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

University of Bedfordshire

A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese University Entrants

Professor C. J. Weir
Centre for Research in English Language Learning and Assessment (CRELLA),
University of Bedfordshire, UK

Table of Contents

Executive Summary

- Rigorous and iterative test design, accompanied by systematic trialing procedures, produced a pilot version of the test which demonstrated acceptable context and cognitive validity for use as an English for academic purposes (EAP) writing test for students wishing to enter Japanese universities.

- 体系的なトライアル実験をしながらテストデザインが綿密に繰り返し修正され、日本の大学受験の英語ライティングテストとして相応しい「テストの内容、背景に関する妥当性」と「認知的妥当性」（context and cognitive validity）のあるパイロット版が作り出された。

- A study carried out on the scoring validity of the rating of the TEAP Writing Test indicated acceptable levels of inter- and intra-marker reliability and demonstrated that receiving institutions could depend on the consistency of the results obtained on the test.

- TEAP ライティングテストの「スコアに関する妥当性」（scoring validity）に関する実験が行われ、評定者間信頼性、評定者内信頼性（inter- and intra-marker reliability）が大学受験として使われるに十分であることが検証され、大学側がこのテストの結果を信頼し得る指標として使用できることが立証された。

- A study carried out on the contextual complexity parameters (lexical, grammatical, and cohesive) of scripts allocated to different bands on the TEAP Writing Test rating scale indicated that there were significant differences between the scripts in adjacent band levels, with the band B1 scripts produced by students being more complex than the band A2 scripts across a broad set of indices.

- 受験者が TEAP ライティングテストで実際に使った「語彙」、「文法」、「文章と文章のつながり」に関する指標が、評価官によって評価されたレベルと整合性があるかどうかが研究された。受験者の実際のライティングサンプルは、検証された全ての指標において、実際に B1 レベルと A2 レベルの間に統計的に有意な差があり、B1 レベルとして評価されたサンプルは A2 レベルとして評価されたサンプルよりも、より複雑で高度なものであった。

Introduction

This report covers the development of an academic English writing test, a core component of the innovative Test of English for Academic Purposes (TEAP) for Japan.

A long-term aim of the TEAP is to have a positive impact on English education in Japan by revising and improving the widely varying approaches to English tests 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 English as a foreign language (EFL) context of Japan.

The TEAP is a collaborative test development project being undertaken by the Eiken Foundation of Japan (Eiken), which administers the EIKEN Test in Practical English Proficiency (EIKEN) to over 2 million test takers a year, and Sophia University, one of the leading private universities in Japan. Professor C. J. Weir and his team at the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire in the UK were contracted to provide specialist assistance to the project. The role of Professor Weir, as a leading expert in language testing and the assessment of academic writing, was to develop test specifications with the Eiken and Sophia University project teams based on the latest research into the assessment of writing for EFL/ESL (English as a foreign/second language) learners and to provide advice on the design of the TEAP Writing Test tasks.

The TEAP was 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 the TEAP are those arising in academic activities conducted in English on (Japanese) university campuses. The TLU domain is defined by Bachman and Palmer (1996, p. 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 Writing Test would thus cover academic contexts relevant to studying at a university in the EFL context of Japan. It is related directly to studying and learning, and not to general, everyday activities or interactions that fall in the personal/private domain.

The TEAP is intended to 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 program 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 exams. The Ministry of Education, Culture, Sports, Science, and Technology (MEXT) in Japan has publicly recommended two levels of English proficiency as goals for high school graduates (MEXT, 2002; MEXT, 2003). MEXT has provided these goals in terms of commonly recognized English proficiency benchmarks, including the EIKEN grade levels. For high school students, MEXT has suggested the EIKEN Grade 2 and Pre-2 levels as appropriate benchmarks (MEXT, 2002; MEXT, 2003). Based on research into the comparability of the EIKEN grades with the Common European Framework of Reference for Languages (CEFR)—an increasingly widely used international proficiency framework developed by the Council of Europe (2001)—these levels of proficiency can be considered relevant to the B1 and A2 levels of the CEFR, respectively (visit http://stepeiken.org/research and see Dunlea & Figueras, 2012).

It was decided through consultation with the main stakeholders that, for the TEAP Writing Test, the main area of interest would be whether students attain a B1 level of writing ability in English, the higher of the two benchmark levels recommended by MEXT. For the purposes of positive washback, it was decided that the TEAP Writing Test should also 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 program from the outset placed the typical test takers at the center 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 ahead to the more demanding TLU domain of the academic learning and teaching context of Japanese universities, it was decided that the test should contain tasks capable of discriminating between students at a B1 level and the more advanced B2 level appropriate to the TEAP TLU, and be able to provide useful feedback for students at this more advanced level of ability.

