英検 公益財団法人 日本英語検定協会
Eiken Foundation of Japan

University of Bedfordshire

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

Dr. Lynda Taylor
Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire, UK

Contents

Executive summary

- This report describes a project to undertake a specification review of the reading and listening components for the Test of English for Academic Purposes (TEAP), a new test of academic English proficiency for university entrance purposes in Japan. The review project was conducted in 2012-13 through a collaboration between the TEAP team at the Eiken Foundation of Japan and an external consultant from CRELLA (Centre for English Language Learning and Assessment) at the University of Bedfordshire, UK.

- 本稿では、日本の大学受験のために開発された新しい英語能力試験、「アカデミック英語能力判定試験（TEAP)」のリーディングとリスニング試験のテスト仕様（specifications）の再検討プロジェクトについて報告する。この再検討プロジェクトは、2012 年から 2013 年にかけて、日本英語検定協会内の TEAP チームと、英国ベッドフォードシャー大学の英語学習・試験センター（CRELLA）所属の外部コンサルタントの共同で行われた。

- The project aims were to review the existing test specifications, task materials and item writer guidelines for the reading and listening components against a recently developed and widely recognised socio-cognitive framework for test development and validation, making recommendations for enhancing the existing documentation and offering suggestions for a possible research agenda for the future.

- このプロジェクトの目的は、既存のリーディング・リスニング試験のテスト仕様（specifications)、タスク材料、及びテスト作成者用ガイドラインを、「社会的・認知的枠組み」に照らし合わせながら再検討することであった。「社会的・認知的枠組み」は、近年開発され広く認知されてきた、テスト開発・妥当性検証のための枠組みである。そして、既存の資料を更に良いものにするために何ができるかを提案し、今後の可能な研究の指針を示した。

- The project generated detailed specification tables for each individual TEAP Reading and Listening task based on Weir's socio-cognitive validity frameworks; these tables identify key contextual and cognitive parameters for each task, together with useful and previously validated empirical measures that are accessible through readily available software.

- このプロジェクトは、Weir の「社会的・認知的妥当性枠組み」に基づき、TEAP リーディング・リスニング試験における個々のタスクに関する、詳細なテスト仕様表を作成した。これらの表は、それぞれのタスクについて、重要な「背景に関するパラメーター」と「認知的パラメーター」を定めている。更に、いくつかのパラメーターに関しては、有益かつ妥当性検証済みの指標を明記することにより、詳細に規定をした。それらの指標は、既存のソフトウェアを使って容易に得ることができる。

- The application of the socio-cognitive frameworks in the validation of other large-scale testing programs facilitated the identification of a taxonomy of explicit cognitive processes relevant to item development for both reading and listening tests. This approach was introduced to the TEAP reading and listening components in order to strengthen the validity argument by allowing for the evaluation of tasks expected in the TLU domain.

- 他の大規模試験における「社会的・認知的枠組み」の適用例を参考にしながら、TEAP リーディング・リスニング試験の項目開発に適する、より明確な認知プロセスの分類法を規定した。これにより、TEAP の目標言語使用領域に TEAP のリーディング・リスニング試験のタスクが、どの程度適しているかを評価することが可能になり、更なる TEAP の妥当性に関する議論に貢献する。

- A list of recommended analysis procedures was compiled for deriving empirical measures, using readily available software, for the reading and listening tasks and items in the TEAP item bank.

- TEAP リーディング・リスニング試験のタスク・項目バンク作成用に、既存のソフトウェアを使って実証的な分析指標を得るための、分析方法のリストが作成された。

- The analyses which were undertaken during the review project by applying available software to the test materials, including the content of the entire reading and listening item banks, provide encouraging empirical evidence for validity claims concerning the current versions of the TEAP Reading and Listening papers, especially with regard to their targeting of the proficiency level(s) of interest and their consistency across multiple forms. The test development team can feel confident that the tests are largely operationalising the test constructs which they were designed to measure.

- この再検討プロジェクトでは、リーディング・リスニング試験の項目バンク内の全ての項目を含むテスト材料の分析を、既存のソフトウェアを使用して行った。その分析結果は、現在の TEAP リーディング・リスニング試験の妥当性に関する実証的証拠を示した。とりわけ、対象としている能力レベルとタスク・項目の難易度の整合性と、複数のテスト・バージョンにおける一貫性が証明された。

- The outcomes of the project are expected to contribute to enhanced standardisation and improved efficiency of the item writing process, enabling an increased number of reading and listening test items to survive the pre-test stage. Furthermore, this publicly available report documents the increasing rigour of the TEAP test development and production processes by basing the test more firmly on a widely used external framework, thus ensuring enhanced accountability to the wider stakeholder community.

- このプロジェクトの結果は、項目作成プロセスを更に標準化し、効率を高めるのに有用であると期待される。このことは、パイロット試験の段階にて、より多くのリーディング・リスニング試験項目が、実際の試験で使用可能だと判断されることに繋がる。更に、この報告書は、TEAP とその土台である「社会的・認知的枠組み」との関係をより強固なものにすることで、今までより一層 TEAP テストの開発と試験作成プロセスが厳密になったことを記した。これらの情報を公表することにより、TEAP は、より広範囲のテストの利害関係者（stakeholder）コミュニティーに対する説明責任を果たしている、と言える。

1 Introduction

This report describes a specification review conducted between June 2012 and March 2013 of the reading and listening components of the Test of English for Academic Purposes (TEAP), a new test of academic English proficiency for university entrance purposes in Japan.

Drawing on Weir's socio-cognitive framework for developing and validating reading and listening tests (Weir, 2005; further elaborated for reading in Khalifa and Weir, 2009, and for listening in Geranpayeh and Taylor, eds. 2013), this project examined the reading and listening test tasks originally developed for the TEAP to explore various validity dimensions that underpin their effectiveness as measures of reading and listening ability.

This introductory section provides a brief overview of the aims of the TEAP Reading and Listening tests and of the rationale for undertaking a systematic review of the test task specifications for these components.

1.1 Background to TEAP

The Test of English for Academic Purposes (TEAP), which includes separate papers on four skills[1] (i.e. Reading, Listening, Writing and Speaking), was designed to measure the language ability of Japanese high school students intending to study at Japanese universities. While specifically taking into account the needs of students intending to study at Sophia University, which is a partner in the development of the test, from the outset the test has been intended to have the potential for wider application beyond one institution. A longer-term aim of the TEAP is to have a positive impact on English language education in Japan by revising and improving the widely varying approaches to English testing used in university admissions and by serving as a model of the English skills needed by Japanese university students to study at the university level in the EFL (English as a Foreign Language) context of Japan.

The TEAP is a collaborative test development project being undertaken by the Eiken Foundation of Japan (Eiken), which administers the EIKEN English proficiency tests to over 2 million test takers a year, and Sophia University, one of the leading private universities in Japan. Following the involvement of Professor C.J. Weir in the TEAP Writing project and of Dr Fumiyo Nakatsuhara in the TEAP Speaking project, Dr Lynda Taylor, also on the faculty staff at CRELLA (the Centre for Research in English Language Learning and Assessment) at the University of Bedfordshire in the UK, was contracted to serve as an external consultant to the TEAP team at Eiken and to provide specialist assistance in a proposed review of the Reading and Listening test specifications and test materials.

The specific role of the external consultant in this latest TEAP project (June 2012 – March 2013) was to review the existing test specifications, task materials and item writer guidelines for the reading and listening components against an established and widely recognised socio-cognitive framework for test development and validation, making recommendations for enhancing the existing documentation and offering suggestions for a possible research agenda in the future. The more detailed rationale for this project is outlined in Section 1.2.

TEAP is intended to evaluate the preparedness of high school students to understand and use English when taking part in typical learning activities at Japanese universities. The Target Language Use (TLU) tasks relevant to TEAP are those arising in academic activities conducted in English on Japanese university campuses. The "TLU domain" is defined by Bachman and Palmer (1996:44) as a "set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize." The TEAP Reading and Listening papers are therefore intended to cover academic contexts relevant to studying at university in the EFL context of Japan. Topics and tasks are

[1] The reading and listening tests are offered as a combined test which provides separate scale scores for each skill. The writing and speaking tests are optional components of the testing program.

related directly to studying and learning, rather than general, everyday activities or interaction that fall in the personal/private domain.

TEAP is a test of academic English proficiency which it is envisaged will be used for the purpose of university admissions, and as such, results must be able to discriminate between an appropriate range of student ability levels. At the same time, the programme is intended to make a positive contribution to English language learning and teaching in Japan by providing useful feedback to test takers beyond the usual pass/fail decisions associated with Japanese university entrance examinations. Following consultation with the main test stakeholders and consistent with guidelines published by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (MEXT, 2002, MEXT, 2003; MEXT, 2011), the key focus is a level of proficiency relevant to the B1 level of reading and listening ability as defined in the Common European Framework of Reference (CEFR) (Council of Europe, 2001), while measuring proficiency across the A2 to B2 levels of the CEFR. The A2 and B1 levels cover the levels of proficiency recommended by MEXT as goals for high school graduates[2], while the B2 level takes account of the more advanced level of language proficiency required in the TLU domain for TEAP (see 3.2 for more detail). It was felt that bringing the CEFR into the test design from the early stages would facilitate stakeholders' understanding of the test scores and task requirements. It should also be useful to report scores not only as scale scores but in terms of bands which can indicate to test takers their approximate level in terms of some external criterion, and the CEFR offered possibilities here.

The TEAP tests were designed to be able to provide useful feedback to students at the A2 level of proficiency, as this is one of the benchmark levels of ability recommended by MEXT, and one that is probably closer to reality for a large number of high school students. In this way, the TEAP programme from the outset placed the typical test takers at the centre of the test design, both in terms of what can realistically be expected of high school students and in terms of providing useful feedback. At the same time, in order to look forward to the more demanding TLU domain of the academic learning and teaching context of Japanese universities, it was decided that the tests should contain tasks capable of discriminating between students at a B1 level and the more advanced B2 level appropriate to the TEAP TLU domain, and thus be able to provide useful feedback for students at this more advanced level of ability. The test tasks and items in the TEAP Reading and Listening papers are thus designed to broadly span the CEFR A2, B1 and B2 levels. This approach is consistent with the decisions that were made when developing the TEAP Writing and Speaking tests.

As mentioned above, a long-term aim of TEAP is to foster a positive impact on English education in Japan. As described in Sasaki's (2008) summary of the 150-year history of English language education and assessments in Japan, greater emphasis is now placed on the teaching of listening and speaking skills as *practical communication abilities.* To achieve the goal of equipping students with practical communication abilities, some innovations in English education have been made in recent years, such as the inclusion of a listening component in the National Center Test for University Admissions administered by the National Centre for the University Entrance Examination from 2005. As already noted, the TEAP project has from the outset placed importance on creating positive washback, and the TEAP development team strongly hoped that the introduction of more detailed and transparent test specifications for the standardised TEAP Reading and Listening Tests would aid public understanding of how reading and listening abilities are being assessed in the TEAP and what the test scores mean for the context in which the test is used.

[2] MEXT has proposed the Grade 2 and Pre-2 levels of the EIKEN set of tests as appropriate English proficiency goals for high school graduates (MEXT, 2002; MEXT, 2003; MEXT, 2011). Based on research into the comparability of the EIKEN grades with the CEFR, Grade 2 and Grade Pre-2 can be considered relevant to the B1 and A2 levels of the CEFR, respectively (Dunlea, 2009; Dunlea, 2010; Dunlea & Figueras, 2012). The B1 level, then, represents the upper level of proficiency suggested by MEXT as a goal for high school graduates.

1.2 Rationale for reviewing the test task specifications for Reading and Listening

Until recently, development of the test content for TEAP Reading and Listening relied primarily upon the expert judgment of trained and experienced item writers and content specialists, working according to instructions provided in task-specific manuals that provide a holistic description of the test specifications.

The overarching aim of the 2012-2013 specification review project was to apply Weir's (2005) general socio-cognitive framework for test development and validation (together with more recent versions of this developed specifically for second language reading and listening assessment) to the existing TEAP Reading and Listening tests. The intention was to obtain a much more detailed and explicit description of each test task. The detailed task-specific descriptions were designed to identify contextual parameters and empirical measures relevant to each reading and listening task, as well as cognitive processes relevant to second-language reading and listening useful both for the validation of the tests and in the item writing and review process. It was anticipated that an explicit description of the most relevant contextual parameters, empirical measures, and cognitive processes for each task would provide a practical and effective quality-control tool, accessible to both item writers and editors, for ensuring that test tasks and items produced are verifiably consistent with the specifications. Additionally, basing the task descriptions on key parameters of Weir's socio-cognitive frameworks provides a strong theoretical foundation and ensures that terminology is consistent with an external and widely used framework. A similar descriptive approach had already been successfully developed and implemented for the TEAP Writing and Speaking tests (see Weir 2014 and Nakatsuhara 2014).

