard-setting panel focused on the advanced-level Grade 1 and Pre-1 tests. The standard setting for the reading tests of both examinations was undertaken as part of a three-day standard setting event which also addressed the listening, vocabulary, and writing components. As the three research questions for this study are focused only on the validation of the reading test components of the EIKEN tests, the subsequent description will focus on the instruments, procedure and results for the standard setting aimed at reading. However, the discussion of procedures and the order of training, etc will necessarily touch on the other components as these were integrated into the standard-setting event.

5.3.2 Participants

Given the typical uses and interpretations, the TLU domain, and the typical test takers described for Grade 1 and Grade Pre-1 in Chapter 1, the following criteria were identified as minimum requirements for the recruitment of participants for Panel 1.

- 3 years teaching English at university level in Japan
- English ability sufficient to deal with high-level English test items and all CEFR related training material
- Knowledge of and experience using EIKEN tests
- Ability to take part in all stages of the workshop

The criteria were adopted taking into consideration Jaeger's recommendation (1991, p. 4) that "expert judges should be well experienced in the domains of expertise we demand of them." The judges would be required to bring to bear knowledge and expertise from several areas, including experience with the context of higher education in Japan and the learners who constitute a

large part of the typical test takers for Grade 1 and Grade Pre-1, as well as knowledge of the content and format of the EIKEN tests, in addition to knowledge of the CEFR. Recent moves to promote the development of standards-based curriculum goals in both higher education as well as secondary school in Japan have led to an increased focus on the CEFR, particularly in relation to developing performance level descriptors for schools in the form of can-do descriptors. Nonetheless, at the time of the recruitment for both panels, the CEFR was not required for any official purposes of certification by MEXT and was not used to define benchmarks of English ability in official documentation (e.g. MEXT 2003, 2011). This meant that it would be difficult to recruit participants who would be familiar with the CEFR and also have extensive experience in the context of Japan. It was thus decided that requiring knowledge of the CEFR in advance would place too great a restriction on the potential pool of participants. It was decided instead to prioritize experience and knowledge of the EIKEN testing system and the higher education context in Japan, and to focus training during the event on developing familiarity with the CEFR.

Initially, it was intended to recruit even numbers of native speakers of English (NS) and non-native speakers of English (NNS) in order to create sub groups within the main panel which would provide the ability to conduct further internal validity checks by allowing for comparison of the standards set by the two different groups. However, this was not possible to achieve, and the final panel consisted of 13 judges, 10 of whom were native speakers of English and only 3 were non-native speakers of English. The number of participants falls within the range of participants noted in the literature on standard setting in Chapter 2.

All participants had experience of the Grade 1 and Pre-1 exams by serving in one or more of the following capacities: a) EIKEN speaking test examiners for official administrations, 2) working as EIKEN item writers on a commission basis, and 3) sitting on editorial review panels which met several times a year to review and critique test content for use in future live examinations. Table 5.1 gives a breakdown of the professional experience of the judges in relation to the higher education sector in Japan. As can be seen, all of the panel participants met the

minimum criteria in terms of professional experience, with the average level of experience being 17.8 years. Table 5.2 gives an overview of the number of participants who also had experience in other educational sectors. As a group, the panelists are a very experienced group of educators with a wealth of experience in the university sector in Japan. The majority of panelists also have experience teaching in the business English sector, in company training programs, etc., which gives them important expertise in another major area of application for the EIKEN Grade 1 and Pre-1 examinations.

Table 5.1 Years of experience teaching at the university level in Japan

R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
32	18	5	13	8	35	17	23	33	8	22	15	3

Table 5.2 Number of participants with experience in other education sectors

Elementary school	Junior high school	High school	Business (company training etc)
2	3	7	10

5.3.3 Instruments

The main materials utilized for the standard-setting event for Panel 1 are listed in Table 5.3. The self-study preparation booklet represents a significant deviation from the recommended procedures in the Manual (Council of Europe, 2009), and will thus be explained in some detail before an overview of the remaining materials is given.

Table 5.3 Materials used for Panel 1 standard-setting event

Purpose / timing	Description of instrument
Training: prior to event	A self-study preparation booklet providing information the CEFR, and the standard setting methods to be used
Training: during event	Reading items supplied by the Council of Europe to demonstrate reading tasks aligned with CEFR
Data collection during event	A retired reading test for each grade
	Empirical feedback (proportion correct values for the items in the live administration in which they were used).
	Separate rating forms for each method for each grade
Data collection: end of event	A questionnaire collecting background information as well as data for evaluating procedural validity

5.3.3.1 The self-preparation training booklet

Participants had been recruited by prioritizing experience and knowledge in areas of expertise other than the CEFR and standard setting in order to maximize the available pool of participants. The intention was to deal with familiarization with the CEFR and training in standard setting itself during the actual event. However, it soon became clear that the time available for training would be limited. Potential participants were asked about their availability prior to acceptance (one of the criteria for recruitment), and it became clear that the longer the event lasted, the fewer qualified participants would be available to commit to the entire process, as outlined in Table 5.4. The final panel would be expected to deal with two complete First-Stage tests during the standard-setting event, which as described in Chapter 1, include grammar and vocabulary, reading, listening and writing components. Although the original intention had been to spread these activities over a four-day period, it was decided to condense this into a three-day session to incorporate as many of the qualified participants as possible. Day 1 and 2, during

which all 13 qualified participants would be able to attend, would focus on the vocabulary, reading, and listening components of the tests, all of which would utilize the same test-centered standard-setting methods. Day 3 would focus on writing, utilizing a different, examinee-centered standard-setting method which was derived from the Paper-Selection method (for descriptions of this method see Loomis & Bourque, 2001; Tannebaum & Wiley, 2005; and Tannenbaum & Wiley, 2008).

Table 5.4 Participants availability over the planned standard-setting event

Workshop Day	Day 1	Day 2	Day 3	Day 4
Judges able to attend	13	13	12	10

Given the reduced time to deal with training during the actual standard-setting event, it was decided to address familiarization activities through a self-study preparation booklet. A 34-page booklet was prepared which covered the following areas (see Appendix H for the table of contents of the booklet):

- Administrative information (maps, schedules, etc)
- The purpose of the project
- An introduction to the CEFR, assuming no prior knowledge and including self-study familiarization tasks
- An outline of the standard-setting methods to be used

The booklet included information on the background and development of the CEFR, including a discussion of the various debates and criticisms surrounding the use of the CEFR in test development. However, the primary purpose was to provide familiarization with the CEFR prior to the actual standard-setting event, in particular the Illustrative Scales spread through chapters 3, 4., and 5 of the CEFR, as these would function as the Performance Level Descriptors (PLDs) to define performance for the standard setting panels. The Manual (Council of Europe, 2009) stresses the importance of familiarization, However, before proceeding to standard setting on the target examination,

familiarization needs to be followed by training with tasks and items already calibrated to the CEFR to illustrate how the CEFR illustrative scales are operationalized in practice. Given the limited face-to-face time available for the event, it was thus decided to incorporate the training with calibrated examples into the standard-setting event, and adapt the familiarization activities from chapter 3 of the Manual (Council of Europe, 2009) for use in the preparation booklet.

The activities described in the Manual were trialed with a focus group consisting of internal EIKEN content specialists. Those activities considered most relevant to the EIKEN tests to be linked during the standard setting and which were most amenable to self-study based on feedback from the focus group were selected for use in the booklet. A draft of the booklet was prepared by the author and feedback was obtained from the same focus group participants before a final version was produced. A total of eight tasks were adapted for use in the booklet. Table 5.5 provides an overview of these tasks, the scale in the CEFR that each task focuses on, a brief description of the activity, and the location in the CEFR of the scale or scales used in the task. For the purposes of the booklet, all relevant scales to be used during the standard setting event were included as appendices, though not all of these were the focus of familiarization tasks. Participants were thus able to review their own responses to tasks that required reordering of descriptors, etc. The appendices, however, were sealed when the booklet was sent to the participants, and participants were asked not to open the sealed section until after they had attempted the tasks.

Table 5.5 Self-assessment tasks in preparation booklet

Tasks	Focus	Description of activity
Tasks 1 & 2	Global Scale	Reflection, using scale to consider level of own learners, summarizing significant level features for B1, B2, C1
Tasks 3 & 4	Self-assessment grid	Rating own level, reviewing (and if appropriate revising) level descriptions made in Task 2
Tasks 5 & 6	Illustrative scales for listening	Re-ordering of jumbled descriptors within each scale, raters put descriptors in level they think appropriate
Task 7	Overall Reading scale	Reordering jumbled descriptors from Overall Reading Scale
Task 8	Overall Listening and Overall Reading	Comparing Overall Listening and Reading scales, noting any significant differences between key words and definitions in the two scales

5.3.3.2 Exemplars of listening and reading items at different CEFR levels

The Council of Europe provides examples of reading and listening items that have been linked to the CEFR (Council of Europe, 2005). The examples are supplied by examination boards with accompanying documentation noting the level the task is calibrated to and any associated content specification. The tasks are provided on a CD ROM which can be obtained from the Council of Europe on request[7]. For the purposes of training with exemplar items before undertaking actual standard setting of EIKEN tests, five tasks each were chosen from the tasks provided by the Council of Europe. The tasks were chosen to span the range of levels from A1 to C1, representing the levels targeted by the EIKEN tests.

5.3.3.3 EIKEN test items and rating forms for standard setting

For the purposes of carrying out the actual standard setting, one complete first-stage test for both Grade 1 and Grade Pre-1 was used. The retired tests had been administered as live tests in nation-wide examinations and followed the same structure and test specifications described in the overview of EIKEN

[7] Details are provided on the *CEFR and language examinations: a toolkit* webpage: http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

examinations in Chapter 1. As explained in Chapter 1, a first-stage test contains vocabulary, reading, listening and writing components, all of which were to be the focus of standard setting during the three-day event.

Separate rating forms were created for each standard-setting method employed. Examples of the rating forms used for both the Basket method and the modified Angoff method are included as Appendix I. The rating forms included the wording for the judgment task for each method at the top of the form. Empirical feedback was presented to candidates in the form of a separate sheet with a list of items and the proportion correct achieved in the live administration of that item.

5.3.3.4 Procedural validity questionnaire

Evidence to evaluate and support the procedural validity of the standard setting was collected through a questionnaire adapted from Cizek and Bunch (2007). As already noted, it was expected that participants would not have a high level of familiarity with either the CEFR or standard-setting, and so in addition to adapting the questions suggested in Cizek and Bunch (2007), questions were added to ascertain the degree of familiarity participants had with both the CEFR and standard setting prior to taking part. Participants were then asked to evaluate the usefulness of the preparation booklet, including the CEFR familiarization tasks, in helping them to establish the necessary understanding of the CEFR in order to take part effectively in the standard-setting panel. Questions relating to the preparation booklet and the confidence that participants had in both their understanding of the CEFR and their ability to apply the standard-setting methods used were considered particularly important from the perspective of procedural validity.

5.3.4 Procedure

Appendix J provides an overview of the schedule of activities, in the order in which they were conducted during the three-day standard-setting event for Panel 1. As explained above, the panel was asked to address standard setting for all

components in the First-Stage exams, including the reading component which is the focus of this study. As procedures for components were to some extent integrated, it is relevant to provide an overview of the full three-day event, not only those sections which applied directly to reading. In addition to planning the events, selecting methodology, and preparing all materials including the preparation booklet, the author acted as coordinator throughout the three-day event, chairing the discussion, leading the training sessions and ensuring the schedule and procedures for standard-setting were followed.

The self-study booklet also contained brief explanations of the standard-setting methods to be applied, and the purpose of the standard-setting event. Day one of the standard-setting event began with an open discussion of the participants' comments and impressions of the CEFR, focusing in particular on the familiarization tasks they had undertaken. In keeping with the tone set by the booklet, participants were not pushed to supply their responses, but instead invited to offer their views on the tasks that were undertaken. This approach proved generally fruitful, with participants willing to share their views and own results regarding the tasks (see Appendix K for results of the participants' views on the opportunity for discussion). On the first day of the event, the discussion was generally broadly focused, and the discussion of the CEFR was brought to a close by agreeing on a list of key words and features which could be considered to distinguish the descriptions of levels in the Global Scale. On the second and third days of the event, the opening discussion sessions focused specifically on the skills to be rated on that particular day. For the reading test, this meant focusing on the scales relevant to reading, including the Overall Reading scale and other sub-scales thought relevant to the reading tasks in the EIKEN tests.

Training was then undertaken with items that had been calibrated to the CEFR by other examination boards. On Day 1, five listening tasks were used, and on Day 2, five reading tasks were used. The items were selected to cover a range of CEFR levels, but were not presented in their order of difficulty. Participants were first asked to consider one item, and answered the item from the perspective of a test taker before being invited to suggest the appropriate CEFR level. Participants discussed the rationale for their decision, and were encouraged to

provide a specific descriptor from the relevant CEFR scales to support their decision. The CEFR level at which the item was calibrated was then provided, along with an explanation of the rationale for that calibration by the examination board which provided it. This process was repeated for the remaining items.

Training with the calibrated items had two goals: 1) it provided exemplars of how descriptors were being operationalized in practice by European examination boards; 2) estimating the CEFR levels of the items was used as a way of introducing the judgment task and procedure for the Basket method, which relies on allocating items to the correct "basket," in this case a CEFR level. At the end of the training session, the description provided in the self-study booklet was recapped to explain how items allocated to each level in the Basket method are aggregated in order to determine the cutscore points between levels During training on the first day with listening items, more time was taken to discuss the process of standard setting itself. On the second day, as the concepts underlying the judgment tasks for the standard-setting procedures remain substantially the same, less time was needed to focus on explaining this element during training.

After training with items calibrated to the CEFR, the panel proceeded to undertake standard setting with EIKEN Grade 1 items. For the vocabulary, reading and listening components, the process was essentially the same. Participants first took the relevant component, answering all items in that component in the same way and under the same conditions as test takers. Participants were asked not to estimate the level of the item until they had first completed the test. Participants were then asked to review each item, and to record their judgments on the rating sheet provided.

Following the rating of items using the Basket method, participants were then introduced to the modified Angoff procedure. An important part of the training procedure was in discussing the wording of the judgment task, focusing in particular on the definition of a minimally competent test taker. For Grade 1, the focus was on a test taker minimally competent at the C1 level, as this was the level considered most relevant to Grade 1. Given the accepted cognitive demand of making probability judgments, two commonly employed modifications noted in Cizek and Bunch (2007, p. 85) were employed to reduce the burden of the task

on participants. The first modification involves the conceptualization of probability in the judgment task, making this more concrete by referring to "100 candidates minimally competent" at a particular level, and asking how many of those candidates will correctly answer the item. This is considered to provide a more accessible judgment task than asking for the probability of a single minimally competent candidate correctly answering the item. The second modification concerns the estimate of probability itself, asking for the judgments in increments of 10 (0, 10, 20, 30, etc), rather than asking for specific estimates, for example 92 test takers out of 100 (Cizek & Bunch, 2007; O'Sullivan, 2015b, Tannenbaum & Wiley, 2004, 2008).

After discussion of the judgment task to be employed for the modified Angoff method, participants were asked to re-rate the same items they had rated for the Basket method. They were not required to take the items again under the same conditions as test takers, as they had already become very familiar with the items during the previous round of rating. Participants had access to their Basket method judgments during the rating process. This was considered important to ameliorate the problems expected with employing the cognitively demanding modified Angoff method in conjunction with PLDs which were a new and unfamiliar standard. Having access to the Basket method ratings was thus deliberate and part of the strategy to use this method as a primer to build an understanding of how the items related to the PLDs in a more general way before asking for probability estimates of how many minimally competent candidates as defined by the PLDs would successfully answer the item.

Following the modified Angoff round of rating, participants were provided with empirical feedback on the difficulty of the items in the form of the proportion of test takers answering the item correctly in a live administration of the examination. Participants were advised that the empirical feedback was for reference only, and that they were not required to take it into account. Following the advice by Hurtz and Auerbach (2003) on the importance of providing instructions on the meaning of feedback in the form of *p* values, participants were also cautioned about the difference in interpreting the proportion correct figures, which come from a live test population spanning a range of abilities, and their

own estimates for the modified Angoff approach, which are based on an imaginary sample of 100 test takers at exactly the same level of ability. The proportion correct figures were, however, derived from a typical administration of both grades, with large numbers of test takers as described in Chapter 1. The *p* values did, then, give the participants an accurate picture of how the items performed in relation to the typical test-taker population for each grade in Japan, which was a standard that the judges were all familiar with as both experienced university educators in Japan and experts with experience of the EIKEN exams.

No discussion was undertaken at this point, and participants were advised to review the empirical data individually. In a meta-analysis of the impact of modifications including discussion on applications of the Angoff method, Hurtz and Auerbach (2003) note that discussion when an explicit performance level descriptor standard has been agreed does reduce variability, but also leads to higher cutoffs. However, they also caution that the reason for the trend towards higher cutoffs with this modification is not entirely clear, nor whether the resulting movement towards a higher score is actually a more valid one. They also note that group dynamics can be dominated by a small number of vocal participants which may also impact on the development of group consensus. Given the ambiguity around the usefulness of group discussion in forging consensus, and the practical time constraints under which the standard setting was taking place, it was decided to eliminate this phase. Bearing in mind Berk's criteria of practicability, it was decided the procedure employed, without discussion, achieved the best balance between procedural validity in the application of the methods and efficient use of the short time available for the panel.

The procedures outlined resulted in the effective inclusion of two rounds of judgments, which follows Reckase' (2001) guidelines to incorporate multiple rounds to allow for raters to act on the inclusion of feedback in the standard standard-setting process. However, the rating forms for Grade 1 and Pre-1 did not give provision for recording two rounds of judgments. Instead raters were asked to make any changes to their rating forms before submission and to include the final rating they wished to submit. This too was a practical consideration, as the

final judgment of the Angoff method, following the provision of feedback, was from the outset intended as the most robust measure of the standard-setting process, based on the literature. It was thus decided to only collect this rating. This procedure was modified for Standard Setting Panel 2, which allowed for an evaluation of the amount of change between rounds for that panel, which was not possible for Panel 1 due to the rating form used

5.3.5 Results

5.3.5.1 Standard-setting results

Table 5.6 provides the results for both standard-setting methods for the Grade 1 Reading items. As explained in Section 5.2, the decision was made a priori to base the cutscores on the results derived from the Angoff method. The Basket method was intended as a primer to help judges estimate the level of EIKEN Reading items in relation to the CEFR before attempting the probability judgments in the Angoff method. The main results of interest, then, in terms of determining the cutscore are those of the Angoff method. The results for the Basket method are supplied to provide a source of evidence for the evaluation of external validity. Table 5.6 also contains information on the standard error of the cutscore (SE_c), which is important for the discussion of internal validity.

For each rater, two figures are reported for the Angoff method. The first is the percentage correct that a test taker needs to achieve on the Grade 1 Reading component in order to be considered minimally competent at the C1 level of the CEFR. This, then, is the cutscore necessary for crossing from B2 to C1, and is derived by taking the mean of the probability judgments across all items in the reading component for each rater. The second figure is this percentage expressed as the number of items which need to be answered correctly out of the total number of items in the reading component. The final cutscore is the mean of cutscores for all individual participants. Referring to Table 5.6, the cutscore for Grade 1 is thus 58.6 percent, or 9.4 items correct out of the total of 16 items in the Reading component.

Table 5.7 provides the results for Grade Pre-1. The interpretation is the same as Grade 1, with the results of the Angoff method providing the basis for

deriving the cutscore for determining how many items a test taker needs to answer correctly in order to be considered minimally competent at the B2 level. For Grade Pre-1, the cutscore is 60 percent, or a raw item count of 9.6 items out of the total number of 16 items in the Grade Pre-1 Reading component.

It is important to note the difference in interpretation of the results of the two standard-setting methods employed in this study for setting the score that test takers are required to achieve in order to demonstrate attainment of a particular performance standard. For the Angoff method, the normal procedure is to average the probability judgments across all items for each participant, and then average the individual cutscores derived for each rater. According to the judgment task for the modified Angoff method used in this study, this reflects the minimum score required to be considered minimally competent at the C1 level for Grade 1 and the B2 level for Grade Pre-1. The judgment task for the Basket method, however, requires judges to allocate items to a CEFR level. The numbers shown in Table 5.6 for Grade 1 are the cumulative number of items (or percentage of the total number of items) that have been allocated by each participant to all levels up to and including the level below the level of interest. For Grade 1, this is the cumulative number of items in all levels up to and including B2 (the level below C1). Accordingly, the number of items actually required to cross the level distinction from B2 to C1 would be the cumulative number of items up to and including B2 plus 1 (see Kaftandjieva, 2010, pp. 61-62). As the average number of items allocated to all levels up to and including B2 is 9.8 (or 61.5% of the total number of 16 items), the actual number of items required to be classified as C1 according to judgments derived by the Basket method for Grade 1 in this study would be 11 (9.8, rounded to the nearest whole number plus 1). For Table 5.7 for Grade Pre-1, the numbers for the Basket method represent the cumulative number of items allocated to all levels up to and including B1 (the level below B2).

Table 5.6 standard-setting results for Grade 1 Reading

	Angoff		Basket	
Rater	Percent	No. of Item	Percent	No. of Items
R1	56.9%	9.1	62.5%	10
R2	51.9%	8.3	31.3%	5
R3	68.8%	11	56.3%	9
R4	52.5%	8.4	68.8%	11
R5	46.3%	7.4	81.3%	13
R6	70.6%	11.3	75.0%	12
R7	64.4%	10.3	12.5%	2
R8	43.1%	6.9	56.3%	9
R9	52.5%	8.4	37.5%	6
R10	65.6%	10.5	93.8%	15
R11	61.9%	9.9	93.8%	15
R12	65.0%	10.4	62.5%	10
R13	61.9%	9.9	68.8%	11
Mean	58.6%	9.4	61.5%	9.8
SD	8.7%	1.4	23.6%	3.8
SE_c		0.4		1.0

Table 5.7 Standard-setting results for Grade Pre-1 Reading

Rater	Angoff		Basket	
	Percent	No. of Item	Percent	No. of Items
R1	60.6%	9.7	81.3%	13
R2	54.4%	8.7	31.3%	5
R3	70.0%	11.2	75.0%	12
R4	55.6%	8.9	62.5%	10
R5	51.9%	8.3	87.5%	14
R6	65.6%	10.5	62.5%	10
R7	61.3%	9.8	6.3%	1
R8	50.6%	8.1	68.8%	11
R9	52.5%	8.4	31.3%	5
R10	69.4%	11.1	100.0%	16
R11	57.5%	9.2	93.8%	15
R12	68.1%	10.9	75.0%	12
R13	61.9%	9.9	56.3%	9
Mean	60.0%	9.6	63.9%	10.2
SD	6.8%	1.1	27.1%	4.3
SE_c		0.3		1.2

As described in Chapter 1, some task types in the Grade 1 and Grade Pre-1 Reading and Listening components are weighted differently. Table 5.8 provides the number of items, the total possible raw sore and the total possible weighted raw score for each component of the First-Stage tests for Grades 1 and Pre-1. The cutscore for the Reading component, expressed as the required weighted raw score, is thus 15.2 out of 26 for Grade 1, and 15.6 out of 26 for Grade Pre-1. An overview of the weighted and unweighted cutscores derived by the Angoff method for all components of the First Stage tests based on the judgments of Panel 1 is provided in Table 5.9.

Table 5.8 Overview of test structure and scoring for Grades 1 and Pre-1

Grade 1	Vocabulary	Reading	Listening	Writing	Total
Weighted score	25	26	34	28	113
Raw Score	25	16	27	28	96
No. of items	25	16	27	1	69
Percent	22%	23%	30%	25%	100%
Grade Pre-1	**Vocabulary**	**Reading**	**Listening**	**Writing**	**Total**
Weighted score	25	26	34	14	99
Raw Score	25	16	29	16	86
No of items	25	16	29	1	71
Percent	25%	26%	34%	14%	100%

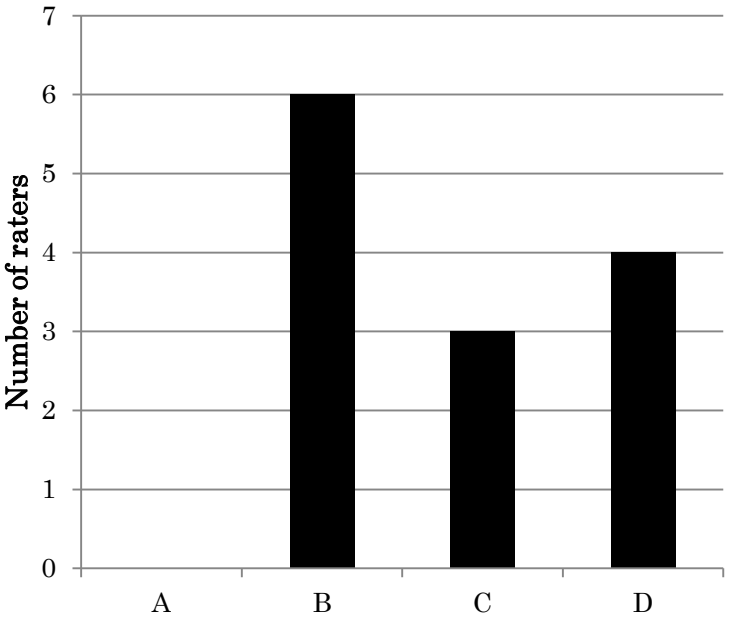
Table 5.9 All First Stage cutscores for Grade 1 and Pre-1

Grade 1	Percent	Raw	Weighted
Vocabulary	58.7%	14.7	14.7
Reading	58.6%	9.4	15.2
Listening	61.9%	16.7	21.0
Writing	83.1%	23.3	23.3
First Stage Overall	65.7%	64.0	74.2
Grade Pre-1	**Percent**	**Raw**	**Weighted**
Vocabulary	58.3%	14.6	14.6
Reading	60.0%	9.6	15.6
Listening	57.8%	16.8	19.7
Writing	86.6%	13.3	12.1
First Stage Overall	62.6%	54.3	61.9

5.3.5.2 Procedural validity

Evidence to evaluate and support the procedural validity of the standard setting was collected through a questionnaire adapted from Cizek and Bunch (2007). Figure 5.1 shows the participants' familiarity with the CEFR prior to agreeing to

take part in the panel. As predicted, there was a low level of familiarity with the CEFR. No participants had actually studied the CEFR in detail prior to agreeing to participate in the event, and just over half of the participants expressed that they had either not heard of the CEFR or if they had heard of it, were not familiar with its aims or contents.



A. I had read the CEFR and was familiar with its aims and contents, including the Common Reference Levels.
B. I was familiar with the aims of the CEFR, but had not studied it in detail.
C. I had heard of the CEFR but was not familiar with its aims or contents.
D. I had not heard of the CEFR.

Figure 5.1 Prior knowledge of CEFR

Table 5.10 shows the responses to four statements designed to elicit information on the level of knowledge and experience participants had of standard setting prior to taking part in the panel. Although it was intended that all participants would respond *yes* or *no* to all four statements, this section seemed to create some confusion, and not all participants answered all of the four questions. Seven participants indicated that they had had experience of acting as a judge/rater on standard-setting panels. However, one of those participants also responded negatively to both statements regarding being familiar with standard setting, indicating that there may have been some confusion regarding the

interpretation of what would constitute a standard-setting panel. As the questionnaire was collected at the end of the standard-setting event, there was no opportunity to investigate responses in more depth. However, the overall trend, as with knowledge of the CEFR, matched the predicted state of knowledge, with just under half of participants indicating that they were not familiar with the concept of standard setting prior to taking part in the panel, and only two participants indicating that they were familiar with the standard-setting methods employed in this study.[8]

Table 5.10 Experience with standard setting

Questions	Yes	No
I had had experience acting as a judge/rater on standard setting panels.	7	5
I had had experience organizing standard setting panels.	3	7
I was familiar with the concept of standard setting, and had heard of one or more of the methods which were used in the STEP project.	2	8
I was familiar with the concept of standard setting, but had not heard of any of the methods used in the STEP project.	5	6

Table 5.11 shows an overview of questions relating to aspects of procedural validity of the standard-setting event, including the activities designed to help build an understanding of the CEFR and the standard-setting methods employed. Participants were asked to respond on a likert-type scale, indicating the degree of agreement with each of the 15 statements. The four possible choices were *strongly disagree*, *disagree*, *agree*, *strongly agree*, The responses were converted to a four-point numerical scale for analysis with *strongly disagree* equal to one and *strongly agree* equal to four. The means of the responses on the four-point scale are shown in Table 5.11. A breakdown of the numbers of

[8] One candidate responded "yes" to both statements on familiarity. As there was no opportunity to confirm with this participant which of the responses was correct, responses by this candidate to both familiarity questions were treated as missing.

participants selecting each response category for each question is shown in Appendix K as figures K1 to K15.

Table 5.11 Questions on procedural validity for Panel 1

	Questions	No. of Raters	Mean
Q1	The preparation booklet gave me a clear understanding of the purpose of the project.	13	3.5
Q2	The explanations and tasks in the preparation booklet helped me understand the structure of the CEFR and the Common Reference Levels.	13	3.4
Q3	The group discussion of the CEFR at the start of the workshop aided my understanding of the CEFR and the Common Reference Levels.	13	3.3
Q4	The time provided for the discussion was adequate.	13	3.1
Q5	There was an equal opportunity for everyone to contribute his/her ideas during the discussion.	13	3.3
Q6	The training tasks with the items supplied by the Council of Europe were useful.	13	3.4
Q7	The time provided for training with the Council of Europe items was adequate.	13	3
Q8	The explanation of the Basket Method was adequate and I felt able to undertake the rating tasks for the listening, reading, and vocabulary items.	13	3.2
Q9	The explanation of the Modified Angoff Method was adequate and I felt able to undertake the rating tasks for the listening, reading, and vocabulary items.	13	3
Q10	The time provided for rating the EIKEN listening, reading, and vocabulary items was adequate.	13	3.4
Q11	The feedback on item difficulty of the EIKEN listening, reading, and vocabulary items was useful.	13	3.2
Q12	The explanation of the Examinee Paper Selection Method used for rating the EIKEN writing samples was adequate and I felt able to undertake the rating task.	12	3.1
Q13	The time provided for rating the EIKEN writing samples was adequate.	12	3.4
Q14	The facilities and food service were adequate and helped create a productive and efficient working environment	13	3.8
Q15	During the workshop I felt I had adequate opportunities to present my opinions and was able to ask questions when I was not sure of how to proceed	13	3.5

The first two questions relate directly to the effectiveness of the self-study preparation booklet. The majority strongly agreed that the booklet gave them a clear understanding of the purpose of the project. For Q2, regarding how effective the explanations and tasks were in building an understanding of the CEFR, fewer participants strongly agreed, but all agreed to some degree that the booklet was effective. Questions 3, 4, and 5 focus on the discussion of the CEFR during the actual event. The majority of participants once again indicated that they were generally satisfied with the amount of time and the conduct of the discussion. For Q4, two participants did not agree that the discussion helped their understanding of the CEFR, but the remaining 11 participants all responded positively to this statement, indicating that they found the discussion useful in building an understanding of the CEFR.

Questions 6 and 7 also indicate a high level of agreement that the selection and use of the training items provided by the Council of Europe was useful. Of crucial importance for establishing procedural validity is the confidence participants have in carrying out the judgment task following training. Question 8 and 9 address this issue specifically, asking about the adequacy of the explanation of the standard-setting methods and the confidence participants had in carrying out the judgment tasks. The responses are overwhelmingly positive, but there is not unanimous agreement, with two participants disagreeing with the statement about the Basket method, and three disagreeing in relation to the Angoff method. Only one participant felt that the time allocated for the rating of items was not sufficient. Similarly with the use of empirical feedback data, the majority of participants found the presentation of the data useful, but two participants did not agree with this statement. Questions 11 and 12 relate to the Paper Selection method used for the evaluation of writing samples. Although the writing component is not directly related to the focus of this dissertation, results from the questionnaire for writing are included for reference as it is relevant to consider the reaction of the participants to the standard-setting event as a whole in the discussion of procedural validity. All participants agreed that the facilities were adequate and, importantly, all felt that they had adequate opportunity to present their ideas and ask questions during the standard-setting event.

5.3.5.3 Internal validity

Three aspects of internal validity are examined in detail. Firstly the precision of the cutscores is examined through an evaluation of the two main sources of error affecting the classification of test takers; the standard error of the cutscore and error associated with the testing instrument. The relative severity and the consistency of the raters is then examined through the application of the multi-facet Rasch model to the rating data.

The precision, or replicability, of the cutscore is considered an important criterion for the evaluation of standard setting (Cizek & Bunch, 2007; Cohen, Kane, & Crookes, 1999; Jaeger, 1991: Kaftandjieva, 2010; Tannenbaum & Wiley, 2008). Cizek & Bunch (2007, p. 60) include this aspect in the internal validity category in their three-way categorization of evaluation criteria.

Cutscores derived through standard-setting methods such as the Angoff method are estimates of the cutscore derived by averaging the individual estimates of the standard-setting judges. As Jaeger (1991, 5) notes, "We can consider the mean standard that would be recommended by an entire population of qualified judges to be a population parameter. The mean of the standards recommended by a sample of judges can, likewise, be regarded as an estimate of this population parameter." The standard error of the mean for the cutscore can thus be calculated and this provides an estimate of the precision, or replicability, of the cutscore (Cizek & Bunch, 2007; Jaeger, 1991; Kaftandjieva, 2010; Tannenbaum & Wiley, 2008). Although different authors use different terminology, with Jaeger (1991) and Cizek & Bunch (2007) simply referring to the standard error of the mean, Tannenbaum & Wiley (2008) referring to the standard error of the judges, and Kaftandjieva (2010) and Cohen, Kane, and Crookes (1999) referring to the standard error of the cutscore, the same formula is used:

Formula 5.1 $SE_c = s_x/\sqrt{n}$

where SE_c refers to the standard error of the cutscore, s_x is the standard deviation of the mean of the estimates for individual judges, and n is the number of judges participating in the study. The resulting standard error provides an estimate of how much the cutscore is likely to vary in "replications of the procedure under similar conditions, with a different (though equivalent) group of

participants" (Cizek & Bunch, 2007, p. 300).

Referring back to the overview of standard-setting results for Grade 1 and Grade Pre-1 in Table 5.6 and Table 5.7, we see that the SE_c for the cutscore for the modified Angoff method for Grade 1 is 0.4 on the raw score scale, and for Grade Pre-1 0.3 on the raw-score scale. Cizek & Bunch (2007, p. 301) demonstrate how the SE_c allows us to calculate confidence intervals to estimate the amount of expected variability in the cutscores across replications of the procedure with the same method using the same number of judges and the same procedure. In order to calculate confidence intervals for the cutscore estimates, we can apply the same basic procedure used to calculate confidence intervals for test scores using the standard error of measurement (SEM). Bachman (2004, p. 173) defines three probability levels, .68, .95, and .95, as the most commonly used and provides the following formula for calculating score confidence intervals from SEM using the z-scores associated with each level of probability:

- $CI_{.68}$=X$\pm$1.00SEM (.68, or 68 percent confidence level)
- $CI_{.68.95}$=X$\pm$1.96SEM (.95, or 95 percent confidence level)
- $CI_{.68.99}$=X$\pm$2.58SEM (.99, or 99 percent confidence level)

where X is the test score and SEM is the standard error of measurement on the test score scale. Replacing X with the cutscore derived through the modified Angoff procedure and SEM with SE_c allows us to calculate similar confidence intervals for the cutscores. Table 5.12 shows the upper and lower ranges of the cutscores for each confidence level. To put the table in perspective, the cutscore for the Grade 1 Reading component is likely to fluctuate between estimates of 9 and 9.8 on the 16-point raw score scale in repeated replications 68 percent of the time. Bearing in mind that the passing score required for both grades is set at 70 percent, which equates to 11.2 or a rounded raw score of 11 on the 16-item Reading component, it can be seen that even at the strictest confidence level of 99 percent, the cutscore for candidates minimally competent at CEFR C1 for Grade 1 and CEFR B2 for Grade Pre-1 will not exceed the passing score required for the First Stage of each grade.

Table 5.12 Confidence intervals for Grades 1 and Pre-1 reading cutscores

Grade	Range	68%	95%	99%
Grade 1	lower	9	8.6	8.4
	upper	9.8	10.2	10.4
Grade Pre-1	lower	9.3	9	8.8
	upper	9.9	10.2	10.4

Guidelines for considering the impact of SE_c on the classification of candidates have been offered by several authors. These guidelines are premised on the assumption that there are in fact two sources of error, independent of each other, which impact on the classification of test takers: error associated with the sampling of judges setting the cutscore, SE_c, and measurement error associated with the items or tasks constituting the test, SEM (Cizek & Bunch, 2007; Cohen, Kane, & Crookes, 1999; Jaeger, 1991; Kaftandjieva, 2010). The impact of both of these sources of error can be evaluated by the following equation from Jaeger (1991):

Formula 5.2 $SE_{tot} = \sqrt{SE_c^2 + SEM^2}$

where SE_{tot} is the total error due to both sources. Jaeger (1991) and Cohen, Kane, and Crookes (1999) provide guidelines for acceptable levels of SE_c in relation to the total error. Based on equation 5.2, the increase in error due to SE_c can be shown to be 3 percent, or 1.03 times the SEM for the test when SE_c is less than $\frac{1}{4}$ the value of SEM, 5 percent or 1.05 times SEM when SE_c is $\frac{1}{2}$ the value of SEM, and 12 percent or 1.12 times SEM when $\frac{1}{3}$ the value of SEM (Cohen, Kane, & Crookes, 1999; Jaeger, 1991; Kaftandjieva, 2010). Based on these estimates, Jaeger recommends adopting the stricter criteria of $SE_c \leq \frac{1}{4}$SEM for evaluating the precision of the standard setting, while Cohen, Kane and Crookes (1999) suggest that such a strict level may be unrealistic in many situations and adopt the more liberal criterion of $SE_c \leq \frac{1}{2}$SEM. Kaftandjieva (2010), in her comprehensive review of standard-setting methods in relation to criterion-referenced tests, recommends adopting $SE_c \leq \frac{1}{3}$SEM as an acceptable

compromise between maximizing the precision of cutscore estimates while striking a balance with the practical realities of standard setting.