1.1 Background Review

An initial background survey was conducted by one of the Eiken project team members, Kazuaki Yanase (in Japanese). The survey examined the new Ministry of Education curriculum guidelines for high school regarding writing instruction, as well as examining writing tasks in approved text books and university entrance exams. It provided valuable information for understanding trends in the Japanese education sector relevant to the TEAP. It mapped out the broad Japanese contextual parameters relating to the potential test takers within which the development team needed to operate. The test taker was at the heart of the test design process *ab initio*.

As part of the further preparatory work undertaken prior to deciding on formats for the writing task, Weir was asked to work with Eiken colleagues to review the assessment literature on academic writing and thereby establish what might be the most appropriate writing task(s) and associated criteria for use in the TEAP test, and to present these findings for discussion at a focus group meeting for key project stakeholders in Japan.

Validity Considerations

The starting point was a felt need to develop a comprehensive understanding of the different types of writing tasks students will encounter in universities to ensure the *cognitive* and *context* validity of any writing tasks used in the TEAP Writing Test. Context validity for a writing task addresses the particular performance conditions, the setting under which it is to be performed (such as the purpose of the task, time available, length, specified addressee, and known marking criteria as well as the linguistic demands inherent in the successful performance of the task) (Weir, 2005, p. 19). Cognitive validity is concerned with the extent to which the writing tasks employed succeed in eliciting from candidates a set of processes which resemble and are representative of those employed by a proficient writer in a real-world academic writing event (Shaw & Weir, 2007, Chapter 3).

A number of large-scale studies that had been conducted on the types of writing that are required of students at the tertiary level were reviewed. Most of these research studies were conducted with the purpose of test development; for instance, the Test of English as a Foreign Language (TOEFL) (Bridgeman & Carlson, 1983; Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996) and the Test of English for Educational Purposes (TEEP) (Weir, 1983).

In an important study as part of the International English Language Testing System (IELTS) research program, Moore and Morton (2005) compared the parameters of real-life academic writing tasks and EAP test tasks. They compared the IELTS Academic Writing Task 2, an impromptu, argumentative essay task, with a corpus of 155 assignment tasks at both undergraduate and post-graduate levels across 79 academic departments in two Australian universities. The results show that while essay (58%) appears the most common academic task type, the IELTS Task 2 is unlike the “university essay” in two critical ways. Firstly, regarding the use of source texts, real-life assignments require students to transform content from *multiple* new information sources, while the IELTS Writing Task 2 only requires test takers to use their prior knowledge in response to a single line prompt, which in all likelihood makes the IELTS task far more of a knowledge-telling than a knowledge-transformation task.

Bereiter and Scardamalia (1987a, 1987b) provided two models, knowledge telling and knowledge transforming, that demonstrate the different composing processes of good and bad first-language writers which were able to account for the various research findings. Hyland (2002, p. 28) provides a clear description of both:

- A knowledge-telling model addresses the fact that novice writers plan less often than experts, revise less often and less extensively, and are primarily concerned with generating content from their internal resources. Their main goal is simply to tell what they can remember based on the assignment, the topic, or the genre.
- A knowledge-transforming model suggests how skilled writers use the writing task to analyze problems and set goals. These writers are able to reflect on the complexities of the task and resolve problems of content, form, audience, style, organization, and so on within a content space and a rhetorical space, so that there is a continuous interaction between developing knowledge and text. Knowledge transforming thus involves actively reworking thoughts so that in the process not only the text, but also ideas, may be changed (Bereiter & Scardamalia, 1987).

They also found that the two most common language functions required in academic writing tasks—description and summarization—are not usually required in the IELTS test task.

The important conclusion from the research literature was that a knowledge-transforming, integrated reading-into-writing task type can address construct validity concerns better than the more common, independent writing-only, knowledge-telling task type (Shaw & Weir, 2007; Plakans, 2008; Weigle, 2002).

After surveying the research literature on potential task types, Weir proposed that the test development team should seriously consider variants of the *summary task,* which had fallen out of favor in the mid-1970s in the UK but was now being reintroduced by examining boards globally because of the felt need from the late 20th century onwards to mirror the real-life demands that are made on students in an academic context in order to enhance test validity. Weir argued strongly that reading-into-writing summary task types have both context and cognitive validity (Weigle, 2002; Shaw & Weir, 2007, Chapters 3 and 4; Plakans, 2009, 2010); i.e., they represent closely the cognitive processing and knowledge-base requirements of real-life writing activities beyond the test.