The following products from the specification review project were anticipated:

- Detailed specification tables for each reading and listening task based on Weir's socio-cognitive validity frameworks, identifying the most relevant contextual parameters for each task, including useful and previously validated empirical measures accessible through readily available software
- An explicit list of cognitive processes relevant to second language reading and listening which can be used as a basis for evaluating the processes targeted by TEAP tasks in relation to TLU tasks.
- A list of recommended analysis procedures for deriving empirical measures for tasks and items in the TEAP item bank using readily available software
- A written report outlining the approach adopted for the review procedures and the consultant's role in the process, including a list of relevant references.

Outcomes from applying these products in the longer term were expected to be:

- Increased standardisation and improved efficiency of the item writing process, leading to an increased number of test items surviving the pre-test stage
- Increased test rigour and enhanced accountability to stakeholders as a result of basing the test more firmly and explicitly on a widely recognised external framework.

2 Details of the review process and procedures

2.1 Scoping the review

Following a series of email exchanges as well as a face-to-face meeting in May 2012, the Project Proposal for the specification review was drawn up by the Eiken team in June 2012 in consultation with Dr. Lynda Taylor, the external consultant selected for the project. In June-July 2012 the Eiken team prepared a comprehensive pack of documentation which was supplied to the external consultant. The pack included: current TEAP test booklets (including Listening test CD), item writer manuals, TLU description, and specification tables.

2.2 Reviewing the test materials

The external consultant reviewed all the TEAP materials supplied by Eiken during August 2012 and began the process of drafting an initial template for developing specification tables for the Reading test. This

process involved:

- Reviewing the existing design and purpose statements for the TEAP Reading and Listening tasks
- Scrutinising the Listening and Reading sample test materials, including working through the test tasks in real time
- Cross-referencing the tasks (i.e. texts and items) to the documentation for item writers (e.g. item writer manuals, TLU domain paper)
- Reviewing recent PowerPoint presentations that explained the rationale for various aspects of the tests.

The initial conceptualisation and design of a generic task specification template was informed by a review of the socio-cognitive framework for conceptualising reading test validity proposed in Khalifa & Weir (2009:5, and see Appendix 1), as well as by the detailed content of Khalifa & Weir's Chapters 3 and 4 in their volume relating to Cognitive and Context Validity. Account was also taken of previous work to develop similar task specification tables for the TEAP Writing test project (Weir 2014).

2.3 Producing detailed draft specification tables for reading and listening tasks

Once a generic template had been drafted as described in 2.2 above, this was used by the external consultant to draw up a set of preliminary specification tables for the Reading test tasks which was sent to Eiken in late August 2012 for comment. This preliminary draft consisted of a one-page table or grid for each of the 6 reading tasks: Tasks R1, R2A, R2B, R2C, R3A and R3B. Each grid contained generic content relating to a reading test task together with an indication of those task-specific features that were considered relevant to each individual reading text and its accompanying task/items (based upon the expert judgement of the external consultant).

The preliminary draft of the specification tables for the set of 6 reading tasks was reviewed in Japan by the TEAP development team. It was then discussed during a 2-hour Skype meeting between the UK and Japan held in September 2012. Some modifications were made to the format and content of the specification grids in line with the feedback received from the Eiken team.

This consultative and iterative approach was continued over the next few weeks to further develop the specification tables for the 6 tasks in the TEAP Reading paper, helping to refine the description of the relevant contextual parameters. The approach was extended to produce a similar set of draft specification tables for the set of 5 tasks in the TEAP Listening paper.

Drafts of both the Listening and the Reading tables were reviewed in a series of discussions within the TEAP team at Eiken, and the tables were progressively refined through a series of 1-to-2-hour Skype meetings with the external consultant in the UK (held in November 2012, January and March 2013). Final versions of the specification tables were agreed and these are included Appendices 3 and 4 to this report.

2.4 Providing a list of analysis procedures for analysing the item bank contents

The progressive and iterative development of the specification tables for both Reading and Listening also enabled a set of empirical measures to be determined, making use of easily available software packages, which the external consultant advised might provide useful empirical indices for some of the contextual parameters. These analytical procedures were then applied by the TEAP team not only to the sample materials, but also to the complete content of the TEAP item bank of Reading and Listening materials.

2.5 Reviewing results from the analyses

The statistical results of these analyses were reviewed at two Skype meetings in March 2013. This stage confirmed the list of most useful analysis procedures for deriving empirical measures for items in the TEAP item bank, particularly the appropriate cut-off levels for selecting items and the measures that should be entered in the specification tables.

2.6 Preparation of final project report

As the final stage of the project, the external consultant completed a full report during March and April 2013 which described the overall process and specific procedures, together with a list of relevant literature references. This report was reviewed by the commissioning team at Eiken and feedback was invited on the content and format. All comments were carefully considered by the external consultant and taken into consideration in revising the project report to produce a final version which could be made publicly available.

3 General issues and comments

3.1 The contribution of the socio-cognitive framework

The socio-cognitive frameworks for validating reading and listening tests originally presented in Weir (2005) provided an important reference point for this specification review project. O'Sullivan and Weir (2011:20) described the socio-cognitive approach as "the first systematic attempt to incorporate the social, cognitive and evaluative (scoring) dimensions of language use into test development and validation."

Weir (2005) provided versions of the framework adapted for each of the four skills, and the frameworks for reading and listening were subsequently applied and refined in Khalifa & Weir (2009) and in Geranpayeh & Taylor (eds. 2013). Taylor (2011:25-28) provides a useful overview of the benefits of using these frameworks. The frameworks for reading and listening (as shown in Appendices 1 and 2) represent a principled and coherent approach to gathering validation evidence for reading and listening tests. The framework comprises *context* validity and *cognitive* validity which should ideally be established before the test becomes operational (i.e. *a priori* validation), and *scoring* validity, *consequential* validity and *criterion-related* validity which are usually examined and reported after the test event (i.e. *a posteriori* validation), or once the test is operational. It is particularly valuable that the framework conceptualises different aspects of validity in terms of temporal sequencing, thus offering test developers a clear plan of what validity evidence should be collected at what stage.

An important part of the TEAP specification review project has been the identification of appropriate models of cognitive processing involved in second language reading and listening that are useful for test validation and item specification and development. One of the important contributions of the socio-cognitive approach to validation has been to highlight the importance of test tasks approximating, as far as is possible under the constraints imposed by the testing situation, the cognitive processes that can be expected to be used when completing real-life language use tasks (Weir, 2005; O'Sullivan & Weir, 2010). The application of the Socio-cognitive frameworks for test validation to large-scale testing programs in relation to reading (Khalifa & Weir, 2007) and listening (Geranpayeh & Taylor, eds., 2013) has resulted in an explicit and defined taxonomy of processes useful for both evaluating the cognitive validity of test tasks and for item development. The specifications tables developed for this project have drawn directly on the processing models described in Khalifa & Weir (2007) and in Geranpayeh & Taylor (eds., 2013). The final sections of the specifications tables contain the list of cognitive processes which have been shown to be relevant to reading and listening. The processes which are expected to be operationalised by a particular task are highlighted for each of the TEAP test tasks. The models described in Khalifa & Weir (2007) and Geranpayeh (eds., 2013) represent to a certain extent a hierarchy of difficulty in terms of the cognitive load imposed by different processes, or levels of reading and listening. As such, they have proven useful in defining criterial differences between tests designed to measure at different levels of proficiency. Identifying the processes relevant to item development for each particular kind of TEAP task will, it is hoped, aid in the consistent production of test items targeting levels of proficiency as defined in the test specifications. Of course, empirical item difficulty has been repeatedly shown to be the result of the interaction of a range of contextual and cognitive features. As such we need to be cautious in associating any particular feature with item difficulty, and as Field (2013) notes, any one particular item is likely to

require several levels of processing. The specification tables have followed Field's caution (ibid) that, "What is of interest is the highest level of processing at which an item requires a test taker to engage." This principle has been followed in the specification tables. In some cases only higher-level processes are highlighted. This does not mean that the lower level processes are not engaged. They may be required in order to access the text and complete the task. What is being indicated is that the highlighted processes are the highest level of processing being anticipated and explicitly targeted by the items used in that particular task. Further research will of course be necessary to confirm the successful elicitation of some of the targeted processes (see the recommendations for future research studies in Appendix 8). An additional benefit of utilizing explicit models of cognitive processes is to improve the transparency of test specifications, helping to communicate the aims of the test more clearly to test users. Improving transparency in this way will be an important part of facilitating positive washback, which as has already been noted, is an important goal of the TEAP program.

3.2 The relevance of the CEFR

The importance and relevance of the CEFR to the TEAP development as a whole was referred to above in Section 1.1. Given the widespread awareness and use of the CEFR around the world today as a framework of reference for language learning, teaching and assessment, it was felt that bringing the CEFR into the test design from the early stages would facilitate stakeholders' understanding of task requirements and test scores, as well as indicate to test takers their approximate level in terms of some externally recognised criterion.

The TEAP tests were designed to be able to provide useful feedback to students at the A2 level of proficiency, as this is one of the recommended benchmark levels of ability and is probably closer to reality for a large number of high school students. At the same time, in order to look forward to the more demanding TLU domain of the academic learning and teaching context of Japanese universities, it was decided that the tests should contain tasks capable of discriminating between students at a B1 level and the more advanced B2 level appropriate to the TEAP TLU domain. The test tasks and items in the TEAP Reading and Listening papers are thus designed to broadly span the CEFR A2, B1 and B2 levels.

Several features of the TEAP Reading and Listening tests are worth mentioning at this point. It is clear that different tasks were originally designed to be appropriate for eliciting different levels of performance. The intention is that tasks gradually increase in difficulty (in terms of their cognitive demands), beginning with tasks designed to be accessible to A2/B1 level candidates and leading on to tasks aimed at higher proficiency levels, specifically the B2 level, thought to be appropriate for the TLU domain of university undergraduate classes. This structure is consistent with the overall aims of the test to provide useful feedback to students across these ability levels. A2 level candidates may indeed find the B2-level tasks inaccessible, but useful feedback should still be available to these students.

Attempts to reflect a measure of alignment with the CEFR are, however, not without their challenges. Although the TEAP Reading and Listening tests are designed to assess across the A2-B2 range, pre-test results showed that too many items tended to fall in the middle of that range. All operational TEAP Reading and Listening test sets are constructed from items in the item bank which have known statistical properties, and results are equated onto a common scale using Rasch analysis. The comparability of difficulty of test sets is thus ensured. An important operational consideration is to ensure that the item bank contains sufficient numbers of items across the levels of proficiency being targeted by the TEAP tests. In order to generate a proper balance of items, a more explicit way of identifying features of texts/items that relate specifically to each of the CEFR levels was needed. For this reason, the specification grids for each task have been developed to identify relevant features of the input texts (e.g. readability) and – as far as possible - the items (e.g. the cognitive processing activated) which correspond to each CEFR level targeted by the task. As has already been noted, empirical item difficulty results from the interaction of a number of contextual and cognitive features, and predicting difficulty in relation to the CEFR has been shown to be a complex issue (Alderson et al, 2006). An interesting ongoing research project for TEAP would be to

investigate the relationship between CEFR levels predicted on the basis of contextual and cognitive features and actual empirical difficulty.

3.3 Choice of measures for analyzing test material

The measures selected for analyzing the existing TEAP reading and listening test material represented those which are easily available through automated software packages that are freely accessible online and which do not require a sophisticated level of technical/statistical expertise, either to run the package or to interpret its outcomes. The four principal criteria used to guide the selection of analytical measures for this project were as follows:

- Use of measures which have been used in published studies and which thus permit comparison of results (e.g. with other widely recognized English proficiency tests such as IELTS and the Cambridge English examinations); this was especially important for deciding on final maximum/minimum levels for the specifications to guide future item writing and editing
- Use of measures which are readily available and interpretable for ongoing item writing and editing
- Use of measures which will be useful and accessible for test users themselves
- Use of measures that are easily obtainable through automatic analysis software and so do not require human judgment of individual items/texts.

Paul Nation's *Range* software (Nation, 2006) is freely available from Victoria University, New Zealand, and this analysis package is widely used and referenced for vocabulary research, especially research regarding lexical levels. *Compleat Lexical Tutor* (Cobb n.d.) offers another set of useful tools for analyzing vocabulary level, range and diversity mapped to the British National Corpus (BNC20) levels. Since the latter uses the same lists produced by Paul Nation for *Range* as the basis for its online vocabulary profiler tool, it was decided to adopt the *Range* option.

A significant advantage of using *Range* is its ability to analyze batches of specified texts at once, producing output files for all input files specified. Results are automatically output to text files, thus negating the need to copy and paste results from the screen online. For the TEAP specification review project, this greatly reduced the time needed to run an analysis for multiple texts from the item bank and it thus offered the most efficient approach given the timescale and resources available. Unlike *Vocabprofile* in *Compleat Lexical Tutor* (or *TextInspector*), *Range* does not, however, provide indices of lexical density and this might be a measure worth exploring further in future research.