Test statistics for the live administrations of the EIKEN Grade 1 and Grade Pre-1 First Stage test forms used in this standard-setting study were discussed earlier in Chapter 4 (see Table 4.1 in Chapter 4). As noted in Chapter 4, the test statistics provided for each form were calculated on the unweighted response data for the combined selected-response components of Vocabulary, Reading and Listening. In order to evaluate the magnitude of the standard error of the cutscore using the various criteria that have been suggested, it will thus be necessary to look at the SE_c for the combined cutscore for the Vocabulary, Reading and Listening components and to compare the SE_c for the combined cutscore of the selected response components to the SEM calculated across the same three components, using the same raw-score scale. The cutscores for each component, along with the aggregated total cutscore for the First Stage test for each grade, were shown earlier in Table 5.9. The cutscores for the other selected-response components have been derived in the same way as for reading; the probability judgments for each participant across all items in a component are averaged to arrive at a cutscore estimate for that component for each participant, and the mean of the participants individual cutscore estimates is treated as the cutscore for that component. The SE_c for the combined cutscore aggregated across cutscores for the three selected-response components is shown in Table 5.13 The aggregated cutscore estimate for the three components for each participant is shown, followed by the mean of these estimates, the standard deviation, and the SE_c. Table 5.14 then shows the level of SEM at each of the three criterion evaluation levels recommended by Jaeger (1991), Cohen, Kane, & Crookes (1999), and Kaftandjieva (2010) recommended above, with an indication of whether the SE_c meets the criteria (YES), or not (NO). As can be seen from Table 5.14, for Grade 1 the SE_c meets the criteria set by Kaftandjieva (2010) and Cohen, Kane, and Crookes (1999), but falls just short of the strictest criterion set by Jaeger (1991). For Grade Pre-1, the SE_c for the aggregated cutscore estimate for the Vocabulary, Reading, and Listening components meets all three of the criteria, including the strictest level of $SE_c \leq \frac{1}{4}$SEM.

Table 5.13 Aggregated cutscores for Vocabulary, Reading, and Listening

Rater	Grade 1	Grade Pre-1
R1	38.2	40.4
R2	34.3	38.1
R3	44.2	45.3
R4	38.2	38.3
R5	36.3	35.7
R6	46.4	44.7
R7	41.4	40.7
R8	38.9	39.4
R9	37.3	34.9
R10	46.8	47.1
R11	40.7	41.3
R12	47.4	44.3
R13	39.8	42.2
Mean	40.8	40.9
SD	4.2	3.7
SE_c	1.2	1.0

Table 5.14 Evaluation criteria for precision: SE_c compared to SEM

	Grade 1	≤ criteria	Grade Pre-1	≤ criteria
SEM	3.6		3.8	
$\frac{1}{4}$SEM	0.9	NO	1.0	YES
$\frac{1}{3}$SEM	1.2	YES	1.3	YES
$\frac{1}{2}$SEM	1.8	YES	1.9	YES

A multi-facet Rasch model (MFRM) analysis of the modified Angoff standard-setting judgment data for Reading was conducted using FACETS

(Linacre, 2014). MFRM has been widely applied to the analysis of performance assessments in language testing, as it provides a means of estimating measures for the various facets or variables affecting performance assessment within a common frame of reference and on a common metric measured in units referred to as logits (for overviews of MFRM and its use in language assessment research, see Bachman, 2004; Eckes, 2011; McNamara, 1996). FACETS not only provides a means of estimating the relative severity of raters, but also accounts for this severity in the final estimates of difficulty for tasks and items and the estimates of ability for test takers, with these estimates converted to the metric of the rating scale employed in the form of *fair averages* estimated for each of the relevant facets (Linacre, 2014). MFRM further provides estimates of the internal consistency of rater judgments, allowing for an evaluation of the quality of those judgments (Linacre, 2014). MFRM has been applied to standard setting (e.g. Engelhard, 2000; Engelhard & Stone, 1998; Lumley, Lynch, & McNamarra, 1994) and has been applied to standard setting in relation to linking exams to the CEFR by O'Sullivan (2008) and Eckes (2009). For this study, a two-facet analysis, with raters and test items as facets was conducted. In the case of standard setting, the response data consists of the raters' probability judgments for the items. This data was converted to a rating scale with possible ratings of 0-10 (in which a 10 percent probability judgment is treated as a rating of 1, a 20 percent judgment as 2, etc).

Figure 5.2 and Figure 5.3 show the facet maps for Grade 1 and Grade Pre-1 respectively. The facet map shows the position of the raters and items on a common logit scale, with the logit scale shown in the Measure column. The position of each of the raters is shown in the Raters column, and the items rated in the Items column. Note that as there are no test takers in this data set, no facet was positively oriented, meaning that the interpretation of the rater severity and item difficulty measures accords with the normal default position in a FACETS analysis: a higher logit measure means more severity for raters and more difficulty for items. For example, for Grade 1, R8 is the most severe of the raters and Q19 is the most difficult item. Referring to the Rater Measurement report in Table 5.15 for Grade 1, R8 gave the lowest average rating of the proportion of 100 minimally

competent C1-level candidates who would correctly answer items on the G1 test, and Q19 had the lowest average estimate of the proportion of C1-level candidates who would answer each item correctly.

In terms of severity, the majority of raters fall within ± 1 logits. This range of severity estimates has been suggested as a tolerable level of rater severity variability in relation to performance assessments (Van Moere, 2006; Taylor & Galaczi, 2012). Nonetheless four raters fall outside this range, with two being more severe and two being more lenient. As already noted, FACETS takes account of the severity of raters and the difficulty of items in allocating final measures, with these then converted into the same metric as the rating scale in the form of *fair average* estimates (Linacre, 2015). The mean of the fair averages for the raters is 5.67 compared to 5.66 for the mean of unadjusted observed ratings. Bearing in mind that the rating data represented 10 percent increments in the estimation of the proportion of C1-level candidates who would correctly answer the items, the fair average, adjusted for the differential difficulty of items and severity of raters, represents a cutscore of 58.7 percent compared to the unadjusted cutscore of 58.6 percent.

For Grade Pre-1, the range of severity is much narrower, with only 3 raters falling just outside the ±1 logit range. The mean of fair averages for rater estimates is 5.99, producing a cutscore of 59.9 percent compared to the mean of observed estimates for raters of 6.00, which produces a cutscore of 60 percent.

In terms of internal consistency, the infit and outfit Mean Square statistics give an estimate of the degree of fit of the observed responses to the responses predicted by the Rasch model. These statistics are used to estimate the degree of consistency of the rater judgments. The infit mean square is usually reported rather than the outfit mean square, as it focuses on "the degree of fit in the most typical responses in the matrix" and is thus less susceptible to a few unpredictable outlying responses than the outfit mean square (McNamara, 1996, p. 172). Englehard and Stone (1998) describe the interpretation of fit statistics in evaluating the quality of judges' ratings for standard setting. A higher fit statistic represents *misfit*, or unpredictability in the data. Levels of misfit greater than 1.5 would be an indication that those raters are not rating the items in the same

relative order of difficulty (Engelhard & Stone, 1998, p. 185). Misfit is usually considered more problematic than *overfit*, or low infit mean squares, which represent response patterns that are *too* predictable (Myford & Wolfe, 2004). Nonetheless, overfit can be an indication of other problematic rater behaviours, such as the *halo effect* or *central tendency* (Engelhard & Stone, 1998). This study uses the commonly employed criteria of 0.5–1.5 as the acceptable range of the infit mean square for examining rater consistency (e.g. Lunz, Wright, & Linacre, 1990; Engelhard & Stone, 1998; O'Sullivan, 2008; Eckes, 2011). It is worth noting that Myford and Wolfe (2004) suggest that fit statistics in the range of 1.5 and 2.0 may still represent useful rater response in many low-stakes situations, and Taylor and Galaczi (2012) describe using this range for identifying problematic raters in training and standardization exercises.

Referring to the Rater Measurement Report for Grade 1, only one rater showed misfit: R12 with an infit mean square of 3.57. For Grade Pre-1, two raters show levels of misfit marginally above the 1.5 criterion: R12 with 1.61 and R3 with 1.58. Although outside the criteria, both of these figures fall within the extended 1.5-2.0 range mentioned above. A follow-up analysis was conducted to investigate the impact of dropping these raters. For Grade 1, dropping R12 from the analysis results in a fair average of 5.82, or a cutscore of 58.2 percent. For Grade Pre-1, dropping R12 and R13 derives a fair average of 5.83, or a cutscore estimate of 58.3 percent.

```
G1 Reading Angoff Method FACETS Analysis 24/06/2016 10:33:21
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,S) Yardstick (columns lines low high extreme)= 0,5,-2,3,End
+-------------------------------------+
|Measr|-Raters   |-Items         |Scale|
|-----+----------+---------------+-----|
|   3 +          +               + (9) |
|     |          |               |     |
|     |          |               |     |
|     |          |               |     |
|     |          |               |     |
|   2 + R8       +               + --- |
|     |          |               |     |
|     | R5       | Q39           |     |
|     |          | Q37           |     |
|     |          |               |  7  |
|   1 + R2       + Q34  Q35      +     |
|     | R4   R9  | Q40  Q41      |     |
|     |          | Q36  Q38      |     |
|     |          | Q26           | --- |
|     | R1       |               |     |
*   0 *          * Q28  Q29  Q32 *     *
|     |          | Q27  Q31  Q33 |  6  |
|     | R11  R13 |               |     |
|     |          |               |     |
|     | R12  R7  |               |     |
|  -1 + R10      + Q30           + --- |
|     |          |               |     |
|     | R3       |               |     |
|     | R6       |               |  5  |
|     |          |               |     |
|  -2 +          +               + (3) |
|-----+----------+---------------+-----|
|Measr|-Raters   |-Items         |Scale|
+-------------------------------------+
```

Figure 5.2 Facet Map for Grade 1 Reading

```
GP1 Reading Angoff Method FACETS Analysis 24/06/2016 11:10:07
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1A,2A,S) Yardstick (columns lines low high extreme)= 0,5,-2,2,End
+-----------------------------------------+
|Measr|-Raters   |-Items               |Scale|
|-----+----------+--------------------+-----|
|   2 +          +                    + (9) |
|     |          |                    |     |
|     |          |                    |     |
|     |          |                    |  7  |
|     | R8       |                    |     |
|   1 + R5   R9  + Q37  Q39           +     |
|     |          |                    | --- |
|     | R2   R4  |                    |     |
|     | R11      | Q33                |     |
|     |          |                    |  6  |
*   0 * R1       * Q32  Q38  Q40      *     *
|     | R13  R7  | Q27  Q28  Q34  Q41 |     |
|     |          | Q31                |     |
|     |          | Q26  Q30  Q35  Q36 | --- |
|     | R6       | Q29                |     |
|  -1 + R12      +                    +     |
|     | R10  R3  |                    |  5  |
|     |          |                    |     |
|     |          |                    |     |
|     |          |                    | --- |
|  -2 +          +                    + (3) |
|-----+----------+--------------------+-----|
|Measr|-Raters   |-Items              |Scale|
+-----------------------------------------+
```

Figure 5.3 Facet Map for Grade Pre-1 Reading

Table 5.15 Rater Measurement Report: G1 Reading for Modificed Angoff

G1 Reading Angoff Method FACETS Analysis 24/06/2016 10:33:21
Table 7.1.1 Raters Measurement Report (arranged by lmN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
69	16	4.31	4.31	2.03	.29	.79	-.5	.78	-.6	1.51	.63	17.7	14.4	8	R8
74	16	4.63	4.65	1.63	.28	1.33	1.0	1.36	1.0	.44	.63	15.6	17.8	5	R5
83	16	5.19	5.22	.91	.29	.31	-2.6	.31	-2.6	1.65	.74	26.6	23.2	2	R2
84	16	5.25	5.28	.83	.29	1.41	1.1	1.39	1.1	.59	.23	22.9	23.6	4	R4
84	16	5.25	5.28	.83	.29	.55	-1.4	.56	-1.3	1.40	.71	28.6	23.6	9	R9
91	16	5.69	5.70	.26	.29	.26	-2.9	.26	-2.9	1.75	.69	28.6	26.0	1	R1
99	16	6.19	6.19	-.42	.30	.24	-3.1	.24	-3.1	1.76	.68	23.4	26.1	11	R11
99	16	6.19	6.19	-.42	.30	1.11	.4	1.11	.4	.89	.64	27.1	26.1	13	R13
103	16	6.44	6.44	-.76	.30	.59	-1.3	.58	-1.3	1.46	.76	25.0	25.1	7	R7
104	16	6.50	6.51	-.85	.30	3.57	4.7	3.47	4.6	-1.73	.31	17.2	24.7	12	R12
105	16	6.56	6.57	-.94	.30	.66	-1.0	.64	-1.0	1.34	.25	23.4	24.3	10	R10
110	16	6.88	6.89	-1.40	.31	1.07	.3	1.08	.3	.97	.43	17.7	21.6	3	R3
113	16	7.06	7.08	-1.69	.32	.52	-1.5	.51	-1.6	1.51	.65	17.7	19.5	6	R6
93.7	16.0	5.86	5.87	.00	.29	.95	-.5	.95	-.6		.57			Mean (Count: 13)	
13.4	.0	.84	.83	1.12	.01	.84	2.1	.82	2.0		.18			S.D. (Population)	
13.9	.0	.87	.87	1.17	.01	.88	2.1	.85	2.1		.19			S.D. (Sample)	

Model, Populn: RMSE .29 Adj (True) S.D. 1.08 Separation 3.67 Strata 5.23 Reliability (not inter-rater) .93
Model, Sample: RMSE .29 Adj (True) S.D. 1.13 Separation 3.83 Strata 5.45 Reliability (not inter-rater) .94
Model, Fixed (all same) chi-square: 187.5 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.3 d.f.: 11 significance (probability): .42
Inter-Rater agreement opportunities: 1248 Exact agreements: 280 = 22.4% Expected: 284.2 = 22.8%

Table 5.16 Rater Measurement Report: GP1 Reading

```
GP1 Reading Angoff Method FACETS Analysis 24/06/2016 11:10:07
Table 7.1.1  Raters Measurement Report  (arranged by lmN).

+-------------------------------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Corr.  | Exact Agree.  |                    |
| Score  Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtBisi| Obs %  Exp %  | Nu Raters          |
|------------------------------+--------------+----------------------+------+-------+---------------+--------------------|
|    81     16   5.06   5.06 |  1.17   .29 | 1.16   .5  1.13   .4 |  .95 |  .42  | 18.8   22.2   |  8 R8              |
|    83     16   5.19   5.18 |  1.01   .28 | 1.36  1.1  1.36  1.1 |  .51 |  .55  | 19.3   23.5   |  5 R5              |
|    84     16   5.25   5.25 |   .93   .28 | 1.18   .6  1.14   .5 |  .83 |  .23  | 20.3   24.1   |  9 R9              |
|    87     16   5.44   5.44 |   .70   .28 | 1.07   .3  1.06   .2 | 1.01 |  .51  | 27.1   25.6   |  2 R2              |
|    89     16   5.56   5.57 |   .55   .28 | 1.15   .5  1.15   .5 |  .85 |  .55  | 24.0   26.5   |  4 R4              |
|    92     16   5.75   5.76 |   .32   .28 |  .48 -1.7  .48 -1.7 | 1.52 |  .21  | 30.2   27.4   | 11 R11             |
|    97     16   6.06   6.07 |  -.07   .28 |  .71  -.8  .70  -.8 | 1.29 |  .09  | 28.1   27.9   |  1 R1              |
|    98     16   6.13   6.13 |  -.15   .28 |  .66  -.9  .66  -.9 | 1.35 |  .79  | 26.6   27.8   |  7 R7              |
|    99     16   6.19   6.19 |  -.23   .28 |  .51 -1.5  .52 -1.5 | 1.44 |  .77  | 30.7   27.7   | 13 R13             |
|   105     16   6.56   6.56 |  -.71   .29 |  .49 -1.6  .49 -1.6 | 1.52 |  .33  | 26.0   25.9   |  6 R6              |
|   109     16   6.81   6.80 | -1.04   .29 | 1.61  1.5 1.63  1.6 |  .37 |  .40  | 17.2   23.7   | 12 R12             |
|   111     16   6.94   6.92 | -1.20   .29 |  .55 -1.4  .55 -1.4 | 1.46 |  .35  | 20.8   22.4   | 10 R10             |
|   112     16   7.00   6.98 | -1.29   .29 | 1.58  1.5 1.58  1.5 |  .37 |  .15  | 17.2   21.7   |  3 R3              |
|------------------------------+--------------+----------------------+------+-------+---------------+--------------------|
|  95.9   16.0   6.00   5.99 |   .00   .29 |  .96  -.1  .96  -.2 |      |  .41  |               | Mean (Count: 13)   |
|  10.5    .0    .66    .65 |   .83   .00 |  .40  1.2  .40  1.2 |      |  .21  |               | S.D. (Population)  |
|  10.9    .0    .68    .68 |   .86   .00 |  .42  1.3  .42  1.2 |      |  .22  |               | S.D. (Sample)      |
+-------------------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .29  Adj (True) S.D. .78  Separation 2.72  Strata 3.96  Reliability (not inter-rater) .88
Model, Sample: RMSE .29  Adj (True) S.D. .81  Separation 2.85  Strata 4.13  Reliability (not inter-rater) .89
Model, Fixed (all same) chi-square:  107.4  d.f.: 12  significance (probability): .00
Model, Random (normal) chi-square:  10.9  d.f.: 11  significance (probability): .45
Inter-Rater agreement opportunities: 1248  Exact agreements: 294 =  23.6%  Expected:  313.2 =  25.1%
```

5.3.5.4 External validity

Cizek & Bunch (2007. p. 60) include "agreement of cutscores across replications using other standard-setting methods" as one potential source of external validity evidence, echoing Kane's (2001b) call for replication with different methods as a powerful source of validity evidence. Two standard-setting methods were used in this study, which provides us with the opportunity to examine the consistency of the cutscores.

Referring back to Table 5.6 and Table 5.7, it is possible to compare the cutscores for the Basket and Angoff methods. As described in Section 5.3.5.1, to derive the minimum score necessary for classification as C1 for Grade 1 and B2 for Grade Pre-1, it is necessary to add one raw score point to the cutscores for the Basket method. This would give minimum required scores of 11 for both grades (the Basket method rounded to the nearest whole score point plus 1). Both cutscores are thus slightly higher than the equivalent Angoff cutscores but fall exactly on the 70 percent passing score (equal to 11 raw score points on both 16-item reading components). Cutscores from both methods, then, show consistency in the interpretation that test takers who score at or above the passing score will have demonstrated sufficient ability to be considered at the C1 level of reading for Grade 1 and at the B2 level of reading for Grade Pre-1, and thus provide some support for the reasonableness of the decisions in relation to RQ3, as described in section 5.2.2.

Cizek & bunch (2007) and Tannenbaum and Wiley (2004, 2008) use the standard deviation of the cutscores as an indication of the variability in rater judgments, with Tannenbaum and Wiley treating a reduction in the standard deviation across rounds as an indication of greater consensus amongst raters. In the present study, there is a noticeable reduction in the variability of cutscores for the Angoff method compared to the Basket method, with the standard deviation for Grade 1 dropping from 3.8 to 1.4 and for Grade Pre-1 from 4.3 to 1.1. The reduction in variability of ratings cannot be ascribed clearly to the choice of standard-setting method, as the methods were not counter-balanced, meaning the method facet is confounded with the order in which the methods were applied. Nonetheless, the greater consensus as indicated by the reduction in standard

deviation for the Angoff method provides some support for the approach taken of adopting the Basket method as a primer to help participants first conceptualize the items in terms of CEFR levels before attempting the more conceptually demanding, but also statistically robust, judgment task in the Angoff method.

5.4 Standard setting panel 2

5.4.1 Introduction

A second standard-setting panel focused on Grades 2, Pre-2, and 3. The rationale for splitting the standard setting into several panels has already been discussed in detail in section 5.2.1.1. Separating the panels also provided the opportunity to apply lessons learned from the standard-setting event for Grades 1 and Pre-1. As with Grades 1 and Pre-1, the standard setting for Reading was undertaken as part of a single standard-setting event combining all components in the First Stage tests for these grades.

5.4.2 Participants

The criteria for participants was modified from Panel 1 to take account of the typical uses and interpretations, the TLU domain, and the typical test takers for Grades, 3, Pre-2 and 2. As with Panel 1, it was decided to prioritize knowledge of the grades of the EIKEN tests that would be the focus of the standard setting and knowledge and experience of the educational sectors in which the tests were mostly widely used over knowledge of the CEFR. Participants were required to meet the following criteria:

- 3 years teaching English at junior and/or senior high schools in Japan
- English ability sufficient to deal with the English test items. As the tests were posited to range from A1 to B1, the minimum required level of proficiency was set at equivalent to B2 (as it was anticipated that the participants would be local educators and non-native speakers of English, all CEFR-related materials were prepared in both English and Japanese)
- Knowledge of and experience using EIKEN tests
- Ability to take part in all stages of the workshop

Table 5.17 Teaching experience of Panel 2 participants in years

Educational Sector	Elementary	Junior HS	Senior HS	University / Jr College	Technical College
R1			6	8	4
R2		5	20		
R3		33			
R4		25	5		
R5		20	20		
R6		5	25		1
R7		22	22	12	
R8		29	1		
R9		13	13		
R10		28			
R11		29		1	
R12	1	27	2		
R13		1	6	14	

The final panel consisted of 13 judges, all of whom had Japanese as their first language. All participants had direct experience of the relevant EIKEN grades by serving as EIKEN speaking test examiners for official administrations and/or sitting on editorial review panels which meet several times a year to review and critique test content for use in future live examinations. Table 5.17 gives a breakdown of the professional experience of the judges in relation to the various educational sectors in Japan. As can be seen, all of the panel participants met the minimum criteria in terms of professional experience in the key secondary school sector. Of the participants, 12 had experience teaching in junior high schools, while 10 had experience teaching in high schools. Only one candidate had experience in only one of these sectors, R1 who had taught at the high school level but not the junior high school level.

5.4.3 Instruments

The same range of materials used for training and data collection for Panel 1 were adapted by the author for use with Panel 2 (see Table 5.3 in Section 5.3.3). The rating forms were amended to allow for two rounds of judgments to be recorded

for each method (see Section 5.4.3 below for details on the changes to procedure). The forms for the Basket method were further modified to allow for judgments of below A1. While this had not been necessary for the Grade 1 and Pre-1 tests, as the tests used by Panel 2 would be dealing with A1, A2, and B1 it was thought necessary to add this category, particularly for Grade 3.

The same range of sample reading and listening items from European examination boards provided by the Council of Europe, along with their content specifications and justification for level placement, were used for training with items calibrated to the CEFR during the event. The items selected had proved robust in illustrating the connection between CEFR descriptors, content specification and level estimation. The focus of discussion and the order in which the items were presented was modified to take account of the different level focus for Panel 2. The questionnaire was translated into Japanese and questions adapted to take account of the different test content for Panel 2.

As with Panel 1, the self-study preparation booklet played a crucial role in providing familiarization with the CEFR prior to the actual event. Some modifications to the content of the booklet were made to take account of the different level focus for Panel 2, and all content was translated into Japanese. The CEFR level descriptors were provided in both their English and Japanese translations, using the Japanese translation by Yoshijima and Ohashi (吉島 大橋 2004) published by Asahi Publishing, which is listed on the Council of Europe website in the list of translated versions of the CEFR.[9]

As can be seen in Table 5.18, the range of familiarization tasks was also modified to take account of feedback from Panel 1 in order to focus attention on those aspects which participants in Panel 1 had indicated were most salient for familiarization. The questions in tasks 4 and 6 in particular were designed to provide a basis for the discussion of the CEFR at the beginning of the event, and to elicit opinions regarding salient key words and concepts

[9] http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

Table 5.18 CEFR familiarization tasks adapted for use in Panel 2 booklet

Tasks	Focus	Description of activity
Tasks 1 & 2	Global Scale	Reflection, using scale to consider level of own learners, summarizing significant level features for A1, A2, B1
Tasks 3	Overall Reading scale	Reordering jumbled descriptors from Overall Reading Scale
Tasks 4	Illustrative scales for Reading	Examine illustrative scales for listening: • Describe which scales are most relevant for listening in relation to school textbooks and EIKEN Listening items.. • Summarize key features across Global, Overall Listening and separate scales for Listening.
Task 5	Overall Listening	Reordering jumbled descriptors from Overall Listening Scale
Task 6	Illustrative scales for Listening	Examine illustrative scales for listening: • Describe which scales are most relevant for listening in relation to school textbooks and EIKEN Listening items.. • Summarize key features across Global, Overall Listening and separate scales for Listening. • Describe similarities and difference between level descriptions for Reading and Listening

5.4.4 Procedure

The procedure followed essentially the same pattern and schedule as Panel 1, with the first day beginning with an overall discussion of the CEFR. As with Panel 1, the Basket method was used as an introduction to the process of standard setting and conceptualizing the EIKEN items in terms of the CEFR level descriptors. The, procedure then moved on to the more conceptually demanding Angoff method.

The schedule for the standard setting was amended slightly to take account of the different test content and the lessons learned from Panel 1 in terms of the time required. A review of availability had demonstrated that the event would only be able to span a maximum of two days for Panel 2 to cover standard setting for all three grades. However, Panel 1 had demonstrated that the self-study booklet gave people a good foundation for the training at the event. It was found that after the initial round with the Basket method, including answering the items

under the same conditions as test takers before standard setting, the follow-up rounds employing the Angoff method progressed generally smoothly and very rapidly. Also, the First Stage Tests of Grades 3, Pre-2, and 2, as described in Chapter 1, do not include a constructed response writing component but instead include selected response reordering tasks as semi-direct tests of writing. It was thus not necessary to include a third day for training with the Paper Selection method as was necessary for Panel 1.

The most significant change to the procedure was to require participants to record two rounds of judgments for both methods, with the second round judgments being made after the provision of empirical feedback following the first round of Angoff judgments. Panel 1 participants had also been allowed to make changes to their ratings following the provision of feedback, but this was done by overwriting their original judgments, and only the final rating was collected. As a measure of internal validity, it was decided to explicitly collect those round 1 and round 2 judgments on the rating forms. Given that time was still of a premium in order to cover all components for all three grades in two days, no discussion was included after the first round, and as with Panel 1, participants were told that they were able to change their ratings for round 2 if they desired, but were not required to. They were also cautioned on the interpretation of *p* values derived from a heterogeneous sample of test takers in a live administration compared to the "100 test takers minimally competent" at a particular CEFR level which formed the basis of their standard-setting judgments.

5.4.5 Results

5.4.5.1 Standard-setting results

Tables 5.19, 5.20, and 5.21 provide an overview of the cutscores for Reading for Grades 2, Pre-2, and 3 respectively. The tables include both Round 1 and Round 2 judgments. The cutscores are presented both as the percentage correct and the raw score (number of items) which a test taker would need to score to be classed as minimally competent at the appropriate CEFR level (i.e. B1 for Grade 2, A2 for Grade Pre-2, and A1 for Grade 3). As with Panel 1, a decision was made a priori

to use the final (round 2) cutscores derived from the Angoff method as the basis for standard setting in order to evaluate the claims regarding the link between each EIKEN grade and a particular CEFR level. To place the Reading test in context, Table 5.22 recaps the breakdown of the number of items and the item weighting and contribution of each section to the overall score. For Grades 2, Pre-2, and 3 all items are equally weighted. As with Panel 1, the standard setting for Reading was carried out as part of an integrated event combining training and standard setting for all components of the First Stage tests. To help interpret the reasonableness of the Reading cutscores and effectiveness of the standard setting for Reading, as with Panel 1, an overview of cutscores set for each component and aggregated for an overall First Stage cutscore for each grade is provided in Table 5.23. The passing score for Grades 2, Pre-2, and 3 is set at 60 percent for the First Stage test. Although the tests are based on a compensatory, rather than conjunctive model, for the purposes of evaluating the reasonableness of the standard setting for the Reading component, it is also necessary to identify the raw score equivalent of 60% for Reading, which for Grade 2 and Pre-2 will be 12 items and for Grade 3 will be 9 items.

The cutscore for Reading for Grade 2 falls just below the overall required passing score, with 58.3 percent or the equivalent raw score of 11.7 items. For both Grade Pre-2 and Grade 3, the Reading cutscores are slightly higher than the required passing score, with Grade 3 being the highest at 68.9 percent, or 10.3 items correct, which compares with 9 items required to be answered correctly to reach the passing score level of 60 percent. The trend is replicated in the overall cutscores for the First Stage shown in Table 5.23, with Grade 2 falling just below the passing score. Although the aggregated cutscores for Grades Pre-2 and 3 remain over the passing-score level, aggregated over all components the difference is reduced, becoming 62.5 percent for Grade Pre-2 and 66.5 percent for Grade 3.

Table 5.19 Grade 2 Reading Cutscores for Reading

	Round 1	Round 1	Round 2	Round 2
	Percent	Score	Percent	Score
R1	60.0	12.0	59.0	11.8
R2	63.0	12.6	61.5	12.3
R3	60.5	12.1	60.0	12.0
R4	58.0	11.6	57.5	11.5
R5	65.5	13.1	68.0	13.6
R6	63.0	12.6	61.0	12.2
R7	64.5	12.9	64.0	12.8
R8	49.5	9.9	50.5	10.1
R9	53.0	10.6	56.0	11.2
R10	45.0	9.0	46.5	9.3
R11	72.5	14.5	64.5	12.9
R12	46.0	9.2	45.0	9.0
R13	64.5	12.9	64.5	12.9
mean	58.8	11.8	58.3	11.7
SD	8.2	1.6	7.1	1.4
SE c	2.3	0.5	2.0	0.4

Table 5.20 Grade Pre-2 cutscores for Reading

	Round 1	Round 1	Round 2	Round 2
	Percent	Score	Percent	Score
R1	55.0	11.0	60.0	12.0
R2	62.0	12.4	62.0	12.4
R3	62.0	12.4	65.5	13.1
R4	61.5	12.3	63.5	12.7
R5	77.5	15.5	77.5	15.5
R6	60.0	12.0	62.0	12.4
R7	69.5	13.9	72.5	14.5
R8	46.5	9.3	51.5	10.3
R9	62.0	12.4	65.5	13.1
R10	74.0	14.8	73.5	14.7
R11	74.5	14.9	69.5	13.9
R12	47.0	9.4	48.5	9.7
R13	73.0	14.6	73.0	14.6
mean	63.4	12.7	65.0	13.0
SD	10.0	2.0	8.5	1.7
SE c	2.8	0.6	2.4	0.5

Table 5.21 Grade 3 cutscores for Reading

	Round 1	Round 1	Round 2	Round 2
	Percent	Score	Percent	Score
R1	70.0	10.5	68.7	10.3
R2	77.3	11.6	74.7	11.2
R3	74.7	11.2	72.7	10.9
R4	69.3	10.4	68.0	10.2
R5	80.0	12.0	80.7	12.1
R6	72.7	10.9	70.7	10.6
R7	77.3	11.6	74.7	11.2
R8	57.3	8.6	58.7	8.8
R9	70.7	10.6	68.7	10.3
R10	64.7	9.7	64.0	9.6
R11	74.0	11.1	70.0	10.5
R12	44.0	6.6	48.0	7.2
R13	78.0	11.7	76.7	11.5
mean	70.0	10.5	68.9	10.3
SD	9.9	1.5	8.4	1.3
SE c	2.8	0.4	2.3	0.4

Table 5.22 Structure and scoring of Grade 2, Pre-2, and 3 First Stage

Grade 2	V & G	Reading	Listening	Writing	Total
Weighted score	20	20	30	5	75
Raw Score	20	20	30	5	75
No. of items	20	20	30	5	75
Percent	27%	27%	40%	7%	100%
Grade Pre-2	**V & G**	**Reading**	**Listening**	**Writing**	**Total**
Weighted score	20	20	30	5	75
Raw Score	20	20	30	5	75
No. of items	20	20	30	5	75
Percent	27%	27%	40%	7%	100%
Grade 3	**V & G**	**Reading**	**Listening**	**Writing**	**Total**
Weighted score	15	15	25	5	60
Raw Score	15	15	25	5	60
No. of items	15	15	25	5	60
Percent	25%	25%	42%	8%	100%

Table 5.23 Cutscores for all components for Grades 2, Pre-2, and 3

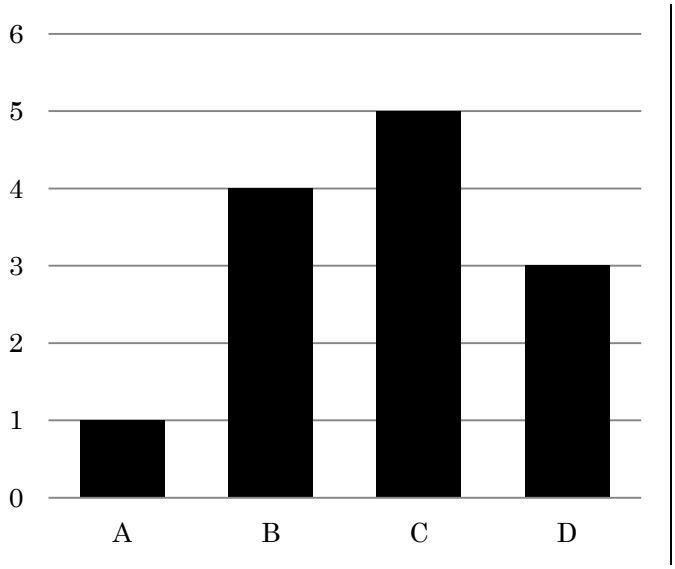
	Round 1		Round 2	
Grade 2	**Percent**	**Raw Score**	**Percent**	**Raw Score**
Vocabulary	55.6	11.1	55.3	11.1
Reading	58.8	11.8	58.3	11.7
Listening	64.6	19.4	64.3	19.3
Writing	53.2	2.7	54.2	2.7
First Stage Overall	59.9	44.9	59.6	44.7
Grade Pre-2	**Percent**	**Raw**	**Percent**	**Raw**
Vocabulary	65.4	13.1	64.0	12.8
Reading	63.4	12.7	65.0	13.0
Listening	63.2	18.9	61.2	18.4
Writing	59.2	3.0	54.6	2.7
First Stage Overall	63.6	47.7	62.5	46.9
Grade 3	**Percent**	**Raw**	**Percent**	**Raw**
Vocabulary	65.7	9.9	65.2	9.8
Reading	70.0	10.5	68.9	10.3
Listening	64.7	19.4	66.5	20.0
Writing	58.9	2.9	62.6	3.1
First Stage Overall	65.7	42.7	66.5	43.2

5.4.5.2 Procedural validity

As with Panel 1, participants were asked about their knowledge and experience of the CEFR and standard setting (Figures 5.19 and 5.20 respectively[10]), The same general trends are evident, with the majority of participants describing themselves as unfamiliar with the contents of the CEFR (8 participants out of 13 choosing options C or D for Panel 2,). The majority of participants in Panel 2 were not familiar with the concept of standard setting at all (8 out of 13), with most of

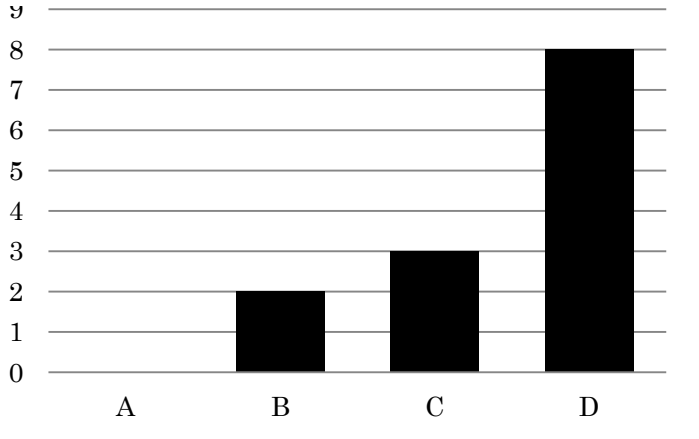
[10] Some confusion over the question on experience with standard setting was noted with Panel 1. The question was amended to take account of this.

those who had heard of the concept being unfamiliar with the methods used in this study (3 out of 8 participants).