Cognitive Validity

Cognitive psychologists working at the discourse level from the 1970s onwards have argued that comprehension naturally involves a form of summarization (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). In summary writing tasks, the reader has to establish the main ideas in a text(s), extract them and reduce them to note form, and then rewrite the notes in a coherent manner in their own words. The summarization of main ideas at the text level is one of the more demanding levels of processing activity which represents real-life language use for university students. Yu (2008, pp. 522-523) offers an impressive list of references in support of the use of summarization in teaching and testing.

Weir argued that the summarization of a single text effectively tests the important advanced-level reading skill of creating a text-level representation, a vital element of academic study, in an authentic manner (Khalifa & Weir, 2009, Chapter 3). No other task type can mirror this high-level cognitive processing skill as effectively, and it is pleasing to note that the recommendation that summarization should be reintroduced into Cambridge ESOL examinations made in Khalifa and Weir (2009, Chapter 8) is to be implemented in future versions of the writing papers in some of these examinations. The TOEFL® iBT (Internet-based test) and Pearson Test of English Academic (PTE Academic) have recently included integrated reading-into-writing tasks in their current international EAP examinations. Accordingly, summarization of a single text was advocated as a suitable format for what we shall call Task A in this project, which was intended to be a task both accessible to and achievable by students at a B1 level of proficiency.

Weir also proposed considering an additional writing task that he saw as essential for assessing high-level writing skills. Processing not just one text at the discourse level but multiple texts (verbal and nonverbal) is seen as the critical requirement of academic study. Based on a direct analysis of writing tasks in 38 faculties, Horowitz (1986a, 1986b) identified the seven most common task types in tertiary-level study, and, among these types, synthesis of multiple sources was the most popular across faculties. Horowitz argued that there is a fundamental discrepancy between real-life tasks and most writing-test tasks in terms of the relationship of the text produced to other texts. In other words, in most test tasks, candidates are not required to synthesize (reorder, combine, remove) ideas from various sources as students do in real life. This finding indicated the importance of writing from multiple sources in academic writing, and a task replicating this was advocated for what we shall call Task B in the TEAP project, a task intended to provide feedback on students at the more advanced B2 level.

There is a growing interest, especially in the field of academic literacy, in what goes on beyond text comprehension when readers read multiple texts. The ability to address intertextuality and the synthesis of information across multiple texts is described as taking place at the B2 level of the CEFR (Council of Europe, 2001). In educational settings, reading and obtaining information from multiple sources is a desirable skill, and a rich understanding of an event or concept requires awareness of different perspectives (see, for example, Stahl, Hynd, Britton, McNish, & Bosquet, 1996). This awareness is achieved by constructing links between the information presented in different texts.

Lacroix (1999) suggests that in the comprehension of multiple texts in a particular domain, a coherent, condensed structuring of the multiple text information may require two distinct levels of macrostructural processing. She suggests that the process of macrostructure construction (extracting important information) outlined in Kintsch and van Dijk (1978)—which involves identification and construction of a hierarchy for units of information through the application of transformational macro rules of deletion, generalization, and integration—accounts well for the comprehension of a single text, but may not be adequate to account for how mental representations can be combined coherently across multiple texts. Lacroix (1999) suggests that an additional process of macrostructural organization (structuring selected information) is necessary for the connection of several text representations through higher-level semantic links.

Stromso and Braten (2002, p. 211) similarly argue that the "discourse synthesis" of multiple texts in a specific domain involves "composing a new text by selecting, organizing, and connecting content from more than one source text." The need for an intertextual model, sometimes referred to as a "documents model" (Perfetti, Rouet, & Britt, 1999), to account for the production of integrated representations of multiple texts is supported in the work of Britt and Sommers (2004), Goldman (1997, 2004), Hartmann (1995), Perfetti (1997), Perfetti, Rouet, & Britt (1999), Spivey and King (1989), and Stahl, Hynd, Britton, McNish, & Bosquet (1996). As Goldman (2004, p. 344) succinctly puts it, "the information across texts is part of a larger whole not necessarily specified in any one of the texts."

Britt & Sommers (2004, p. 318) point to the higher cognitive demands this imposes on the reader: since texts are not normally written to be read in conjunction with other texts, they lack explicit links to facilitate the integration of information across texts, and the demands on the reader to form a macrostructure are higher than when reading a single text with its own intratextual cohesion.

Processing in multiple text reading has clear resonances with findings from research into the cognitive processing that takes place in writing tasks involving knowledge transformation (Scardamalia & Bereiter, 1987) where the selection, connection, and organization of information from source texts constitute the first cognitive components in the writing process (see Shaw & Weir, 2007). Researchers such as Spivey and King (1989) have shown how competent students interweave texts when writing research papers by utilizing source material deemed to have intertextual importance.