Coh-Metrix is another freely available computational software package which analyzes linguistic data in terms of its lexical, syntactic and cohesive features. Several *Coh-Metrix* indices were initially considered as possible measures for this project. For example, *number of higher order constituents per sentence* is a potentially useful marker of syntactic complexity. Some other measures (e.g. *celex logarithm for content words*, *concreteness, mean number of modifiers per noun phrase, mean number of words before main verb*) were considered less useful taking into account criteria 1, 2 and 3 above, though they may prove useful to revisit in any future analysis if resources permit. A new version of the software – Coh-Metrix 3.0 – has recently become available, and the new and adapted measures the program offers may well be worth exploring for future studies of the TEAP input texts.

Some of the linguistic analysis functions in *Word* were also considered viable (e.g. word length, sentence length, readability indices), together with the publicly available experimental *TextInspector* tools, created by Professor Stephen Bax at CRELLA, which can offer valuable help in analyzing various lexical, syntactic and discourse features of text. It is important to note that some tools for linguistic analysis, e.g. readability indices and type-token ratios, have limited usefulness when applied to shorter texts (e.g. below 100 or 200 words) and thus are really only worth using with the longer texts for Part R3A and R3B. For example, *Flesch-Kincaid Grade Level* readability values are only appropriate for Parts R2C, R3A and R3B since stable and meaningful results generally require at least 200 words of text.

The test material for Reading task 1 was analyzed according to 3 different categories: the stem as a set, the targets as a second set, and the distractors as a third set. In this task the focus is on vocabulary knowledge, so it was relevant to analyze the vocabulary level of targets and options with the following caveats:

- Since stems and targets/options are very short, it made little sense to analyze individual items, so these were combined into a single text file for each of the 3 categories.
- Phrasal verbs were removed from targets and distractors as the difficulty of these levels is not transparently associated with frequency level.

For R2A, the main reading input is in the stems and options. For each item, a text file was produced which included the question and 4 options for that item.

For Reading tasks R2B, R2C, R3A and R3B and for Listening tasks R1, R1B R1C, R2A and R2B it was decided to focus on analyzing only the input texts, not the stems and options, taking into account issues of rationale as well as resource limitations:

- Parameters for stems/options are already set in terms of word length, thus are controlled for at the writing and editing stages.
- Since vocabulary levels, readability, etc. are likely to have the most impact on the accessibility of the input texts, identifying more closely defined parameters for these was felt likely to be most useful for criteria 2 and 3 above; also, for most studies these are the measures reported for texts used in test materials (criterion 1 above).
- Item difficulty is a complex interaction of more than just vocabulary features; e.g. for listening tasks the interaction of features such as position of target information in the input text, overlap between options and input texts, etc. may influence difficulty; the type of cognitive process targeted is likely to be a more pertinent predictor of difficulty, though this requires human judgment and could usefully be targeted as the focus of future research into the TEAP listening tasks (see Field 2013 for recent and comprehensive discussion of cognitive processing in listening tests).

Although the analytical measures referred to above have a direct practical application for reading test tasks, it is important to note that they are likely to be less appropriate for use with listening test materials, for several reasons. First, many of the text analysis tools were developed based on written text and for application to written text; they have not been properly validated for use with spoken language. Secondly, words are realized very differently in speech and writing, and so are apprehended differently. The same learner may recognize and understand the word in written text but not in continuous speech, due to variation in pronunciation, or to its location within the stream of speech. Thus for analysis purposes it may be inappropriate to consider it the 'same' word.

Other speech-appropriate measures exist for analyzing listening input, but these are often time-consuming and expensive, requiring manual analysis undertaken by raters with specialist expertise, and are therefore not yet widely used by test providers. *Mean length of utterance*, *propositional density* and *complexity* were all considered as desirable measures but all require human judgment and were rejected as viable options for this project given the limited time and resources available[3]. Though analysis measures for this study were restricted to those that could be calculated with automatic software, future studies might be able to take greater account of these measures.

Despite the limitations of using analytical tools designed for written text with listening materials, it is still useful for item writers and test constructors to run the listening tapescripts through easily accessible vocabulary analysis software in order to gain some idea of relative level, to control for level and to maintain consistency across forms. This will also enable a better understanding of the listening construct and the interaction of task features.

[3] Words per minute/second had already been calculated for all item banked items

The final list of measures utilized in the specification tables, the software used to calculate the measures, and the sections of the reading and/or listening tests to which they were applied are listed in Table 1[4]. In addition to analyzing texts associated with individual items in the item bank, the two vocabulary measures were applied to three complete test sets for both reading and listening (see section 3.4 for details).

Table 1

Measure	Software	Reading	Listening
AWL coverage	Range GSL	All	All
BNC14 word level coverage	Range BNC	All	All
Readability (Flesch-Kincaid Grade Level)	Coh-Metrix	R3A, R3B	
Sentence length (number of words)	Coh-Metrix	R2B, R2C, R3A, R3B	All
Text length (total number of words)	Coh-Metrix	R2B, R2C, R3A, R3B	All
Speech rate (words per minute).	N/A		All

It was anticipated that the final recommended criterial levels for each measure to be included in the specs for future publication to test takers and for use in item writing would be based upon the empirical results of analyzing the item bank contents in conjunction with reviewing the results for the same measures obtained for external criterion measures (e.g. values for Cambridge English reading and listening tests reported in Khalifa and Weir, 2009, and in Geranpayeh and Taylor (eds) 2013), as well as values reported by Green et al, 2009, in their study investigating university reading texts and IELTS. The Green et al study was considered particularly important with regard to Academic Word List (AWL) values, given that the Cambridge English AWL values are relatively low due to the tests' focus on assessing general English rather than EAP. Detailed values for all the measures listed above can be found in the task-specific specification grids in Appendices 3 and 4.

3.4 Lexical content and levels

The word-frequency lists developed by Paul Nation (2006) from an analysis of the well-established British National Corpus (BNC) provide a useful way of estimating the number of word families a test taker needs to know in order to be able to read and comprehend 95% of running words in a text. The lists are divided into 14 levels which each contain 1000 word families. The BNC-14 lists are provided with the relevant version of Range. In fact, output from Range using the lists allocates words to one of 16 levels: one of the 14 frequency levels noted above, Level 15 which consists of proper nouns, or Level 16 which is a list of commonly used interjections, etc. Any words not allocated to one of these categories would be put in the "not-on-the-lists" category.

The cumulative percentage of running words in a text covered by successive levels of the lists is displayed by Range and can be used to calculate the coverage of text by various levels of the BNC14 lists. As explained further in 4.7.5, this study used a criterion of 95% coverage of running words to identify the number of words necessary for a reader, or listener, to access and understand a text. The analysis followed the method described by Nation (2006) and included proper nouns in the calculation of the cumulative coverage needed to reach the 95% criterion, as proper nouns are assumed to pose a low learning and processing

[4] As noted previously, a number of measures available in Coh-Metrix were considered but not chosen for final use in the specification tables due to a lack of clear criteria for comparison. However, data on all measures for which Coh-Metrix provides feedback were obtained for item types to which Coh-Metrix was applied in Table 1. This data will provide a rich resource for ongoing review and validation.

burden (Nation, 2006)[5]. The result for this measure displays the number of word families necessary to reach the 95% criterion, e.g. 3000, 4000, etc. This measure is useful from the perspectives of both selection criteria 2 and 3 (see section 3.3 above), i.e. it is easy to identify words that need to be reviewed and possibly paraphrased or explained, and also easy to identify learning goals for test uses as the word lists on which the levels are based are easily available and widely used in research and teaching/learning.

Following Chujo & Oghigian (2009) and the pilot study on EIKEN/TEAP vocabulary levels, it was decided to analyse reading test sets (the 3 live tests) as a whole text, to explore the level needed to read and understand 95% of running words in a complete reading test. This analysis included input texts and questions, plus options. Only the 3 complete test sets used in live administrations were used for this part of the analysis.

Despite the earlier caveats regarding the lexical analysis of listening material, the vocabulary indices for the listening texts were considered appropriate for the following reasons:

- It might reasonably be assumed that the same principle of higher frequency being associated with higher learnability applies to listening as much as to reading; the higher the frequency of a word, the more likely a learner is to encounter it, and the greater the chance of reaching a critical input threshold for noticing a new word as well as for recycling and reinforcing a newly acquired form. The vocabulary level values were intended to be used to compare texts from different listening sections to one another as well as to arrive at general linguistic measures for maintaining consistency within sections of the listening test.
- The BNC14 lists from Paul Nation (2006) are based upon the spoken corpora of the BNC so there is some confidence that these frequency levels are also relevant to the listening context; consistent with the previous point, more frequent words are likely to be easier/more recognizable than less frequent words in listening as they would be in reading.
- Though based on written corpora, in the absence of a listening-specific academic corpus the AWL still provides a useful basis for vocabulary likely to be encountered in the academic context during lectures, etc.; as such, it offers at least some principled interim measure of the lexical level of listening texts for the TEAP TLU domain (as well as some measure of the content validity of texts, though there exists as yet no external criterion to refer to comparable with the 10% level of AWL coverage for written academic texts).

4 Results and discussion: Reading tasks

Appendix 3 shows the finalised specification grids for the set of 6 tasks in the TEAP Reading test. The grids are underpinned by the theoretical framework for conceptualising reading test validity which is shown in Appendix 1 – particularly in terms of the cognitive processing and contextual features of a reading test task. In addition, the specification grids report the various analytical measures which are considered to be relevant and useful for item writers and test constructors as a practical means of targeting texts, items and tasks at an appropriate level and across the intended proficiency range. The reader is referred to the grids in Appendix 3 for detail of the analytical values which are recommended for each reading task to match it to the targeted level.

The intention is that 4 broad types of reading should be represented and sampled across the Reading test as a whole: *careful local, careful global, expeditious local, expeditious global* (Weir and Khalifa, 2008; Khalifa and Weir, 2009). The genre, length and linguistic characteristics of each reading text are chosen in accordance with the stated target language use domain. Reading task difficulty is determined by a complex interaction of features that includes not only contextual parameters but also the types of cognitive

[5] Khalifa & Schmitt, (2010) use an alternative approach in which proper nouns are first cleaned from the text, and the cumulative coverage is calculated on the cleaned texts.

processing which the task is believed to elicit. The level of cognitive demand is designed to increase from Part R1 to Part R3B.

This section addresses each reading task in turn, highlighting any specific features of note for that task (4.1 to 4.6). Features that are held in common across the set of 6 reading tasks (e.g. timing, domain relevance, lexical level) are discussed in 4.7 under the heading of General Comments.

4.1 Part R1

This task has a vocabulary focus and is aimed to range across the A2/B1/B2 levels.

The stem for Part R1 is considered as the input text. There are 2 sets of quantitative measures for the vocabulary, etc. in Part R1 items, one for the input texts and one for the target and the distractors. Treating the stem separately as the input text permits easier definition of abstractness, etc. for this element which is important for eliciting the target in context. It was agreed that Part R1 can include some *fairly abstract* items though most will be *mostly concrete*.

Vocabulary targets will in principle be constrained within the 5,000 word limit and should target vocabulary depth (varied uses of the same lexical items) not just breadth. A word beyond the 5,000 word limit may be targeted if it can be demonstrated that it is relevant and common to the TLU, and that test takers can reasonably be expected to have knowledge of that word. All stems should also be within the 5,000 word limit. (See full discussion of vocabulary level recommendations in section 4.7.5 below.)

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Careful reading: local; word recognition; lexical access; syntactic parsing; establishing propositional meaning* (see Appendix 3).

4.2 Part R2A

This task is aimed at the A2/B1 level. All items require reading across the stem, alternatives and the graph to identify the correct answer. Global reading for this task includes the graph. This is quite a complex task, requiring a number of processes.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Expeditious reading: local; expeditious reading: global; careful reading*: *local; word recognition; lexical access; syntactic parsing; establishing propositional meaning* (see Appendix 3).

4.3 Part R2B

This task is aimed at the A2/B1 level. With regard to cognitive processing, this task is intended to elicit the creation of a *text-level representation* as well as some measure of *expeditious reading* (though see 4.7.1 below). Though there is a limited amount of content, it is expected there will be enough structural organization of propositions within an overall textual framework (e.g. as a notice) to require the formation of text-level representation. Items targeting local expeditious reading are typically included. For example, some texts may include a list of bullets and the item will require test takers to scan for specific information across the bullet points. Such items are intended to be appropriate for A2/A2+ level test takers.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Careful reading*: *local; careful reading: global; word recognition; lexical access; syntactic parsing; establishing propositional meaning; building a mental model; creating an intertextual representation* (see Appendix 3).

4.4 Part R2C

This task is aimed across the A2/B1/B2 levels. It is intended to stimulate local expeditious reading as one of the processes and, in a similar way to Part R2B, some items target the lower, A2 level and require scanning for specific information. Further validation studies are needed to confirm this hypothesis, but it seems likely that test takers employ both reading types with this item, i.e. read the text carefully first, then, after encountering the questions, return to the passage and scan for specific information, depending what the item calls for.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Expeditious reading: global; careful reading*: *local; careful reading: global; word recognition; lexical access; syntactic parsing; establishing propositional meaning; inferencing; building a mental model; creating an intertextual representation* (see Appendix 3).