A. I had read the CEFR and was familiar with its aims and contents, including the Common Reference Levels.
B. I was familiar with the aims of the CEFR, but had not studied it in detail.
C. I had heard of the CEFR but was not familiar with its aims or contents.
D. I had not heard of the CEFR.

Figure 5.4 Knowledge of CEFR for Panel 2



A. I had had experience acting as a judge/rater on standard setting panels.
B. I was familiar with the concept of standard setting, and had heard of one or more of the methods which were used in the EIKEN project.
C. I was familiar with the concept of standard setting, but had not heard of any of the methods used in the EIKEN project.
D. I was not familiar with the concept of standard setting.

Figure 5.5 Knowledge of standard setting for Panel 2

Table 5.24 lists the questions on the procedural validity questionnaire for Panel 2, with the mean score on the 4-point likert-type scale[11]. The questions parallel those used for Panel 1, but were translated into Japanese for use with Panel 2 (Table 5.24 shows the English translations. Questions 12 and 13 differ slightly from those asked of Panel 1, as Grades 2, Pre-2 and 3 contain no constructed response writing component. Questions 12 and 13 address the approach taken to standard setting with the items used to test writing through indirect, selected-response item types (see Chapter 1 for an explanation of the item types used). Appendix L, figures L1 to L15, gives the individual response breakdowns for each question.

The trend once again mirrors the pattern seen in Panel 1, with the majority of participants responding positively regarding the procedures, time, facilities, and their confidence in their ability to carry out the judgment tasks after training. Importantly the role of the preparation booklet in building an understanding of the goals of the project and familiarity with the CEFR and Common Reference Levels was endorsed by a majority of participants.

Several differences to the pattern of responses by Panel 1 judges are, however, worth noting. A sizeable number of participants in Panel 2 felt the time for discussion of the CEFR was too short (6 participants for Q4), and that they were not able to participate fully in that discussion (4 participants for Q5). The largest number of negative responses was related to Q12, for which 9 participants did not agree that the approach to standard setting for the indirect writing items was clear.

As already noted, Grades 2, Pre-2, and 3, do not have a constructed-response Writing component. The rationale for the inclusion of items which target aspects of writing through indirect, selected-response sentence construction tasks was discussed in Chapter 1. However, it was recognized from the outset that these items would prove problematic in terms of standard setting and the CEFR. As the items did not elicit stretches of the test takers' actual writing,

[11] Note the order of the responses in the Japanese version: option A was the equivalent of Strongly Agree, and option D Strongly Disagree. For the analysis of averages, the options were assigned the same values as for Panel 1, with Strongly Disagree being equal to 1 and Strongly Agree equal to 4.

it was not possible to easily compare them to the CEFR level descriptors from the Illustrative Scales for writing, or the examples of writing performance at CEFR levels supplied from the CEFR. This was discussed explicitly with the panel, and the limitations of the indirect form of testing writing were discussed.

As the items all involve reordering of jumbled sentences from within a short passage or printed dialogue, the tasks were treated as integrated tasks combining aspects of reading ability with elements of syntactic knowledge necessary for composition. At Grades Pre-2 and 2, there is also an element of cohesion, as the sentences to be reordered are embedded within longer sections of text. Panelists were instructed to refer to both the scales for reading relevant to the reading text employed in each item, but also to make reference to the Writing Assessment Criteria from the Manual (Council of Europe, 2009), particularly the aspects of Range, Coherence and Accuracy described in the scales. Panelists made the same probability judgments as for all other selected-response items in the tests. The panelists proved able to carry out the judgment task effectively in relation to determining the probability of test takers minimally competent at the appropriate CEFR level successfully completing the task. However, the negative responses to Q12 reflect the discussion that took place during the event and the difficulty of relating indirect tests of writing to the can-do based level descriptors for performance in real-world language use situations used in the CEFR,

Table 5.24 Procedural validity questionnaire for Panel 2

	Questions	No. of Raters	Mean
Q1	The preparation booklet gave me a clear understanding of the purpose of the project.	13	3.0
Q2	The explanations and tasks in the preparation booklet helped me understand the structure of the CEFR and the Common Reference Levels.	13	3.2
Q3	The group discussion of the CEFR at the start of the workshop aided my understanding of the CEFR and the Common Reference Levels.	13	3.3
Q4	The time provided for the discussion was adequate.	13	2.6
Q5	There was an equal opportunity for everyone to contribute his/her ideas during the discussion.	13	3.0
Q6	The training tasks with the items supplied by the Council of Europe were useful.	13	3.2
Q7	The time provided for training with the Council of Europe items was adequate.	13	2.7
Q8	The explanation of the Basket Method was adequate and I felt able to undertake the rating tasks for the listening, reading, and vocabulary items.	12	2.8
Q9	The explanation of the Modified Angoff Method was adequate and I felt able to undertake the rating tasks for the listening, reading, and vocabulary items.	12	3.0
Q10	The time provided for rating the EIKEN listening, reading, and vocabulary items was adequate.	13	3.2
Q11	The feedback on item difficulty of the EIKEN listening, reading, and vocabulary items was useful.	13	3.3
Q12	The explanation of the approach to standard setting for the EIKEN indirect Writing items was adequate and I felt able to undertake the rating task.	13	2.4
Q13	The time provided for rating the EIKEN indirect Writing items was adequate.	13	3.4
Q14	The facilities and food service were adequate and helped create a productive and efficient working environment	13	3.8
Q15	During the workshop I felt I had adequate opportunities to present my opinions and was able to ask questions when I was not sure of how to proceed	13	3.7

5.4.5.3 Internal validity

As with Panel 1, the standard error of the cutscore is used as a measure of evaluating the precision and replicability of the cutscores obtained (see Section 5.3.5.3 for a detailed description of the standard error of the cutscore).

The SE_c for the second round of Angoff judgments for Reading for all three grades is less than half of one raw score point (see Tables 5.19, 5.20, 5.21), indicating a high level of precision in the cutscores, with little expected variability if replicated with different groups of (equivalent) judges. Table 5.25 further provides the confidence intervals around the cutscores, calculated according to the same procedures explained in Section 5.3.5.2.

Table 5.25 Confidence intervals for Reading cutscores for G2, Pre-2, and 3

Grade	Range	68%	95%	99%
Grade 2	lower	10.7	9.7	9.1
	upper	12.7	13.7	14.3
Grade Pre-2	lower	11.8	10.6	9.9
	upper	14.2	15.4	16.1
Grade 3	lower	9.3	8.3	7.7
	upper	11.3	12.9	12.9

As with noted in Section 5.3.5.3, in order to evaluate the contribution of SE_c to the total error by comparing it to the size of SEM for the same test on the same raw-score scale, it will be necessary to utilize the SE_c for the aggregated cutscores across the First Stage Tests[12]. These statistics and the result of the comparison with three commonly cited criteria are noted in Tables 5.26 and 5.27 respectively. Grade 2 meets all three levels of SEM, including the most stringent recommended by Jaeger (1991). Grades Pre-2 and 3 meet the criteria suggested by

[12] As noted, Grades 2, Pre-2, and 3 use indirect selected response items to test aspects of writing. These items are dichotomously scored in the same way as other items in the First Stage Test and are included in the calculation of the test performance statistics shown in Chapter 4, whereas the constructed response Writing components for Grades 1 and Pre-1 were not included in the test performance statistics for those grades.

Kaftandjieva (2010) and Cohen, Kane, and Crooks (1999) but not the most stringent level recommended by Jaeger.

Table 5.26 Overall cutscores and SE_c for Grades 2, Pre-2, 3

Rater	Grade 2		Grade Pre-2		Grade 3	
	Percent	Score	Percent	Score	Percent	Score
R1	55.7	41.8	60.1	45.1	64.6	42.0
R2	60.4	45.3	56.5	42.4	62.2	40.4
R3	57.7	43.3	63.9	47.9	72.3	47.0
R4	56.8	42.6	59.7	44.8	58.5	38.0
R5	65.9	49.4	69.5	52.1	73.5	47.8
R6	65.6	49.2	63.2	47.4	72.5	47.1
R7	62.4	46.8	67.2	50.4	70.8	46.0
R8	53.6	40.2	53.1	39.8	63.1	41.0
R9	53.6	40.2	61.6	46.2	63.4	41.2
R10	64.5	48.4	72.4	54.3	68.9	44.8
R11	60.3	45.2	63.5	47.6	68.0	44.2
R12	52.4	39.3	53.5	40.1	55.7	36.2
R13	66.1	49.6	68.4	51.3	70.8	46.0
mean	59.6	44.7	62.5	46.9	66.5	43.2
SD	5.0	3.8	6.0	4.5	5.7	3.7
SE_c	1.4	1.0	1.7	1.2	1.6	1.0

Table 5.27 Evaluation criteria for precision: SE_c compared to SEM

	G2	≤ criteria	G Pre-2	≤ criteria	G 3	≤ criteria
SEM	3.9		3.7		3.3	
$\frac{1}{4}$SEM	1.0	YES	0.9	NO	0.8	NO
$\frac{1}{3}$SEM	1.3	YES	1.2	YES	1.1	YES
$\frac{1}{2}$SEM	2.0	YEs	1.9	YES	1.7	YES

Panel 2 provided extra data from an internal validity perspective in the form of two rounds of recorded judgments made for the same modified Angoff method. Referring back to Tables 5.19, 5.20, and 5.21 we can see that the two rounds of judgments produced almost identical cutscores, with a maximum difference of 0.3 raw score items for Grade Pre-2. For both Panel 1 and 2, no discussion was built into the procedure between the two rounds due to time constraints and the ambiguity in the literature over the benefits of discussion. Incorporating discussion may have led to more noticeable changes between rounds, but as noted previously, it is not clear whether those changes would be more valid or would be the result of forced consensus towards dominant group members.

Examining the standard deviation for the cutscores, it can be seen that while the difference is small, the trend across all three grades is for lower standard deviation in round two, indicating a slight increase in consensus over the two rounds (Cizek & Bunch, 2007; Tannenbaum & Wiley, 2008). The provision of feedback provided the opportunity for reflection on the items, and all participants indicated in the questionnaire that the feedback was useful, The combination of feedback with the opportunity to amend judgments in a second round may thus have contributed to a slight increase in consensus.

A two-facet MFRM analysis using FACETS (Linacre, 2014) was carried out on the reading judgment data for each grade to evaluate the relative severity and the consistency of the ratings provided by judges, and to provide fair average estimates

of the cutscores after taking into account the relative severity of raters and difficulty of the items. The methodology and interpretation of output from FACETS follow that described for Panel 1 in Section 3.3.5.3.

In terms of severity, a similar pattern to Panel 1 can be observed in the three facet maps shown in Figures 5.6 to 5.8. The majority of judges fall within the ±1 logit, range of severity, which as noted in Section 5.2.5.3, has been suggested as a tolerable range of variation. For Grade 2, two raters fall outside this range at the more severe end of the severity/leniency spectrum, while for Grade Pre-2 three raters fall outside that range—two more severe, one more lenient. For Grade 3 four raters fall outside the range, evenly split between more severe and more lenient.

The rater measurement reports for Grade 2, Grade Pre-2, and Grade 3 (Tables 5.30, 5.31, and 5.32) provide the fair average estimates and the infit mean square estimates of consistency. The mean of fair average estimates for each rater thus provides an estimate of the cutscore adjusted for the relative severity of raters and the difficulty of items. Following the same rationale described for Panel 1, these fair averages can be converted to percentage cutscore estimates, so that the adjusted fair average cutscore for Grade 2, for example, would be 58.3 percent. In terms of consistency, as with Panel 1, the raters demonstrate a high level of consistency, with two raters for Grade 2, two for Grade Pre-2 and one for Grade 3 failing to meet the fit criteria. The levels of misfit are summarized in Table 5.28. For those raters showing misfit, the levels, however tend to be marginal and fall within the 1.5-2.0 range (R11 falls just outside this at 2.08), which Taylor & Galazci (2012) and Myford and Wolfe (2004) suggest will still yield useful results for low-stakes situations such as training for raters.

As with Panel 1, a second analysis was run in which the misfitting raters were dropped. The custscores converted to percentages from the mean of fair average ratings across raters for both analyses for all three grades are shown in Table 5.29. The adjusted fair average cutscores for the main analysis change very little from the mean of observed cutscores (the observed mean cutscores can be seen in the rater measurement reports in Tables 5.30, 5.31, and 5.32, and are identical to the Round 2 Angoff cutscores in Tables 5.19, 5.20, and 5.21. The second analysis, with misfitting raters removed also has little substantive impact on the cutscores.

Table 5.28 Misfitting raters (infit mean square > 1.5)

Grade	Rater	Infit mean square
Grade 2	R3	1.79
	R12	1.97
Grade Pre-2	R1	2.08
	R10	1.54
Grade 3	R3	1.74

Table 5.29 Cutscore estimates (percentages) from fair averages

FACETS Analysis	Grade 2	Grade Pre-2	Grade 3
All judges	58.3	65.8	69.1
Without misfitting judges	59.7	65.0	68.8

G2 Reading Angoff Method FACETS Analysis 15/11/2014 15:23:09
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1A,2A,S) Yardstick (columns lines low high extreme)= 0,3.5,-4,3,End

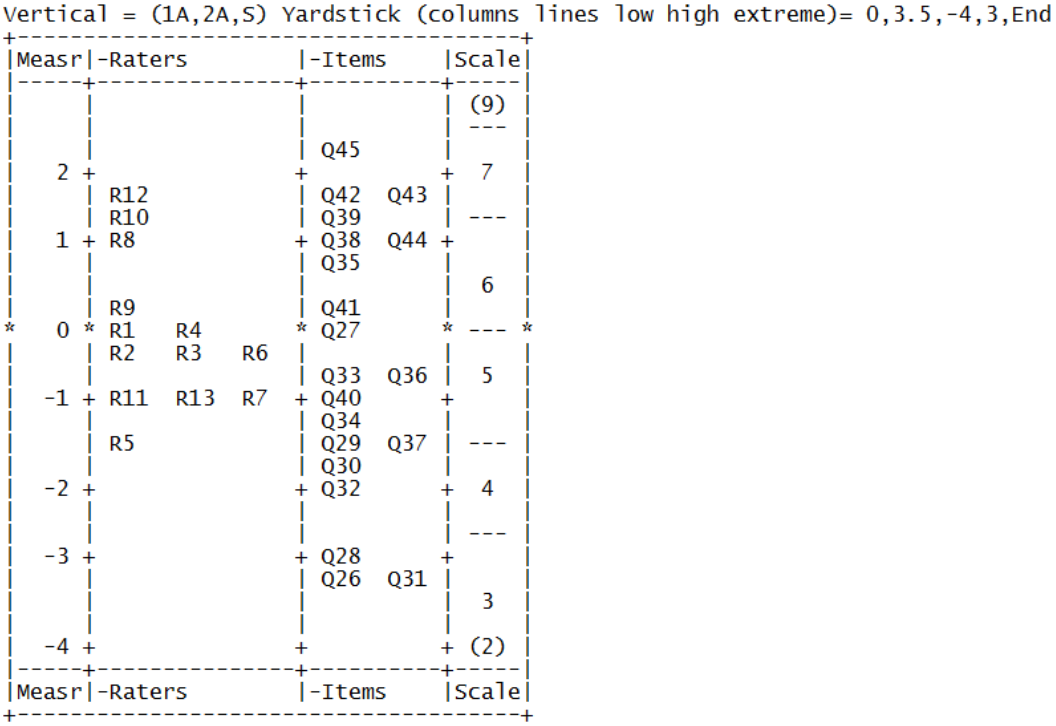


Figure 5.6 Facet map for Grade 2

GP2 Reading Angoff Method FACETS Analysis 14/11/2014 10:40:22
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1A,2A,S) Yardstick (columns lines low high extreme)= 0,3,-4,5,End

Measr	-Raters	-Items	Scale
5			(10)
		Q42	
4			

		Q38	
3	R12	Q45	
	R8		8
		Q36 Q43	
2			
		Q44	
		Q39	---
1	R1	Q28 Q31 Q40 Q41	
	R2 R6	Q33	
	R4		7
* 0 *	R3 R9	*	*
		Q29 Q35 Q37	

-1	R11		
	R7		6
	R10 R13	Q34	
-2		Q30	

	R5		
-3			
			5
		Q32	
-4		Q26 Q27	(3)

Figure 5.7 Facet map for Grade Pre-2

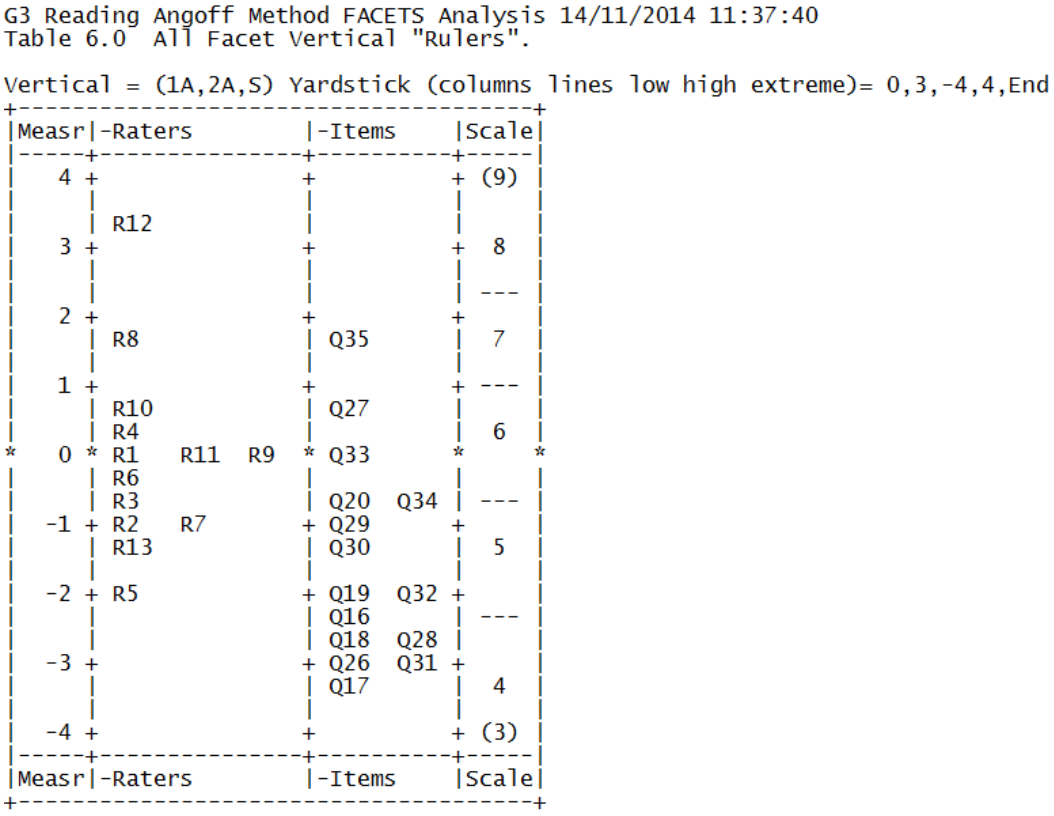


Figure 5.9 Facet map for Grade 3

Table 5.30 Rater measurement report for Grade 2 Reading

G2 Reading Angoff Method FACETS Analysis 15/11/2014 15:23:09
Table 7.1.1 Raters Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
90	20	4.50	4.51	1.75	.26	1.97	2.5	1.95	2.5	-.18	.48	14.6	16.4	12	R12
93	20	4.65	4.66	1.55	.26	.83	-.4	.85	-.4	1.18	.91	12.1	18.3	10	R10
101	20	5.05	5.05	1.03	.26	.76	-.7	.76	-.7	1.24	.83	16.7	23.0	8	R8
112	20	5.60	5.58	.31	.26	1.09	.3	1.09	.3	.87	.86	22.9	27.4	9	R9
115	20	5.75	5.73	.12	.26	.56	-1.5	.60	-1.4	1.48	.88	30.0	28.0	4	R4
118	20	5.90	5.88	-.08	.26	1.05	.2	1.07	.3	1.02	.86	28.8	28.2	1	R1
120	20	6.00	5.98	-.21	.26	1.79	2.1	1.77	2.1	.23	.73	21.3	28.2	3	R3
122	20	6.10	6.08	-.35	.26	.49	-1.9	.51	-1.8	1.47	.88	29.6	28.0	6	R6
123	20	6.15	6.13	-.41	.26	.90	-.2	.89	-.2	1.19	.93	31.7	27.9	2	R2
128	20	6.40	6.39	-.75	.26	.94	.0	.92	-.1	1.10	.83	23.8	26.7	7	R7
129	20	6.45	6.45	-.82	.26	.33	-2.8	.33	-2.8	1.64	.92	27.1	26.3	11	R11
129	20	6.45	6.45	-.82	.26	1.04	.2	1.00	.1	.94	.86	25.8	26.3	13	R13
136	20	6.80	6.84	-1.31	.27	.43	-2.2	.47	-2.0	1.60	.94	26.7	23.0	5	R5
116.6	20.0	5.83	5.83	.00	.26	.94	-.3	.94	-.3		.84			Mean (Count: 13)	
13.7	.0	.68	.68	.90	.00	.47	1.5	.45	1.5		.12			S.D. (Population)	
14.2	.0	.71	.71	.94	.00	.49	1.6	.47	1.5		.12			S.D. (Sample)	

Model, Populn: RMSE .26 Adj (True) S.D. .86 Separation 3.32 Strata 4.76 Reliability (not inter-rater) .92
Model, Sample: RMSE .26 Adj (True) S.D. .90 Separation 3.47 Strata 4.96 Reliability (not inter-rater) .92
Model, Fixed (all same) chi-square: 155.6 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.2 d.f.: 11 significance (probability): .43
Inter-Rater agreement opportunities: 1560 Exact agreements: 373 = 23.9% Expected: 393.2 = 25.2%

Table 5.31 Rater measurement report for Grade Pre-2 Reading

GP2 Reading Angoff Method FACETS Analysis 14/11/2014 10:40:22
Table 7.1.1 Raters Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisI	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
97	20	4.85	4.92	3.09	.30	.66	-1.1	.66	-.8	1.35	.89	11.7	12.7	12	R12
103	20	5.15	5.25	2.54	.30	1.05	.2	1.05	.2	.99	.89	18.3	16.6	8	R8
120	20	6.00	6.07	1.00	.31	2.08	2.6	2.10	2.7	.01	.79	27.5	27.2	1	R1
124	20	6.20	6.26	.62	.31	.51	-1.7	.48	-1.9	1.54	.91	33.8	28.9	2	R2
124	20	6.20	6.26	.62	.31	.62	-1.2	.60	-1.3	1.40	.91	32.1	28.9	6	R6
127	20	6.35	6.41	.34	.31	.52	-1.7	.53	-1.7	1.42	.93	29.2	29.9	4	R4
131	20	6.55	6.62	-.05	.31	1.15	.5	1.08	.3	.98	.91	25.4	30.7	3	R3
131	20	6.55	6.62	-.05	.31	1.14	.5	1.12	.4	.87	.87	30.8	30.7	9	R9
139	20	6.95	7.04	-.84	.32	.90	-.2	1.00	.1	1.02	.81	31.7	30.0	11	R11
145	20	7.25	7.36	-1.46	.33	.62	-1.2	.61	-1.2	1.35	.89	27.5	27.5	7	R7
146	20	7.30	7.42	-1.57	.33	.25	-3.2	.25	-3.2	1.75	.96	30.0	26.9	13	R13
147	20	7.35	7.47	-1.68	.33	1.54	1.5	1.68	1.8	.31	.60	22.1	26.3	10	R10
155	20	7.75	7.87	-2.56	.34	1.11	.4	1.10	.4	.91	.95	15.8	20.1	5	R5
129.9	20.0	6.50	6.58	.00	.32	.94	-.4	.94	-.3		.87			Mean (Count: 13)	
16.4	.0	.82	.83	1.58	.01	.47	1.5	.49	1.6		.09			S.D. (Population)	
17.0	.0	.85	.87	1.65	.01	.49	1.6	.51	1.6		.09			S.D. (Sample)	

Model, Populn: RMSE .32 Adj (True) S.D. 1.55 Separation 4.91 Strata 6.88 Reliability (not inter-rater) .96
Model, Sample: RMSE .32 Adj (True) S.D. 1.62 Separation 5.12 Strata 7.16 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-square: 327.6 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.6 d.f.: 11 significance (probability): .39
Inter-Rater agreement opportunities: 1560 Exact agreements: 403 = 25.8% Expected: 403.6 = 25.9%

Table 5.32 Rater measurement report for Grade 3 Reading

G3 Reading Angoff Method FACETS Analysis 14/11/2014 11:37:40
Table 7.1.1 Raters Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBisi	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Raters
72	15	4.80	4.74	3.34	.35	1.23	.6	1.32	.9	.73	.66	9.4	6.4	12	R12
88	15	5.87	5.81	1.57	.32	.77	-.5	.81	-.4	1.18	.70	17.8	18.9	8	R8
96	15	6.40	6.34	.77	.32	1.00	.1	1.04	.2	.93	.64	17.2	25.3	10	R10
102	15	6.80	6.78	.17	.32	.92	-.1	.98	.0	1.12	.84	30.0	28.6	4	R4
103	15	6.87	6.85	.07	.32	.78	-.5	.81	-.4	1.24	.87	31.1	28.9	1	R1
103	15	6.87	6.85	.07	.32	1.08	.3	1.06	.2	.96	.74	25.6	28.9	9	R9
105	15	7.00	7.00	-.14	.32	.25	-3.0	.24	-3.0	1.77	.91	29.4	29.4	11	R11
106	15	7.07	7.08	-.24	.32	.54	-1.4	.57	-1.3	1.47	.85	30.0	29.5	6	R6
109	15	7.27	7.32	-.56	.33	1.74	1.8	1.57	1.4	.28	.64	26.7	29.4	3	R3
112	15	7.47	7.55	-.89	.34	1.40	1.1	1.29	.8	.67	.86	30.6	28.6	2	R2
112	15	7.47	7.55	-.89	.34	.77	-.5	.76	-.6	1.25	.78	28.3	28.6	7	R7
115	15	7.67	7.78	-1.24	.35	1.29	.8	1.22	.6	.65	.72	23.9	27.1	13	R13
121	15	8.07	8.22	-2.04	.39	.74	-.5	.78	-.4	1.33	.85	22.2	22.1	5	R5
103.4	15.0	6.89	6.91	.00	.33	.96	-.1	.96	-.1		.77			Mean (Count: 13)	
12.1	.0	.81	.87	1.30	.02	.38	1.2	.34	1.1		.09			S.D. (Population)	
12.6	.0	.84	.90	1.35	.02	.39	1.2	.35	1.2		.09			S.D. (Sample)	

Model, Populn: RMSE .33 Adj (True) S.D. 1.25 Separation 3.76 Strata 5.35 Reliability (not inter-rater) .93
Model, Sample: RMSE .33 Adj (True) S.D. 1.31 Separation 3.93 Strata 5.57 Reliability (not inter-rater) .94
Model, Fixed (all same) chi-square: 178.2 d.f.: 12 significance (probability): .00
Model, Random (normal) chi-square: 11.3 d.f.: 11 significance (probability): .42
Inter-Rater agreement opportunities: 1170 Exact agreements: 290 = 24.8% Expected: 298.3 = 25.5%

5.4.4.4 External validity

It was not possible to conduct a comparison of cutscores from different methods as a source of external validity evidence in the same way as was done for Panel 1. The limitations associated with the Basket method are discussed in detail in Kaftandjieva (2010) and also described in the Manual (2009). One particular side product of the judgment task and the way results are aggregated in the Basket method to form cutscores makes the method unsuitable for tests targeting a very narrow range of ability. This is exacerbated when those tests are at the lower end of the CEFR, as is the case for Grades Pre-2 and 3. Kaftandjieva (2010, p. 61) suggests that "the cut scores can be set only when the total number of items belonging to the levels preceding a certain level is different from zero or from the maximum number of test items." Kaftandjieva (2009, 2010) and the Manual (Council of Europe, 2009) point to the tendency for distortion of the cutscores in the Basket method due to the judgment task which forces a yes/no decision which will result in all items for which test takers at the required level have a probability of greater than 50 percent of answering correctly essentially being given a probability of 1, and all items for which the same test takers have a probability of less than 50% of answering correctly being given a probability of 0.

In fact a similar issue has been noted for the Angoff Yes/No method (Cizek & Bunch, 2007; Council of Europe, 2009; Kaftandjieva, 2009). In the case of the Angoff Yes/No method, the resulting distortion will result in a situation in which a test that contains all items closely targeted at one level (e.g. B1), would derive a cutscore of 100%. This is because a B1-level candidate would have a greater than 50 percent chance of answering each item, meaning an accurate judge would assign a score of 1 for all items, which when tallied according to the Angoff Yes/No method, would require a B1-level candidate to achieve 100% in order to pass the test, which as Cizek & Bunch (2007, p. 94) note is "clearly not the intention of the rater or a realistic expectation based on the difficulty of the test."

For the Basket method, the distortion will tend to work in the opposite direction due to the method of aggregating results to derive the cutscore, though the root cause derives from the same method of reducing the difficulty probability judgments to 0/1, or yes/no, decisions. The cutscores for the Basket method are

derived by counting the total number of items allocated to all levels below the level of interest. The minimum score needed to be classified as belonging to the level of interest would thus be the cutscore derived through the basket method plus one item from the level of interest. If a test is targeted at a particular level, for example B1, and all items are constructed according to a well-designed test specification to be at the B1 level of difficulty, then all items would be placed in the B1 "basket" by an accurate rater, as the first level at which test takers can be expected to complete the items correctly would be B1. If there are 10 items in the test, all at B1, the number of items in A1 and A2 would be 0. The cumulative number of items below the level of interest, B1, would thus be 0. The score necessary to be classified as B1 would be the cumulative number of items below B1 plus 1. In other words a test perfectly targeted at B1 would result in a cutscore of 1, which by its nature is as equally absurd as the situation highlighted by CIzek and Bunch (2007) above.

Table 5.33 provides the mean number of Reading items allocated to each CEFR level in Round 2 judgments (the results have been averaged across the judgments of each participant and rounded). The distortion described above is most evident in the two lowest-level tests. The majority of items in both reading tests fall in the target level. For Grade 3 of course, there was no level below A1 clearly defined in the CEFR, but rating forms allowed raters to place items in a *below A1* category if they felt that test takers would be able to correctly answer an item without A1-level ability.

Table 5.33 also shows the cutoffs in terms of the number of items and as a percentage of the total possible score for Reading that would be derived by applying the Basket method. As can be seen, the cutoffs, as expected, are unrealistically low for both Grade 3 and Grade Pre-2 and certainly do not reflect the probability judgments derived through the Angoff method. The reason, ironically, is precisely because the items are well-targeted at the level of interest, and because there are few levels below them into which items could be placed. For Grade 2, however, which is targeted at the intermediate B1 level and has items below, at, and above the target level, the method derives a cutscore which is almost identical to the final cutscore derived through the Angoff method.

Table 5.33 Average number of reading items allocated to CEFR levels in round 2 Bssket method judgments

CEFR :Level	Grade 3	Grade Pre-2	Grade 2
Below A1	0	0	0
A1	13	4	8
A2	2	15	3
B1	0	1	7
B2	0	0	2
C1	0	0	0
C2	0	0	0
Total	15	20	20
Cutscore (items)	**1**	**5**	**12**
Cutscore (percent)	**8 %**	**25%**	**60%**

5.5 External validation study

5.5.1. Introduction

A separate external validation study was designed to address one of the major potential counter-claims to the validity of standard setting carried out to support claims of alignment between the EIKEN tests and the CEFR: Did the local understanding of the CEFR developed for the purposes of standard setting adequately reflect the understanding of the CEFR developed by similar educators in the context of Europe? As noted earlier in the discussion of the selection of standard-setting methods for Panels 1 and 2, Kane (2001b, p. 75), suggests that replicating standard setting not only with different methods, but also "by different researchers, with a different group of participants, under different circumstances" would be a powerful source of external validity evidence. The external validation study was thus an attempt at applying Kane's recommendation to the application of standard setting in the context of the CEFR, and in so doing address one of the major potential threats to the validity of the standard setting carried out through Panels 1 and 2.

In order to provide external evidence to support the validity of the standard setting carried out with local educators for RQ3, a separate research question was formulated:

Do educators in the context of Europe who are experienced at using the CEFR for teaching and assessment demonstrate a similar estimation of the EIKEN tests in relation to the CEFR as was derived through Standard Setting Panels 1 and 2?

The external validation study was undertaken in collaboration with a researcher in Spain who was prepared to collaborate in the recruitment of participants and the administration of EIKEN tests to those participants in that local context. The author designed the methodology for the study and carried out the statistical analyses on results. However, the project was only possible because of the ability of the Europe-based collaborator to secure the help and cooperation of teachers and learners willing to contribute their valuable time, and through her efforts to coordinate the collection of data with those participants in Europe. Such a collaborative approach is, in practice, likely the only way to realize Kane's (2001) call for replication of standard setting as a powerful validity check. At the same time, the lack of this kind of external validation specifically in relation to linking to the CEFR makes this aspect of the overall standard setting an important contribution to the literature. In Section 2.2.3, Chapter 2, the criteria offered by Cronbach was noted for deciding on the evidence to target for validation research. The fourth criteria, *leverage*, or how critical the information is for achieving consensus in the relevant audience, also makes this external validity study a high priority given the questions that have been asked about the relevance of using the CEFR outside of its original European context.

5.5.2 Methodology

It was decided to focus on only one of the EIKEN grades due to the logistical constraints posed by carrying out an external validation study in which not only would the participants and method be different, but the physical context would be in Europe. It was further decided that the external validation study would collect data only on vocabulary and reading items to allow for administration within regular classroom schedules and also to eliminate the need for any special equipment to

administer listening sections. This reduction in scale brings limitations to the generalizability of the results, but practicality needs to be taken into consideration in the planning and implementation of standard setting, as noted by Berk (1986). The study was conceptualized not as the primary source of evidence supporting any claim of relevance between the EIKEN tests and the CEFR, but as a way of adding an extra layer of depth to the body of evidence obtained from the main standard-setting studies.

5.5.2.2 Standard setting method

The claim for external validation, then, given that the focus is on Grade Pre-1, is that a B2-level of proficiency is needed to pass the Pre-1 test. In Section 5.2.2 we suggested criteria of reasonableness in determining acceptable differences between cutscores derived from multiple standard-setting methods. The same approach will be employed here. The results of this study will be taken as adding support to the standard setting carried out for Panel 1 provided that Panel 1 and the external validation study both derive cutscores for classifying test takers as minimally competent at B2 level *that are lower than or very close to the score required to actually pass the Pre-1 test*.

Kane (2001b, p. 75) suggests that if "the Angoff method were used in the original study, the new study might involve an examinee-centered method." Following this approach, for this study, the Contrasting Groups method was chosen As noted in Section 2.5, the distinguishing feature of examinee-centered methods is that judges make judgments about actual test takers. Cizek and Bunch (2007, p. 107) provide the following overview of the Contrasting Groups methodology:

> Participants, who are unaware of examinees' actual test scores, make judgments about each examinee as to their mastery/nonmastery status. . . . Participants' category judgments are used to form distributions of total test scores for each of the two groups. . . . The two distributions are then plotted and analyzed to arrive at a cut score that distinguishes group membership.

Cizek and Bunch (2007, p. 107) note that usually judges "have personal knowledge of individual, real examinee's levels of knowledge or skill with respect to the characteristics assessed." This was the case in Livingston and Zieky (1989), Van Nijlen and Jansenn (2008), and Green, Trimble, and Lewis (2003).

Applications of the Contrasting Groups method show variation in terms of training for judges and the specificity of the judgment task. In Livingston and Zieky (1989), training was limited to a short meeting with the teachers to explain the procedures. In Van Nijlen and Jansenn (2008), the authors state that training was not provided, as the teachers were familiar with the official attainment targets for biology used for primary school students in Flemish schools. In Green et al. (2003) teachers rated their own students in relation to performance level descriptors which were sent to schools.

Bechger, Kujper, and Maris (2009) provide an interesting application of the Contrasting Groups procedure for linking tests for Dutch as a second language to the CEFR. The judges in this study were not familiar with the students they rated, but instead were asked to review the actual spoken and written test performances of students and to rate the students against a rating scale developed from CEFR descriptors. An 80% definition of mastery was used. An examinee could receive a maximum "CEFR sum score" of 10 if both raters answered 'yes' to the judgment question for all 5 descriptors on the rating scale. If an examinee scored 8 or more, he or she was considered to be a master.

5.5.2.1 The judgment task

Following Bechger et al. (2009), an external criterion measure was derived from B2-level descriptors in the CEFR. Teachers in this study would be rating their own students with whom they were thus familiar, but they would also be using a rating measure based on the CEFR to clarify the judgment task. Several changes were made to the procedure outlined in Bechger et al. (2009). Firstly, it was considered unrealistic to ask teachers to rate each student against each CEFR descriptor in the rating scale, so teachers were instructed to form a holistic judgment of 'B2 level' or

'not B2 level' after reviewing the B2 descriptors for reading provided in the rating form.

5.5.3 Participants

It was decided to follow the precedent set in the applications of the Contrasting Groups procedure listed above, in which training was not given and the researchers relied on teachers' accumulated knowledge and experience of the content domain. While not ideal, it was felt this provided the best balance between practicality, given the limits on how much time participants could devote to the project, and deriving robust results. To maximize the knowledge and experience of teachers to compensate for the lack of training, the following criteria was set for judges;

- They should be experienced EFL teachers with knowledge of the content and purpose of the CEFR
- They should have experience evaluating or judging their students in relation to the levels expressed in the CEFR
- They should be familiar with the reading ability of the learners they rate.

The Europe-based researcher was able to secure the participation of teachers working in language schools in Spain. These teachers were preparing students for B2-level exams. The CEFR was commonly used in training, curriculum planning, and as a basis for testing in these schools, meaning these teachers could be assumed to be familiar with the CEFR, and of course they would be familiar with their students. The students were preparing for B2-level exams, so it was expected that there would be learners who had reached this level, and so could be considered 'masters' in the terminology of the Contrasting Groups procedure, and learners who had not yet reached this level, who would be classed as 'non-masters.'

5.5.4 Instruments

Teachers were provided with a rating form which included detailed instructions for the judgment task, a list of the B2 reading descriptors in Table 5.34, and a form for

writing the name of each student and three alternatives for judging each student's level (Below B2, B2, Cannot judge).

Table 5.34 Source of B2-level reading descriptors used for rating scale

Name of CEFR scale	Number of descriptors
Overall Reading Comprehension	1
Reading Correspondence	1
Reading for Orientation	2
Reading for Information & Argument	3
Reading Instructions	1

The same Grade Pre-1 vocabulary and reading test items from the same test form used for test-centered standard setting in Japan were used for this project. The test booklet for the vocabulary and reading sections was reproduced exactly as it appeared in the live administration and in the original standard setting in Japan. Two questionnaires were also prepared, one for the teachers involved and one for the students.

5.5.5 Procedure

The Europe-based collaborator met individually with the teachers to explain the procedures and materials. Each teacher administered the reading test and student questionnaires during a normal class period. Different classes took the tests on different days depending on class schedules. Teachers rated their students separately and filled out the questionnaires without reference to the test results. Each student was rated once by his or her classroom teacher. Results from the test answer sheets, teacher judgment forms, and questionnaires were collated and input into Excel format by the European collaborator and sent to Japan for analysis by the author.

5.5.6 Results

The final number of participants, both students and teachers, is shown in Table 5.35. The 10 classes were distributed across four geographically distinct language schools located in different parts of one region in Spain.

Table 5.35 Number of participants

Number of classes	Number of teachers	Number of students
10	6	170

Five of the six teachers returned their questionnaires. Table 5.36 shows the teachers' ages and years of experience in various educational sectors. Table 5.37 shows the teachers' degree of familiarity with the CEFR. Of the 154 students who responded to the questionnaire question on gender, 116 were female and 38 were male. The average age was 31, while the youngest was 16 and the oldest 80 (based on 149 responses to this question).

Table 5.36 Teachers' experience (years in different educational sectors)

	Age	Language school	Secondary	Company classes	Other
T1	48	20	2	5	
T2	40	12	1		
T3	48	21		4	2
T4	41				18
T5	38	5	7	1	

Table 5.37 Degree of familiarity with the CEFR

Degree of familiarity with CEFR	Number of teachers
I had read the CEFR and was familiar with its aims and content, including the Common Reference Levels.	1
I had experience using the Common Reference Levels to classify students in the classes I teach, but had not received any specialized training on how to interpret the levels in the CEFR	1
I had experience using the Common Reference Levels to classify students in the classes I teach and had received specialized training on how to interpret the levels in the CEFR	3

The results for Vocabulary and Reading have been combined for use with the Contrasting Groups method as it is important to be able to create frequency distributions which provide a reasonable spread of test takers across the score scale in order to distinguish differences between the mastery and non-mastery groups in a meaningful way. Table 5.38 presents the descriptive statistics for the 170 students on the 41 vocabulary and reading comprehension items. Table 5.39 presents the results for each of the three categories into which the teachers classified the students: B2 level (masters), Below B2 (non-masters), or the cannot judge category.

Table 5.38. Descriptive statistics (test)

Items	41	Reliability (α)	.79
Number of Test takers	170	SEM	2.30
Mean (raw score)	30.75	Mean item facility	.75
SD	4.96	Mean item discrimination	.33
Min	14		
Max	40		

Table 5.39. Test results for each classification category

	All	B2	Below B2	Cannot judge
Mean	30.75	32.83	27.25	30.40
Mode	32	32	27	30
Median	32	33	27	30
SD	4.98	3.86	4.93	3.81
Min	14	15	14	23
Max	40	40	36	36
N	170	101	59	10

Having classified students into groups of masters who were considered to be at a B2 level of proficiency, and non-masters who were classified as being below B2 level, it was now necessary to estimate the appropriate score boundary between these two groups. In their overview of the Contrasting Groups procedure, Cizek and Bunch (2007) describe the most commonly used ways for doing this:

1) Use the midpoint between the means of the two groups.
2) Use the midpoint between the medians of the two groups.
3) Plot the point of overlap between the two score distributions.
4) Use logistic regression to find the raw score point at which examinees classified as non-masters first reach a 50% chance of being classified as masters. Cutscores scores derived using each of these procedures are shown in Table 5.40. The cutscores are shown as the raw-score number of items-correct and also as a percentage of the total possible raw score for the combined Vocabulary and Reading components for Grade Pre-1, with both rounded to one decimal place.

Table 5.40 Comparison of cut-off scores

Method	Raw score	Percent	
Mean of means	30.0	73.2	
Midpoint of medians	30.0	73.2	
Overlap of distribution plots	28.0	68.3	
Logistic regression	30.3	73.9	

The calculation of cutscores using the third and fourth procedures is explained more fully below. Cizek and Bunch (2007) point out that score distributions may often be jagged, with multiple points of overlap which make it difficult to identify the appropriate cutscore. One strategy is to use smoothing procedures (Cizek & Bunch, 2007; Livingston & Zieky, 1982). Both Cizek and Bunch (2007) and Livingston and Zieky (1982) note a number of alternatives for doing this. The analysis has followed the suggestion in Cizek and Bunch (2007) for employing the smoothing procedures available in Microsoft Excel. Figure 1 below shows the smoothed distributions.

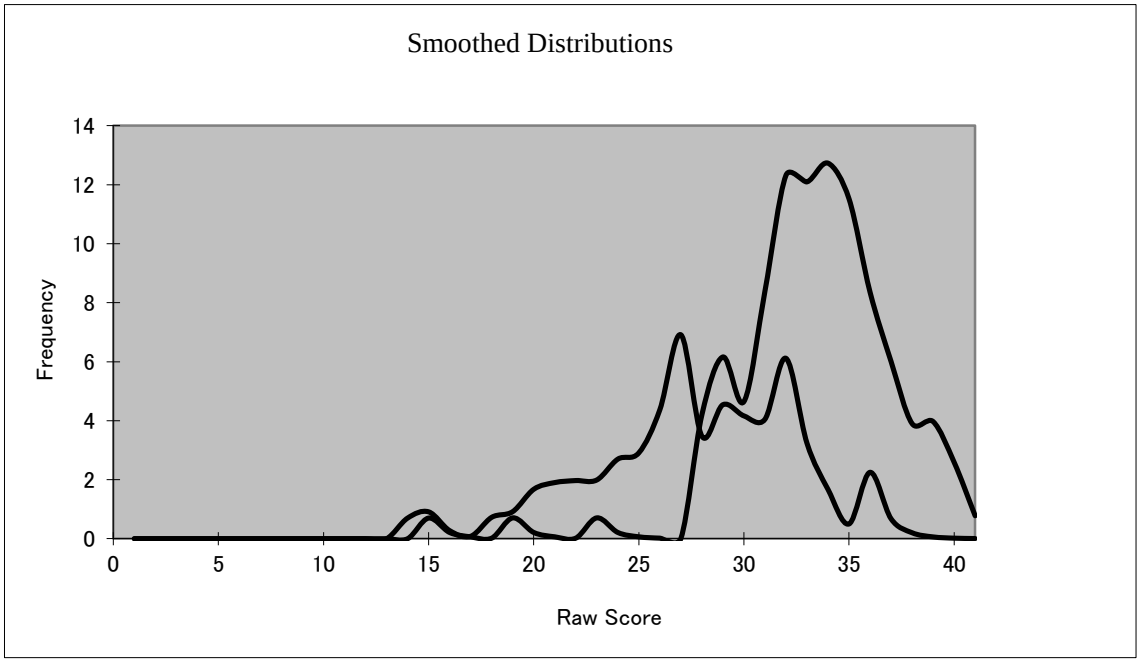


Figure 5.39 Smoothed distributions

The logistic regression was carried out using SPSS version 18. The results are displayed in Tables 5.41 and 5.42. Following Cizek and Bunch (2007, p. 112), the cutscore was obtained using equation 5.3 to calculate the raw score point at which the probability that a student in the Below B2 category has a 50% chance of being classified as belonging to the B2 category.

Formula 5.3 $50=-8.802+(.307)x$

Table 5.41 Model summary for logistic regression

STEP	-2 Log Likelihood	Cox & Snell R-Square	Nagelkerke R-Square
1	158.478	.278	.380

Model Chi Square (1)=52.174, *p<.01

Table 5.42 Results of logistic regression: variables in equation

		B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Step 1a	Raw score	.307	.055	31.022	1	.000	1.359
	Constant	-8.802	1.688	27.197	1	.000	.000

Based on Table 5.40, we can identify two probable cutscores for classifying test takers as B2: 28 (based on plotting the point of overlap between the smoothed distributions) or 30 (rounding down each of the 1st, 2nd, and 3rd procedures). Table 5.43 provides information on the number of misclassifications that would occur at each cutscore, taking the teacher level placements as the benchmark criterion. The percentage is based on the total number of students placed into B2 or Below B2 categories: a total of 160 test takers. The higher cutscore derives a slightly higher overall degree of misplacement. Table 5.43 further breaks this down into the percentage of false negatives (students classified as B2 by their teachers but whose test scores fall below the cutscore) and false positives (students classified as Below B2 by their teachers but whose test scores fall above the cutscore). The higher

cutscore has slightly fewer false positive classifications, but also results in a jump in false negatives. As Cizek and Bunch (2007) note, deciding on which type of misclassification to reduce when adjusting cutscores in operational circumstances is a policy decision which will depend on the relative impact of the different types of errors. For the purposes of this external validation study, the lower cutscore does show a tendency to reduce overall misclassification and ensure the greatest degree of accurate placement for students identified as being at the B2 level of ability.

Table 5.43. Decision tables for classification decisions at 2 cut-off points

	Cut-off=28			**Cut-off=30**		
	Cannot judge	Below B2	B2	Cannot judge	Below B2	B2
Below cut-off	1	31	3	4	38	16
Cut-off or higher	9	28	98	6	21	85
Total	10	59	101	10	59	101
Misclassification (total)	19%			23%		
False negative	18%			13%		
False positive	2%			10%		

5.5.4 Conclusions

As already noted, utilizing multiple standard-setting methods brings with it the ambiguity of multiple cutscores. For the purposes of this study, it will be useful to return to the original principal of evaluating the reasonableness of decisions as described in Section 5.2. The concern here is with evaluating the reasonableness of the standard-setting results in relation to the claim of alignment between Grade Pre-1 and the CEFR B2-level to which it is posited to be relevant. In particular, the purpose of the external validation study was to explore the effectiveness of the methods

adopted in Panel 1 and 2 to build an understanding of the CEFR level descriptors as PLDs for the standard setting. The cutscores will be examined from the perspective of acceptable differences, looking first at the comparison with the cutscore set for the same components by Panel 1, and then at the cutscores in relation to the passing score required for Grade Pre-1.

Referring back to Table 5.9, it can been seen that the combined cutscore for Vocabulary and Reading for Grade Pre-1 from round 2 Angoff method judgments is 59 percent or 24.2 raw score points. The two possible cutscores indicated through this external validation study are thus both higher than the final cutscore form Panel 1. One interpretation of this could be that the difficulty of the test in this particular context is seen as slightly lower, requiring candidates to achieve a higher score to demonstrate the same B2-level of ability. The passing score required for Grade Pre-1 is 70 percent. For the 41 items in the combined Vocabulary and Reading components this would come to a raw score (rounded to the nearest whole-score point) of 29. The two cutscores suggested by this external validity study fall on either side of this passing mark, supporting the most important part of the claim for which evidence was gathered from Panel 1: students passing the Grade Pre-1 test can be considered to have a B2-level of ability in terms of the CEFR level descriptors.

While differences have been identified between the cutscores set by the different panels and methods, this, as noted, is to be expected. Some discrepancy in the strength of the relationship between Grade Pre-1 and the B2 level of the CEFR is evident. Results from Panel 1 would suggest that a borderline passing candidate for the test would have demonstrated a strong B2 level of proficiency as the cutscore for CEFR level placement was lower than the passing mark. The results for the external validity standard-setting study on the other hand point to a more stringent interpretation of the PDLs in one European context, with a candidate achieving the passing score of 70 percent on the test being minimally competent at the B2 level. However, the underlying relationship between the B2 level of ability as defined by the CEFR and the ability required to pass the Pre-1 test is supported by both studies.

Referring back to the research question identified at the beginning of Section 5.5, the results do give some confidence that the procedures used to build familiarity with the CEFR, and the training and standard-setting procedures

employed in Panel 1, did indeed achieve an interpretation of the PLDs for B2 which was relatively consistent with "educators in the context of Europe who are experienced at using the CEFR for teaching and assessment." Given the integration of procedures and training within each Panel, and the application of similar approaches across both panels, such as the use of the self-study preparation booklets, this in turn does allow us some cautious endorsement of the procedures used across both Panel 1 and 2. It also offers support for the assumption, on which both Panel 1 and 2 are premised, that it is indeed possible to build a consensus interpretation of the CEFR level descriptors for standard-setting panels in EFL contexts outside Europe that is relatively consistent with that held by educators in Europe.

Chapter 6 Conclusion

6.1 Introduction

The primary goal of this study was to employ the latest developments in the theory and practice of language testing validation to gather evidence in support of the uses and interpretations of an established, large-scale EFL testing program in Japan. Using an explicit model of validation to guide the collection, analysis, and evaluation of data, the study aimed to contribute to the creation of a comprehensive, clear, and coherent validity argument for the testing program. Due to the scale of the testing program, this study focused only on the vocabulary, grammar, and reading sections of the First Stage tests. The study was intended to be comprehensive in terms of the amount of data collected in order to make definitive claims about the key areas of interest identified by the validation model.

Core features at the heart of the validation model underpinning the study were identified to drive the collection of evidence related to contextual and cognitive features of the tests, the scoring validity of the test, and criterion-related aspects of validity in the form of a relationship to an external criterion of language proficiency. A key feature of the testing program which was investigated is the way it is structured as a set of seven level-specific tests, known as grades. Each grade targets a different level of proficiency, and the grades are posited by the test developer to increase in difficulty in a clear and meaningful progression through a common frame of reference relevant to a coherent construct of EFL proficiency. The study thus distilled the core of the socio-cognitive model into three research questions designed to focus on and validate this key aspect of the testing program in terms of the meaningfulness, empirical distinction, and relevance of the stepped progression in proficiency.

6.2 Conclusions

6.2.1 Research Question 1: contextual and cognitive validity parameters

The investigation of criterial contextual and cognitive features supports the claim that the levels of proficiency are indeed distinct in coherent and meaningful ways. At the same time, the results across both expert judgment and automated textual analysis measures demonstrate clearly the importance of building *a profile* of criterial features. No one measure can serve the purpose of clearly defining criterial differences across texts and tasks targeting all of the different grades. It is the interaction and combination of relevant criterial features that will allow us to build a useful, transparent, relevant and interpretable profile of features, and it is such a profile that offers the greatest potential for test task specification.

It is also important to note that, not unsurprisingly, the greatest distinctions were identified between those grades targeting broader levels of proficiency associated for example with distinctly different CEFR levels. The distinctions between Grade 2, Grade Pre-1, and Grade 1 (B1, B2, and C1 respectively in terms of the CEFR), particularly in terms of vocabulary profiles and lexical threshold levels, were often much clearer than the distinctions amongst the lower levels targeting CEFR A1 and A2. The broad distinction between *below B*1 and *B1 and above* seems to be clearly supported. Some features did not produce statistically significant results for adjacent levels, although in general the trends across levels all changed in the directions expected. These distinctions do make sense in terms of the varied uses and interpretations for which the different grades are designed and applied. As noted in Chapter 1, the lower level grades are used primarily within formal educational contexts, and particularly for Grades 5, 4, and 3, are targeted at much smaller steps in terms of learning goals. In Chapter 5, it was suggested that these three grades are situated *within* A1, representing A1.1, A1.2, and A1.3 respectively. It is not surprising then, that many of the broader measures of criterial features were not able to capture the finer distinctions between these levels.

At the same time, the upper levels of Grades Pre-1 and Grade 1 have clearly demonstrated criterial features suitable to the advanced levels of the CEFR. In terms of vocabulary in particular, the lexical threshold levels for the long reading texts indicated a lexical threshold of between 5000 and 6000 word families for Grade

Pre-1 and 7000 for Grade 1. Both figures are above the lexical threshold required to reach the 95 percent coverage criterion for adequate comprehension of authentic texts recommended by Ravenhorst-Kalovski. The slightly lower vocabulary levels required for these grades when considering all task types across the First Stage tests, not just long reading passages, would still be appropriate for this coverage criterion. In terms of other linguistic features of the texts used for these grades, including AWL coverage, Grades Pre-1 and 1 showed features reflecting the levels in the literature for examinations such as IELTS and the authentic university texts analyzed by Green et al (2010). The criterial features then, not only support the distinctions between the grades, the primary focus of RQ1, but also support a claim of relevance to the kinds of texts used in the TLU domain to which these grades are intended to generalize. The intermediate B1 level of Grade 2, and to a lesser extent the A2-level Grade Pre-2, demonstrate a pattern of being important transition levels, particularly with the expert-judgment criteria such as topics and level of abstractness. The more achievement-test focused grades targeting smaller steps on the proficiency scale within the A1 band show characteristics broadly consistent with tests targeting this level in an EFL context. At the same time, some measures capable of distinguishing between broader levels of proficiency on the CEFR were not always able to distinguish between theses grades targeting focused steps more closely connected to the formal educational context, indicating that these grades would benefit from the use of finer-grained indices, particularly for lexical resources, to help flesh out specifications.

.

6.2.2 Research Question 2: empirical difficulty of levels

For RQ2, the various methods of evaluating the differences observed have all clearly underscored that the empirical difficulty of items designed for and used in the seven EIKEN grades are distinctly different and progress in the order of difficulty intended by the test developers. A key aspect of the vertical scaling has been the adoption of a data collection design which allowed for the retrospective calibration of large numbers of items administered in live tests across all seven grades over multiple years. The scale of this undertaking should not be underestimated. In many situations, due to the complexity of linking plans, it would only be possible to carry out such

scaling under specially administered experimental designs, which greatly reduces the generalizability of the results. The procedures employed for this study have thus facilitated the evaluation of the empirical difficulty of operational tests using large numbers of test taker responses, and this contributes to the robustness of the results.

At the same time, in order to achieve these results, the methodology employed for the vertical scaling has made pragmatic decisions across a range of the variables noted in the literature on vertical scaling studies. It is widely recognized that vertical scaling studies are design-dependent, and yet there is no commonly agreed framework or process for prioritizing the potentially different results generated through the various choices possible (Kolen & Brennan, 2004; Harris, 2007; Young, 2006). It is thus imperative that the choices made must make sense in terms of the particular goals of the context for which vertical scaling is being conducted, and most importantly must be documented and applied consistently. The principles underlying the choices made in this study have been clearly documented for future review and evaluation.

6.2.3 Research Question 3: criterion related validity

The external criterion measure of proficiency selected as the basis for investigating RQ3 was the Common European Framework of Reference (Council of Europe, 2001). The selection of this descriptive framework of proficiency was made with reference to the four subsidiary criteria identified in Chapter 1: relevance, transparency, interpretability, and comparability. Standard setting was only carried out for Grades 3 and above. As noted in Chapter 5, an a priori evaluative judgment had been made that the lower grades, 5, 4, and 3, would fall within the A1 band, with Grades 5 and 4 in particular representing focused, smaller steps in terms of learning goals more closely connected to a formal learning environment. Such a perspective is indeed encouraged by the CEFR (Council of Europe, 2001, pp. 31-33). Following the approach suggested by the CEFR, Grades 5, 4, and 3 were considered to be relevant to a branching A1.1, A1.2, A1.3 distinction.

Reviewing the information collected over all three standard-setting panels, including the external validity study carried out in Europe, it is clear that the major distinctions on the CEFR scale posited as being relevant for the EIKEN grades hold.

The Grades progress in clearly distinct ways beginning with Grade 3 at the upper end of the A1 band, with Grade Pre-2 being considered relevant to A2, Grade 2 to B1, Grade Pre-1 to B2 and Grade 1 to C1. For the two lower grades, 3 and Pre-2, the cutscores for determining entry into the relevant CEFR levels fell just above the pass mark used to determine certification at these grades. From Grade 2 upwards, the cutscores fell under the pass marks for certification. The cutscores set for Grades Pre-1 and Grade 1 by the first standard-setting panel indeed indicated that test takers achieving certification at these levels would in fact have already demonstrated a strong performance at the relevant CEFR level. The standard setting procedures demonstrated that panels of relevant experts with training and familiarity with the CEFR were able to set plausible cutscores when evaluating the tests against the CEFR, and generally felt confident in their ability to do so once training had been provided.

6.2.4 The interaction between RQ1, RQ2, and RQ3

The investigation of each research question has involved a multi-faceted approach to data collection and analysis *within t*he methodology for that question. What is striking from the separate discussion of the results for each question above is the clear demonstration that no one research question on its own would provide definitive, or even extensively useful, information for the purposes of justifying the uses and interpretations of the tests, communicating the possible interpretations to test users, or incorporating the results into ongoing test validation and development procedures. These goals are only achieved by *integrating* the discussion of the results from each of the research questions into a coherent evaluation of how these aspects interact. Indeed, the discussion of the results above often required reference to the results and methodology of *other* research questions to make any substantive claims or judgments. For example, the criterial contextual and cognitive features of each grade, such as the lexical threshold levels discussed above, carry much more interpretative meaning when associated with the relevant CEFR levels. It is then possible to see the substantive links between the grades, the criterial features that distinguish them, and their relevance to the wider TLU domain. The relationship to a widely used proficiency framework allows us to discuss these grades and the

appropriacy of their uses and interpretations, for example for entrance requirements to English medium universities, in relation to other benchmark measures of proficiency utilized for the same purposes. It then becomes possible to compare the key criterial features of these different benchmark measures of proficiency targeted at similar levels and usages. The three research questions, which target the aspects suggested by O'Sullivan and Weir (2011) as constituting the core of a validity argument, clearly interact to provide substantive meaning to interpretations of test performance across the EIKEN grades.

6.2.5 Subsidiary goals

6.2.5.1 Subsidiary goal 1: applying international standards to a local context

In Chapter 1, Section 1.1, three important subsidiary goals were set for this study, and these will now be reviewed briefly below. The first of these was to consider the efficacy of applying international standards of best practice to a language testing program designed primarily for a particular local context. The specific contextual and cognitive parameters posited by the model for describing reading test tasks were able to provide sufficient data to answer RQ1 and the criterial features were able to distinguish profiles across the grades of the test. Reviewing the range of benchmarks for best practice and published results for studies employing similar indices also proved instructive, and developing similar benchmarks for the EIKEN tests on these measures did not provide major hurdles in terms of any particular features of the local context. This indicates that "international standards" can indeed be interpretable and applicable to local contexts. Perhaps this is not surprising when one considers that, for example, a large part of the ILTA Guidelines were in fact drafted in Japan by the Japan Language Testing Association (ILTA, 2007). This underscores that the distinction between international and local can easily become blurred, particularly in the case of a large scale proficiency testing program such as EIKEN. It is also possible to look at this question from the direction posed by the subtitle of this study; *demonstrating locally designed tests meet international standards*. When considering the range of evidence types and the measures used to operationalize them, this locally designed testing program has demonstrated technical properties and criterial features comparable to the international benchmarks for those measures

that the literature review and study design suggested. A locally designed test can not only meet these standards,, but can actually provide important contributions to setting international levels of best practice. Some aspects of this study clearly demonstrated this, for example the comprehensive tagging of criterial features of items, the use of innovative vertical scaling methodology, and the use of external validation in the linking to the CEFR.

6.2.5.2. Subsidiary goal 2: evaluating the socio-cognitive model

The second important subsidiary goal was to evaluate the usefulness of the socio-cognitive model itself. This point is addressed in the discussion of 6.2.4 above and also within the Implications below. Several points regarding interpreting and applying the model, however, will be discussed here. Firstly, it was noted in Chapter 1 that O'Sullivan and Weir (2011) have suggested that the relationship between the components, particularly consequences, needs to be revisited, and O'Sullivan (2011, 2012, 2015a) has questioned the temporal sequencing of the model. Interestingly, Shepard (1993, p.. 427) makes a similar point regarding a potentially fixed temporal interpretation of Messick's progressive matrix, noting that "the separate rows in Messick's table, however, make it appear as if one would resolve scientific questions of test score meaning and then proceed to consider value issues." She notes that this was not Messick's intention, but that the visual presentation of the model invites this misinterpretation. This points to a wider issue in terms of how to incorporate the necessary elements of a model within such a visual representation that necessarily entails simplification, without risking such misinterpretations. The reworking of the visual representation of the model in O'Sullivan and Weir (2001, Figure 1.2) and O'Sullivan (2011, 2012, and 2015a) in fact runs a similar risk. O'Sullivan and Weir (2011) are right to suggest that consequences do indeed need to be considered at all stages, not just as an a posteriori step. In reference to his reworked presentation of the model (2011, 2012, 2015a), O'Sullivan (2012, p. 82) suggests that "consequence is not ignored in the model but, like the target language use domain, it should be reflected in every decision made in the development process." If the aspect of consequences is not incorporated visually into the model, however, there is the serious risk of it being ignored in an overly simplistic interpretation, much in the

same way as Shephard (1993) warned regarding Messick's four boxes. Another potential shortcoming of O'Sullivan's later representations (2011, 2012, 2015a) has been the removal of cognitive processes as a criterial feature of *the test tasks* under the test system, and repositioning it under the *test taker*. While this reflects the centrality of the test taker, who will be carrying out any actual processing when interacting with test tasks, it fails to explicitly emphasize the importance of designing test tasks that would operationalize and elicit the appropriate processes as intended. As this study has shown, it is possible to design tagging criteria for test tasks to operationalize the cognitive processing suggested by the model of reading in Khalifa and Weir (2009). Equally as shown in O'Sullivan and Dunlea (2015) and Taylor (2014) it is possible to build cognitive processing into explicit test specification, and this specification can be validated, for example in the study by Brunfaut and McCray (2015).

In relation to the general temporal flow of the original model, in practice the parts of the model are more likely to require a flexible and interactive relationship involving data collection and analysis across various components in conjunction, leading to revision of parts of the testing system and further data collection. Rather than an a priori and a posteriori stage with components of the model allocated to one or the other in a linear fashion, it may be useful to adopt Kane's (2013) description of a *development stage* and *appraisal stage*, with all aspects of the model being applicable to both stages. In the development stage, in particular, the process is integrated and iterative. As this study shows, the aspects of the core elements of the model are intimately intertwined, and whether one is dealing with a large-scale, retrospective analysis in which all forms of data, including scoring data are either in hand or collected in tandem, such as this study, or the development of a new test based on the model (for example, Nakatsuhara, 2014; O'Sullivan, 2015a; Weir, 2014), the different aspects of data collection will need to be temporally integrated. Nonetheless, this study has also emphasized that the contextual and cognitive aspects of the model are in fact essential to understanding and defining the construct, and in this respect, even in a retrospective validation exercise such as this one, treating these aspects as the foundation of validation has proved useful. The role of *domain modeling, domain analysis, and domain description* in Chapelle et al (2008) and

Mislevy et al (2003) also underscores the central importance of having a clear and explicit definition of what a test is intended to measure. In this respect, this study does in fact lend support to Weir's (2005a) contention that these aspects must form the driving force in designing, validating, and interpreting the meaning of test scores, particularly with regard to targeting level-specific tests such as the EIKEN grades at distinct levels of proficiency.

This study would also suggest another point of difference with O'Sullivan and Weir (2011) and O'Sullivan (2011, 2012, 2015a) in the way these papers have suggested the reconceptualization of criterion related validity as an element of scoring validity. In this study, investigating and interpreting the relationship to the CEFR has proven a crucially important aspect of the holistic, integrated discussion of the meaningfulness and usefulness of the EIKEN grades for the purposes for which they are intended. While the investigation of this relationship certainly overlaps with scoring validity concerns, the methodology involves a great deal more, including both qualitative and quantitative data collection and analysis. This study suggests that the process of carrying out the process of validation, and interpreting the results, would benefit from criterion-related studies receiving separate and specific attention, rather than being subsumed under scoring issues, as this risks this important area being overlooked in any oversimplification of the model.

O'Sullivan's more recent discussion of the model (2015c) suggests a useful way of overcoming some of these issues. In this revised presentation of the model, the test system comprising a descriptive taxonomy for defining both the test tasks (including the crucial aspects of both contextual and cognitive validity parameters) and the scoring system are visually presented as being situated within a wider context. The test taker, along with all key stakeholders, is presented as also being part of this context, within which stakeholders and test system are viewed as having an interactive relationship, with consequences and impact operating in both directions—from stakeholders to the test system, and from the test system to stakeholders. This offers a useful way of incorporating consequences in the model without restricting it to a temporal afterthought. It also offers a way of incorporating visually into the model a crucial step in the validation process which in its previous visual presentations has remained more implicit than explicit: reference to evaluating

criterial features of the TLU domain tasks using the same framework as for test tasks. The need to extend the process of validation to this crucial phase is discussed further below under Implications, and it is suggested that visually referencing the TLU and context of use, as O'Sullivan (2015c) does, would facilitate this step.

While models must not remain static, and the iterative application and evaluation of them adds depth and refines their applicability, it is also important during that review and change process not to lose sight of key aspects. The socio-cognitive model was chosen for this study based partly on the growing body of documented application which provides a common frame of reference across studies and aids comparability. Future revisions and re-interpretations need to maintain the link through terminology and conceptual presentation to the foundation studies which have been carried out so far in order to maintain the benefits of the model noted in relation to the criteria of relevance, transparency, interpretability and comparability.

6.2.5.3 Subsidiary goal 3: applicability to operational test development

The scale of data collection managed for the study emphasizes that the pragmatic choices in terms of measures selected for defining contextual and cognitive validity parameters is indeed useful for application on a large scale to ongoing test development and production. The measures selected covered a reduced range of features compared to some other studies employing the socio-cognitive model, but this was due to the specific intention to choose measures which would meet the four criteria of relevance, transparency, interpretability, and comparability, and in particular be accessible to item writers and test production staff. The use of ongoing focus groups to interactively review and refine the tagging criteria helped to ensure the application of those criteria on a large scale. The study has demonstrated a core set of procedures and measures which are amendable to large scale use, can clearly define criterial features of the different grades, and have relevance to interpreting the substantive meaning of the grades in terms of proficiency. The integration of the aspects of validity demonstrated by the three research questions offers the potential to clearly define features which are accessible to item developers and reviewers to ensure the ongoing development of test tasks appropriate for the intended levels and

which are comparable in terms of both criterial features and empirical difficulty across different forms of the same grade. The content of these empirically distinct levels can also be described and communicated in terms of a widely used framework of proficiency, and this in turn gives confidence that the content and difficulty distinctions are not just arbitrary but are indeed relevant to wider interpretations of proficiency.

6.3 Limitations

The limitations are intimately connected to several of the design choices made to facilitate the distinctive advantages posited for the study, and to some extent, they are flipsides of the same coin. As noted above, the scale and scope of data collection and analysis undertaken for each research question have been central to the generation of robust results generalizable to the operational testing program. That scale underpins the distinctiveness of the contribution of this study to the literature on applications of the socio-cognitive model. But it is also this very scale which has imposed limitations.

In terms of the ultimate goal of contributing to a comprehensive, clear, and coherent validity argument of the EIKEN testing program, practical considerations required this study to limit its focus to one component of the testing program, rather than addressing all four skills. The focus of data collection focused on the aspects identified as core components of a validity argument, but this entailed not addressing other aspects, most importantly the issue of washback and consequences. As noted in Chapter 1, positive impact has been an explicit aim and assumption of the testing program from the outset, and so must be included as an important element of validation research to establish the extent to which such claims can be upheld. This study has demonstrated how the core aspects of validity addressed through the three research questions can and indeed must be integrated to support meaningful interpretations. At the same time, the study has not been able to address the integration of this information with the investigation of impact, nor has it been able to do this across all skill components for all grades in order to construct the fuller validity argument which is required. In terms of the research questions, the scale of

the study has also imposed particular constraints for the methodology employed for each. As noted in Section 6.2.4.2 above, creating a fuller validity argument will entail creating a richer description of the target language use domain tasks posited as being relevant to each grade, utilizing the same kinds of features employed for the description of the tasks in the tests. This study has focused on building a very detailed picture of the test tasks, but future work will need to more clearly demonstrate the relevance of the test tasks to TLU domain tasks which share the same features.

In terms of vertical scaling, the principle limitation is the use of grammar and vocabulary items for creating overlapping linking items in the common-item, non-equivalent groups design, and also for facilitating the horizontal equating back to the vertical scale for previously administered live data sets. The rationale for the use of these items is explained in Chapter 4. The point made in the literature review are reiterated here that vocabulary and grammar have consistently been shown to be good predictors of proficiency across all skills, especially reading. Nonetheless, using common items consisting of a range of task types more representative of the full spectrum of tasks in the test may have an impact on the resulting vertical scale. As noted above, however, there is no consensus model for vertical scaling, and the methodology was chosen to balance theory with pragmatism and facilitate the large-scale retrospective calibration of past data. The methodology has been fully documented, which will enable future, ongoing validation of the results obtained from this large-scale study.

The investigation of the CEFR faced constraints that are in fact inherent in many standard-setting situations. Principally the constraint on time for participants in the panels means decisions need to be taken in terms of the focus and balance of activities. These decisions need to ensure participants are able to undertake the necessary familiarization and also have enough time during the sessions to engage in discussion and the actual processes of standard setting. This was compounded in the case of the first standard-setting panel in particular by the lack of familiarity of participants with the CEFR. The solutions adopted in this study, including the use of self-access booklets to aid preparation for the meetings, went some way toward ameliorating these issues, as the feedback from participants demonstrated—and in

fact may provide useful suggestions for others facing the same constraints. The external validity study carried out in Europe also adds weight to the claim that it is possible to achieve a suitable understanding of the CEFR for the purposes of standard setting with panels in Japan. This external validity study, however, due to restrictions on resources and its nature as an additional, but not the main, source of evidence, was also necessarily limited in scope, covering only Grade Pre-1. Face-to-face standard setting is by its nature a necessarily small scale affair. This limitation is not specific to this study, but is a general limitation of most applications of standard setting. The number of judges in the panels in this study certainly complies with recommendations and reported best practice in the literature. Nonetheless, it would be useful to investigate approaches to standard setting which may allow the contribution of larger numbers of participants and the collection of larger numbers of judgments of a wider span of test forms.

6.4 Implications

The study is the most comprehensive collection and analysis of evidence in relation to the EIKEN testing program ever undertaken, a program which is large scale and clearly important in the educational and social context in which it has been developed and used. In that respect the study has provided clear and compelling evidence in support of key assumptions underlying the testing program. The separate Grades do represent empirically distinct levels and these levels can be described both in terms of measureable criterial features and in reference to a widely used descriptive framework of proficiency. The utility and relevance of the tests for the purposes for which they are used is to some extent already implicitly supported by its widespread acceptance, and accountability and transparency are facilitated by the public release of test materials. Nonetheless, this study marks the first time that these claims can be supported empirically through a detailed framework of theory-driven data collection and analysis, and demonstrated comprehensively by the integrated interpretation of the results of that analysis. For reading, vocabulary, and grammar components of the First Stage tests, it has thus clearly established a sound basis for the key claim of the testing program: that the seven grades do mark important stages

in a coherent, common frame of reference defining a common construct of EFL proficiency. Having provided a principled justification for this core aspect of the program, the study thus allows, and indeed entails, research to focus now on two ongoing parallel strands. The approach used here for reading can now be employed to investigate how all components fit within this coherent EFL proficiency construct. At the same time, research should also look in more detail *inside* each proficiency level, holding up a magnifying glass, as it were, to individual grades, and components within those grades, to utilize the tools provided by this study to more closely define, review, and where necessary revise features relevant to the specific uses and interpretations of each grade.