4.5 Part R3A

This task is aimed at the B1/B2 levels. It is designed to require careful global reading. The item gaps focus on discourse markers and features associated with establishing textual cohesion by marking the relationship between micro-propositions and macro-propositions. The task bears similarities with what has been referred to as a 'discourse cloze'. Targets should require careful reading across sentences to understand the logical flow and relationship of the propositions.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Expeditious reading: global; careful reading: global; word recognition; lexical access; syntactic parsing; establishing propositional meaning; inferencing; building a mental model* (see Appendix 3).

4.6 Part R3B

This task aims to elicit careful global reading. The four dimensions of reading types elaborated in Khalifa & Weir (2007) are broad categories which subsume a number of potential reading processes relevant to item specification. In an investigation of the reading construct of IELTS tests, Weir et al (2009) identified the following range of purposes as belonging to Careful Global reading:

- Establishing accurate comprehension of explicitly stated main ideas and of explicitly stated main idea or supporting details across sentences
- Making propositional inferences
- Establishing how ideas and details relate to each other in a whole text
- Establishing how ideas and details relate to each other across texts.

Establishing comprehension of main ideas and supporting details across sentences is a commonly employed focus of item development in reading tests. Along with propositional inferences, these two categories will form an important focus of item development for Part 3B. The third purpose, establishing how ideas and details relate to each other in a whole text, has been identified as an important higher-level reading purpose relevant to academic contexts, but has proven more difficult to operationalize in test items. Item types which may be used to elicit this important reading process may include:

- Asking test takers to select the best overall title for the text – which requires test takers to have read the whole text and integrated the various micro- and macro-propositions into an overall textual representation;
- Asking test takers how the writer's argument or stance changes over the course of a text – which requires test takers to have read all of the text to know which order the different stages occurred.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Expeditious reading: global; careful reading:*

global; word recognition; lexical access; syntactic parsing; establishing propositional meaning; inferencing; building a mental model (see Appendix 3).

4.7 General comments

4.7.1 Time constraints in the reading test

In the original design and development of the TEAP Reading test, an overall time (70 minutes) was set for completing all the tasks within it. Pre-pilot and piloting stages had shown the overall time allocated for the test to be sufficient for test takers. As is the case for many tests of reading, no time allocation or recommendation is currently given to students when completing the individual tasks. Scrutiny of the response data following the pretesting, combined with feedback questionnaires from test takers, suggests that test takers are not under time pressure when completing the tasks. This is consistent with the design intention that students with B2 ability should be able to employ *careful reading* of the texts, including the longer R3A and R3B texts. It is considered appropriate if students with reading ability below B2 may be pressed for time and not able to complete all of the tasks. However, the current approach makes it more difficult to ensure that students are activating some of the intended reading processes reflecting *expeditious reading* such as skimming, scanning and search reading, i.e. processes which are less likely to be fully elicited if there is no time pressure on the reader. Future small-scale validation studies could perhaps undertake recall protocols and interviews with test takers to explore more closely the reading processes that are actually being employed for each reading test task, thus providing useful validation evidence for claims as well as possible insights for any future revision of the test.

4.7.2 Domain relevance for the reading texts

Given the test's intended purpose and context of use, it was agreed that only public and educational domains are relevant to TEAP and that this classification system should serve as guidance to item writers. Texts from the occupational domain are therefore not included in the test. It was also acknowledged that some texts, such as newspaper articles, may originate in the public domain but that they become particularly relevant to TEAP because they are brought into the educational domain, for example to be used in the classroom as learning materials or source materials for teaching and learning.

4.7.3 Nature of information in terms of concreteness/abstractness

The nature of information in terms of the abstractness dimension uses the four-level scale developed for the CEFR Content Analysis Grids developed by Alderson et al (2006). This distinction has since been applied to the analysis of reading items by Khalifa and Weir (2007) and the analysis of EFL tests in Taiwan (Wu, 2012). A working definition of the four levels of this dimension had been developed for internal EIKEN content analysis manuals, and this definition of abstractness was adopted as a useful working definition. Adjusting items on this dimension may be one way to manipulate difficulty and create more B2-level items, but it was agreed it was not appropriate to have only *mostly abstract* items at the B2 level.

4.7.4 Understanding implicit and explicit meaning

Inference is considered to be an important academic skill; it involves the joining of related propositions within a text even if the relationship between propositions is not explicitly marked. The ability to make inferences is integral to forming a global, textual representation, but the necessary parts for forming that representation still reside within the text (rather than being dependent on background or general knowledge). Given that inferencing is fundamental to successful reading comprehension (as not all necessary or relevant information can be included in a text), it is generally accepted that it takes place at all levels of ability – it is not simply a high level skill. Bridging or necessary inferences are a suitable candidate for testing (but not pragmatic or elaborative inferences which tend to be more personal and idiosyncratic, typically shaped by individual cultural or experiential knowledge). For TEAP reading tasks, the approach will tend to entail the need to make inferences by synthesizing propositions from different parts of a text and understanding the attitude and tone of the writer even if this is not explicitly stated. The information is

always intra-textual and items will not require test takers to bring in outside knowledge to answer any question.

4.7.5 Setting a general vocabulary level for the TEAP test

The overall aim of the lexical analyses for the TEAP reading materials was to be able to make a general recommendation for test takers in the public specifications regarding vocabulary levels for the tests as a whole, i.e. the vocabulary size needed to understand adequately a typical TEAP Reading Test (based upon the number of word families needed to know 95% of running words in a text). The word families are calculated from the BNC14 lists, a set of vocabulary lists publicly available and for which documentation and validation evidence exist (Nation, 2006). Though both Hu & Nation (2000) and Nation (2006) recommended 98% coverage as the criterion, this is a difficult level to achieve in practice, especially with short texts where a small number of low-frequency words are likely to have a disproportionate effect. Furthermore, a 98% criterion allows little leeway for using currently common words which might be 'low frequency' simply as an artefact of the BNC corpus (now more than 20 years old) and the methodology for creating lists. Nation (personal communication, March 2013) confirmed the reasonableness of adopting the 95% criterion for the reasons outlined above.

Although a number of frequency lists derived from BNC data and other corpora are available (some lemma-based, some based on word families), the BNC14 lists were chosen as they have been validated using a number of spoken and written genres. Text coverage figures are available from these studies against which to compare results from BNC14 analyses of test tasks. The BNC14 lists remain an explicit, transparent and readily available resource that has been shown to be consistent and comprehensive in its coverage and in the rank of words allocated to each level. The BNC14 lists have also been used in studies of texts relevant to the TEAP TLU domain, including Chujo & Oghigian (2009) investigating how many words are needed to cover 95% of the words in TOEFL, TOEIC and EIKEN tests. The BNC14 lists replace the earlier GSL 1,000 and 2,000 measures in the specifications to ensure that the TEAP can be constructed to focus primarily upon the CEFR B1-B2 levels. A comparative analysis by the Eiken team showed that 95% of the GSL base words are contained within the first 4 BNC levels and almost 98% within the first 5,000, meaning that the lexical levels specified in terms of the BNC14 will cover the vast majority of words in the GSL. Thus the GSL remains a relevant learning resource, while not being sufficient in itself to provide the vocabulary needed to read with ease at the B1-B2 level in TEAP.

Once the vocabulary-learning goals have been recommended, the test developers' responsibility is to maintain appropriate vocabulary levels across the test sets used in live administration.

Based upon the analyses undertaken in the specification review project, the vocabulary size judged necessary to access 95% of running words in TEAP tests amount to 4,000-5,000 words (i.e. word families in the first 5 levels of BNC14 Levels of Nation, 2006). These values are based upon the results of analyzing the 3 live tests and also take into account the results of analyses published for other tests at comparable levels and designed for similar purposes (Green et al, 2009; Khalifa & Schmitt, 2010). An upper range of 5,000 words is recommended for accessing B2-level, extended texts (e.g. R3A and R3B). A lower level of 4,000 words is recommended for accessing (95%) of B1-level, shorter texts. In the future, all test sets will be analyzed to ensure that vocabulary size constraints are met at the test level.

Since the task-specific analyses revealed a greater range of variation with higher levels needed to cover 95% of some texts, a slightly more flexible approach is recommended for the individual task-based texts. At the task level: vocabulary levels of texts for R2A, R2B, L1A, L1B and L1C should be within the 4,000 word limit; for R1A, R3A, R3B, L2A and L2B, texts should be within the 5,000 word limit. Individual input texts may exceed the 5,000 word level provided that the test set overall does not require a greater vocabulary size than 5,000 words. A higher level will be acceptable if:

- It can be demonstrated that although words are low frequency (i.e. higher level) on the BNC14 lists, they are nowadays common and test takers can be expected to know them (e.g. Internet, email)

- An attempt at paraphrasing would result in awkwardness
- The words can be shown by content specialists to be supported by the context of the text, and test takers can be expected to understand the meaning of the words from context
- No extra-textual, specialised content knowledge would be needed to understand the vocabulary items.

Since 'academic' words are directly relevant to the TLU domain, it is recommended that the 550 words on the Academic Word List (AWL) should all be legitimate candidates for inclusion in the TEAP tests, even if some of these words are beyond the 5,000 word limit. Given its relevance, the test developers advise that the entire AWL should constitute a manageable and desirable learning goal for test takers. At the test level, the aim will be for the AWL to account for in the region of 7% of overall words in the Reading test (in the region of 5% for the Listening test to reflect the difference between written and spoken language – see more on this below). This is based on comparison to IELTS (Green et al, 2009). Analyses of vocabulary levels in the reading texts used for the pilot version of the TEAP reading test showed a 5% coverage level for the AWL (Dunlea, 2010). Information from field trialing indicated that more cognitively demanding items were required, and adjustments to the item writing manuals made in response to the results of piloting resulted in a slight increase in AWL coverage, which it is posited is more in line with the TLU definition and test purpose. Studies have shown that the AWL covers approximately 10% of unmodified academic texts (Coxhead, 2000; Green et al, 2009). However, this may be too great a burden for high school students in an EFL context, particularly for a test designed with B2 as an upper range. A higher level of AWL words in a text may be accepted if the words are understandable from the context and relevant to the content of the text. The rationale for the position of TEAP regarding the AWL is explained more fully below.

The source texts for the AWL cover a wide variety of academic disciplines. The inclusion criteria prioritized range over simple frequency (a word had to occur at least 10 times in each of the four main areas, and in 15 or more of the 18 subject areas – Coxhead, 2000). Almost two-thirds of the texts were sourced in New Zealand but all were written for an international audience and covered academic journals as well as text books. Source texts were also taken from the Learned and Scientific section of the Brown corpus and the Lancaster/Oslo-Bergen corpus (Coxhead, 2000). Although the language variety / geographical bias may have had some effect, the source texts can be regarded as relatively international (if somewhat dated). At the same time, the Green et al (2009) study replicated the original findings of 10% coverage for a corpus of texts used in one university, which would seem to validate the original study's findings and support the usefulness of the AWL across contexts.

A brief search for articles specifically investigating the AWL coverage of spoken corpora did not return any studies with this focus. The Eiken team conducted a small-scale analysis of academic lectures from the MICASE corpus at Michigan University to investigate differences of AWL coverage for spoken academic contexts. The lectures ranged from 7,000-12,000 words and covered a range of disciplines. Results corresponded with data in the TEAP item bank and support the decision to set a lower AWL coverage for listening items (see Appendix 5).

In the future, AWL coverage at task and test level will be routinely monitored by the test developers to check the percentage levels. There may be scope for a small-scale research study in the future to explore the issues associated with the AWL in greater depth with the TEAP test population.

In addition to the work associated with the AWL, the specification review project undertook a supplementary study to investigate the appropriacy of the TEAP vocabulary levels for typical test takers (high school students in the EFL context in Japan). This involved an automated analysis of the content of high school text books used for teaching English in Japan. Text books were grouped into the series in which they belonged. Typically a series will be designed to cover the three years of high school education. Different series of text books are offered by publishing companies for use in one of the variety of English language subjects for which MEXT provides guidelines in the Courses of Study. Individual text books within a series were collapsed together to form one large text file representing the entire series. The logic for this

methodology was that the total vocabulary covered by the entire series, rather than just one text book aimed at one particular year of study, would more accurately represent the vocabulary to which a student whose school had selected a particular series would be exposed during the three years of his or her high school education. The results showed that 3,000 to 4,000 words from BNC14 lists would be sufficient to cover 95% of almost all 37 series of text books (see Appendix 6). The results support the TEAP guidelines in the following way. MEXT recommends EIKEN Grade Pre-2 (CEFR A2) and EIKEN Grade 2 (CEFR B1) as benchmarks for high school graduates. The TEAP test posits 3,000 to 4,000 words as sufficient for the B1-level reading tasks. Typical high school students would thus have been exposed to vocabulary sufficient to read 95% of B1-level texts in the TEAP test. However the TEAP also raises the bar for students who wish to acquire a B2 level of performance (in anticipation of more advanced academic language use in the university TLU domain). For such students an extra learning goal (transparent and realizable) is recommended of a further 1,000 word families beyond what is likely to be encountered in high school text books.