The evidence noted above, and the contribution that it makes towards a coherent justification of the uses and interpretations of the EIKEN tests, has been essentially a retrospective stock-taking exercise. The substantial scale of the data collection has derived very robust results that provide a large degree of confidence in the interpretations relevant to each grade. However, an important goal of this study was to also look forward, and to use the results of the study to inform the clear specification of test tasks at each level. The measures used in this study, particularly the criterial contextual and cognitive parameters derived through automated analyses tools and human judgments, provide the means for doing this. The measures have demonstrated that they can be employed on the scale required for ongoing operational use. What is most important of course is that the measures have proven useful for distinguishing between the profiles of criterial features for each grade, and equally importantly have been demonstrated to have clear relevance to outside benchmarks referenced in a wide body of literature. The study has demonstrated that these measures can, and indeed should, be used in explicit specifications to ensure the ongoing and consistent production of test tasks targeting the levels intended by the test developer. A further important potential of such explicit test specification is to facilitate the very washback that has been an integral part of the test design and aims from the program's beginning. Explicit description of the criterial features relevant to each level, particularly for aspects such as vocabulary, can provide clear learning goals and inform both formal educational contexts and facilitate autonomous learning.

To illustrate this point, consider the Grade 2 test, which the study has demonstrated is relevant to a B1 level on the CEFR. This grade was clearly shown to be a pivotal level in terms of transitioning from the more restricted lower grades to the more advanced Grades relevant to B2 and C1 on the CEFR. This interpretation lends support to claims by Yanase (2009) of the pivotal nature of this level for EFL learners, and of the importance for learners at this level to expand the nature of their exposure to and use of the target language in order to move through this transition zone. The extra detail of what constitutes the profile of criterial features of reading texts at this level, and what distinguishes them from levels below and above has the potential to greatly elucidate the kind of reading materials and the kind of reading tasks and activities that learners need to engage in in order to move beyond this level. While this brief discussion has focused on Grade 2, the same principle holds true for the other grades. The integration of information across the three research questions derived from this study has the potential to not only inform test task specification but to help elucidate criterial features of proficiency relevant to those levels which can facilitate learning and teaching.

The study has demonstrated that the socio-cognitive model provides a clear and powerful framework within which to design and implement a comprehensive validation research agenda for language testing. Suggestions for how the model can be adapted to take account of recent discussions were provided above in Section 6.2.4. One of the clear advantages of the socio-cognitive model that was noted in Chapter 2 was the body of literature documenting its application. It is this body of work that makes the model relevant to the four subsidiary criteria. What distinguishes this study from previous applications is the scale of the data collection and analysis. Measures were deliberately selected from previous studies to create a core group of measures with wide currency and application that can be relatively easily applied in an operational context to large amounts of data. This has facilitated the collection of data on an unprecedented scale for applications of the socio-cognitive model. While the studies in Khalifa and Weir (2009) are comprehensive, the expert analysis of parameters there was largely restricted to a smaller number of test forms considered representative of typical test tasks. For this study, *all* operational test forms across all seven grades for all years from the latest

major revision were tagged for the expert judgment criteria. The use of automated analysis tools utilized a body of texts which included one full test form from every administration across all years since the most recent revision. Although Weir et al (2013) carried out an extensive historical analysis using such tools, this was restricted to one level of the Cambridge tests, CPE. This study has analyzed a comprehensive corpus of reading texts from Grade 4 through 1, and for vocabulary analysis for all grades across the years noted above.

In terms of vertical scaling, Brown et al (2012) and Wu (2012) carried out useful studies to demonstrate the efficacy of this approach to linking level specific EFL tests on a common framework. However, those studies employed experimental data collection designs using small numbers of test forms outside the operational testing program, due to the constraints under which each study was being carried out. This study has employed an innovative approach to overcoming the inability to incorporate vertical links within the operational test forms to retrospectively calibrate multiple test forms across all grades across multiple years, creating a robust and powerful common metric on which to compare the empirical difficulty of the test forms.

In terms of the criterion-referenced aspect of validity, the approach to linking to the CEFR has demonstrated innovative ways of overcoming the practical constraints noted under the limitations section. In particular this study has demonstrated how Kane's (2001b) exhortation to validate standard setting through replication with different methods and participants can be achieved in practice, within an already extensive standard-setting agenda. In relation to linking to the CEFR, Kane's recommendation is particularly relevant, and yet there are few documented cases of linking studies taking up the call to obtain external validity evidence by investigating if the local interpretation of the CEFR established in one standard-setting project holds across panelists, methods, and national contexts for the same test. This study has clearly demonstrated a plausible approach to doing so, and furthermore underscored the usefulness of carrying out such studies.

6.5 Final thoughts

The EIKEN testing program, as noted in Chapter 1, has been in large-scale, operational use in Japan for decades. The tests have evolved within that context, and constitute an important part of the educational and societal fabric of Japan in relation to language learning and teaching. At the same time, recent changes have meant that the EIKEN tests are now used for some purposes outside that original context, and local educators also want to interpret the results from the tests not just in terms of previously held local interpretations but with reference to descriptions of proficiency with wider currency. The tests have been developed and maintained by teams of production staff working in close concert with important stakeholders and local educators. To some extent the approach to test construction and development could be characterized as lying toward the connoisseurship side of the connoisseurship and empiricism dichotomy suggested by Green et al (2010). What is notable is that even without explicit reference to many of the empirical measures employed here to define criterial contextual and cognitive features, the teams producing the tests have achieved a very high standard in terms of maintaining clear criterial distinctions within and across grades. The processes employed clearly achieved very consistent results. What is equally clear, however, is that without the approach to empiricism employed in this study, it would be impossible to confidently *demonstrate* the degree of consistency and quality of that work. It is often suggested that Ebel (1951) emphasized the creative nature of item writing, suggesting that it was an art rather than a science (Haladyna, Downing and Rodriquez, 2002). In reference to the results seen in this study for the EIKEN tests, it is suggested that rather than art, *artisanship i*s a more appropriate term for the skilled craftsmanship that has contributed to the production of the tests to such high standards. At the same time, it is the science of language testing which has made it possible to elucidate and evaluate the efficacy of that skilled craftsmanship. The many tools which the field now has at its disposal in the application of that science provide a means of ensuring the consistent production of test tasks capable of measuring a clearly defined and meaningful construct of English language proficiency.

In Chapter 2 it was suggested that the socio-cognitive model provides a way of operationalizing Messick's conceptualization of the evidential basis for the

validation of score interpretation and use, and overcoming the ambiguities and limitations that have been noted for the unified approach to validity he proposed. The model has provided a coherent way of defining a realistic, comprehensive, and targeted research agenda for the validation of this large-scale EFL testing program. It has provided a coherent methodology for collating, organizing and evaluating the results of that research agenda, and indeed to "touch all the bases" in Messick's terms. While this study was not designed to cover all of those bases, the model nonetheless clearly identifies a road map for what other aspects need to be included in order to do so. The study has thus provided a clear demonstration of the efficacy of the socio-cognitive model to help design an agenda to answer the question of *how much of what is needed to justify the uses and interpretations of a language test*?

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Jamie Dunlea Signature: 

Date: 22 December 2015

Appendix A Structure of the EIKEN First Stage tests

Grade 1

Time allotted: Reading & Writing（100 minutes）／ Listening（30 minutes）

Skill area	Task Label [13]	Task	Format	Items	Notes
Vocab	W1	Sentence completion	25 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	25	Multiple-choice (Four printed options) [a] Items weighted to 2 points each
Reading	W2	Gap fill (passages)	2 passages from which several words or phrases have been omitted	6	
	W4	Q&A based on passages	2 passages followed by questions	6[a]	
			1 extended passage followed by questions	4[a]	
Writing	W9	200-word composition on a given topic		1[b]	Handwritten essay [b] 28 points
Listening	L1	Q&A based on dialogues	10 recorded conversations/discussions followed by questions	10	Multiple-choice (Four printed options) [a] Items weighted to 2 points each
	L3	Q&A based on monologues	5 recorded announcements, advertisements, news stories, or short lectures followed by questions	10	
	L8	Real-life Listening	5 recorded announcements, advertisements, news stories, or short lectures followed by questions (Examinees read a short description of the situation, and the question, before listening.)	5[a]	
	L9	Q&A based on long interview	1 recorded interview followed by questions	2[a]	

[13] The task labels reflect the internal labelling structure used by the test development teams at Eiken. The *W* is a translation of the Japanese *hiki-shiken*, or *written test*, and subsumes tasks in the grammar and vocabulary, reading and writing components, while *L* denotes tasks in the *listening* components.

Grade Pre-1

Time allotted: Reading & Writing（90 minutes）／ Listening（25 minutes）

Skill area	Task Label	Task	Format	Items	Response type
Vocab	W1	Sentence completion	25 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	25	Multiple-choice (Four printed options) [a] Items weighted to points each
Reading	W2	Gap fill (passages)	2 passages from which several words or phrases have been omitted	6	
	W4	Q&A based on passages	3 passages followed by questions	10[a]	
Writing	W9	100-word to a letter or e-mail on a given topic		1[b]	Response to letter or e-mail [b] 28 points
Listening	L1	Q&A based on dialogues	10 recorded conversations/discussions followed by questions	10	Multiple-choice (Four printed options) [a] Items weighted to points each
	L3	Q&A based on monologues	5 recorded announcements, advertisements, news stories, or short lectures followed by questions	10	
	L8	Real-life Listening	5 recorded announcements, advertisements, news stories, or short lectures followed by questions (Examinees read a short description of the situation, and the question, before listening.)	5[a]	

Grade 2
Time allotted: Reading & Writing（75 minutes）／ Listening（25 minutes）

Skill area	Task Label	Task	Format	Items	Response type
Grammar & Vocabulary	W1	Sentence completion	20 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	20	Multiple-choice (Four printed options)
Reading	W2	Gap fill (passages)	2 passages from which several words or phrases have been omitted	8	
	W4	Q&A based on passages	3 passages followed by questions	12	
Writing	W9	Word reordering	A five-word section of a short text is removed. The words are arranged below the text in a scrambled order.	5	After putting words into correct order, examinees indicate which words should appear in 2nd and 4th positions.
Listening	L1	Q&A based on dialogues	15 recorded conversations/discussions followed by questions	15	Multiple-choice (Four printed options)
	L3	Q&A based on monologues	5 recorded announcements, advertisements, news stories, or short lectures followed by questions	15	

Grade Pre-2
Time allotted: Reading & Writing（75 minutes）／ Listening（25 minutes）

Skill area	Task Label	Task	Format	Items	Response type
Vocab & Grammar	W1	Sentence completion	20 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	20	Multiple-choice (Four printed options)
Reading	W10	Gap fill in dialogues	6 short texts (dialogues) from which one or two phrases have been omitted	8	
	W2	Gap fill in passages	2 passages from which several words or phrases have been omitted	5	
	W4	Q&A based on passages	2 passages followed by questions	7	
Writing	W7	Word reordering	A five-word section of a short text is removed. The words are arranged below the text in a scrambled order.	5	After putting words into correct order, examinees indicate which words should appear in 2nd and 4th positions.
Listening	L2	Conversation completion	Examinees listen to short conversations and choose the best response to complete the last turn of the conversation.	10	Multiple-choice (Three recorded options)
	L1	Q&A based on dialogues	10 recorded conversations/discussions followed by questions	10	Multiple-choice (Four printed options)
	L3	Q&A based on monologues	10 recorded stories or explanations followed by questions	10	

Grade 3

Time allotted: Reading & Writing（40 minutes）／ Listening（27 minutes）

Skill area	Task Label	Task	Format	Items	Response type
Vocab & Grammar	W1	Sentence completion	15 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	15	Multiple-choice (Four printed options)
Reading	W3	Gap fill in dialogues	5 short texts (dialogues) from which one or two phrases have been omitted	5	
	W4	Q&A based on passages	Poster, advertisement, or memo	2	
			Letter or e-mail	3	
			Passage	5	
Writing	W6	Word reordering	A sentence is provided from which six words have been removed and are scrambled. Examinees reorder the words to complete the sentence. A Japanese translation of the sentence is also provided.	5	After putting words into correct order, examinees indicate which words should appear in 2nd and 4th positions.
Listening	L7	Conversation completion	Examinees listen to short conversations and choose the best response to complete the last turn of the conversation. An illustration provides contextual information about the situation.	10	Multiple-choice (Three recorded options)
	L1	Q&A based on dialogues	10 recorded conversations followed by questions. All conversations are heard twice.	10	Multiple-choice (Four printed options)
	L3	Q&A based on monologues	10 recorded stories or explanations followed by questions. All stories/explanations are heard twice.	10	

Grade 4

Time allotted: Reading & Writing（35 minutes）／ Listening（25 minutes）

Skill area	Task Label	Task	Format	Items	Response type
Vocab & Grammar	W1	Sentence completion	15 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	15	Multiple-choice (Four printed options)
Reading	W3	Gap fill in dialogues	5 short texts (dialogues) from which one or two phrases have been omitted	5	
	W4	Q&A based on passages	Poster, advertisement, or memo	2	
			Letter or e-mail	3	
			Passage	5	
Writing	W6	Word reordering	A sentence is provided from which five words have been removed and are scrambled. Examinees reorder the words to complete the sentence. A Japanese translation of the sentence is also provided.	5	After putting words into correct order, examinees indicate which words should appear in 2nd and 4th positions.
Listening	L7	Conversation completion	Examinees listen to short conversations and choose the best response to complete the last turn of the conversation. All conversations are heard twice. An illustration provides contextual information about the situation.	10	Multiple-choice (Three recorded options)
	L1	Q&A based on dialogues	10 recorded conversations followed by questions. All conversations are heard twice.	10	Multiple-choice (Four printed options)
	L3	Q&A based on monologues	10 recorded stories or explanations followed by questions. All stories/explanations are heard twice.	10	

Grade 5

Time allotted: Reading & Writing（25 minutes）／ Listening（20 minutes）

Skill area	Task Label	Task	Format	Items	Response type
Vocab & Grammar	W1	Sentence completion	15 short texts (one or two sentences/dialogues) from which one word or phrase has been omitted	15	Multiple-choice (Four printed options)
Reading	W3	Gap fill in dialogues	5 short texts (dialogues) from which one or two phrases have been omitted	5	
Writing	W6	Word reordering	A sentence is provided from which four words have been removed and are scrambled. Examinees reorder the words to complete the sentence. A Japanese translation of the sentence is also provided.	5	After putting words into correct order, examinees indicate which words should appear in 1st and 3rd positions.
Listening	L7	Conversation completion	Examinees listen to short conversations and choose the best response to complete the last turn of the conversation. All conversations are heard twice. An illustration provides contextual information about the situation.	10	Multiple-choice (Three recorded options)
	L1	Q&A based on dialogues	10 recorded conversations followed by questions. All conversations are heard twice.	10	Multiple-choice (Four printed options)
	L4	Matching	10 illustrations are provided in the test booklet. Examinees listen to three short statements for each illustration and choose the statement that best describes the action or situation in the illustration. All statements are heard twice.	10	Multiple-choice (Three recorded options

Appendix B Table of Contents from Tagging Manual

CONTENTS

Appendix C Pie Charts showing use of topics in First Stage tests

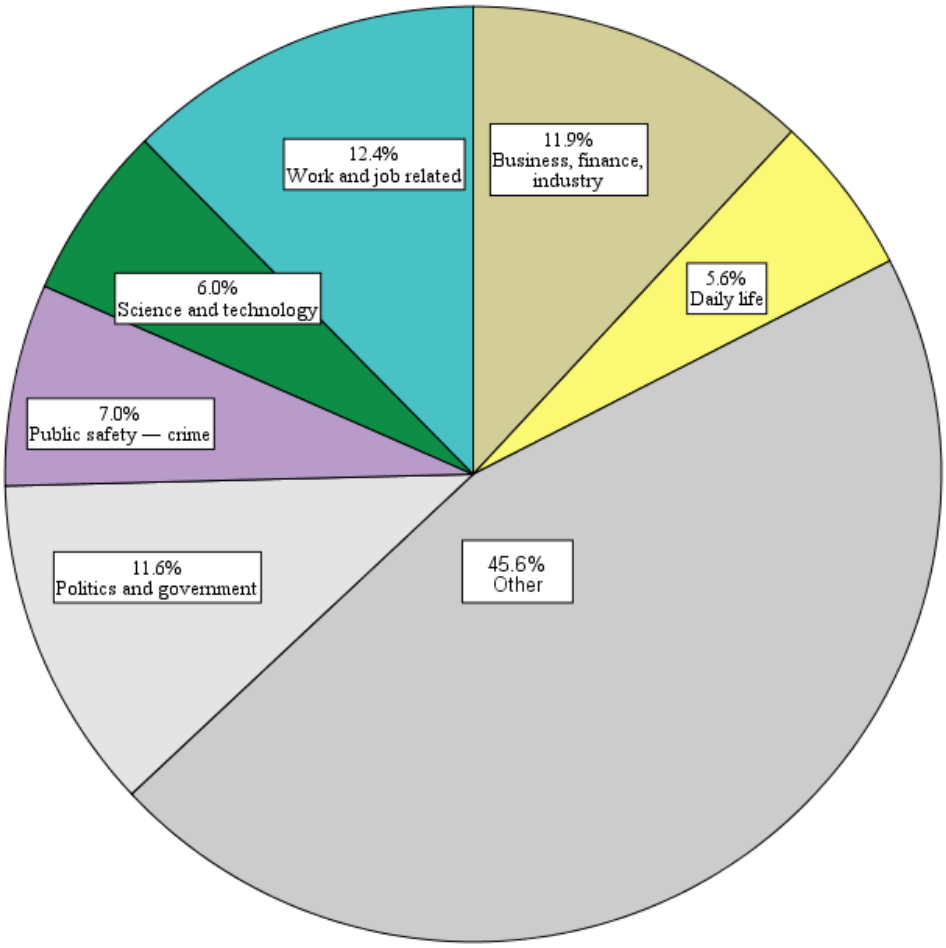


Figure C1 Topics in Grade 1 for all vocabulary and reading sections

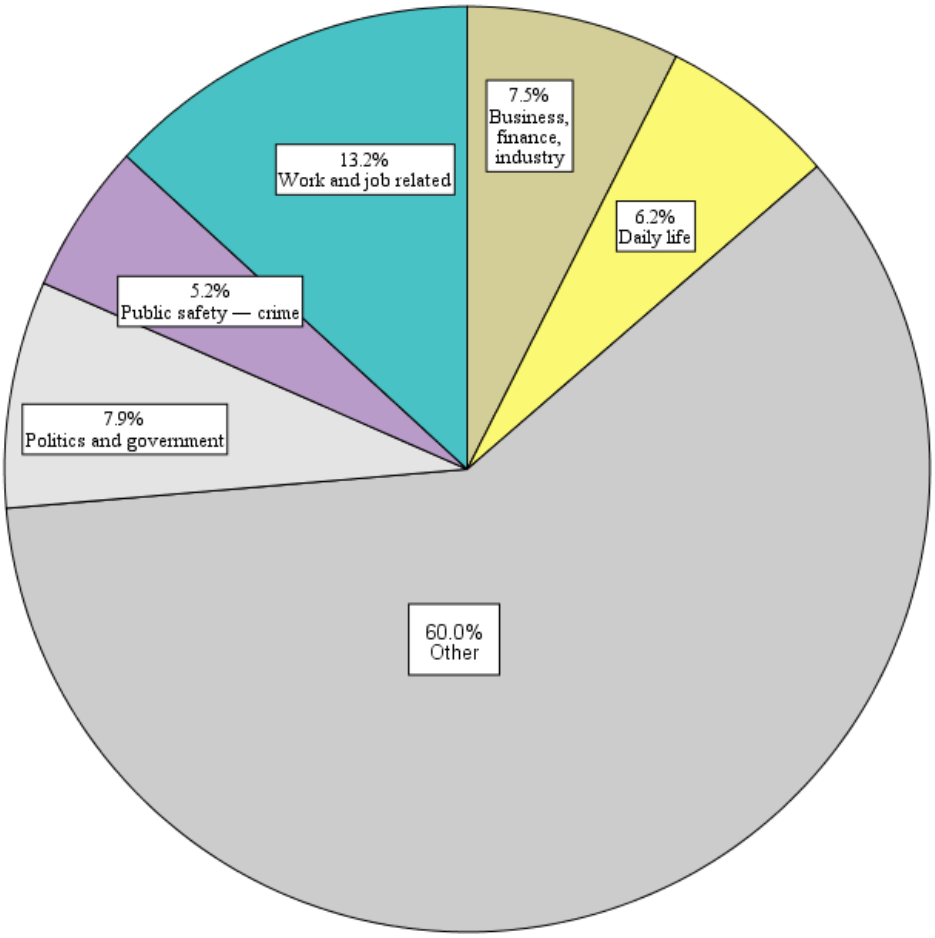


Figure C2　Topics in Grade Pre-1 for all vocabulary and reading sections

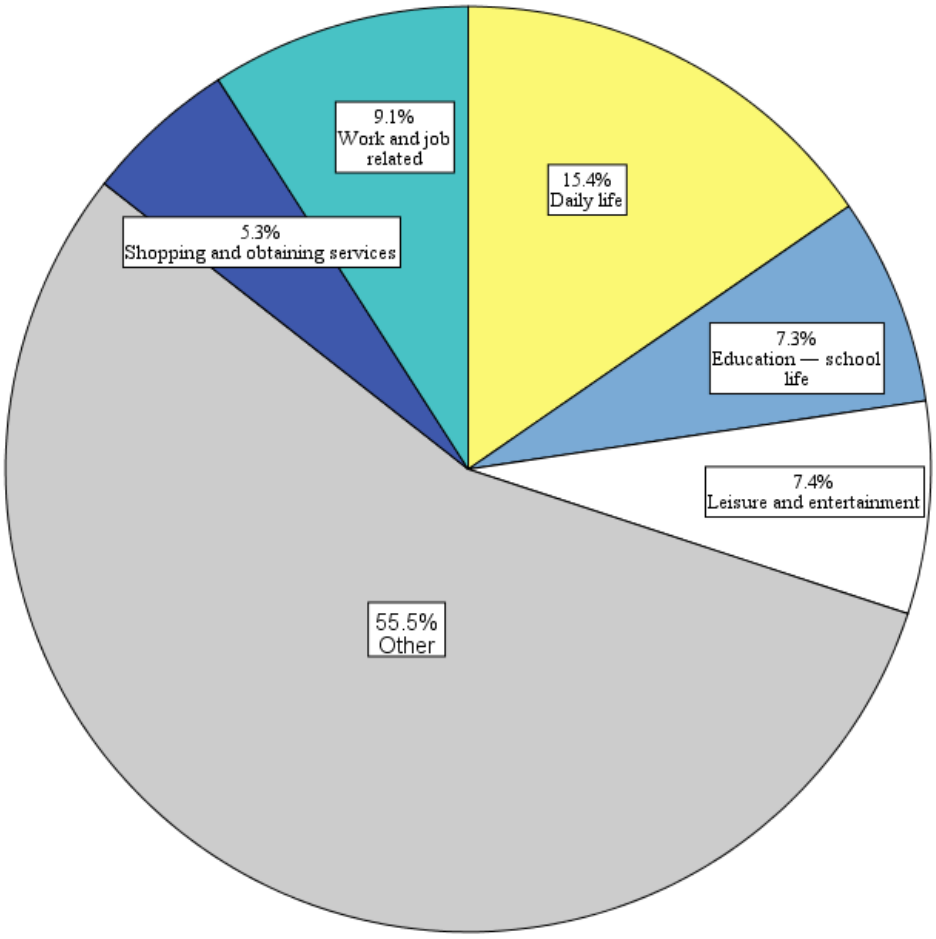


Figure C3 Topics in Grade 2 for all grammar, vocabulary and reading sections

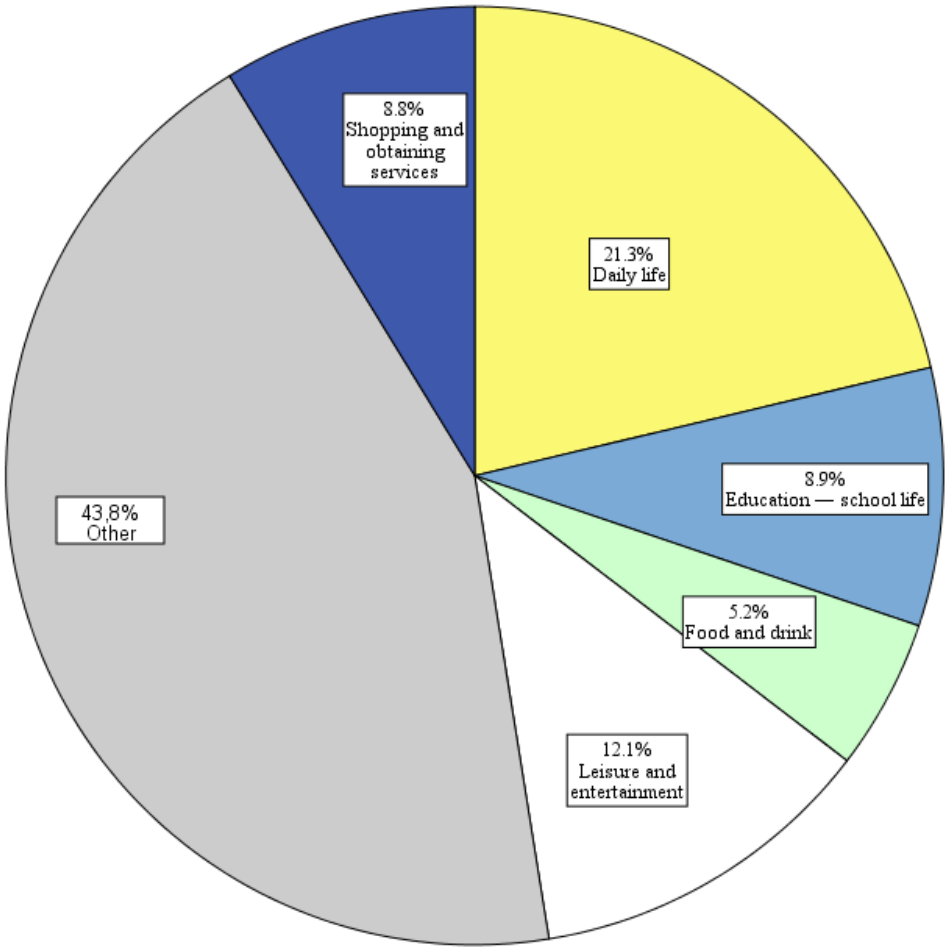


Figure C4 Topics for Grade Pre-2 for all grammar, vocabulary and reading sections

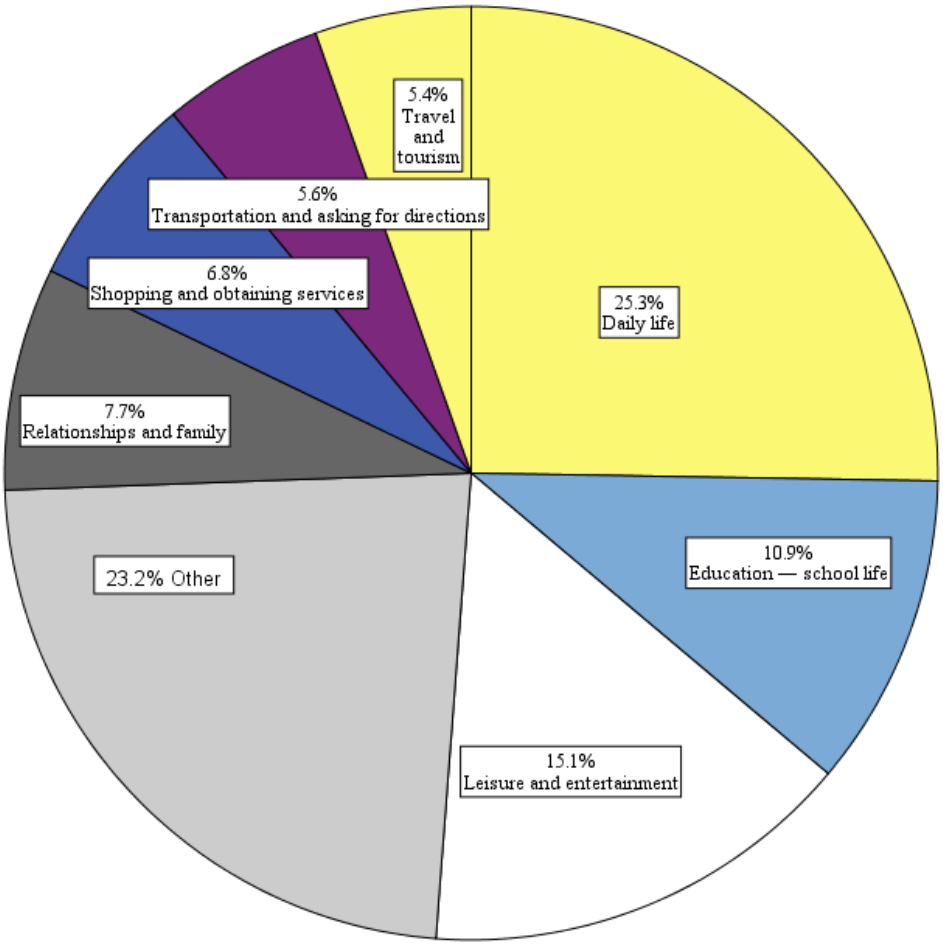


Figure C5 Topics in Grade 3 for all grammar, vocabulary and reading sections

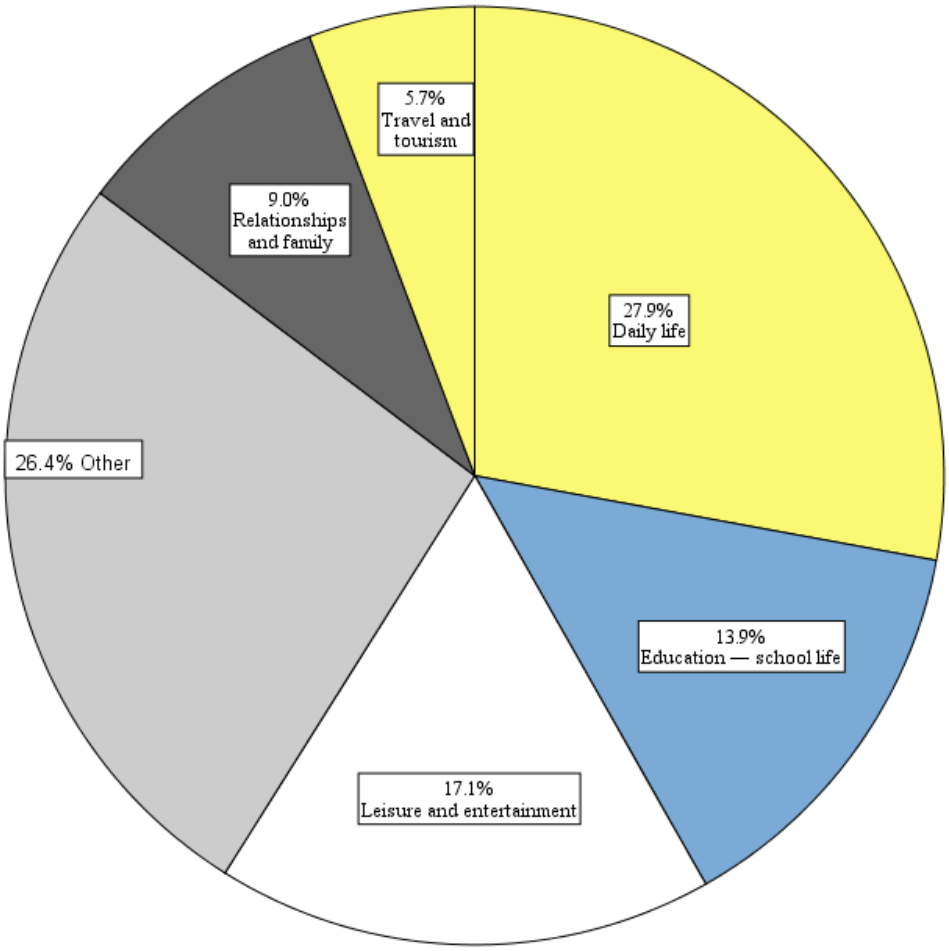


Figure C6 Topics for Grade 4 for all grammar, vocabulary and reading sections

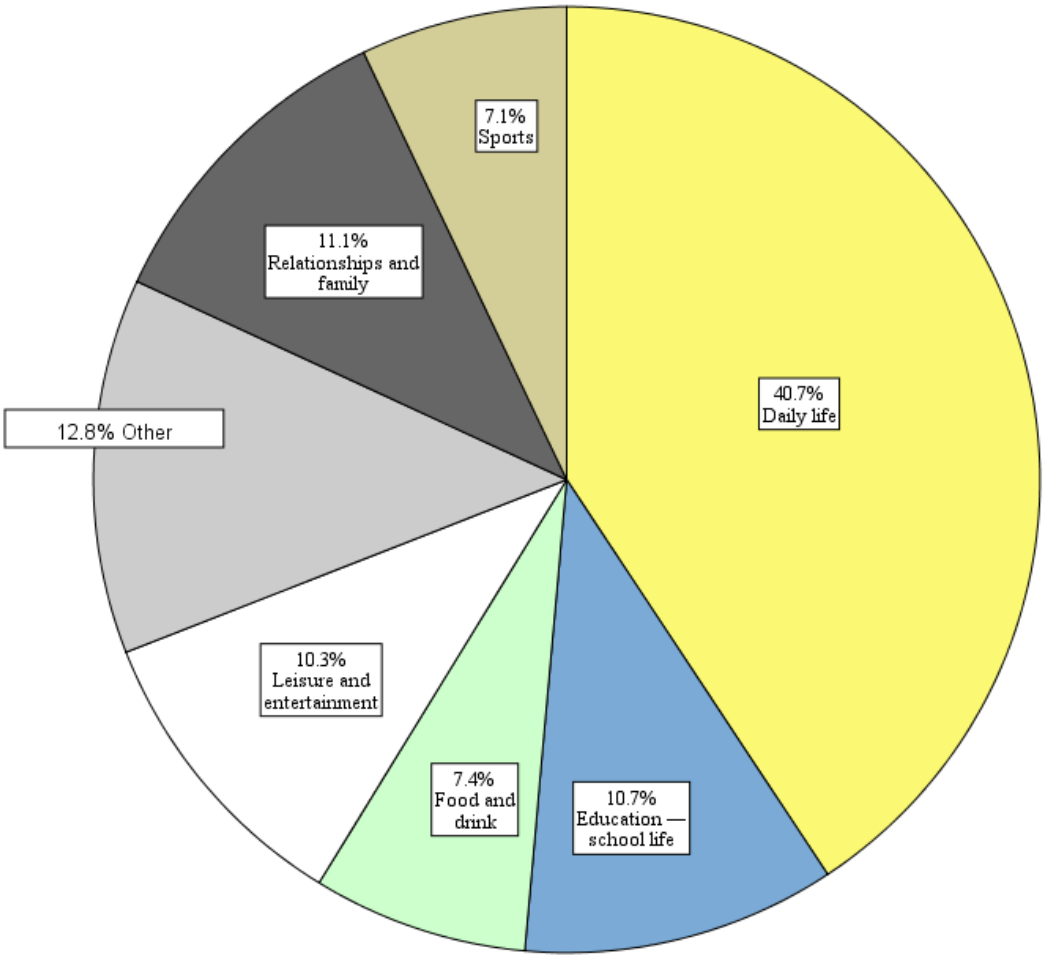


Figure C7 Topics in Grade 5 for all grammar, vocabulary and reading sections

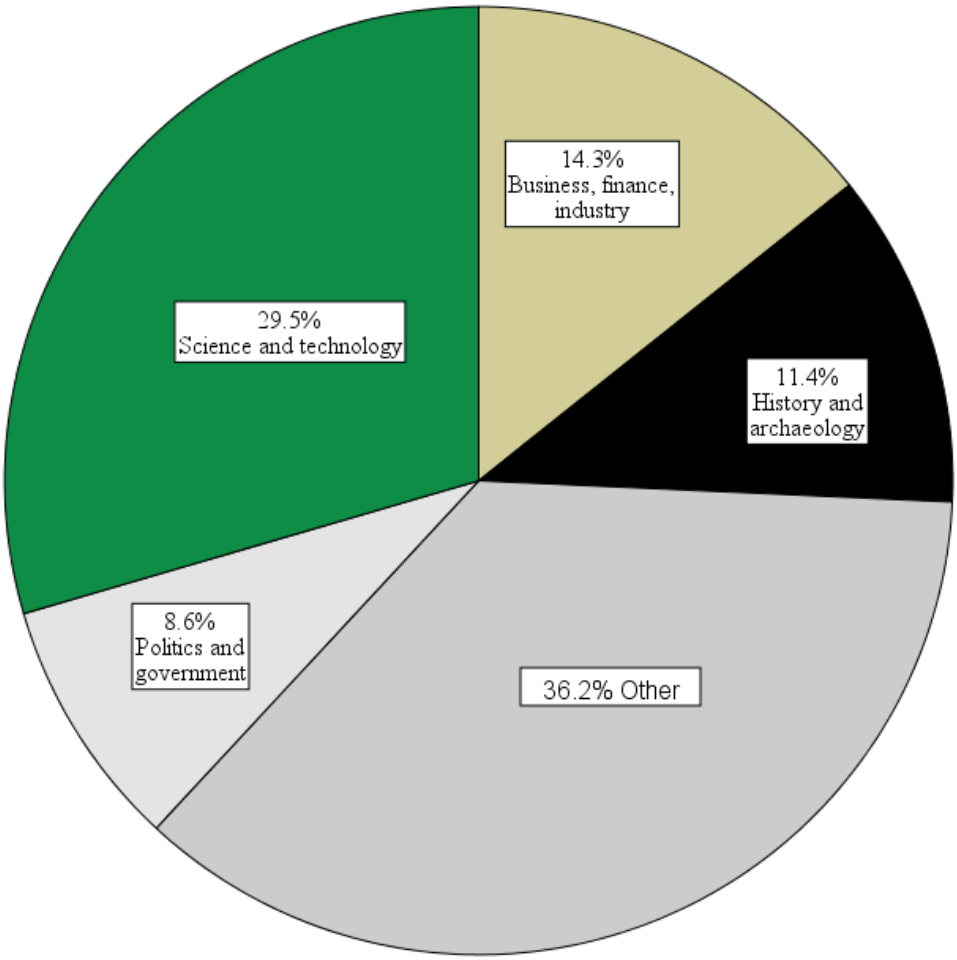


Figure C8 Topics for Grade 1 for long reading passages (W2, W4)

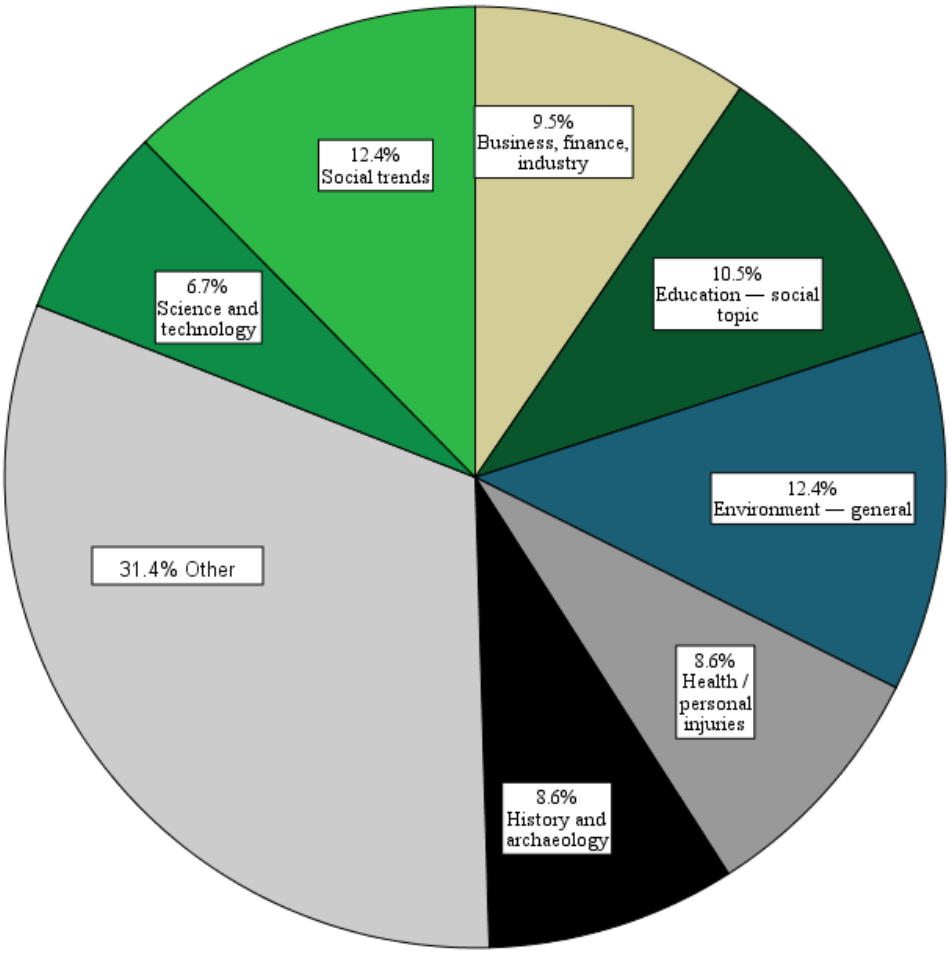


Figure C9 Topics for Grade Pre-1 for long reading passages (W2, W4)

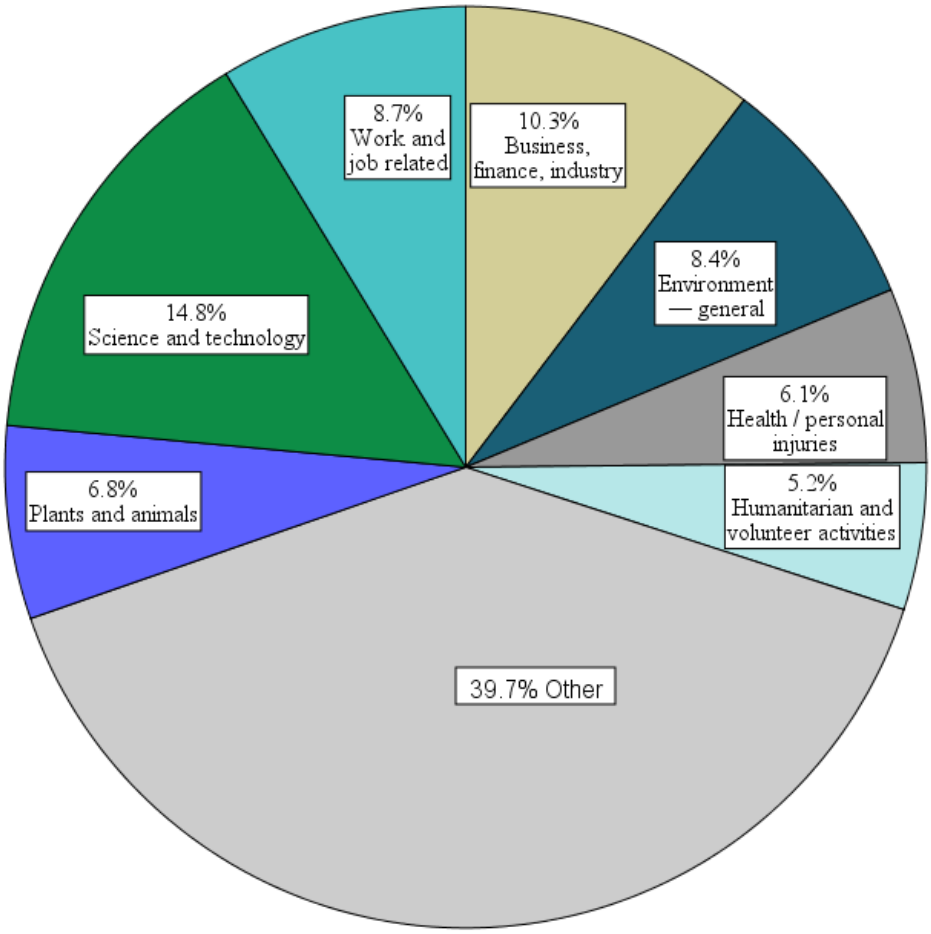


Figure C10 Topics for Grade 2 for long reading passages (W2, W4)

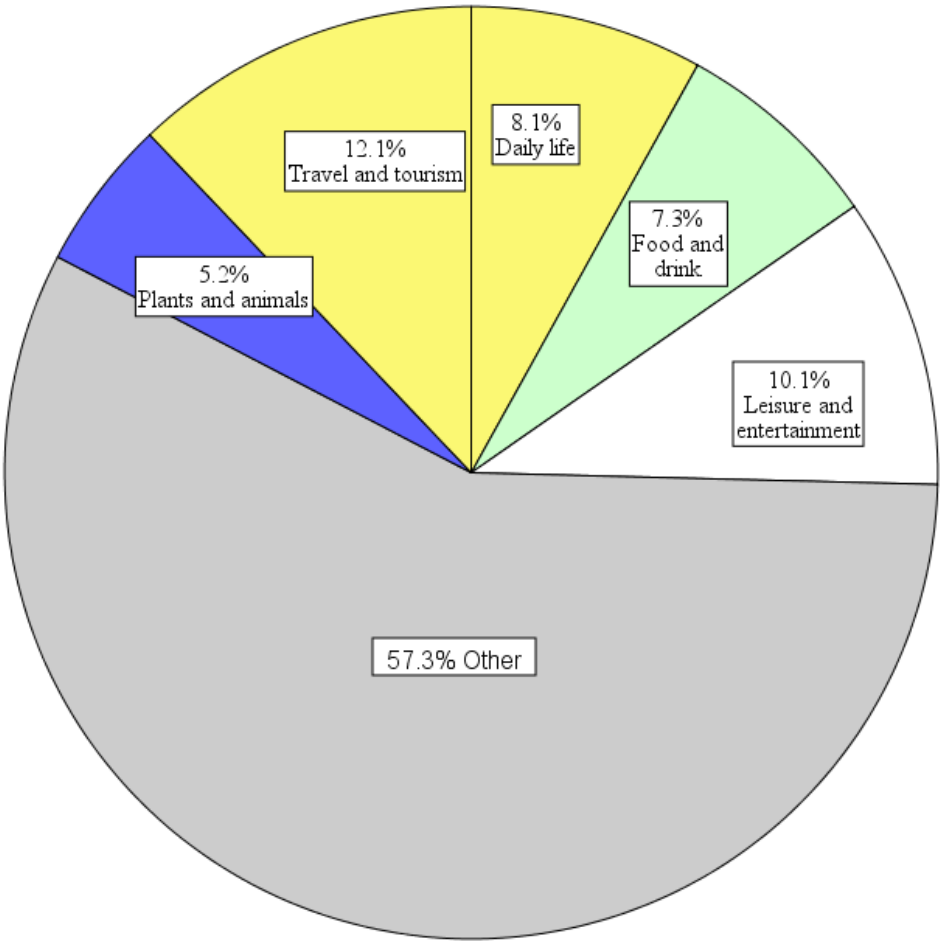


Figure C11 Topics for Grade Pre-2 for long reading passages (W2, W4)

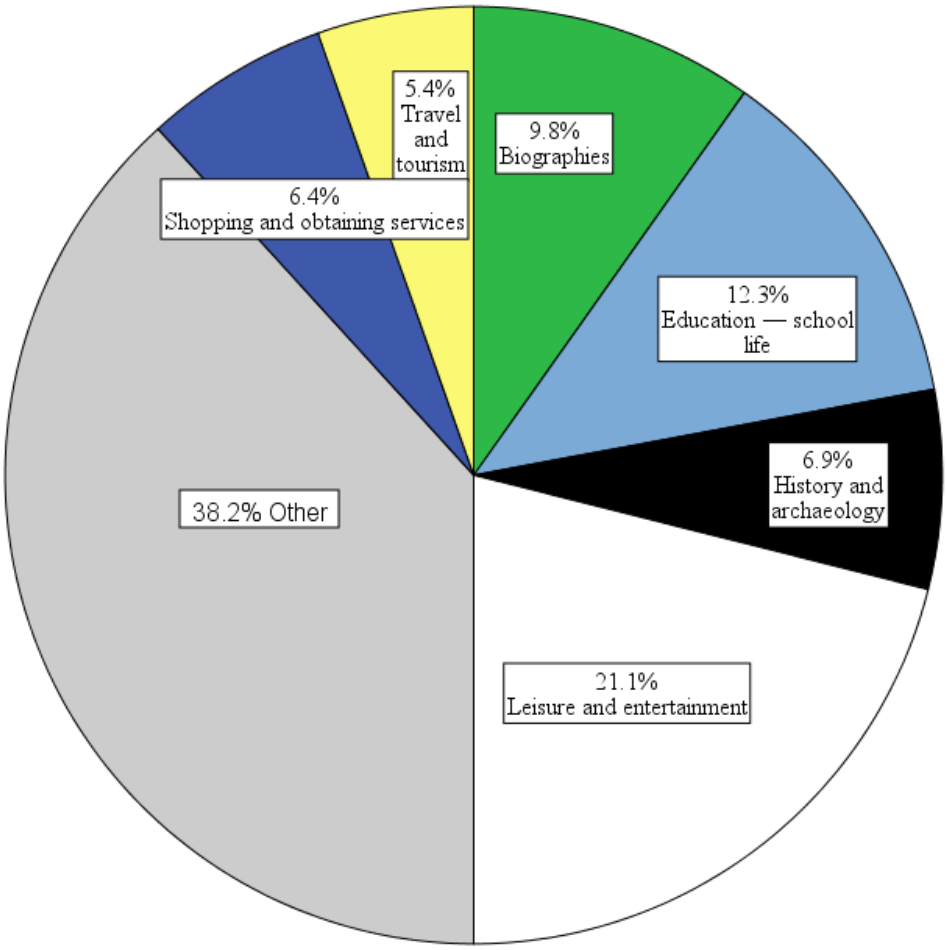


Figure C12 Topics for Grade 3 for long reading passages (W2, W4)

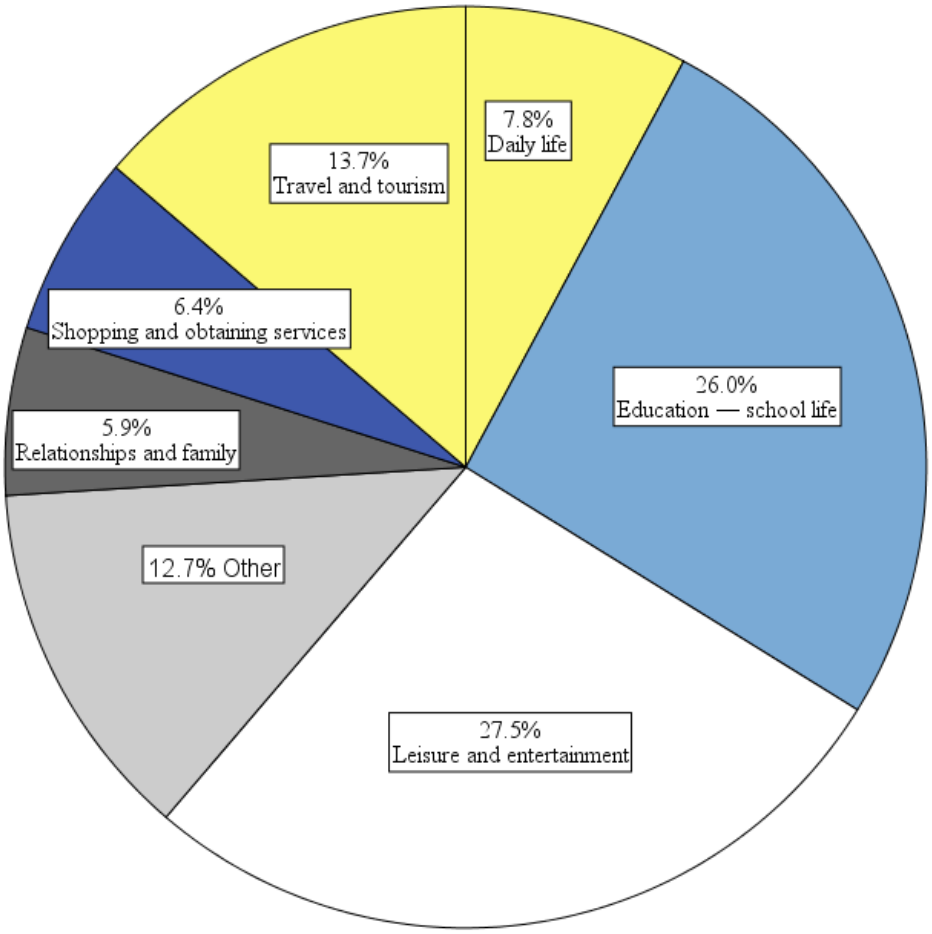


Figure C13 Topics for Grade 4 for long reading passages (W2, W4)

Appendix D Use of all topics in First Stage tests

Table D1 Use of all topics across all non-listening sections

Topic	G1	GP1	G2	GP2	G3	G4	G5
Arts and literature	2.2%	1.6%	2.3%	0.9%	0.1%	0.0%	0.2%
Biographies	1.4%	1.1%	0.4%	0.6%	1.1%	0.0%	0.0%
Business, finance, industry	11.9%	7.5%	4.8%	1.1%	0.1%	0.0%	0.0%
Culture and customs	1.0%	0.5%	1.1%	0.9%	0.7%	0.3%	0.0%
Daily life	5.6%	6.2%	15.4%	21.3%	25.3%	27.9%	40.7%
Descriptions of places and buildings	1.1%	2.5%	1.3%	1.5%	1.2%	0.5%	0.2%
Dreams and future plans	0.2%	0.5%	0.8%	0.7%	0.9%	0.6%	0.0%
Education— college life	2.5%	2.7%	1.7%	1.6%	0.2%	0.2%	0.0%
Education — school life	1.6%	2.5%	7.3%	8.9%	10.9%	13.9%	10.7%
Education — social topic	0.8%	2.4%	0.5%	0.1%	0.1%	0.1%	0.0%
Education — training and learning	0.2%	1.4%	1.0%	1.1%	1.6%	1.1%	0.2%
Environment — energy	0.5%	0.3%	0.8%	0.0%	0.0%	0.0%	0.0%
Environment — general	1.6%	3.7%	1.7%	0.4%	0.2%	0.0%	0.0%
Food and drink	0.8%	1.6%	3.1%	5.2%	4.8%	4.4%	7.4%
Health	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%

Topic	G1	GP1	G2	GP2	G3	G4	G5
Health / personal injuries	4.6%	4.6%	3.6%	4.0%	2.6%	0.8%	0.4%
History and archaeology	2.1%	2.5%	0.7%	0.4%	0.7%	0.0%	0.0%
Humanitarian and volunteer activities	1.1%	1.1%	1.1%	0.2%	0.7%	0.3%	0.1%
Leisure and entertainment	3.3%	3.2%	7.4%	12.1%	15.1%	17.1%	10.3%
Media	1.9%	1.3%	0.6%	0.3%	0.0%	0.0%	0.0%
Not clear	0.0%	0.0%	1.1%	0.4%	0.3%	0.0%	0.0%
Other	1.0%	1.0%	0.6%	0.4%	0.0%	0.0%	0.0%
Personal finances	2.5%	1.9%	1.7%	0.5%	0.4%	0.0%	0.1%
Pets	0.3%	0.3%	1.1%	1.4%	0.9%	0.5%	2.2%
Plants and animals	1.3%	1.7%	2.1%	1.2%	1.1%	0.7%	1.4%
Politics and government	11.6%	7.9%	1.8%	0.3%	0.0%	0.1%	0.0%
Public safety — accidents and natural disasters	1.7%	4.1%	1.2%	1.3%	0.1%	0.0%	0.0%
Public safety — crime	7.0%	5.2%	1.5%	0.4%	0.0%	0.0%	0.0%
Relationships and family	2.9%	1.9%	3.2%	3.3%	7.7%	9.0%	11.1%
Science and technology	6.0%	2.7%	3.7%	0.6%	0.4%	0.0%	0.1%
Shopping and obtaining services	1.4%	3.0%	5.3%	8.8%	6.8%	4.6%	3.2%
Social trends	1.1%	2.5%	0.4%	0.4%	0.0%	0.0%	0.0%
Sports	4.0%	2.7%	3.5%	3.9%	2.4%	4.4%	7.1%

Topic	G1	GP1	G2	GP2	G3	G4	G5
Transportation and asking for directions	0.3%	1.1%	3.8%	4.5%	5.6%	3.9%	2.3%
Travel and tourism	1.6%	2.7%	3.5%	5.0%	5.4%	5.7%	0.1%
Weather	0.6%	0.6%	0.8%	1.9%	1.5%	2.9%	1.8%
Work and job related	12.4%	13.2%	9.1%	4.6%	1.3%	0.8%	0.6%

Table D2 Use of all topics in longer reading comprehension passages (W2, W4)

Topics	G1	GP1	G2	GP2	G3	G4
Arts and literature	1.0%	1.0%	0.6%	0.4%	0.0%	0.0%
Biographies	1.9%	1.0%	2.3%	4.4%	9.8%	0.0%
Business, finance, industry	14.3%	9.5%	10.3%	3.6%	0.5%	0.0%
Culture and customs	3.8%	2.9%	3.2%	4.4%	4.4%	2.9%
Daily life	0.0%	1.9%	0.3%	8.1%	4.4%	7.8%
Descriptions of places and buildings	0.0%	3.8%	3.2%	3.2%	2.0%	0.0%
Dreams and future plans	0.0%	0.0%	0.0%	0.4%	0.0%	0.5%
Education — college life	0.0%	0.0%	0.6%	2.0%	1.0%	0.0%
Education — school life	0.0%	0.0%	1.0%	2.8%	12.3%	26.0%
Education — social topic	3.8%	10.5%	1.9%	1.2%	0.0%	0.0%
Education — training and learning	1.0%	1.0%	1.0%	1.6%	2.9%	2.0%
Environment — energy	2.9%	1.9%	4.5%	0.4%	0.0%	0.0%
Environment — general	2.9%	12.4%	8.4%	2.0%	1.0%	0.0%
Food and drink	1.0%	1.9%	1.9%	7.3%	2.5%	1.0%
Health	0.0%	1.0%	0.0%	0.0%	0.0%	0.0%
Health / personal injuries	4.8%	8.6%	6.1%	3.2%	0.0%	0.0%
History and archaeology	11.4%	8.6%	2.3%	2.0%	6.9%	0.0%
Humanitarian and volunteer activities	1.9%	1.9%	5.2%	1.6%	3.4%	2.5%
Leisure and entertainment	1.0%	0.0%	0.6%	10.1%	21.1%	27.5%
Media	3.8%	1.9%	1.0%	1.2%	0.0%	0.0%
Not clear	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Personal finances	0.0%	0.0%	0.0%	0.8%	0.0%	0.0%
Pets	0.0%	1.0%	1.6%	2.0%	3.9%	1.5%
Plants and animals	1.9%	3.8%	6.8%	5.2%	2.9%	0.0%
Politics and government	8.6%	2.9%	1.0%	0.0%	0.0%	0.0%
Public safety — accidents and natural disasters	0.0%	1.0%	1.0%	0.8%	0.0%	0.0%

Topics	G1	GP1	G2	GP2	G3	G4
Public safety — crime	1.0%	0.0%	1.6%	0.0%	0.0%	0.0%
Relationships and family	0.0%	0.0%	0.6%	3.6%	2.5%	5.9%
Science and technology	29.5%	6.7%	14.8%	2.0%	2.0%	
Shopping and obtaining services	0.0%	0.0%	3.9%	0.4%	6.4%	6.4%
Social trends	3.8%	12.4%	1.9%	1.6%	0.0%	0.0%
Sports	0.0%	1.0%	1.3%	3.6%	0.5%	1.5%
Transportation and asking for directions	0.0%	0.0%	1.9%	2.4%	0.5%	
Travel and tourism	0.0%	1.9%	0.0%	12.1%	5.4%	13.7%
Weather	0.0%	0.0%	0.3%	0.4%	0.0%	0.0%
Work and job related	0.0%	0.0%	8.7%	4.8%	3.9%	1.0%

Appendix E Post-hoc tests for one-way ANOVAs conducted on five linguistic features of long reading texts

Average Sentence Length		Mean Diff (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Grade 4	Grade 3	-2.77238*	0.490	0.000	-4.244	-1.300
	Grade Pre-2	-6.59190*	0.522	0.000	-8.155	-5.029
	Grade 2	-8.11048*	0.550	0.000	-9.757	-6.464
	Grade Pre-1	-12.71857*	0.738	0.000	-14.947	-10.490
	Grade 1	-14.55619*	0.924	0.000	-17.377	-11.736
Grade 3	Grade 4	2.77238*	0.490	0.000	1.300	4.244
	Grade Pre-2	-3.81952*	0.455	0.000	-5.183	-2.456
	Grade 2	-5.33810*	0.487	0.000	-6.801	-3.875
	Grade Pre-1	-9.94619*	0.692	0.000	-12.059	-7.834
	Grade 1	-11.78381*	0.888	0.000	-14.521	-9.047
Grade Pre-2	Grade 4	6.59190*	0.522	0.000	5.029	8.155
	Grade 3	3.81952*	0.455	0.000	2.456	5.183
	Grade 2	-1.519	0.519	0.059	-3.073	0.036
	Grade Pre-1	-6.12667*	0.715	0.000	-8.296	-3.958
	Grade 1	-7.96429*	0.906	0.000	-10.742	-5.187
Grade 2	Grade 4	8.11048*	0.550	0.000	6.464	9.757
	Grade 3	5.33810*	0.487	0.000	3.875	6.801
	Grade Pre-2	1.519	0.519	0.059	-0.036	3.073
	Grade Pre-1	-4.60810*	0.736	0.000	-6.831	-2.385
	Grade 1	-6.44571*	0.922	0.000	-9.263	-3.629
Grade Pre-1	Grade 4	12.71857*	0.738	0.000	10.490	14.947
	Grade 3	9.94619*	0.692	0.000	7.834	12.059
	Grade Pre-2	6.12667*	0.715	0.000	3.958	8.296
	Grade 2	4.60810*	0.736	0.000	2.385	6.831
	Grade 1	-1.838	1.045	0.504	-4.977	1.302
Grade 1	Grade 4	14.55619*	0.924	0.000	11.736	17.377
	Grade 3	11.78381*	0.888	0.000	9.047	14.521
	Grade Pre-2	7.96429*	0.906	0.000	5.187	10.742
	Grade 2	6.44571*	0.922	0.000	3.629	9.263
	Grade Pre-1	1.838	1.045	0.504	-1.302	4.977

Average Syllables per Word		Mean Diff (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Grade 4	Grade 3	-.06952*	0.022	0.031	-0.135	-0.004
	Grade Pre-2	-.11000*	0.021	0.000	-0.174	-0.046
	Grade 2	-.15238*	0.020	0.000	-0.211	-0.094
	Grade Pre-1	-.30619*	0.028	0.000	-0.391	-0.221
	Grade 1	-.36143*	0.023	0.000	-0.429	-0.294
Grade 3	Grade 4	.06952*	0.022	0.031	0.004	0.135
	Grade Pre-2	-0.040	0.021	0.427	-0.105	0.024
	Grade 2	-.08286*	0.020	0.002	-0.142	-0.024
	Grade Pre-1	-.23667*	0.028	0.000	-0.322	-0.151
	Grade 1	-.29190*	0.023	0.000	-0.360	-0.224
Grade Pre-2	Grade 4	.11000*	0.021	0.000	0.046	0.174
	Grade 3	0.040	0.021	0.427	-0.024	0.105
	Grade 2	-0.042	0.019	0.255	-0.100	0.015
	Grade Pre-1	-.19619*	0.028	0.000	-0.281	-0.112
	Grade 1	-.25143*	0.022	0.000	-0.318	-0.185
Grade 2	Grade 4	.15238*	0.020	0.000	0.094	0.211
	Grade 3	.08286*	0.020	0.002	0.024	0.142
	Grade Pre-2	0.042	0.019	0.255	-0.015	0.100
	Grade Pre-1	-.15381*	0.027	0.000	-0.235	-0.073
	Grade 1	-.20905*	0.021	0.000	-0.271	-0.147
Grade Pre-1	Grade 4	.30619*	0.028	0.000	0.221	0.391
	Grade 3	.23667*	0.028	0.000	0.151	0.322
	Grade Pre-2	.19619*	0.028	0.000	0.112	0.281
	Grade 2	.15381*	0.027	0.000	0.073	0.235
	Grade 1	-0.055	0.029	0.415	-0.142	0.032
Grade 1	Grade 4	.36143*	0.023	0.000	0.294	0.429
	Grade 3	.29190*	0.023	0.000	0.224	0.360
	Grade Pre-2	.25143*	0.022	0.000	0.185	0.318
	Grade 2	.20905*	0.021	0.000	0.147	0.271
	Grade Pre-1	0.055	0.029	0.415	-0.032	0.142

Lexical Diversity MTLD		Mean Diff (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Grade 4	Grade 3	-7.272	3.385	0.289	-17.503	2.959
	Grade Pre-2	-23.75143*	2.796	0.000	-32.133	-15.369
	Grade 2	-30.67952*	4.562	0.000	-44.661	-16.698
	Grade Pre-1	-67.12238*	4.082	0.000	-79.572	-54.673
	Grade 1	-79.09810*	6.914	0.000	-100.568	-57.628
Grade 3	Grade 4	7.272	3.385	0.289	-2.959	17.503
	Grade Pre-2	-16.47952*	3.604	0.001	-27.304	-5.655
	Grade 2	-23.40762*	5.097	0.001	-38.757	-8.058
	Grade Pre-1	-59.85048*	4.673	0.000	-73.871	-45.830
	Grade 1	-71.82619*	7.278	0.000	-94.113	-49.539
Grade Pre-2	Grade 4	23.75143*	2.796	0.000	15.369	32.133
	Grade 3	16.47952*	3.604	0.001	5.655	27.304
	Grade 2	-6.928	4.727	0.688	-21.308	7.452
	Grade Pre-1	-43.37095*	4.265	0.000	-56.282	-30.459
	Grade 1	-55.34667*	7.023	0.000	-77.052	-33.641
Grade 2	Grade 4	30.67952*	4.562	0.000	16.698	44.661
	Grade 3	23.40762*	5.097	0.001	8.058	38.757
	Grade Pre-2	6.928	4.727	0.688	-7.452	21.308
	Grade Pre-1	-36.44286*	5.585	0.000	-53.168	-19.718
	Grade 1	-48.41857*	7.895	0.000	-72.259	-24.578
Grade Pre-1	Grade 4	67.12238*	4.082	0.000	54.673	79.572
	Grade 3	59.85048*	4.673	0.000	45.830	73.871
	Grade Pre-2	43.37095*	4.265	0.000	30.459	56.282
	Grade 2	36.44286*	5.585	0.000	19.718	53.168
	Grade 1	-11.976	7.627	0.623	-35.122	11.170
Grade 1	Grade 4	79.09810*	6.914	0.000	57.628	100.568
	Grade 3	71.82619*	7.278	0.000	49.539	94.113
	Grade Pre-2	55.34667*	7.023	0.000	33.641	77.052
	Grade 2	48.41857*	7.895	0.000	24.578	72.259
	Grade Pre-1	11.976	7.627	0.623	-11.170	35.122

Lexical Diversity VOCD		Mean Diff (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Grade 4	Grade 3	0.802	3.471	1.000	-9.586	11.190
	Grade Pre-2	-18.96905*	3.720	0.000	-30.104	-7.834
	Grade 2	-22.95190*	4.669	0.000	-37.045	-8.858
	Grade Pre-1	-49.49714*	3.791	0.000	-60.849	-38.145
	Grade 1	-56.27762*	5.405	0.000	-72.707	-39.848
Grade 3	Grade 4	-0.802	3.471	1.000	-11.190	9.586
	Grade Pre-2	-19.77143*	3.641	0.000	-30.675	-8.868
	Grade 2	-23.75429*	4.606	0.000	-37.681	-9.827
	Grade Pre-1	-50.29952*	3.713	0.000	-61.426	-39.173
	Grade 1	-57.08000*	5.350	0.000	-73.374	-40.786
Grade Pre-2	Grade 4	18.96905*	3.720	0.000	7.834	30.104
	Grade 3	19.77143*	3.641	0.000	8.868	30.675
	Grade 2	-3.983	4.797	0.960	-18.420	10.454
	Grade Pre-1	-30.52810*	3.947	0.000	-42.338	-18.718
	Grade 1	-37.30857*	5.515	0.000	-54.019	-20.599
Grade 2	Grade 4	22.95190*	4.669	0.000	8.858	37.045
	Grade 3	23.75429*	4.606	0.000	9.827	37.681
	Grade Pre-2	3.983	4.797	0.960	-10.454	18.420
	Grade Pre-1	-26.54524*	4.852	0.000	-41.135	-11.956
	Grade 1	-33.32571*	6.195	0.000	-51.897	-14.755
Grade Pre-1	Grade 4	49.49714*	3.791	0.000	38.145	60.849
	Grade 3	50.29952*	3.713	0.000	39.173	61.426
	Grade Pre-2	30.52810*	3.947	0.000	18.718	42.338
	Grade 2	26.54524*	4.852	0.000	11.956	41.135
	Grade 1	-6.780	5.563	0.825	-23.615	10.055
Grade 1	Grade 4	56.27762*	5.405	0.000	39.848	72.707
	Grade 3	57.08000*	5.350	0.000	40.786	73.374
	Grade Pre-2	37.30857*	5.515	0.000	20.599	54.019
	Grade 2	33.32571*	6.195	0.000	14.755	51.897
	Grade Pre-1	6.780	5.563	0.825	-10.055	23.615

Flesch-Kincaid Grade Level		Mean Diff (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Grade 4	Grade 3	-1.91286*	0.331	0.000	-2.903	-0.923
	Grade Pre-2	-3.87762*	0.335	0.000	-4.881	-2.874
	Grade 2	-4.96571*	0.305	0.000	-5.878	-4.053
	Grade Pre-1	-8.58762*	0.472	0.000	-10.025	-7.150
	Grade 1	-9.94381*	0.457	0.000	-11.333	-8.554
Grade 3	Grade 4	1.91286*	0.331	0.000	0.923	2.903
	Grade Pre-2	-1.96476*	0.358	0.000	-3.037	-0.892
	Grade 2	-3.05286*	0.331	0.000	-4.043	-2.062
	Grade Pre-1	-6.67476*	0.489	0.000	-8.156	-5.194
	Grade 1	-8.03095*	0.475	0.000	-9.466	-6.596
Grade Pre-2	Grade 4	3.87762*	0.335	0.000	2.874	4.881
	Grade 3	1.96476*	0.358	0.000	0.892	3.037
	Grade 2	-1.08810*	0.335	0.027	-2.092	-0.084
	Grade Pre-1	-4.71000*	0.492	0.000	-6.199	-3.221
	Grade 1	-6.06619*	0.478	0.000	-7.509	-4.623
Grade 2	Grade 4	4.96571*	0.305	0.000	4.053	5.878
	Grade 3	3.05286*	0.331	0.000	2.062	4.043
	Grade Pre-2	1.08810*	0.335	0.027	0.084	2.092
	Grade Pre-1	-3.62190*	0.473	0.000	-5.060	-2.184
	Grade 1	-4.97810*	0.458	0.000	-6.368	-3.588
Grade Pre-1	Grade 4	8.58762*	0.472	0.000	7.150	10.025
	Grade 3	6.67476*	0.489	0.000	5.194	8.156
	Grade Pre-2	4.71000*	0.492	0.000	3.221	6.199
	Grade 2	3.62190*	0.473	0.000	2.184	5.060
	Grade 1	-1.356	0.583	0.207	-3.100	0.387
Grade 1	Grade 4	9.94381*	0.457	0.000	8.554	11.333
	Grade 3	8.03095*	0.475	0.000	6.596	9.466
	Grade Pre-2	6.06619*	0.478	0.000	4.623	7.509
	Grade 2	4.97810*	0.458	0.000	3.588	6.368
	Grade Pre-1	1.356	0.583	0.207	-0.387	3.100

Appendix F Descriptive statistics for metadiscourse markers

Table F1 Descriptive Statistics for 11 Metadiscourse Markers

Attitude Marker	Mean	Mdn	Max	Min	SD
Grade 4	0.03%	0.00%	0.62%	0.00%	0.14%
Grade 3	0.24%	0.00%	1.20%	0.00%	0.31%
Grade Pre-2	0.37%	0.32%	0.98%	0.00%	0.36%
Grade 2	0.46%	0.29%	1.42%	0.00%	0.38%
Grade Pre-1	0.41%	0.38%	1.04%	0.00%	0.33%
Grade 1	0.32%	0.38%	0.78%	0.00%	0.19%
Code gloss	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	0.06%	0.00%	0.63%	0.00%	0.18%
Grade 3	0.46%	0.39%	1.54%	0.00%	0.43%
Grade Pre-2	0.68%	0.65%	1.34%	0.00%	0.44%
Grade 2	0.66%	0.58%	1.13%	0.00%	0.37%
Grade Pre-1	0.36%	0.40%	0.95%	0.00%	0.28%
Grade 1	0.32%	0.20%	0.98%	0.00%	0.27%
Emphatic	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	0.33%	0.00%	1.88%	0.00%	0.50%
Grade 3	0.29%	0.35%	1.18%	0.00%	0.36%
Grade Pre-2	0.45%	0.32%	2.63%	0.00%	0.57%
Grade 2	0.41%	0.29%	1.43%	0.00%	0.37%
Grade Pre-1	0.32%	0.38%	0.61%	0.00%	0.19%
Grade 1	0.58%	0.55%	1.55%	0.19%	0.35%
Endophoric	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	0.06%	0.00%	0.65%	0.00%	0.19%
Grade 3	0.11%	0.00%	1.20%	0.00%	0.29%
Grade Pre-2	0.08%	0.00%	0.66%	0.00%	0.18%
Grade 2	0.04%	0.00%	0.54%	0.00%	0.13%
Grade Pre-1	0.12%	0.00%	0.62%	0.00%	0.18%
Grade 1	0.06%	0.00%	0.20%	0.00%	0.09%
Evidential	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	1.75%	1.88%	3.47%	0.61%	0.95%
Grade 3	0.22%	0.00%	1.89%	0.00%	0.46%
Grade Pre-2	0.49%	0.31%	2.68%	0.00%	0.69%
Grade 2	0.53%	0.29%	1.99%	0.00%	0.47%
Grade Pre-1	0.73%	0.61%	2.00%	0.00%	0.50%
Grade 1	0.66%	0.40%	1.59%	0.00%	0.49%