Regarding vocabulary guidelines, the following points should be made clear:

- Word-family based frequency lists do not take account of multi-word lexical items for which the meaning of the whole is not transparent from the individual parts.
- The definition of word families used for the BNC14 lists is an appropriate indication of vocabulary knowledge for advanced level learners (Nation, 2006), but may be too broad for lower level learners as explained above.
- The vocabulary lists are corpus-dependent. Although some American English texts were included in the corpora to make the BNC14 lists, there is still a bias towards British English.
- The age of the corpora used to compile the lists means that some words which are common in modern contexts have a very low frequency on the BNC14 (e.g. Internet). When expert judgment and empirical evidence from other sources is available (e.g. usage in online corpora) to support the use of such words, exceptions may be made to the overall vocabulary guidelines. Nonetheless, at the test level (where the total word-count is much higher), the default should be to maintain the vocabulary guidelines of the test specifications.

Despite the caveats, it should be emphasized that the BNC14 lists were constructed using the spoken subcorpora from the BNC. Validation research has demonstrated that the levels show consistent, comprehensive coverage of a range of texts from both written and spoken genres (Nation, 2006).

5 Results and discussion: Listening tasks

Appendix 4 shows the finalised specification grids for the set of 5 tasks in the TEAP Listening test. The grids are underpinned by the theoretical framework for conceptualising listening test validity which is shown in Appendix 2 – particularly in terms of the cognitive processing and contextual features of a listening test task. In addition, the specification grids report the various analytical measures which are considered to be relevant and useful for item writers and test constructors as a practical means of targeting texts, items and tasks at an appropriate level and across the intended proficiency range. The reader is referred to the grids in Appendix 4 for detail of the analytical values which are recommended for each listening task to match it to the targeted level. Speech rate is considered in terms of words per minute (wpm) and the issues associated with this feature are discussed more fully in section 5.6.1 below.

5.1 Part L1A

Part L1A is a short dialogue aimed at the A2/B1 level. This part aims to target a careful understanding of propositions contained within the dialogue. The number of words per sentence will be maintained at an approximate average of 10, with a speech rate average of 150 wpm.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted

cognitive processes identified in the task specification table: *Listening for main idea/important information/key message; listening for detailed/specific information; decoding acoustic/phonetic (and visual) input; lexical search; syntactic parsing; establishing propositional meaning; constructing a meaning representation; constructing a discourse representation* (see Appendix 4).

5.2 Part L1B

Part L1B is a short monologue aimed at the A2/B1 level. The number of words per sentence will be maintained at an approximate average of 15-16, with a speech rate average of 150 wpm.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Listening for main idea/important information/key message; listening for detailed/specific information; decoding acoustic/phonetic (and visual) input; lexical search; syntactic parsing; establishing propositional meaning; constructing a meaning representation; constructing a discourse representation* (see Appendix 4).

5.3 Part L1C

Part L1C is a short monologue aimed at the A2/B1 level. The number of words per sentence will be maintained at an approximate average of 15-16, with a speech rate average of 150 wpm.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Listening for main idea/important information/key message; listening for detailed/specific information; decoding acoustic/phonetic (and visual) input; lexical search; syntactic parsing; establishing propositional meaning; constructing a meaning representation; constructing a discourse representation* (see Appendix 4).

5.4 Part L2A

Part L2A is a longer dialogue aimed at the B1/B2 level. The number of words per sentence will be maintained at an approximate average of 10, with a speech rate average of 150 wpm.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Listening for main idea/important information/key message; listening for detailed/specific information; Listening to infer opinion/attitude/intention; decoding acoustic/phonetic (and visual) input; lexical search; syntactic parsing; establishing propositional meaning; constructing a meaning representation; constructing a discourse representation* (see Appendix 4).

5.5 Part L2B

Part L2B is a longer monologue aimed at the B1/B2 level. The number of words per sentence will be maintained at an approximate average of 16-18, with a speech rate average of 150 wpm.

Examples of the task which were reviewed by the consultant were judged to be eliciting the targeted cognitive processes identified in the task specification table: *Listening for main idea/important information/key message; listening for detailed/specific information; Listening to infer opinion/attitude/intention; decoding acoustic/phonetic (and visual) input; lexical search; syntactic parsing; establishing propositional meaning; constructing a meaning representation; constructing a discourse representation* (see Appendix 4).

5.6 General comments

Since issues concerning domain relevance, nature of information, implicitness/explicitness and setting vocabulary level were addressed in sections 4.7.2 – 4.7.5 and apply to listening in the same way as they do for reading, they will not be repeated here. With regard to time constraints, given that listening tasks are completed online in real time and in lock-step with the recording, the issue does not really apply.

5.6.1 Rationale for speech rate

It is recognized that wpm is a very rough measure, as the variable length of words can impact on the time needed to produce them. For example, a monologue with longer, less frequent words (e.g. likely to be encountered in an academic lecture) may have a much lower speech rate measured in wpm than a monologue containing shorter, more frequent words, even though there may be less difference in the actual rate of articulation. Syllables per second is sometimes recommended as a more accurate measure, but is more time-consuming to calculate so was not used in this study. As Buck (2001) notes, despite the drawbacks of the wpm measure, it remains widely used because of its practicality. As the measures used in this study were also designed to be used in ongoing item writing and review, it was decided to use the wpm measure as it will be easier to implement in ongoing item development. The limitations for this measure need to be acknowledged, however, and recommendations will remain approximate guidelines only.

Table 2 provides an overview of some of the data for speech rate which has been published. The impact of word-length on wpm rates as noted by Buck (2001) was mentioned above. This feature may explain some of the higher wpm rates seen across the Cambridge main suite tests. The general vocabulary level and AWL levels of these tests, including FCE, is generally lower than TEAP, possibly reflecting a more general rather than academic/CALP TLU domain for these tests. The words being used may also be higher frequency and thus shorter, resulting in a higher wpm for texts used in these tests. The ranges recommended for TEAP avoid wpm rates crossing over into the level which studies have indicated may impact on comprehension test scores. Both Griffith (1992) and Robinson et al (1997) report statistically significant differences in test scores when comparing the effect of listening texts for the slower rate (approximately 130) compared to the average rate (approximately 188). Maintaining the average rate of 150 wpm will allow the TEAP developers to avoid the effect of speech rate interfering with test scores and ensure comparability (and fairness) across test forms.

While the rates in the specification tables may be slower than some of the reported results in Table 2, the rate selected for TEAP listening texts falls within the range for lectures noted for native speakers (NS) by Robinson et al (1997) and it exceeds the average noted by Tauroza et al (quoted in Buck, 2001) for lectures to non-native speakers (NNS). The TEAP TLU domain is the EFL context of Japan, in which both NS and NNS interlocutors, particularly lecturers and teachers, will be aware of the level of students and can be expected to make some accommodations. For this reason, the rate designated by Pimsleur et al (1977) can reasonably be considered appropriate and authentic for the TLU domain.

Table 2 Range of speech rates mentioned in literature

130-160	Moderately slow	Pimsleur, Hancock, & Furey (1977)
Below 130	Slow	
Approx. 127	Slow	Griffiths (1992)
Approx. 188	Average	
Approx. 250	Fast	
140	Lectures to NNS	Tauroza & Allison, quoted in Buck (2001)
160	Radio monologues	
190	Interviews	
210	Conversations	
100-180	Range of lectures (NS)	Carver, quoted in Robinson, Sterling, Skinner, & Robbinson (1997)
135	Speed beyond which students (NS)	Ladas, quoted in Robinson, Sterling,

	cannot take lecture notes	Skinner, & Robbinson (1997)
200	Typical medium-paced conversational rate	Field (2012)
150.6 (2.51 wps)	Average rate for KET (A2)	
167.4 (2.79 wps)	Average rate for PET (B1)	
207.6 (3.46 wps)	Average rate for FCE (B2)	

6 Conclusions and further recommendations

This report has described the aims, process and outcomes of the specification review project for the TEAP Reading and Listening tests. The report constitutes part of the *a priori* test validation activity which Weir (2005) advises is essential for the sound development of any test.

The test specification tables for the TEAP Reading and Listening papers which have emerged from the review project succeed in making much more explicit the cognitive and contextual parameters of the reading and listening tasks for the benefit not only of the item writers and editors but also the wider test stakeholder community. For example, the tables will now make it much easier for the TEAP developers to provide more explicit information to test takers about what kind of reading approaches and processes are intended to be elicited by the different tasks in the Reading test. This is consistent with the approach to eliciting positive washback recommended by Green (2013).

The analyses which have been undertaken during the review project by applying available software to the test materials, including the content of the entire reading and listening item banks, provide encouraging empirical evidence for validity claims concerning the current of the TEAP Reading and Listening papers, especially with regard to their targeting of the proficiency level(s) of interest and their consistency across multiple forms. The test development team can feel confident that the tests are largely operationalising the test constructs which they were designed to measure.

On-going research and validation studies will nevertheless be important as the TEAP tests move into an operational phase in the near future so that a sound and comprehensive validity argument can be assembled and maintained, and can be used to inform the continuing evolution of the test over time (see Appendix 8).

Acknowledgements
Grateful acknowledgements are due to the members of the TEAP project team at Eiken who assisted with this collaborative review project. Without their enthusiasm and ongoing commitment over a relatively short timeframe, this project could not have been completed. Special thanks go to Jamie Dunlea, Todd Fouts, Shinnosuke Morita, and Yusuke Okuwaki. Thanks are also due to my colleagues at CRELLA who kindly gave me their feedback and advice at various stages of this project.

References

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.

Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Brown, J.D., & Yamashita, S.O. (1995). English language entrance examinations at Japanese universities: 1993 and 1994. In Brown, J.D, & Yamashita, S.O. (Eds.), *Language teaching in Japan*. Tokyo. JALT.

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

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.

Carver, R.P. (1982). Optimal rate of reading prose. *Reading Research Quarterly*, 18, 56-88.

Cobb, T. n.d. Web Vocabprofile / BNC-20 Version 3.2. Retrieved from http://www.lextutor.ca/vp/bnc/

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

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

Dunlea, J. (2010). Using word frequency lists to investigate the vocabulary used in a pilot version of a new entrance exam. Paper presented at the 14th Japanese Language Testing Association Conference, Toyohashi, Japan.

Dunlea, J. (2010).「英検と CEFR の関連性について Part 2」 [About the relationship between EIKEN and the CEFR, Part 2].「Eiken 英語情報 1・2 月号」 [Eiken Eigo Joho, January/February Edition.] Retrieved from http://www.eiken.or.jp/eiken/group/result/pdf/report_02.pdf

Dunlea, J. (2009).「英検と CEFR の関連性について Part 1」[About the relationship between EIKEN and the CEFR, Part 1].「Eiken 英語情報 11・12 月号」[Eiken Eigo Joho, November/December Edition.]. Retrieved from http://www.eiken.or.jp/eiken/group/result/pdf/report_02.pdf

Elliott, M. and Wilson, J. (2013). Context validity. In A. Geranpayeh and L. Taylor (eds.), *Examining Listening: Research and practice in assessing second language listening* (pp. 152-241). Cambridge: UCLES/Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh and L. Taylor (eds.), *Examining Listening: Research and practice in assessing second language listening* (pp. 77-151). Cambridge: UCLES/Cambridge University Press.

Heatley, A., Nation, I.S.P. and Coxhead, A. (2002). *RANGE and FREQUENCY programs*. http://www.vuw.ac.nz/lals/staff/Paul_Nation

Green, A. (2014). The Test of English for Academic Purposes (TEAP) Impact Study:

Report 1 - Preliminary Questionnaires to Japanese High School Students and Teachers. Eiken Foundation of Japan. Internal report scheduled for publication in 2014.

Green, A., Unaldi, A., & Weir, C. (2009). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, 27(3), 1-21.

Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, 26(2), 385-395.

Hirsch, D. & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2), 689-696.

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Khalifa, K. & Weir, C. J. (2009). *Examining Reading: Research and practice in assessing second language reading.* Cambridge: UCLES/Cambridge University Press.

Khalifa, H, & Schmitt, N. (2010). A mixed-method approach towards investigating lexical progression in Main Suite Reading test papers'. *Cambridge ESOL: Research Notes* 41: 19-25.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal*, 28(1), 77-96.

Ladas, H.S. (1980). Note taking on lectures: An information-processing approach. *Educational Psychologist*, 15, 44-53.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds) *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.

MEXT (2002). *Japanese Government Policies in Education, Culture, Sports, Science and Technology 2002.* Retrieved on April 17, 2012 from http://www.mext.go.jp/b_menu/hakusho/html/hpac200201/hpac200201_2_015.html.

MEXT (2003). *Action plan to cultivate "Japanese with English abilities".* Retrieved on March 7, 2007 from http://www.mext.go.jp/b_menu/houdou/15/03/03033101/001.pdf.

MEXT (2008). *The course of study for upper secondary school.* Retrieved on May 1, 2010 from http://www.mext.go.jp/a_menu/shotou/new-cs/index.htm.

MEXT (2011). Five proposals and specific measures for developing English proficiency in international communication: provisional translation. Commission on the Development of Foreign Language Proficiency. Retrieved from http://www.mext.go.jp/english/elsec/1319701.htm.