Table F1 continued

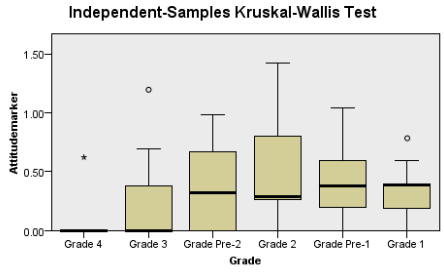
Hedge	Mean	Mdn	Max	Min	SD
Grade 4	0.32%	0.00%	2.98%	0.00%	0.72%
Grade 3	0.86%	0.77%	2.39%	0.00%	0.67%
Grade Pre-2	0.96%	0.99%	2.60%	0.00%	0.64%
Grade 2	1.14%	1.09%	3.13%	0.27%	0.72%
Grade Pre-1	1.05%	1.00%	2.27%	0.00%	0.65%
Grade 1	1.10%	0.99%	2.88%	0.00%	0.67%
Logical Connective	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	3.76%	3.18%	6.21%	1.99%	1.41%
Grade 3	5.06%	4.91%	8.88%	2.79%	1.48%
Grade Pre-2	5.04%	5.23%	7.43%	1.70%	1.31%
Grade 2	3.89%	3.63%	6.65%	2.14%	1.26%
Grade Pre-1	4.09%	4.09%	6.02%	2.66%	0.86%
Grade 1	4.47%	4.21%	6.85%	3.16%	0.94%
Person Marker	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	1.01%	0.65%	2.89%	0.00%	0.91%
Grade 3	0.02%	0.00%	0.39%	0.00%	0.09%
Grade Pre-2	0.03%	0.00%	0.32%	0.00%	0.10%
Grade 2	0.20%	0.00%	1.13%	0.00%	0.35%
Grade Pre-1	0.24%	0.00%	0.97%	0.00%	0.30%
Grade 1	0.29%	0.00%	1.16%	0.00%	0.41%
Relation Marker	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	0.87%	0.68%	2.00%	0.00%	0.73%
Grade 3	0.05%	0.00%	1.15%	0.00%	0.25%
Grade Pre-2	0.18%	0.00%	1.64%	0.00%	0.48%
Grade 2	0.11%	0.00%	0.56%	0.00%	0.19%
Grade Pre-1	0.16%	0.00%	0.83%	0.00%	0.24%
Grade 1	0.21%	0.19%	0.98%	0.00%	0.26%
Sequencing	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	1.70%	1.86%	3.33%	0.00%	0.94%
Grade 3	1.10%	1.12%	2.52%	0.00%	0.66%
Grade Pre-2	0.40%	0.32%	1.61%	0.00%	0.45%
Grade 2	0.60%	0.54%	2.23%	0.00%	0.59%
Grade Pre-1	0.38%	0.22%	1.25%	0.00%	0.34%
Grade 1	0.37%	0.19%	1.37%	0.00%	0.42%
Topic Shift	**Mean**	**Mdn**	**Max**	**Min**	**SD**
Grade 4	0.03%	0.00%	0.68%	0.00%	0.15%
Grade 3	0.02%	0.00%	0.38%	0.00%	0.08%
Grade Pre-2	0.14%	0.00%	0.96%	0.00%	0.26%
Grade 2	0.09%	0.00%	0.57%	0.00%	0.16%
Grade Pre-1	0.09%	0.00%	0.38%	0.00%	0.12%
Grade 1	0.03%	0.00%	0.20%	0.00%	0.07%

Table F2 Skewness and Kurtosis results for 11 metadiscourse markers

Grade		*Attitude marker*	*Code gloss*	*Emphatic token*	*Endophoric*	*Evidential*	*Hedge*
G1	Kurtosis	0.655	0.798	1.734	-1.572	-1.101	1.507
	Skew	0.164	1.125	1.297	0.765	0.535	1.033
GP1	Kurtosis	-0.672	-0.635	-0.587	1.835	0.486	-0.662
	Skew	0.648	0.076	-0.343	1.499	0.751	0.342
G2	Kurtosis	0.239	-1.075	1.210	11.749	3.507	1.285
	Skew	0.740	-0.304	0.955	3.436	1.604	1.123
GP2	Kurtosis	-1.292	-0.939	11.565	5.226	4.003	0.771
	Skew	0.387	-0.003	3.008	2.343	1.918	0.441
G3	Kurtosis	3.264	0.435	1.315	10.716	8.660	0.509
	Skew	1.588	0.813	1.329	3.120	2.779	0.845
G4	Kurtosis	21.000	7.678	3.322	7.572	-1.373	9.175
	Skew	4.583	2.987	1.766	2.976	0.117	2.877

Grade		*Hedge*	*Logical connective*	*Person marker*	*Relational marker*	*Sequencing*	*Topic Shift*
G1	Kurtosis	1.507	0.663	0.013	2.570	0.320	3.139
	Skew	1.033	0.987	1.200	1.520	1.122	2.202
GP1	Kurtosis	-0.662	-0.108	0.316	1.686	0.561	-0.310
	Skew	0.342	0.517	1.063	1.495	0.980	0.814
G2	Kurtosis	1.285	-0.685	3.632	1.658	2.785	2.265
	Skew	1.123	0.445	2.052	1.640	1.614	1.662
GP2	Kurtosis	0.771	1.095	7.599	6.699	1.730	3.941
	Skew	0.441	-0.587	2.978	2.776	1.401	2.049
G3	Kurtosis	0.509	0.830	21.000	21.000	-0.267	21.000
	Skew	0.845	0.865	4.583	4.583	0.518	4.583
G4	Kurtosis	9.175	-1.124	-0.611	-1.297	-0.656	21.000
	Skew	2.877	0.639	0.595	0.242	-0.332	4.583

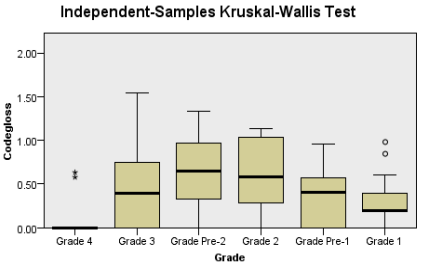
Appendix G Non-parametric test results for metadiscourse markers



Total N	126
Test Statistic	29.637
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

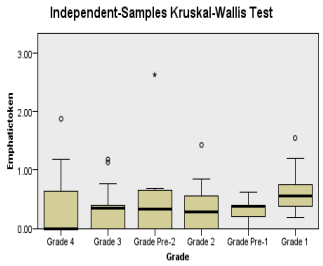
Attitude markers



Total N	126
Test Statistic	37.865
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

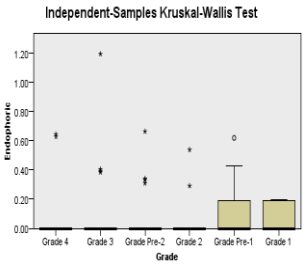
Code glosses



Total N	126
Test Statistic	10.889
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.054

1. The test statistic is adjusted for ties.
2. Multiple comparisons are not performed because the overall test does not show significant differences across samples.

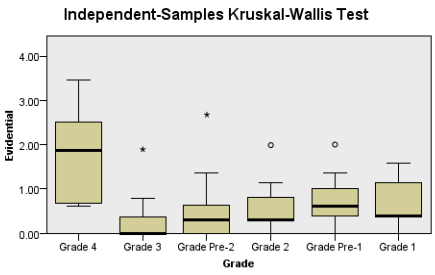
Emphatics



Total N	126
Test Statistic	7.884
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.163

1. The test statistic is adjusted for ties.
2. Multiple comparisons are not performed because the overall test does not show significant differences across samples.

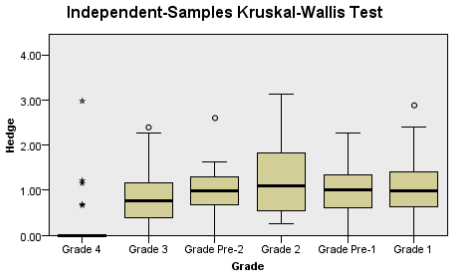
Endophorics



Total N	126
Test Statistic	46.734
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

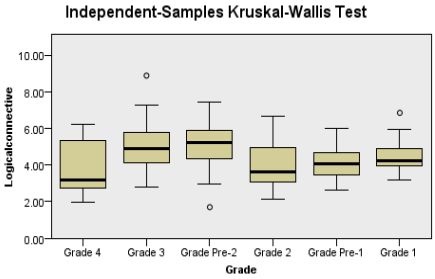
Evidentials



Total N	126
Test Statistic	24.382
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

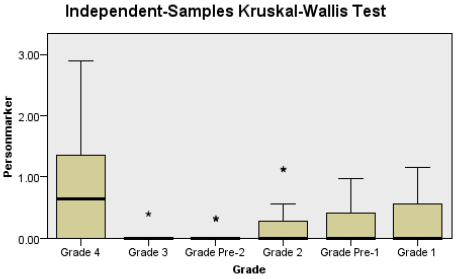
Hedges



Total N	126
Test Statistic	19.488
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.002

1. The test statistic is adjusted for ties.

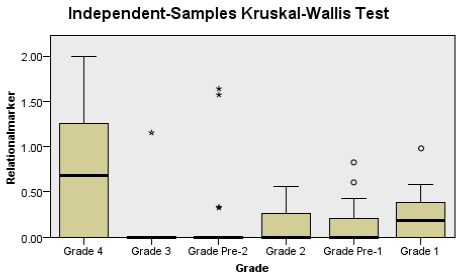
Logical Connectives



Total N	126
Test Statistic	36.846
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

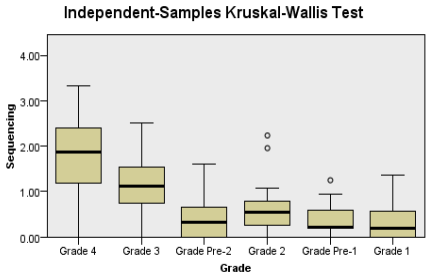
Person Markers



Total N	126
Test Statistic	33.136
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

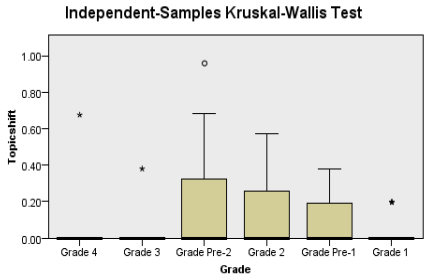
Relation Markers



Total N	126
Test Statistic	41.567
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

1. The test statistic is adjusted for ties.

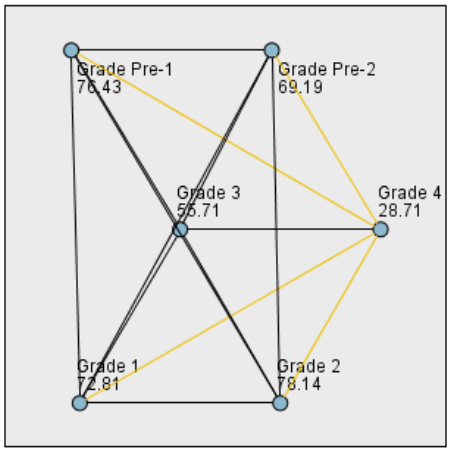
Sequencing



Total N	126
Test Statistic	13.242
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.021

Topic shift

Pairwise Comparisons of Grade



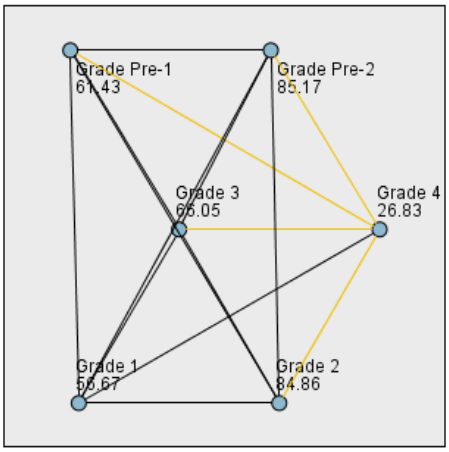
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 4-Grade 2	-49.429	10.933	-4.521	.000	.000
Grade 4-Grade Pre-1	-47.714	10.933	-4.364	.000	.000
Grade 4-Grade 1	-44.095	10.933	-4.033	.000	.001
Grade 4-Grade Pre-2	-40.476	10.933	-3.702	.000	.003
Grade 4-Grade 3	-27.000	10.933	-2.470	.014	.203
Grade 3-Grade 2	-22.429	10.933	-2.051	.040	.603
Grade 3-Grade Pre-1	-20.714	10.933	-1.895	.058	.872
Grade 1-Grade 2	5.333	10.933	.488	.626	1.000
Grade 3-Grade 1	-17.095	10.933	-1.564	.118	1.000
Grade Pre-2-Grade Pre-1	-7.238	10.933	-.662	.508	1.000
Grade 1-Grade Pre-1	3.619	10.933	.331	.741	1.000
Grade Pre-1-Grade 2	1.714	10.933	.157	.875	1.000
Grade Pre-2-Grade 1	-3.619	10.933	-.331	.741	1.000
Grade Pre-2-Grade 2	-8.952	10.933	-.819	.413	1.000
Grade 3-Grade Pre-2	-13.476	10.933	-1.233	.218	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Attitude Markers

Pairwise Comparisons of Grade



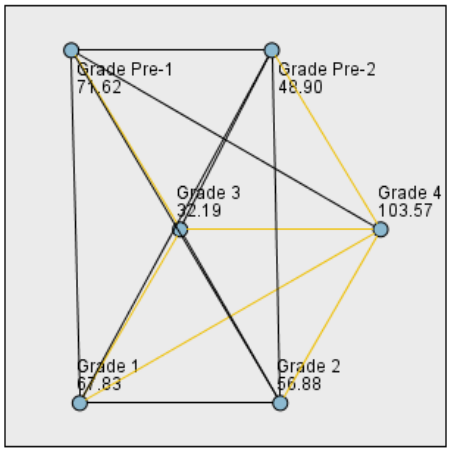
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 4-Grade 2	-58.024	11.088	-5.233	.000	.000
Grade 4-Grade Pre-2	-58.333	11.088	-5.261	.000	.000
Grade 4-Grade 3	-39.214	11.088	-3.537	.000	.006
Grade 4-Grade Pre-1	-34.595	11.088	-3.120	.002	.027
Grade 4-Grade 1	-29.833	11.088	-2.691	.007	.107
Grade 1-Grade Pre-2	28.500	11.088	2.570	.010	.152
Grade 1-Grade 2	28.190	11.088	2.542	.011	.165
Grade Pre-1-Grade Pre-2	23.738	11.088	2.141	.032	.484
Grade Pre-1-Grade 2	23.429	11.088	2.113	.035	.519
Grade 1-Grade 3	9.381	11.088	.846	.398	1.000
Grade 3-Grade 2	-18.810	11.088	-1.696	.090	1.000
Grade 1-Grade Pre-1	4.762	11.088	.429	.668	1.000
Grade 2-Grade Pre-2	.310	11.088	.028	.978	1.000
Grade Pre-1-Grade 3	4.619	11.088	.417	.677	1.000
Grade 3-Grade Pre-2	-19.119	11.088	-1.724	.085	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Code Glosses

Pairwise Comparisons of Grade



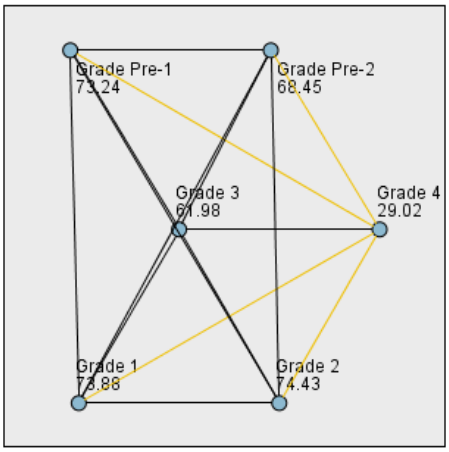
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 2-Grade 4	46.690	11.193	4.171	.000	.000
Grade 3-Grade 4	71.381	11.193	6.377	.000	.000
Grade Pre-2-Grade 4	54.667	11.193	4.884	.000	.000
Grade 3-Grade Pre-1	-39.429	11.193	-3.523	.000	.006
Grade 1-Grade 4	35.738	11.193	3.193	.001	.021
Grade 3-Grade 1	-35.643	11.193	-3.184	.001	.022
Grade Pre-1-Grade 4	31.952	11.193	2.855	.004	.065
Grade 3-Grade 2	-24.690	11.193	-2.206	.027	.411
Grade Pre-2-Grade Pre-1	-22.714	11.193	-2.029	.042	.636
Grade 2-Grade 1	-10.952	11.193	-.978	.328	1.000
Grade 1-Grade Pre-1	3.786	11.193	.338	.735	1.000
Grade Pre-2-Grade 1	-18.929	11.193	-1.691	.091	1.000
Grade 2-Grade Pre-1	-14.738	11.193	-1.317	.188	1.000
Grade Pre-2-Grade 2	-7.976	11.193	-.713	.476	1.000
Grade 3-Grade Pre-2	-16.714	11.193	-1.493	.135	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Evidentials

Pairwise Comparisons of Grade



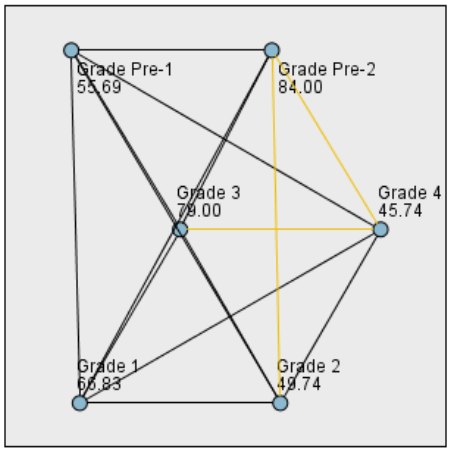
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 4-Grade 1	-44.857	11.230	-3.994	.000	.001
Grade 4-Grade 2	-45.405	11.230	-4.043	.000	.001
Grade 4-Grade Pre-1	-44.214	11.230	-3.937	.000	.001
Grade 4-Grade Pre-2	-39.429	11.230	-3.511	.000	.007
Grade 4-Grade 3	-32.952	11.230	-2.934	.003	.050
Grade 1-Grade 2	.548	11.230	.049	.961	1.000
Grade 3-Grade 1	-11.905	11.230	-1.060	.289	1.000
Grade 3-Grade 2	-12.452	11.230	-1.109	.268	1.000
Grade Pre-2-Grade Pre-1	-4.786	11.230	-.426	.670	1.000
Grade Pre-1-Grade 1	-.643	11.230	-.057	.954	1.000
Grade Pre-1-Grade 2	1.190	11.230	.106	.916	1.000
Grade Pre-2-Grade 1	-5.429	11.230	-.483	.629	1.000
Grade 3-Grade Pre-1	-11.262	11.230	-1.003	.316	1.000
Grade Pre-2-Grade 2	-5.976	11.230	-.532	.595	1.000
Grade 3-Grade Pre-2	-6.476	11.230	-.577	.564	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Hedges

Pairwise Comparisons of Grade



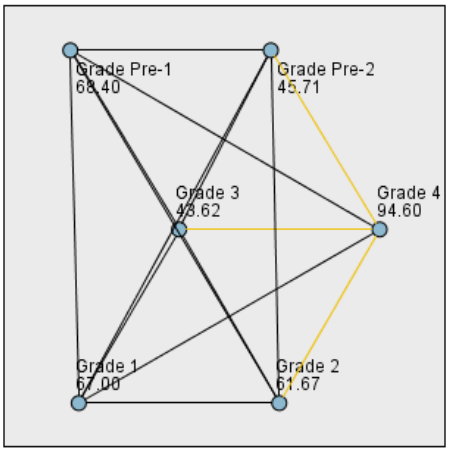
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 4-Grade Pre-2	-38.262	11.269	-3.395	.001	.010
Grade 2-Grade Pre-2	34.262	11.269	3.040	.002	.035
Grade 4-Grade 3	-33.262	11.269	-2.952	.003	.047
Grade 2-Grade 3	29.262	11.269	2.597	.009	.141
Grade Pre-1-Grade Pre-2	28.310	11.269	2.512	.012	.180
Grade Pre-1-Grade 3	23.310	11.269	2.068	.039	.579
Grade 4-Grade 1	-21.095	11.269	-1.872	.061	.918
Grade 2-Grade 1	-17.095	11.269	-1.517	.129	1.000
Grade 1-Grade 3	12.167	11.269	1.080	.280	1.000
Grade 4-Grade 2	-4.000	11.269	-.355	.723	1.000
Grade Pre-1-Grade 1	-11.143	11.269	-.989	.323	1.000
Grade 1-Grade Pre-2	17.167	11.269	1.523	.128	1.000
Grade 2-Grade Pre-1	-5.952	11.269	-.528	.597	1.000
Grade 3-Grade Pre-2	-5.000	11.269	-.444	.657	1.000
Grade 4-Grade Pre-1	-9.952	11.269	-.883	.377	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Logical connectives

Pairwise Comparisons of Grade



Each node shows the sample average rank of Grade.

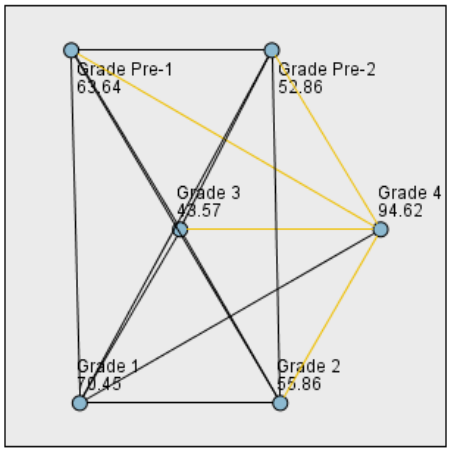
Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 3-Grade 4	50.976	9.657	5.279	.000	.000
Grade Pre-2-Grade 4	48.881	9.657	5.062	.000	.000
Grade 2-Grade 4	32.929	9.657	3.410	.001	.010
Grade 1-Grade 4	27.595	9.657	2.857	.004	.064
Grade Pre-1-Grade 4	26.190	9.657	2.712	.007	.100
Grade 3-Grade Pre-1	-24.786	9.657	-2.567	.010	.154
Grade 3-Grade 1	-23.381	9.657	-2.421	.015	.232
Grade Pre-2-Grade Pre-1	-22.690	9.657	-2.350	.019	.282
Grade Pre-2-Grade 1	-21.286	9.657	-2.204	.028	.413
Grade 3-Grade 2	-18.048	9.657	-1.869	.062	.925
Grade 2-Grade 1	-5.333	9.657	-.552	.581	1.000
Grade 1-Grade Pre-1	1.405	9.657	.145	.884	1.000
Grade 2-Grade Pre-1	-6.738	9.657	-.698	.485	1.000
Grade Pre-2-Grade 2	-15.952	9.657	-1.652	.099	1.000
Grade 3-Grade Pre-2	-2.095	9.657	-.217	.828	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

s

Person markers

Pairwise Comparisons of Grade



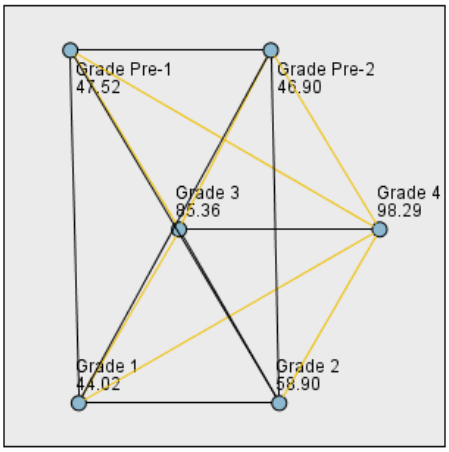
Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 3-Grade 4	51.048	9.783	5.218	.000	.000
Grade Pre-2-Grade 4	41.762	9.783	4.269	.000	.000
Grade 2-Grade 4	38.762	9.783	3.962	.000	.001
Grade Pre-1-Grade 4	30.976	9.783	3.166	.002	.023
Grade 3-Grade 1	-26.881	9.783	-2.748	.006	.090
Grade 1-Grade 4	24.167	9.783	2.470	.013	.202
Grade 3-Grade Pre-1	-20.071	9.783	-2.052	.040	.603
Grade 2-Grade 1	-14.595	9.783	-1.492	.136	1.000
Grade 3-Grade 2	-12.286	9.783	-1.256	.209	1.000
Grade Pre-2-Grade Pre-1	-10.786	9.783	-1.103	.270	1.000
Grade Pre-1-Grade 1	-6.810	9.783	-.696	.486	1.000
Grade Pre-2-Grade 1	-17.595	9.783	-1.799	.072	1.000
Grade 2-Grade Pre-1	-7.786	9.783	-.796	.426	1.000
Grade Pre-2-Grade 2	-3.000	9.783	-.307	.759	1.000
Grade 3-Grade Pre-2	-9.286	9.783	-.949	.343	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Relation markers

Pairwise Comparisons of Grade



Each node shows the sample average rank of Grade.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Grade 1-Grade 4	54.262	11.225	4.834	.000	.000
Grade Pre-1-Grade 4	50.762	11.225	4.522	.000	.000
Grade Pre-2-Grade 4	51.381	11.225	4.577	.000	.000
Grade 1-Grade 3	41.333	11.225	3.682	.000	.003
Grade 2-Grade 4	39.381	11.225	3.508	.000	.007
Grade Pre-2-Grade 3	38.452	11.225	3.426	.001	.009
Grade Pre-1-Grade 3	37.833	11.225	3.370	.001	.011
Grade 2-Grade 3	26.452	11.225	2.357	.018	.277
Grade 1-Grade 2	14.881	11.225	1.326	.185	1.000
Grade 3-Grade 4	12.929	11.225	1.152	.249	1.000
Grade Pre-2-Grade Pre-1	-.619	11.225	-.055	.956	1.000
Grade 1-Grade Pre-1	3.500	11.225	.312	.755	1.000
Grade 1-Grade Pre-2	2.881	11.225	.257	.797	1.000
Grade Pre-1-Grade 2	11.381	11.225	1.014	.311	1.000
Grade Pre-2-Grade 2	-12.000	11.225	-1.069	.285	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Sequencing

Appendix H Table of Contents from Self-Study Preparation Booklet for Standard-Setting Panel 1

CONTENTS

IMPORTANT: PLEASE DO NOT LOOK AT THE APPENDICES UNTIL YOU ARE INSTRUCTED TO DO SO IN THE TASKS.

Appendix I Rating forms reading used for Standard-Setting Panel 1

Judge's Name:	Grade 1 Test Items

Instructions for rating items (Basket)						
Read each reading text and choose the correct answer for each test item. After you have answered each item, answer the following question: *At which CEFR level can a test taker already answer the item correctly?* Put a tick in the appropriate column						
Item	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**
26						
27						
28						
29						
30						
31						
32						
33						
34						
35						
36						
37						
38						
39						
40						
41						
Total						

Judge's name:

Grade 1 Test Items

	Of 100 test takers who are minimally competent at the following CEFR level, how many will answer the item correctly?
Item	C1 (border between B2/C1)
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
Total	

Appendix J Schedule of activities for Standard Setting Panel 1

Day 1 (9:30 am to 6 p.m.)

- Discussion of CEFR and familiarization tasks from booklet
    - Agree on key features defining levels for the Global Scale, focusing on B1, B2, C1
- Training with listening items calibrated to the CEFR provided by the Council of Europe (using BASKET method).
- Rate Grade 1 Listening items (BASKET method)
    - Listen to and answer items as a test taker
    - Listen again and judge at which CEFR level a test taker could answer each item correctly (given access to listening scripts during second listening)
- Explanation of modified Angoff procedure
- Rate he same Grade 1 Listening items again (using the MODIFIED ANGOFF)
- Feedback on Grade 1 Listening items; No discussion, but chance to adjust ratings for both Basket and modified Angoff ratings based on feedback
- Rating forms for both methods for Grade 1 items collected
- Rate Grade Pre-1 Listening items (using BASKET method)
- Recap explanation of Angoff procedure
- Rate Grade Pre-1 Listening items (MODIFIED ANGOFF method)
- Feedback on Grade Pre-1 Listening items; No discussion, but chance to adjust ratings for both Basket and modified Angoff ratings based on feedback
- Rating forms for both methods for Grade Pre-1 items collected

Day 2 (9:30 am to 6 p.m.)

Morning / First part of afternoon session

- Discussion of CEFR, focusing on reading
    - Agree on key words and features for the Global Scale, focusing on B1, B2, C1
- Training with listening items provided by the Council of Europe (using BASKET method).
- Rate Grade 1 Listening items (BASKET method)
    - Answer reading test items as a test taker
    - Review items to judge at which CEFR level a test taker could answer each item correctly (given access to listening scripts during second listening)
- Explanation of modified Angoff procedure
- Rate the same Grade 1 reading items again (using the MODIFIED ANGOFF)
- Feedback on Grade 1 reading items; No discussion, but chance to adjust ratings for both Basket and modified Angoff ratings based on feedback
- Rating forms for both methods for Grade 1 items collected
- Rate Grade Pre-1 reading items (using BASKET method)
- Recap explanation of Angoff procedure
- Rate Grade Pre-1 reading items (MODIFIED ANGOFF method)
- Feedback on Grade Pre-1 reading items; No discussion, but chance to adjust ratings for both Basket and modified Angoff ratings based on feedback
- Rating forms for both methods for Grade Pre-1 items collected

Second part of afternoon session

- Process repeated with Grade 1 and Grade Pre-1 vocabulary items

Day 3 (9:30 am to 5 p.m.)

- Discussion of CEFR focusing on writing
- Training with examples of writing scripts provided by the Council of Europe judged to be at particular CEFR levels by examination boards
    - Explanation of Paper Selection standard-setting method
    - Rating of 30 Grade 1 EIKEN writing scripts
    - Rating of 30 Grade Pre-1 writing scripts
    - Fill out participant background and feedback questionnaire

Appendix K Results of procedural questionnaire for Panel 1

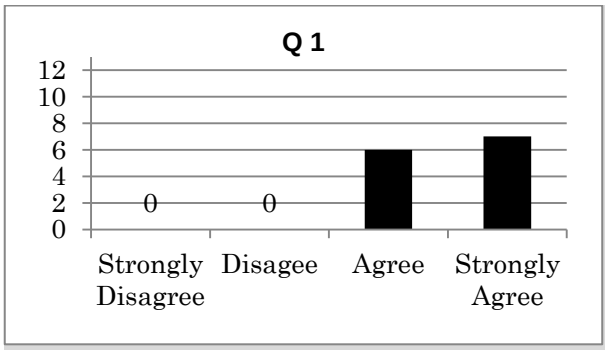


Figure K1 Responses to Question 1

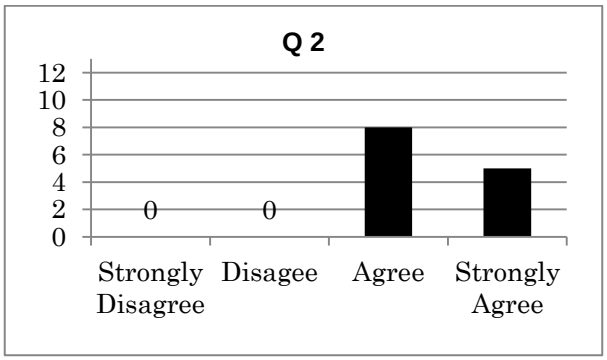


Figure K2 Responses to Question 2

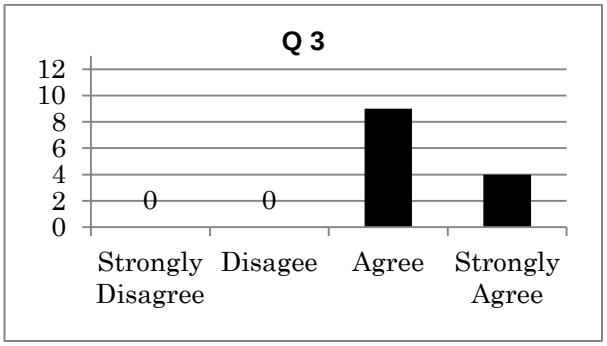


Figure K3 Responses to Question 3

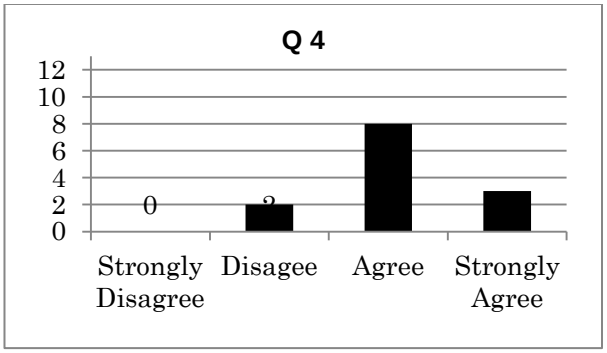


Figure K4 Responses to Question 4

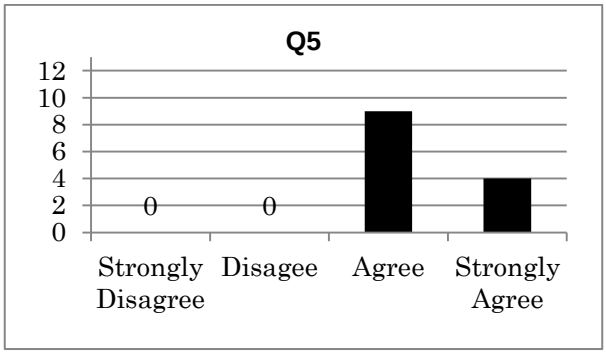


Figure K5 Responses to Question 5

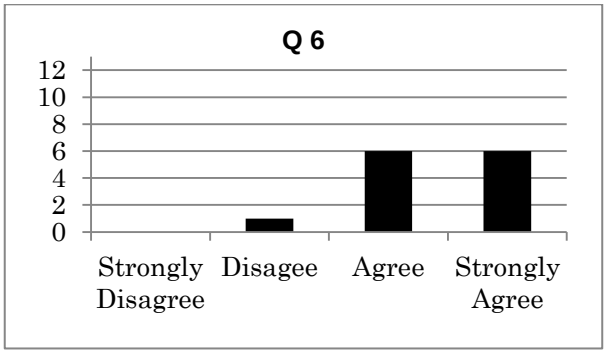


Figure K6 Responses to Question 6

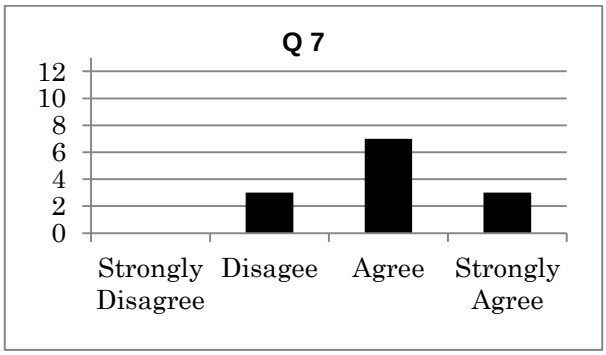


Figure K7 Responses to Question 7

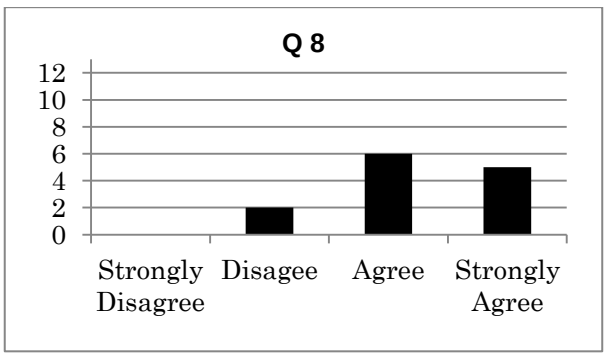


Figure K8 Responses to Question 8

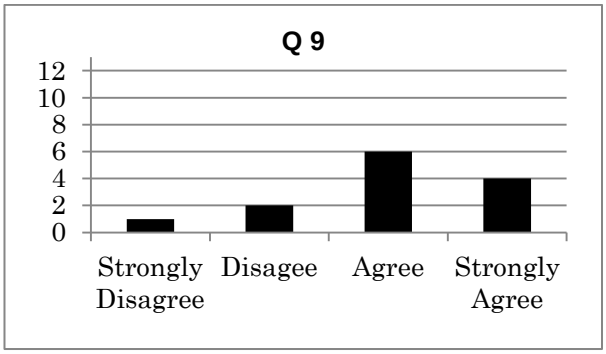


Figure K9 Responses to Question 9

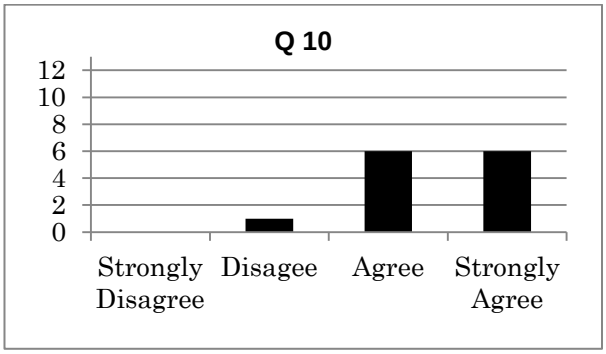


Figure K10 Responses to Question 10

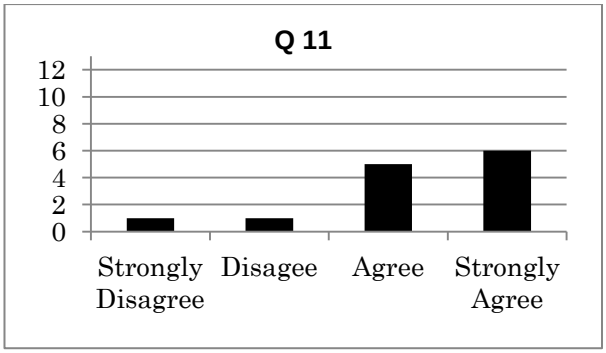


Figure K11 Response to Question 11

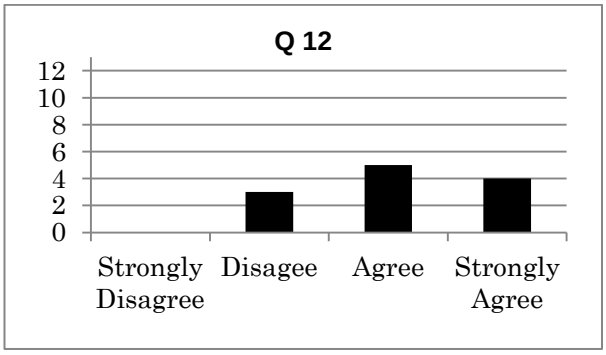


Figure K12 Responses to Question 12

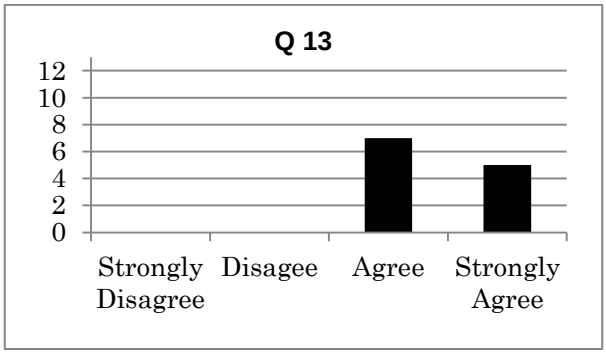


Figure K13 Responses to Question 13

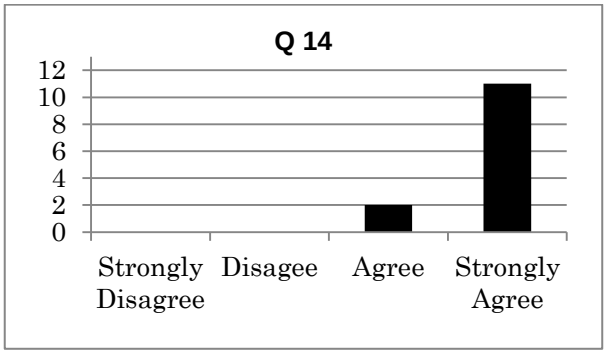


Figure K14 Responses to Question 14

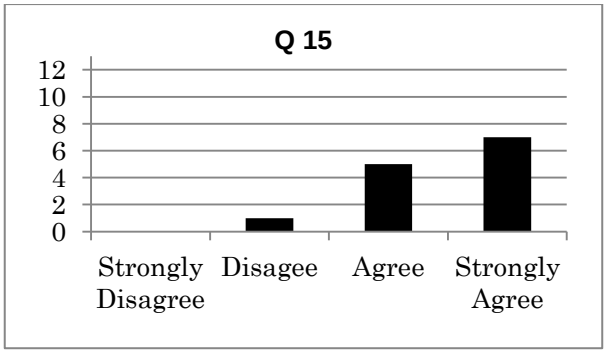


Figure K15 Responses to Question 15

Appendix L Results of procedural questionnaire for Panel 2

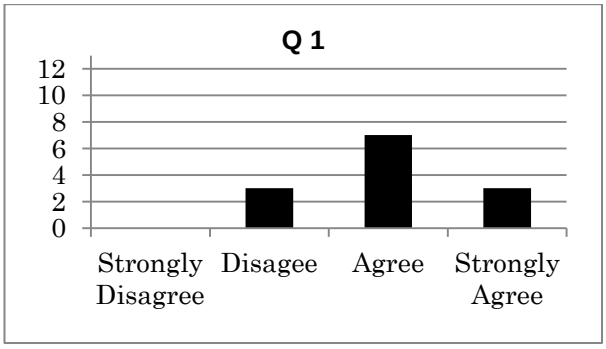


Figure L1 Responses to Question 1 for Panel 2

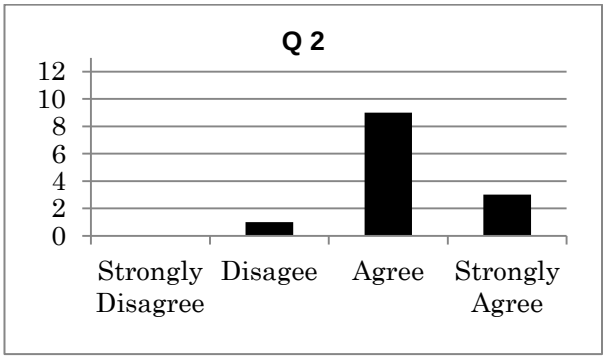


Figure L2 Responses to Question 2 for Panel 2

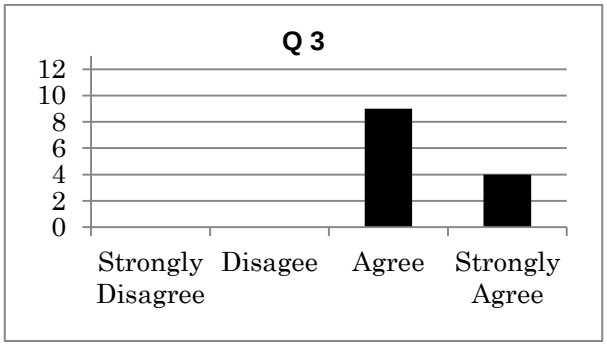


Figure L3 Responses to Question 3 for Panel 2

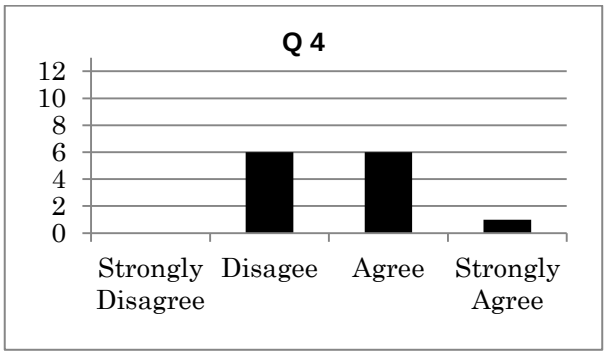


Figure L4 Responses to Question 4 for Panel 2

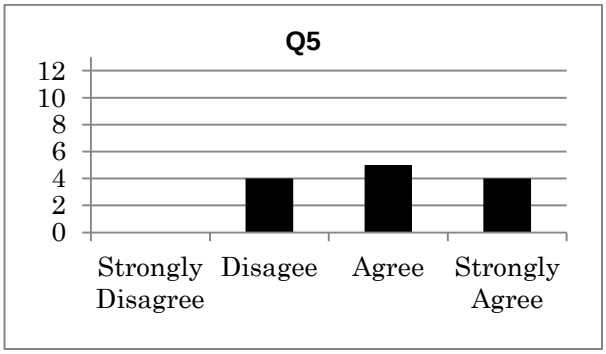


Figure L5 Responses to Question 5 for Panel 2

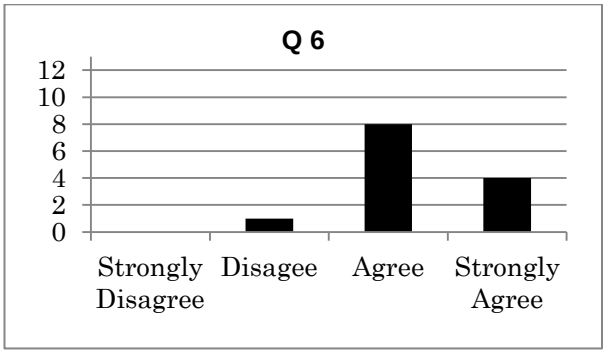


Figure L6 Responses to Question 6 for Panel 2

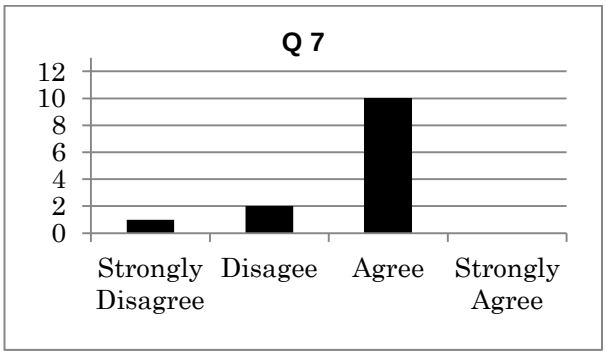


Figure L7 Responses to Question 7 for Panel 2

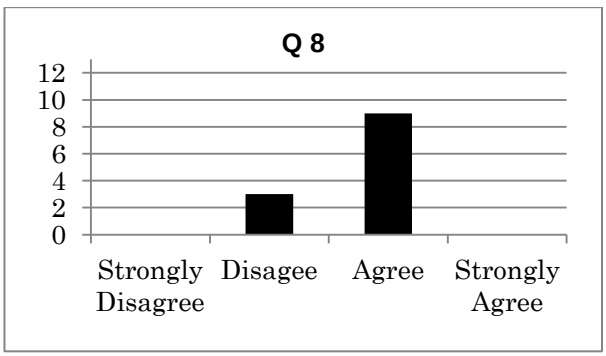


Figure L8 Responses to Question 8 for Panel 2

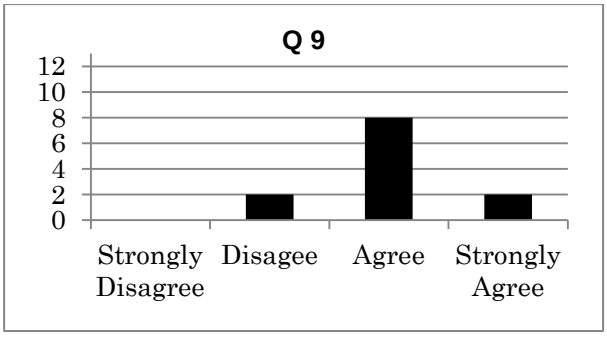


Figure L9 Responses to Question 9 for Panel 2

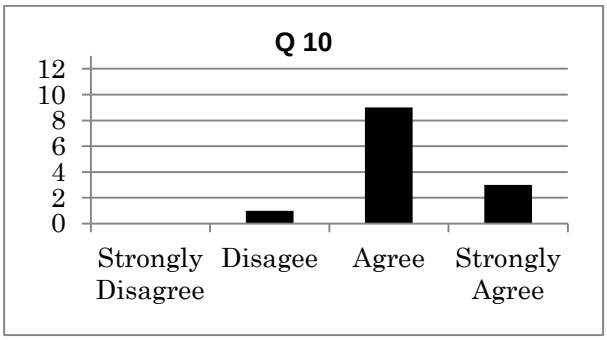


Figure L10 Responses to Question 10 for Panel 2

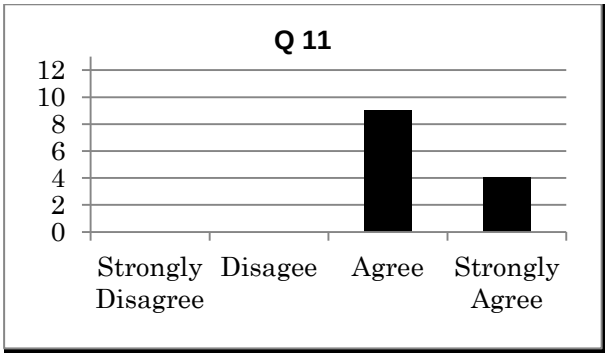


Figure L11 Response to Question 11 for Panel 2

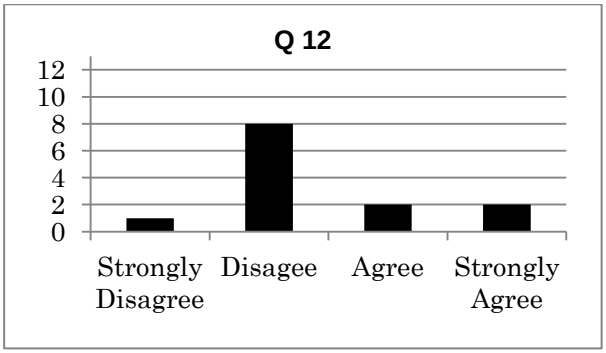


Figure L12 Responses to Question 12 for Panel 2

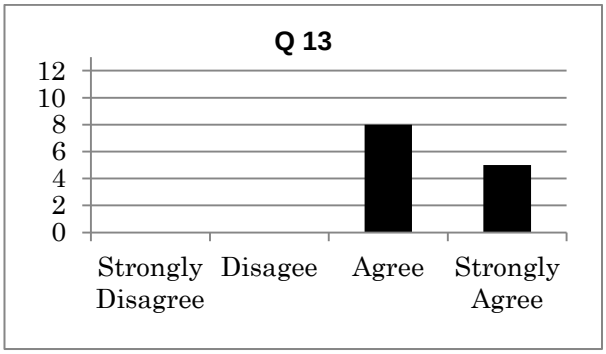


Figure L13 Responses to Question 13 for Panel 2

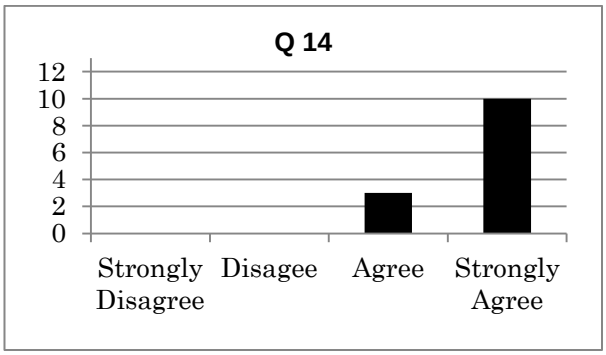


Figure L34 Responses to Question 14 for Panel 2

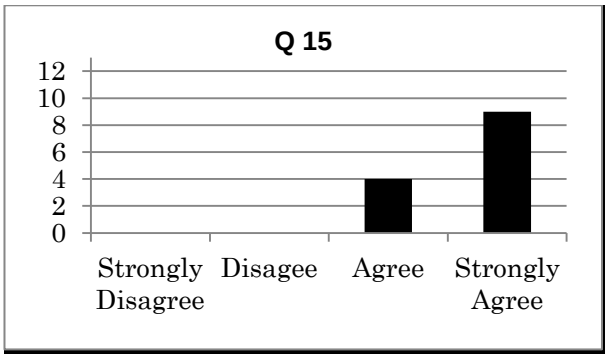


Figure L15 Responses to Question 15 for Panel 2

List of References

Alderson, J.C. (1991). Bands and scores. In J.C. Alderson & B. North (eds.), *Language Testing in the 1990s* (pp. 71–86). London: Modern English Publications / British Council / Macmillan.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J.C. (2005). Editorial. *Language Testing*, 22 (3) 257–26.

Alderson, J C, Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J., Figueras, N., Kuijper, H., Nold, G., Takala, S., & 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.

Almond, R.G., Mislevy, R.J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. New York: Springer.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17(1), 1–42.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Bailey, K.M. (1999). *Washback in language testing*. Princeton, N.J: Educational Testing Service

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465.

Bax, S., & Weir, C. J. (2012). Investigating learners' cognitive processes during a computer-based CAE reading text. *Research Notes,* 47, 3–14.

Bax, S., Waller, D., & Nakatsuhara, F. (2013). *Researching metadiscourse markers in candidates' writing at Cambridge FCE, CAE and CPE levels*. Paper presented at the 2013 BAAL conference, Heriot-Watt University, Edinburgh.

Bechger, T., Kujper, H., & Maris, G., (2009). Standard setting in relation to the Common European Framework of Reference for Languages: the case of the State Examination of Dutch as a Second Language. *Language Assessment Quarterly*, 6(2), 126-150.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion referenced tests. *Review of Educational Research*, 56(1), 137-172.

Bonk, W.J., & Ockey, G.J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.

Briggs, D. C. & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3-14.

Brown, J.D. (2002). Statistics Corner: Questions and answers about language testing statistics: The Cronbach alpha reliability estimate. *Shiken: JALT Testing & Evaluation SIG Newsletter,* 6(1), 17-18.

Brown, J.D. (2008). Testing-context analysis: assessment is just another part of language curriculum development. *Language Assessment Quarterly,* 5(4), 275 -312

Brown, J.D., Davis, J. McE., Takahashi, C., & Nakamura, K. (2012). *Upper-level EIKEN examinations: linking, validating, and predicting TOEFL iBT scores at advanced proficiency EIKEN levels*. Eiken Foundation of Japan. Retrieved from http://www.eiken.or.jp/eiken/group/result/pdf/eiken-toeflibt-report.pdf.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brown, J. D., & Yamashita, S. O. (1995). English language entrance examinations at Japanese universities: 1993 and 1994. In J. D. Brown & S. O. Yamashita (Eds.) *Language Testing in Japan* (pp. 86-100). Tokyo: Japan Association for Language Teaching

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study.* ARAGs Research Reports Online, AR-G/2015/001. London: British Council.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145-170.

Cambridge English (n.d.). *The Cambridge English scale explained.* Retrieved from http://www.cambridgeenglish.org/images/177867-the-methodology-behind-the-cambridge-english-scale.pdf.

Cambridge Michigan Language Assessments (2012). *MELAB 2012 report*. Retrieved from http://www.cambridgemichigan.org/about-us/research/

Camiciottoli, B.C. (2009). Metadiscourse and ESP reading comprehension: An exploratory study. *Reading in a Foreign Language*, 15(1), 28-44.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1 (1), 1-47.

Chapelle, C., Enright, M., & Jamieson, J., (2008). *Building a validity argument for the Test of English as a Foreign Language.* New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing,* 14(1), 3–22.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20 (4), 369-383.

Chujo, K., & Oghigian, K. (2009). How many words do you need to know to understand TOEIC, TOEFL & EIKEN? An examination of text coverage and high frequency vocabulary. *The Journal of Asian TEFL*, 6(2), 121-148.

Cizek, G., & Bunch, M. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage Publications.

Cizek, G., Rosenberg, S., & Koons, H., (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement,* 68 (3), 397-412.

Cizek, G., Bowen, D., & Church, K., (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement,* 70(5) 732–743.

Clauser, B., Harik, P., Margolis, M., McManus, I. Mollon, J., Chis, L., & Williams, S., (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement Quarterly*, 22(1), 1-21.

Cobb, T. (2015). VocabProfile, The Compleat Lexical Tutor. Retrieved May 2015 from http://www.lextutor.ca.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, A., Kane, M., & Crooks, T. (1999). A generalized examinee-centered method for setting standards on Achievement tests. *Applied Measurement in Education,* 12(4), 343-366.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2003). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment: Manual: Preliminary pilot version.* Strasbourg: Council of Europe.

Council of Europe. (2004). *Reference supplement to the preliminary version of the manual for relating language examinations to the Common European Framework of References for Languages: learning teaching, assessment.* Strasbourg: Language Policy Division.

Council of Europe (2005). *Reading and listening items.* Strasbourg, France: Council of Europe. Compact disc.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for Languages: Learning teaching, assessment.* Strasbourg: Language Policy Division.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Crossley, S.A., Allen, D., & McNamara, D.S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1), 84–101.

Crossley, S.A., Allen, D., & McNamara, D.S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research,* 16(1) 89–108.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42* (3), 475–493.

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language testes: A matter of effect. *Language Teaching*, 40, 231-241

Dunlea, J. (2010). The EIKEN can-do list: improving feedback for an English proficiency test in Japan. In L. Taylor & C. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment, Studies in Language Testing volume 31*. Cambridge: Cambridge University Press.

Dunlea, J. (2014). *Investigating the relationship between empirical task difficulty, textual features and CEFR levels*. Paper presented at the 11th EALTA conference, University of Warwick, UK.

Dunlea, J., & Figueras, N. (2012). Replicating results from a CEFR test comparison project across continents. In D. Tsagari and I. Csepes (eds.), *Collaboration in language testing and assessment* (pp. 31-45). New York: Peter Lang

Eckes, T. (2009). Section H: Many-facet Rasch measurement. In Council of Europe, *Reference supplement to the preliminary version of the manual for relating language examinations to the Common European Framework of References for Languages: learning teaching, assessment*. Strasbourg: Language Policy Division.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.

Eiken. (n.d.-a). http://www.eiken.or.jp/eiken/en/eiken-tests/overview/

Eiken. (n.d.-b). http://www.eiken.or.jp/eiken/en/grades/

Eiken. (n.d.-c). http://www.eiken.or.jp/eiken/en/research/

Eiken. (n.d.-d). http://faq.eiken.or.jp/faq/show/499

Eiken. (n.d.-e). http://www.eiken.or.jp/eiken/en/association/history/

Eiken. (n.d.-f). http://www.eiken.or.jp/eiken/en/eiken-tests/administration/

Eiken. (n.d-g). http://www.eiken.or.jp/eiken/en/recognition/

Engelhard, G., & Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, 58(2), 179-196.

European Association for Language Testing and Assessment (2006). *Guidelines for good practice in language testing and assessment*. Retrieved from http://www.ealta.eu.org/guidelines.htm

Educational Testing Service. (2011). *Validity evidence supporting the interpretation and use of TOEFL iBT scores*. TOEFL Research Insight Series Vol. 4. Princeton, NJ: ETS.

Enright, M, Grabe, W, Koda, K, Mosenthal, P, Mulcany-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper*. TOEFL Monograph Series 17. Princeton, NJ: ETS.

Field, A. (2009). *Discovering statistics using SPSS (3rd ed.).* London: Sage.

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005) Relating examinations to the Common European Framework: a manual. *Language Testing*, 22 (3), 1–19.

Fitzpatrick, A. (2008). NCME 2008 presidential address: the impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice*, 27(4), 34-40.

Fouts, M.T. (2013, July). *Building bridges to the international community: the EIKEN example.* In L. Bachman (Discussant), The challenges and issues in developing English language tests in the Asian EFL context. Symposium conducted at the 35th Language Testing Research Colloquium, Seoul, South Korea

Frisbie, D. A. (1988), Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, 7(1), 25–35.

Fulcher G. & Davidson, F. (2007). *Language testing and assessment*. London:

Routledge.

Geranpayeh, A. (2007). Using structural equation modeling to facilitate the revision of high stakes testing; the case of CAE. *Research Notes,* 30, 8-12.

Geranpayeh, A., & Taylor, L. (eds.) (2013). *Examining listening: research and practice in assessing second language listening*. Studies in Language Testing 35. Cambridge: Cambridge University Press.

Green, A., Unaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts for testing reading for academic purposes. *Language Testing*, 27 (3), 1–21.

Green, D., Trimble, C., & Lewis, D. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice,* 22(1), 22-32.

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal*, 26, 5–24.

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3), 309–334.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 89-116). New York: Routledge.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24(4), 355–366.

Hamp-Lyons, L. (1989). Applying the partial credit model of Rasch analysis: Language testing and accountability. *Language Testing, 6*(1), 109–118.

Hardy. M. A., Young, M.J., Qing, Y., Sudweeks, R.R., Bahr, D.L. (2011). *Investigating content and construct representation of a common-item design when creating a vertically scaled test*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice,* 8: 35–41.

Harris, D. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich,

& P. Holland (Eds), *Linking and aligning scores and scale*s. New York: Springer.

Harsch, C., & Hartig, J. (2015). Comparing c-tests and yes/no vocabulary tests as predictors of receptive language skills. *Language Testing*, First published on 10 August 2015. DOI: 10.1177/0265532215594642

Heatley, A., Nation, I.S.P., and Coxhead, A. (2002). RANGE and FREQUENCY programs. retrieved from http://www.vuw.ac.nz/lals/staff/Paul_Nation.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1–11.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 2(2), 141–154.

Hill, Y. Z. (2010). *Validation of the STEP EIKEN test for college admission* (Unpublished doctoral dissertation). Manoa, Hawai'i: University of Hawai'i at Manoa.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–30.

Hurtz, Gr., & Hertz, N. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, 59 (6), 885-897.

Hyland. K. (1999). Talking to students: Metadiscourse in introductory textbooks. *English for Specific Purposes*, 18, 3-26.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Michigan: University of Michigan Press.

Hyland, K. and Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, 25(2). 156 – 177.

International Language Testing Association (2007). *Guidelines for practice.* Retrieved from http://www.iltaonline.com.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: American Council on Education and Macmillan.

Jaeger, R. (1991). Selection of Judges for Standard Setting. *Educational*

measurement: Issues and Practice, 10, (2), 3-10.

Jarvis, S. (2002): Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* 19: 57–84.

Jones, N. (2001). Reliability as one aspect of test quality. *Research Notes,* 54, 2-5.

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.

Kaftandjieva, F. (2009).The Basket method: the bread basket or the basket case of standard setting methods? In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: research perspectives* (pp. 103-109). Arnhem: CITO and EALTA.

Kaftandjieva, F. (2010). *Methods for setting cut-off scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading*. Arnhem: CITO and EALTA.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kane, M. (1998). Choosing between examinee-centered and test centered standard-setting methods. *Educational Assessment*, 5(3), 129-145.

Kane, M. (2001a). So much remains the same: conception and status of validation in standard setting. In G. Cizek (Ed.) *Setting performance standards*. New York: Routledge.

Kane, M. T. (2001b). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice,* 21, 31–41.

Kane, M. T. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, 29(1) 29–36.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.

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.

Kenyon, D., MacGregor D., Dongyang, L., & Cook, H. (2011). Issues in vertical scaling of a K-12 English language proficiency test. *Language Testing,* 28(3), 383–400.

Khalifa, H., & Weir, C J. (2009). *Examining reading: Research and practice in assessing second language reading.* Studies in Language Testing 29. Cambridge: Cambridge University Press.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal, 28* (1), 77-96.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: methods and practices* (3nd ed.). New York: Springer-Verlag.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22 (/1) 15–30.

Linacre, J.M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.

Linacre, J.M. (2014). Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2015). Winsteps Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

Livingston, S., & Zieky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: ETS.

Livingston, S., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, 2(2), 121-141.

Loomis, S. & Bourque, M. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. Cizek (Ed.) *Setting performance standards* (175-217). New York: Routledge.

Lumley, T., Lynch, B K., & McNamara, T. (1994). A new approach to standard

setting in language assessment. *Melbourne Papers in Language Testing*, 19-39.

Lunz, M.E., Wright,B., & Linacre,J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education,* 3, 331-345.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, VOCD-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381-392

McNamara, T (1996). *Measuring second language performance*. Longman: London and New York.

McNamara, T., (2006). Validity in language testing: the challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3(1), 31–51

McNamara, T., & Knock, U. (2012). The Rasch wars: the emergence of Rasch measurement in language testing. *Language Testing*. First published March 7 2012.

McNamara, D.S., Graesser, A., Cai, Z., & Kulikowich, J. (2011). *Coh-Metrix dimensions of text difficulty: Aligning text difficulty with theories of text comprehension.* Paper presented at the 2011 annual meeting of the American Educational Research Association. Retrieved from the AERA Online Paper Repository.

Messick, S. (1986), *The once and future issues of validity: Assessing the meaning and consequences of measurement*. ETS Research Report RR-86-30; Princeton, NJ: ETS

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, 32(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.

Messick, S. (1996) Validity and washback in language testing. *Language Testing*, 13, 241-256.

Milanovic, M., & Weir, C. (2010). Series editors' note. In Martyniuk, W. (Ed) *Aligning Tests with the CEFR; reflections on using the Council of Europe's draft manual (pp. viii-xx).* Cambridge: Cambridge University Press.

Milton, J. (2010). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In Bardel, C., Lindqvist, C. and Laufer, B. (Eds), *L2 Vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis.* Eurosla monographs Series, Volume 2. Online: Eurosla.

Ministry of Education, Culture, Sports, Science, and Technology (2003). *Regarding the establishment of an action plan to cultivate Japanese with English abilities.* Retrieved from http://www.mext.go.jp/english/topics/03072801.htm.

Ministry of Education, Culture, Sports, Science, and Technology. (2011). *Five proposals and specific measures for developing proficiency in English for International communication.* Retrieved from http://www.mext.go.jp/english/elsec/ 1319701.htm

Mislevy R J, Steinberg, L S., & Almond, R G (2003). *On the structure of educational assessments*. CSE Tech. Rep. No. 597. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://www.cse.ucla.edu

Morrow, K. (Ed.). (2004). *Insights from the Common European Framework.* Oxford, England: Oxford University Press.

Mulvey, B. (2001). The role and influence of Japan's university entrance exams: A reassessment. *The Language Teacher,* 25(7*)*, 11-17

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5, 189-227.

Nakatsuhara, F**.** (2014) *A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1 & Study 2*. Retrieved from http://www.eiken.or.jp/teap/ group/report.html

Nation, P., (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59-82.

Negishi, M., Takada, T. & Tono, Y. (2012). A progress report on the development of the CEFR-J. In E. Galaczi and C.J. Weir (Eds.), *Exploring language frameworks: Studies in language testing 36* (137-165). Cambridge: Cambridge University Press.

North, B., (2001). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, B. (2007). *Response by Brian North*. In B. North & T. MacNamara (Chairs), The CEFR in Europe and beyond: challenges and experiences. Symposium conducted at the 4th European Association of Language Testing and Assessment Conference, Sitges. Retrieved from http://www.ealta.eu.org.

North, B., & Jones, N. (2009). *Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling.* Retrieved from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/Standards_language_EN.pdf

North, B., Martyniuk, W., & Panthier, J. (2010). Introduction: The manual for relating examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language Education. In Martyniuk, W. (Ed) *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manaul (pp. 1-17).* Cambridge: Cambridge University Press.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15(2), 217-263.

OECD (2014), PISA 2012 Technical Report, PISA. OECD Publishing.

O'Sullivan, B. (2008). *City & Guilds Communicator IESOL Examination (B2) CEFR linking project: Case study.* Retrieved from: http://cdn.cityandguilds.com/ProductDocuments/International_English/General_English/8984/Additional_documents/8984_Case_study_v1.pdf

O'Sullivan, B. (2010). The City and 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*. Cambridge: Cambridge University Press.

O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. Oxford: Routledge.

O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. *Modern Language Journal, 96*, 71-88.

O'Sullivan, B. (2015a). *Aptis test development approach.* Aptis Technical Report, TR/2015/001. London: British Council.

O'Sullivan, B. (2015b). *Linking the Aptis reporting scales to the CEFR.* Aptis Technical Report, TR/2015/003. London: British Council.

O'Sullivan, B. (2015c). *Operationalising an assessment use argument approach.* Paper presented at the 12th EALTA conference, University of Copenhagen, Denmark.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General technical manual version 1.0.* Aptis Technical Report TR/2015/005. London: British Council.

O'Sullivan, B., & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: theories and practices* (pp. 13-32). Oxford: Palgrave Macmillan.

Paek, I., Young, M. & Yi, Q. (2008). The impact of data collection design, linking method, and sample size on vertical scaling using the Rasch model. *Journal of Applied Measurement*, 9(3) 229-248.

Papageorgio, S. (2007). *Relating the Trinity College London GESE and ISE exams to the Common European Framework of Reference: piloting of the Council of Europe draft manual, final project report.* London: Trinity College London. Retrieved from http://www.trinitycollege.it/ accreditamenti/cefr-report.pdf

Papageorgio, S. (2009). Linking international examinations to the CEFR: the Trinity College London Experience. In W. Martyniuk (Ed), *Aligning Tests with the*

CEFR: Reflections on using the Council of Europe's draft Manual (pp. 145-158). Cambridge: Cambridge University Press.

Plake, B., Impara, J., & Irwin, P. (2000). Consistency of Angoff-based predictions of item performance: evidence of technical quality of results from the Angoff standard-setting method. *Journal of Educational Measurement*, 37(4), 347-355.

Pomplun, M., Omar, MD., & Custer, M. (2004). A comparison of Winsteps and Bilog-Mg for vertical scaling with the Rasch model. *Educational and Psychological Measurement* 64, 600-616,

Raymond, M. R. & Reid, J. B. (2001). Who made thee judge? Selecting and training participants for standard setting. In G. J Cizek (Ed.), *Setting performance standards* (pp. 119-157). New York: Routledge.

Reckase, M.D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard-setting methods. *Educational Measurement: Issues and Practice*, 25(2), 14-17.

Reckase, M. (2010). *Study of best practices for vertical scaling and standard setting with recommendations for FCAT 2.0*. Tallahassee: Florida Department of Education.

Sasaki, M. (2008). The 150-year history of English language assessment in English education in Japan. *Language Testing*, 25 (1) 63–83.

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: Cambridge University Press.

Schmitt, N., Xiangying, J., & Grabe, W. (2011). The Percentage of words known in a text and reading comprehension. *Modern Language Journal*, 95(1), 26-43.

Shaw, S & Weir, C.J. (2007). *Examining writing: Research and practice in assessing second language writing.* Studies in Language Testing 26. Cambridge: Cambridge University Press.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.

Shiotsu, T. (2010). *Components of L2 reading.* Cambridge*:* Cambridge University Press and Cambridge ESOL.

Shiotsu, T. & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of second language reading comprehension test performance. *Language Testing*, 23 (4), 99-128.

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, 249–275.

Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25 (1), 47-55.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework*. (ETS Research Rep. No. RR-05-18; TOEFL Research Rep. No. RR–80). Princeton, NJ: ETS.

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. Princeton, NJ: ETS.

Taylor, L. (Ed.). (2012). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Taylor, L. (2014). *A report on the review of test specifications for the reading and listening papers of the Test of English for Academic Purposes (TEAP) for Japanese University Entrants*. Retrieved from http://www.eiken.or.jp/teap/group/report.html

Taylor, L., & Galaczi, E. (2012). Scoring validity. In Taylor, L. (Ed.), *Examining speaking: research and practice in assessing second language speaking.* Cambridge: Cambridge University Press.

Text Inspector (2015) Text Inspector online analysis tool. Retrieved February 2015

from http://www.textinspector.com.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253.

Tong, Y., & Kolen, M. J. (2008). *Maintenance of vertical scales*. Paper presented at the National Council on Measurement in Education annual conference, New York City.

Tong, Y., & Kolen, M. (2010). Scaling: an ITEMS module. *Educational Measurement: Issues and Practice*, 29(4), 39-48.

Trim, J. (2010). The modern languages programme of the Council of Europe as a background to the English Profile Project. *English Profile Journa*l, 1(1), 1-12.

Van Moere (2006). Validity evidence in a university group oral test. *Language Testing*, 23 (4) 411–440.

Van Nijlen, D., & Jansenn, R. (2008). Modeling judgments in the Angoff and Contrasting-Groups methods of standard setting. *Journal of Educational Measurement*. 45 (1), 45-63.

van Zeeland, H. & Schmitt, N. (2012). Lexical coverage and L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics*. First published December 18 2012.

Watanabe. Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, 13 (3) 318-333.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: research contexts and methods* (pp. 19-36). Mahwah, New Jersey: Lawrence Erlbaum.

Watanabe, Y. (2013). The National Centre Test for university admissions: test review. *Language Testing*, 30(4) 565–573.

Weir, C.J. (2005a). *Language Test Validation: an evidence-based approach*. Oxford: Palgrave.

Weir, C.J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing,* 22, 281–300.

Weir, C.J.. (2014). *A research report on the development of the Test of English for Academic Purposes (TEAP) writing test for Japanese university entrants.* Retrieved from http://www.eiken.or.jp/teap/group/report.html

Weir, C., Hawkey, R., Green, T., & Devi, S. (2009). *The cognitive processes underlying the academic reading construct as measured by IELTS*. British Council/IDP Australia IELTS Research Reports, 9(4), 157–189.

Weir, C.J. , Vidakovic, I., & Galaczi, E. (2013). *Measured constructs: a history of the constructs underlying Cambridge English language (ESOL) examinations 1913-2012*. Cambridge: Cambridge University Press.

Wu, R. Y. F. (2012). *Establishing the validity of the General English Proficiency Test Reading Component through a critical evaluation on alignment with the Common European Framework of Reference*. Unpublished PhD thesis: University of Bedfordshire, Bedfordshire.

Wu, J. R. W., & Wu, R.Y.F. (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. 2014-224)*. Cambridge: Cambridge University Press.

Yanase, K., (2009). 話題・題材の「広がり」と「深み」(Topics: breadth and depth). STEP Eigo Joho. STEP.

Yoshida, K. (1996a). Language testing in Japan: A cultural problem? *The Daily Yomiuri.* (Educational Supplement). January 15, 1996, 15.

吉島茂　大橋理枝　訳・編 (2004)　「外国語教育Ⅱ　外国語の学習、教授、評価のための世六派共通参照枠」朝日出版社

Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum.

Zieky, M. (2001). So much has changed: How the setting of cutscores has changed since the 1980s. In G. J Cizek (Ed.), *Setting performance standards* (pp. 19-52).

New York: Routledge.