Nakatsuhara, F. (2014). A research report on the development of the Test of English for Academic Purposes (TEAP) speaking paper for Japanese university entrants. Internal report scheduled for publication in 2014.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59-82.

O'Sullivan, B. and Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (ed.), *Language Testing: Theories and Practices* (pp. 13-32). Basingstoke: Palgrave.

Pimsleur, P., Hancock, C., & Furey, P. (1977). Speech rate and listening comprehension. In M. K. Burt, H. C.

Robinson, S.L., Sterling, H.E., Skinner, C.H., & Robinson, D.H. (1997). Effects of lecture rate on students' comprehension and ratings of topic importance. Contemporary Educational Psychology, 22, 260-277.

Sasaki, M. (2008). The 150-year of English language assessment in Japanese education. *Language Testing* 25 (1), 63-83.

Tauroza, S., & Allison, D. (1990). Speech rates in British English, *Applied Linguistics*, 11, 90-105.

Van Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics*, Advance access published online. Retrieved from http://applij.oxfordjournals.org/content/early/2012/12/18/applin.ams074.abstract

Weir, CJ, Hawkey, Green, RA, and Devi, S, 2009. 'The relationship between the Academic Reading construct as measured by IELTS and the reading experiences of students in the first year of their courses at a British University', IELTS Research Reports 9, British Council, pp 97-156.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* London: Palgrave Macmillan.

Weir, C. J. (2014). A research report on the development of the Test of English for Academic Purposes (TEAP) writing paper for Japanese University entrants.

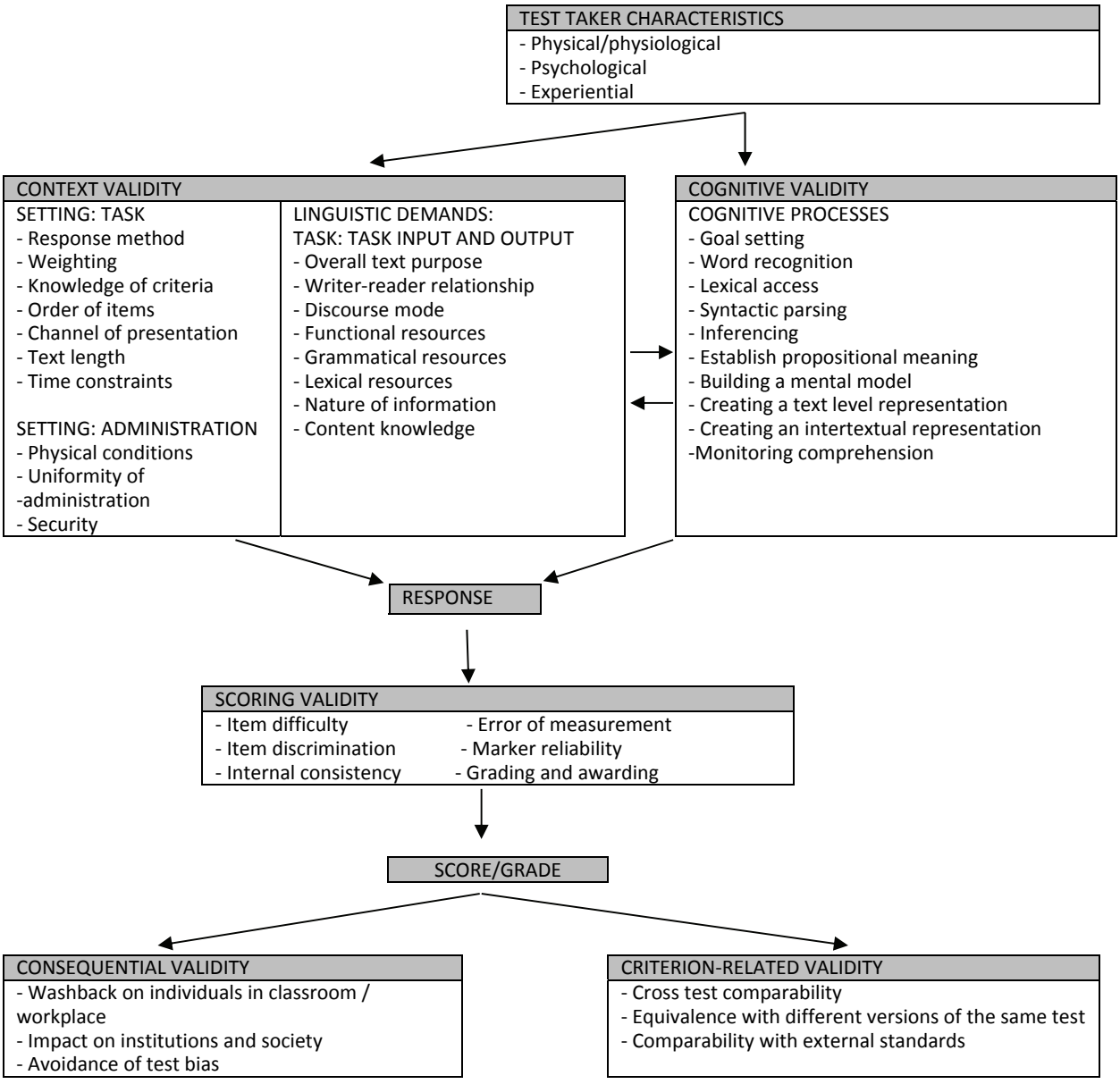
Weir, C. J., & Khalifa, H. (2008). A cognitive processing approach to defining reading comprehension. *Cambridge ESOL: Research Notes* 31: 2-10.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green and Co.

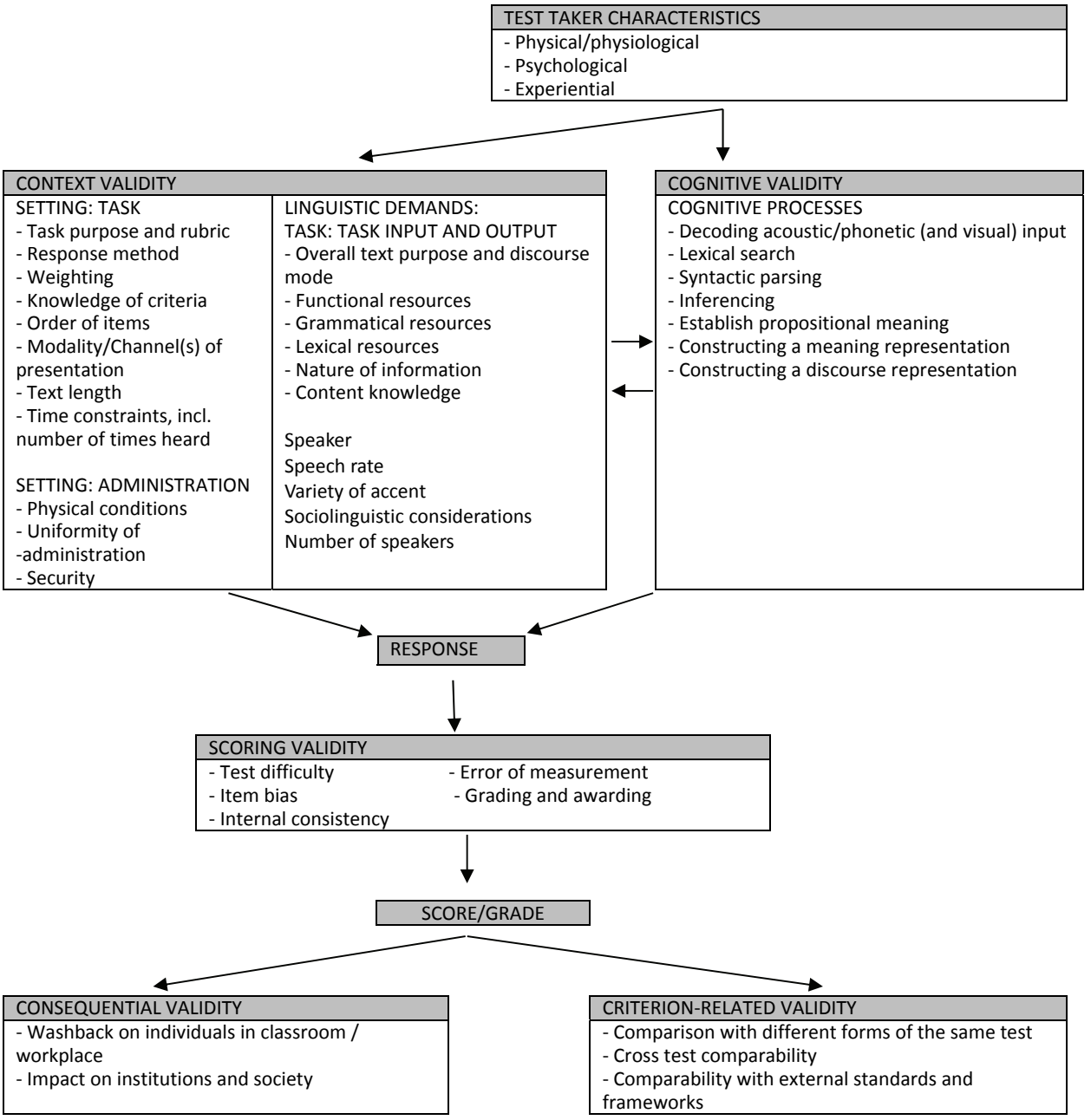
Wu, R., Y., F. (2012). Establishing the validity of the General English Proficiency Test Reading Component through a critical evaluation on alignment with the Common European Framework of Reference. Unpublished Ph.D. thesis: University of Bedfordshire.

[First 14 BNC family lists by P. Nation, VUW New Zealand; 16-20 by T. Cobb, UQAM Canada.]

Appendix 1: The socio-cognitive framework for conceptualising reading test validity (Khalifa and Weir 2009:5, adapted from Weir, 2005:44)



Appendix 2: The socio-cognitive framework for conceptualising listening test validity (Geranpayeh & Taylor, eds., 2013: 28, adapted from Weir, 2005:45)



Appendix 3: Set of task specification tables for the reading paper

PART	R1	TEAP READING TEST SPECIFICATIONS				
Time given for part		70 minutes for whole test (all 6 parts)				
Skill focus		Vocabulary and word usage				
Related TLU task		Language knowledge necessary to comprehend texts of an academic nature which students are likely to encounter in the context of their university studies.				
Test task type		Read a short text from which a word or phrasal verb has been deleted and choose the best word or phrase to fill the gap.				
Instructions to candidates		There are 20 very short reading texts below, and in each text there is a gap. Choose the best word or phrase from among the four choices to fill the gap. Mark your answer on your answer sheet.				
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	20 discrete sentence-based items				
Input reading text: contextual parameters	**Word count**	20-30 words (one or two sentences)				
	Text purpose	Referential	Conative	Emotive	Poetic	Phatic
	Domain	Public			Educational	
	Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
	Rhetorical organisation	Explicit		Both explicit and implicit		Implicit
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete		Fairly abstract	Mainly abstract
	Channel of presentation	Verbal	Non-verbal (i.e. graphs)		Both	
Sentence stem input: Level	**General CEFR level**	A2	B1		B2	
	AWL	Not specified				
	BNC Vocab Level	4-5				
	Words per sentence	Not specified				
Target (key): Level	**General CEFR level**	A2	B1		B2	
	AWL	All AWL words (550 word families) are acceptable as targets				
	BNC Vocab Level	1-5*				
Distractor options: Level	**General CEFR level**	A2	B1		B2	
	AWL	All AWL words (550 word families) are acceptable as distractors				
	BNC Vocab Level	1-5*				
		*Exceptions will be considered for words demonstrably relevant and common to the TLU domain				
Task level		A2 to B2				
Topic (content knowledge)		Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.				
Scoring parameters		Objectively scored dichotomous items, with each item equally weighted.				
Cognitive processing (of visual input)		Based upon model of reading processes in *Examining Reading* (p. 5 and p. 43):				
	Goal setting (i.e. types of reading), incl. processing of stem/options	**Expeditious reading: local** (scan/search for specifics)	**Careful reading: local** (understanding sentence)			
		Expeditious reading: global (skim for gist/search for key ideas/detail)	**Careful reading: global** (comprehend main idea(s)/overall text(s))			
	Word recognition					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

PART	R2A	TEAP READING TEST SPECIFICATIONS

Time given for part	70 minutes for whole test (all 6 parts)					
Skill focus	Reading graphs and charts					
Related TLU task	Interpreting and drawing inferences from visual information such as graphs and charts which students are likely to encounter in the classroom.					
Test task type	Look at information displayed in a graph or chart and choose the best response to answer a question about the graph or chart.					
Instructions to candidates	There are five graphs or charts below. Each graph or chart is followed by a question about it. For each question, choose the best answer from among the four choices and mark your answer on your answer sheet.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	5 discrete items				
Input reading text: contextual parameters	**Word count**	N/A - the reading input is in the form of a graph or chart with title/legend				
	Text purpose	Referential	Conative	Emotive	Poetic	Phatic
	Domain	Public			Educational	
	Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
	Rhetorical organisation	Explicit		Both explicit and implicit		Implicit
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete		Fairly abstract	Mainly abstract
	Channel of presentation	Verbal	Non-verbal (i.e. graphs)		Both	
Stem and options	**General CEFR level**	A2	B1			
	AWL	Not specified				
	BNC Vocab Level	3-4				
	Words per sentence	Not specified				
	Length (in words)	20-25 words for the Situation; 10-15 words for the Question				
Task level	A2 to B1					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of visual input)	Based upon model of reading processes in *Examining Reading* (p. 5 and p. 43):					
	Goal setting (i.e. types of reading), incl. processing of stem/options	**Expeditious reading: local** (scan/search for specifics)	**Careful reading: local** (understanding sentence)			
		Expeditious reading: global (skim for gist/search for key ideas/detail)	**Careful reading: global** (comprehend main idea(s)/overall text(s))			
	Word recognition					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

PART	R2B	TEAP READING TEST SPECIFICATIONS

Time given for part	70 minutes for whole test (all 6 parts)					
Skill focus	Reading advertisements and notices					
Related TLU task	Comprehending important information from notices, announcements, e-mails, etc. which students are likely to encounter on campus and which relate to the context of teaching and learning.					
Test task type	Look at the information displayed in a notice, announcement, or e-mail, etc. and choose the best response to answer a question about it.					
Instructions to candidates	There are five short reading texts (notices, advertisements, posters, etc.) below. Each text is followed by a question. For each question, choose the best answer from among the four choices and mark your answer on your answer sheet.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	5 discrete items				
Input reading text: contextual parameters	**Word count**	50-100 words for each text, including titles, headers, e-mail addresses, etc.				
	Text purpose	Referential	Conative	Emotive	Poetic	Phatic
	Domain	Public			Educational	
	Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
	Rhetorical organisation	Explicit		Both explicit and implicit		Implicit
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract
	Channel of presentation	Verbal	Non-verbal (i.e. graphs)	Both		
Input reading text: level	**General CEFR level**	A2	B1			
	AWL	3-8%				
	BNC Vocab Level	3-4				
	Words per sentence	Not specified				
Task level	A2 to B1					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of visual input)	Based upon model of reading processes in *Examining Reading* (p. 5 and p. 43):					
	Goal setting (i.e. types of reading), incl. processing of stem/options	**Expeditious reading: local** (scan/search for specifics)	**Careful reading: local** (understanding sentence)			
		Expeditious reading: global (skim for gist/search for key ideas/detail)	**Careful reading: global** (comprehend main idea(s)/overall text(s))			
	Word recognition					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

PART	R2C	TEAP READING TEST SPECIFICATIONS

Time given for part	70 minutes for whole test (all 6 parts)					
Skill focus	Reading short texts					
Related TLU task	Comprehending important information at the paragraph level in texts of an academic nature which students are likely to encounter in the classroom.					
Test task type	Read a short expository text and then choose the best response to answer a question about the text (one question per text).					
Instructions to candidates	There are 10 short reading passages below. Each passage is followed by a question. For each question, choose the best answer from among the four choices and mark your answer on your answer sheet.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	10 discrete items				
Input reading text: contextual parameters	**Word count**	Approximately 70 words for each passage				
	Text purpose	Referential	Conative	Emotive	Poetic	Phatic
	Domain	Public			Educational	
	Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
	Rhetorical organisation	Explicit		Both explicit and implicit		Implicit
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete		Fairly abstract	Mainly abstract
	Channel of presentation	Verbal	Non-verbal (i.e. graphs)			Both
Input reading text: level	**General CEFR level**	A2		B1		B2
	AWL	3-8%				
	BNC Vocab Level	3-4				4-5
	Words per sentence	Avg. 15				Avg. 18-20
	Flesch-Kincaid Grade Level	5-8				9-12
Task level	A2 to B2					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of visual input)	Based upon model of reading processes in *Examining Reading* (p. 5 and p. 43):					
	Goal setting (i.e. types of reading), incl. processing of stem/options	**Expeditious reading: local** (scan/search for specifics)			**Careful reading: local** (understanding sentence)	
		Expeditious reading: global (skim for gist/search for key ideas/detail)			**Careful reading: global** (comprehend main idea(s)/overall text(s))	
	Word recognition					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

PART	R3A	TEAP READING TEST SPECIFICATIONS

Time given for part	70 minutes for whole test (all 6 parts)					
Skill focus	Reading extended texts					
Related TLU task	Comprehending text-level information such as logical sequence in longer texts of an academic nature which students are likely to encounter in the context of their university studies					
Test task type	Read longer texts from which several words and phrases have been deleted. Choose the best response to fill each gap. Gaps target discourse-level understanding and require reading across sentences and paragraphs.					
Instructions to candidates	There are two reading passages below. In each passage, there are four gaps. Choose the best word or phrase from among the four choices to fill each gap. Mark your answer on your answer sheet.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	8 discrete items				
Input reading text: contextual parameters	**Word count**	Approximately 270 words				
	Text purpose	Referential	Conative	Emotive	Poetic	Phatic
	Domain	Public			Educational	
	Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
	Rhetorical organisation	Explicit		Both explicit and implicit		Implicit
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete		Fairly abstract	Mainly abstract
	Channel of presentation	Verbal	Non-verbal (i.e. graphs)		Both	
Input reading text: level	**General CEFR level**	B1			B2	
	AWL	3-8%				
	BNC Vocab Level	4-5				
	Words per sentence	Avg. 18-20				
	Flesch-Kincaid Grade Level	8-12				
Task level	B1 to B2					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of visual input)	Based upon model of reading processes in *Examining Reading* (p. 5 and p. 43):					
	Goal setting (i.e. types of reading), incl. processing of stem/options	**Expeditious reading: local** (scan/search for specifics)	**Careful reading: local** (understanding sentence)			
		Expeditious reading: global (skim for gist/search for key ideas/detail)	**Careful reading: global** (comprehend main idea(s)/overall text(s))			
	Word recognition					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)					

PART	R3B	TEAP READING TEST SPECIFICATIONS

Time given for part	70 minutes for whole test (all 6 parts)					
Skill focus	Reading extended texts (including graphs and charts)					
Related TLU task	Comprehending information and ideas in, and drawing inferences from, extended texts of an academic nature which students are likely to encounter in the context of their university studies, including the integration of information from both the text and visual information such as graphs and charts.					
Test task type	Read an extended argumentative or expository text and choose the best response to answer questions about it.					
Instructions to candidates	There are two long reading passages below. Each passage is followed by six questions. For each question, choose the best answer from among the four choices and mark your answer on your answer sheet.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	12 discrete items				
Input reading text: contextual parameters	**Word count**	Approximately 600 words; one text is accompanied by a graph/chart				
	Text purpose	Referential	Conative	Emotive	Poetic	Phatic
	Domain	Public			Educational	
	Discourse mode	Descriptive	Narrative	Expository	Argumentative	Instructive
	Rhetorical organisation	Explicit		Both explicit and implicit		Implicit
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract
	Channel of presentation	Verbal	Non-verbal (i.e. graphs)	Both		
Input reading text: level	**General CEFR level**	B1	B2			
	AWL	3-8%				
	BNC Vocab Level	4-5				
	Words per sentence	Avg. 18-20				
	Flesch-Kincaid Grade Level	8-12				
Task level	B1 to B2					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of visual input)	Based upon model of reading processes in *Examining Reading* (p. 5 and p. 43):					
	Goal setting (i.e. types of reading), incl. processing of stem/options	**Expeditious reading: local** (scan/search for specifics)	**Careful reading: local** (understanding sentence)			
		Expeditious reading: global (skim for gist/search for key ideas/detail)	**Careful reading: global** (comprehend main idea(s)/overall text(s))			
	Word recognition					
	Lexical access					
	Syntactic parsing					
	Establishing propositional meaning (cl./sent. level)					
	Inferencing					
	Building a mental model					
	Creating a text level representation (disc. structure)					
	Creating an intertextual representation (multi-text)	i.e. between visual and text				

Appendix 4: Set of task specification tables for the Listening paper

PART	L1A	TEAP LISTENING TEST SPECIFICATIONS					
Time given for part		50 minutes for whole test (all 5 parts)	*No compulsory time limits are set for individual listening parts; timing in the listening test is determined by the CD. All tasks are single-play.*				
Task description		[Introductory rubric provided (audio only) with instructions on what will happen (10 short conversations) and how to respond but no scene-setting context for the listening material. A question in English is read at the end of each of the 10 short dialogues. After hearing this, examinees have 10 seconds to read the 4 options and choose an answer.]					
Skill focus		Listening to short dialogues					
Related TLU task		Comprehending dialogues between students and persons with whom students are likely to converse in the context of their university studies (e.g., professors, academic advisors, exchange students).					
Test task type		Listen to a short dialogue and choose the best response to answer a question about it. Dialogue and question are heard once.					
Instructions to candidates		In this part, you will hear 10 short conversations. Each conversation will be followed by one question. For each question, you will have 10 seconds to choose the best answer and mark your answer on your answer sheet. The conversations and questions will be played only once. Now, let's begin.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)					
	Items per part	10 discrete listening items/tasks					
Input listening material for each task/item: contextual parameters	**Length**	6-8 spoken turns (approximately 100 words)					
	Discourse purpose	Discursive (discussing for/against)	Expository (factual and causal, e.g. informative/ explanatory)	Argumentative/ Persuasive (promoting a point of view)	Process-Descriptive (describing a staged process)	Analytical (interpretation and criticism, e.g. of film)	
	Domain (topic)	Public		Educational			
	Discourse type	Short monologue	Longer monologue	Short dialogue (2 speakers)	Longer dialogue (2 or 3 speakers)		
	Content/subject knowledge	General				Specific	
	Cultural specificity	Neutral				Specific	
	Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract	
	Channel of presentation	Aural	Visual (verbal)	Visual (non-verbal)			
Input listening material : level	**General CEFR level**	A2	B1				
	AWL	2-6%					
	BNC Vocab Level	3-4					
	Words per sentence	Avg. 10					
	Speech rate (words per minute)	Avg. 150					
Task level		A2 to B1					
Topic (content knowledge)		Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters		Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of audio input from CD + visual input in question booklet)		Based upon model of listening processes in *Examining Listening* (p. 28 and pp. 97-103):					
	Listener goals/objectives (i.e. types of listening focus), inc. processing of stem/options	Listening for gist/overall understanding					
		Listening for main idea/important information/key message					
		Listening for detailed/specific information					
		Listening to infer opinion/attitude/intention					
	Decoding acoustic/phonetic (and visual) input						
	Lexical search						
	Syntactic parsing						
	Establishing propositional meaning						
	Constructing a meaning representation						
	Constructing a discourse representation						

PART	L1B	TEAP LISTENING TEST SPECIFICATIONS					
Time given for part		50 minutes for whole test (all 5 parts)	*No compulsory time limits are set for individual listening parts; timing in the listening test is determined by the CD. All tasks are single-play.*				
Task description		[Introductory rubric provided (audio only) with instructions on what will happen (10 short passages) and how to respond but no scene-setting context for the listening material. A question in English is read at the end of each of the 10 short monologues. After hearing this, examinees have 10 seconds to read the 4 options and choose an answer.]					
Skill focus		Listening to short monologues					
Related TLU task		Comprehending important information from brief lectures and announcements relevant to academic subjects or the university context and interpreting visual information such as graphs and charts which students are likely to encounter in the context of their university studies.					
Test task type		Listen to a short monologue and choose the best response to answer a question about it. Monologue and question are heard once.					
Instructions to candidates		In this part, you will hear 10 short passages. Each passage will be followed by one question. For each question, you will have 10 seconds to choose the best answer and mark your answer on your answer sheet. The passages and questions will be played only once. Now, let's begin.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)					
	Items per part	10 discrete listening items/tasks					
Input listening material for each task/item: contextual parameters	**Length**	Approximately 70 words					
	Discourse purpose	Discursive (discussing for/against)	Expository (factual and causal, e.g. informative/ explanatory)	Argumentative/ Persuasive (promoting a point of view)	Process-Descriptive (describing a staged process)	Analytical (interpretation and criticism, e.g. of film)	
	Domain (topic)	Public		Educational			
	Discourse type	Short monologue	Longer monologue	Short dialogue (2 speakers)	Longer dialogue (2 or 3 speakers)		
	Content/subject knowledge	General				Specific	
	Cultural specificity	Neutral				Specific	
	Nature of information	Only concrete	Mostly concrete	Fairly abstract	Mainly abstract		
	Channel of presentation	Aural	Visual (verbal)	Visual (non-verbal)			
Input listening material: level	**General CEFR level**	A2	B1				
	AWL	2-6%					
	BNC Vocab Level	3-4					
	Words per sentence	Avg. 15-16					
	Speech rate (words per minute)	Avg. 150					
Task level		A2 to B1					
Topic (content knowledge)		Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters		Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of audio input from CD + visual input in question booklet)		Based upon model of listening processes in *Examining Listening* (p. 28 and pp. 97-103):					
	Listener goals/objectives (i.e. types of listening focus), inc. processing of stem/options	Listening for gist/overall understanding					
		Listening for main idea/important information/key message					
		Listening for detailed/specific information					
		Listening to infer opinion/attitude/intention					
	Decoding acoustic/phonetic (and visual) input						
	Lexical search						
	Syntactic parsing						
	Establishing propositional meaning						
	Constructing a meaning representation						
	Constructing a discourse representation						

PART	L1C	TEAP LISTENING TEST SPECIFICATIONS					
Time given for part		50 minutes for whole test (all 5 parts)	*No compulsory time limits are set for individual listening parts; timing in the listening test is determined by the CD. All tasks are single-play.*				
Task description		[Introductory rubric provided (audio only) with instructions on what will happen (5 short passages) and how to respond - but no scene-setting context for the listening material. A question in English is read at the end of each of the 5 short monologues. After hearing this, examinees have 10 seconds to read/evaluate 4 visuals and choose an answer.]					
Skill focus		Listening to short monologues					
Related TLU task		Comprehending important information from brief lectures and announcements relevant to academic subjects or the university context and interpreting visual information such as graphs and charts which students are likely to encounter in the context of their university studies.					
Test task type		Listen to a short monologue and choose the best response to answer a question about it. Questions ask test takers to choose from 4 graphs or charts. Monologue and question are heard once.					
Instructions to candidates		In this part, you will hear 5 short passages. Each passage will be followed by one question. For each question, you will see four graphs or charts in your test booklet. You will have 10 seconds to choose the best graph or chart to answer the question. Mark your answer on your answer sheet. The passages and questions will be played only once. Now, let's begin.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)					
	Items per part	5 discrete listening items/tasks					
Input listening material for each task/item: contextual parameters	**Length**	Approximately 70 words					
	Discourse purpose	Discursive (discussing for/against)	Expository (factual and causal, e.g. informative/ explanatory)	Argumentative/ Persuasive (promoting a point of view)	Process-Descriptive (describing a staged process)	Analytical (interpretation and criticism, e.g. of film)	
	Domain (topic)	Public		Educational			
	Discourse type	Short monologue	Longer monologue	Short dialogue (2 speakers)	Longer dialogue (2 or 3 speakers)		
	Content/subject knowledge	General				Specific	
	Cultural specificity	Neutral				Specific	
	Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract	
	Channel of presentation	Aural	Visual (verbal)	Visual (non-verbal)			
Input listening material: level	**General CEFR level**	A2	B1				
	AWL	2-6%					
	BNC Vocab Level	3-4					
	Words per sentence	Avg. 15-16					
	Speech rate (words per minute)	Avg. 150					
Task level		A2 to B1					
Topic (content knowledge)		Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters		Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of audio input from CD + visual input in question booklet)		Based upon model of listening processes in *Examining Listening* (p. 28 and pp. 97-103):					
	Listener goals/objectives (i.e. types of listening focus), inc. processing of stem/options	Listening for gist/overall understanding					
		Listening for main idea/important information/key message					
		Listening for detailed/specific information					
		Listening to infer opinion/attitude/intention					
	Decoding acoustic/phonetic (and visual) input						
	Lexical search						
	Syntactic parsing						
	Establishing propositional meaning						
	Constructing a meaning representation						
	Constructing a discourse representation						

PART	L2A	TEAP LISTENING TEST SPECIFICATIONS

Time given for part	50 minutes for whole test (all 5 parts)	*No compulsory time limits are set for individual listening parts; timing in the listening test is determined by the CD. All tasks are single-play.*				
Task description	[Short printed description for each situation (1-2 sentences) audio and in booklet; 3 questions and 4 options for each long conversation also printed. Questions also read aloud after dialogue with 10 seconds allowed to answer each one.]					
Skill focus	Listening to long dialogues					
Related TLU task	Comprehending important information in long dialogues between students and persons with whom students are likely to converse in the context of their university studies (e.g., professors, academic advisors, exchange students). Includes both two- and three-person dialogues.					
Test task type	Listen to a long dialogue and choose the best response to answer questions about it. Dialogue and question are heard once.					
Instructions to candidates	In this part, you will hear three long conversations, A, B, and C. Before each conversation, you will hear a short description of the situation. The situation is also printed in your test booklet. Each conversation will be followed by three questions. The questions are also printed in your test booklet. For each question, you will have 10 seconds to choose the best answer and mark your answer on your answer sheet. The conversations and questions will be played only once. Now, let's begin.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	3 discrete listening tasks – each one with 3 test items (i.e. 9 items in all)				
Input listening material for each task/item: contextual parameters	**Length**	14-16 spoken turns (approximately 300 words); scene-setting sentence = 30 words max				
	Discourse purpose	Discursive (discussing for/against)	Expository (factual and causal, e.g. informative/ explanatory)	Argumentative/ Persuasive (promoting a point of view)	Process-Descriptive (describing a staged process)	Analytical (interpretation and criticism, e.g. of film)
	Domain (topic)	Public		Educational		
	Discourse type	Short monologue	Longer monologue	Short dialogue (2 speakers)	Longer dialogue (2 or 3 speakers)	
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract
	Channel of presentation	Aural	Visual (verbal)	Visual (non-verbal)		
Input listening material: level	**General CEFR level**	B1	B2			
	AWL	2-6%				
	BNC Vocab Level	4-5				
	Words per sentence	Avg. 10				
	Speech rate (words per minute)	Avg. 150				
Task level	B1 to B2					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of audio input from CD + visual input in question booklet)	Based upon model of listening processes in *Examining Listening* (p. 28 and pp. 97-103):					
	Listener goals/objectives (i.e. types of listening focus), inc. processing of stem/options	Listening for gist/overall understanding				
		Listening for main idea/important information/key message				
		Listening for detailed/specific information				
		Listening to infer opinion/attitude/intention				
	Decoding acoustic/phonetic (and visual) input					
	Lexical search					
	Syntactic parsing					
	Establishing propositional meaning					
	Constructing a meaning representation					
	Constructing a discourse representation					

PART	L2B	TEAP LISTENING TEST SPECIFICATIONS

Time given for part	50 minutes for whole test (all 5 parts)	*No compulsory time limits are set for individual listening parts; timing in the listening test is determined by the CD. All tasks are single-play.*				
Task description	[Short printed description for each passage (1-2 sentences, 30 words max.) audio and in booklet; questions and options also printed. Questions also read aloud at end of passage with 10 seconds allowed to answer each one.]					
Skill focus	Listening to long monologues (including graphs and charts)					
Related TLU task	Comprehending monologues of an academic nature which students are likely to encounter in the context of their university studies, including the integration of information from both the listening text and visual information such as graphs and charts.					
Test task type	Listen to long monologues and answer 4 questions about each monologue. Monologue and question are heard once.					
Instructions to candidates	In this part, you will hear four long passages, D, E, F, and G. Before each passage, you will hear a short description of the situation. The situation is also printed in your test booklet. Each passage will be followed by four questions. The questions are also printed in your test booklet. For each question, you will have 10 seconds to choose the best answer and mark your answer on your answer sheet. The passages and questions will be played only once. Now, let's begin.					
Characteristics of expected response	**Response format**	Selected response : 4-option multiple choice (marked on answer sheet)				
	Items per part	4 discrete listening passages each with 4 discrete test items (i.e. 16 items in all)				
Input listening material for each task/item: contextual parameters	**Length**	Approximately 220 words				
	Discourse purpose	Discursive (discussing for/against)	Expository (factual and causal, e.g. informative/ explanatory)	Argumentative/ Persuasive (promoting a point of view)	Process-Descriptive (describing a staged process)	Analytical (interpretation and criticism, e.g. of film)
	Domain (topic)	Public		Educational		
	Discourse type	Short monologue	Longer monologue	Short dialogue (2 speakers)	Longer dialogue (2 or 3 speakers)	
	Content/subject knowledge	General				Specific
	Cultural specificity	Neutral				Specific
	Nature of information	Only concrete	Mostly concrete	Fairly abstract		Mainly abstract
	Channel of presentation	Aural	Visual (verbal)	Visual (non-verbal)		
Input listening material: level	**General CEFR level**	B1	B2			
	AWL	2-6%				
	BNC Vocab Level	4-5				
	Words per sentence	Avg. 16-18				
	Speech rate (words per minute)	Avg. 150				
Task level	B1 to B2					
Topic (content knowledge)	Topics will be selected from a broad range of content areas relevant to first-year undergraduate study in the EFL context of Japan. Relevant and appropriate topics will be those which (1) are likely to be encountered in the course of engaging in TLU tasks; and (2) are at an appropriate level of abstraction and do not require specific content or background knowledge. Tools for identifying and evaluating appropriate topics have been developed and incorporated into item writer manuals.					
Scoring parameters	Objectively scored dichotomous items, with each item equally weighted.					
Cognitive processing (of audio input from CD + visual input in question booklet)	Based upon model of listening processes in *Examining Listening* (p. 28 and pp. 97-103):					
	Listener goals/objectives (i.e. types of listening focus), inc. processing of stem/options	Listening for gist/overall understanding				
		Listening for main idea/important information/key message				
		Listening for detailed/specific information				
		Listening to infer opinion/attitude/intention				
	Decoding acoustic/phonetic (and visual) input					
	Lexical search					
	Syntactic parsing					
	Establishing propositional meaning					
	Constructing a meaning representation					
	Constructing a discourse representation					

Appendix 5: Analysis of academic lectures from MICASE corpus

Item ID	AWL%
Principles in Sociology	2.83
Renaissance to Modern Art History	2.9
Biology Of Cancer	4.5
Fantasy In Literature	2.08
Intro To Evolution	4.47
Intro To Physics	2.29
Macroeconomic	7.17

	AWL
Avg.	3.75
Max.	7.17
Min.	2.08

Appendix 6: Analysis of vocabulary in high school text books

Textbook Series*	BNC95%
1	4
2	3
3	4
4	3
5	4
6	4
7	3
8	3
9	3
10	3
11	3
12	4
13	6
14	3
15	4
16	3
17	3
18	3
19	3
20	4
21	3
22	3
23	4
24	4
25	4
26	4
27	3
28	2
29	3
30	5
31	4
32	4
33	7
34	3
35	3
36	6
37	3

	BNC95%
Avg.	3.648649
Max.	2
Min.	7

*Titles are anonymized

Appendix 7: Flesch-Kincaid Grade Level Readability statistics

Flesch-Kincaid Grade Level Readability statistics				
	Min	**Max**	**Mean**	**Source**
Private universities 1994	6.06	12.26	9.83	Brown & Yamashita (1995)
Private universities 2004	9.08	12.14	9.62	Kikuchi (2006)
Public universities 1994	6.76	13.61	9.11	Brown & Yamashita (1995)
Public universities 2004	8.23	15.32	10.98	Kikuchi (2006)
Center Test 1994			9.29	Brown & Yamashita (1995)
Center Test 2004			8.79	Kikuchi (2006)
Third-year high school texts*			8.7	Chujo & Hasegawa (2004)
Cambridge KET	2	7.4	5.5	Khalifa & Weir (2009)
Cambridge PET	5	10.1	7.9	Khalifa & Weir (2009)
Cambridge FCE	5	12.3	8.4	Khalifa & Weir (2009)
IELTS			12.64	Green et al (2009)
University first-year texts			13.66	Green et al (2009)
*Chujo & Hasegawa used the mean of 3 measures that report readability in U.S. grade levels				

Appendix 8: Suggestions for future research

A number of specific recommendations are offered here for possible research and validation studies in the future with reference to the TEAP Reading and Listening papers:

- **Investigate timing issues for individual reading tasks in relation to the cognitive processing elicited** from test takers (i.e. careful vs. expeditious reading): this could be explored through a mixed methods approach involving observation, interview, verbal protocol, and even eye-tracking methodology.

- **Investigate cognitive processing involved in Reading Part 2A:** this could be done by asking a pair of test takers at the same ability level to do the task together and talk about it as they do so (a form of concurrent introspection) followed by a retrospective interview protocol after the event (all of which can be recorded for later analysis).

- **Investigate cognitive processing involved in Reading Part 2B by comparing alternative formats (see 4.3)**. One issue worthy of investigation would be to compare (i) the current approach in which test takers read the text before reading the questions (thus potentially encouraging them to read the text carefully), with (ii) an alternative format in which the questions are placed before the text, in order to signal more clearly the information/reading style which is being elicited.

- **Investigate other vocabulary measures** for the input texts using available automatic software, e.g. lexical density through *Vocabprofile* or *TextInspector.*

- **Monitor possible changes needed to the wordlists**, undertaking regular documentary checks and routine analyses, and updating the wordlists as appropriate in light of linguistic change.

- **Monitor and investigate the AWL coverage** in the TEAP test and any issues arising with the test taker population.

- **Explore the value of the latest version of Coh-Metrix (v3.0)** for analysing lexical, syntactic and discourse features of the TEAP reading texts.

- **Monitor speech rates in the listening tasks** and keep these under review.

- **Extend the analysis of listening input** using measures such as *mean length of utterance*, *propositional density*, etc. as automatic tools or specialist raters become more readily available.