To inform the design of test tasks, the writing processes used in the evaluation of the socio-cognitive validity of Cambridge ESOL examinations by Shaw and Weir (2007) were proposed for consideration by the project team. Shaw and Weir's validation framework was designed for L2 writing assessment and it built on earlier models of writing as well as a more recent model by Field (2004) based on information processing principles in the psycholinguistic area. The framework has demonstrated its practicality in the validation of the Cambridge ESOL tests and the IELTS and Trinity College examinations in the UK as well as the General English Proficiency Test (GEPT) in Taiwan. It was felt to be appropriate for the TEAP project.

Weir suggested that the real-life *cognitive* processes that might be considered in the design of the TEAP writing tasks should include:

- Task representation
- Macro-planning
- Reading source texts
- Selecting
- Connecting
- Organizing
- Micro-planning
- Translating
- Monitoring and revising

In addition, the importance of establishing appropriate contextual parameters was to be discussed with the project team.

Context Validity

Following Khalifa and Weir (2009) and Shaw and Weir (2007), a set of contextual parameters to be addressed by the test developers was also put forward for consideration, as follows.

Setting: Task

- Response format
- Purpose
- Knowledge of marking criteria
- Text length
- Time constraints
- Writer-reader relationships
- Topic

Linguistic Demands: Task Input and Output

- Lexical resources
- Structural resources
- Discourse mode
- Functional resources
- Content knowledge

It was decided that all components of the suggested validation framework would be critically reviewed at a three day focus group meeting for key project staff convened at Sophia University in Japan in July 2010.

1.2 Designing the TEAP: Focus Group Meeting for Key Project Staff Convened at Sophia University in July 2010 and Follow-up

The goal of the meeting was to finalize a draft of the writing test specifications (e.g., task rubrics, level and nature of input texts, intended level of tasks in terms of cognitive and contextual parameters, expected quality/quantity of output, outline of scoring model, and timing) as well as finalizing the task formats to be developed.

Discussion addressed the following points *inter alia*:

- What are the cognitive, contextual, and scoring parameters for the initial test task draft specification?
- How to achieve scoring validity? What are the issues involved in scoring? What needs to be in an overall scoring framework? What rating criteria should be used?
- What kind of EAP writing task would be appropriate (expository, argumentative, etc.)? Which of the alternative task formats provided by Weir are the most suitable for the target audience in Japan?
- How much output is needed for a suitable essay of this type? How much time is necessary?
- How does the amount of text required compare to what we can expect test takers to actually do?
- What percentage of test takers do we actually expect to be able to attempt and then to successfully complete each of the tasks?
- How to incorporate the synthesizing of different sources of information into an essay?
- How to write the rubric (i.e., instructions) in English so the demands of the task are clear (e.g., do we tell them how many paragraphs they need to write, etc.)?

The socio-cognitive model for the validation of writing tests (Shaw & Weir, 2007) which is used by CRELLA in its global test development activities provided a basis for guiding the discussion on the required cognitive and contextual parameters for the test.

A number of existing reading-into-writing test tasks from large-scale EAP tests were critically reviewed by the team at the meeting. They included IELTS Academic Task 1 (nonverbal), PTE Academic – Summarize Written Text (single, short, verbal), TOEFL iBT® Integrated Writing (verbal and listening), GEPT Advanced Writing Task 1 (multiple verbal) and Task 2 (multiple nonverbal), and Trinity Integrated Skills in English (ISE) Level IV Task 1 (multiple verbal and nonverbal) and Task 2 (multiple verbal). The recent importance being given in EAP tests to reading-into-writing tasks was taken on board.

The framework of initial ideas provided as prior input to the meeting by Weir informed a general discussion of overall test design which resulted in a decision that the test should be structured into two levels, with initially two B1-level *Task As* to cover the expected level of most test takers, but also allowing for a B2-level *Task B* to discriminate amongst high-level test takers as appropriate for the future TEAP test-taker population. The second task would be at this higher level but it still should be accessible to candidates at the B1 level.

The meeting looked carefully at the possibilities for Task A and B variants, focusing on their cognitive/context validity aspects. It concentrated particularly on the appropriateness of the task types to the TEAP TLU situation and the appropriateness of the linguistic demands of the tasks.

The meeting agreed that real-life academic writing tasks involve both verbal and nonverbal sources. Following lengthy discussion of the research reported in the literature and a review of well-considered contemporary academic writing tasks, the team felt that the TEAP should include a summarization of a single text to be catered for in Task A and a second writing Task B which would require test takers to write an essay by drawing upon multiple sources, both verbal and nonverbal.

The project team felt that the TEAP tasks should focus on more cognitively demanding topics and themes with a more sophisticated focus appropriate to academic settings, usually with wider social relevance; i.e., a cognitive academic language proficiency (CALP) level rather than the purely personal or the familiar aspects of everyday life, such as shopping or hobbies and likes and dislikes, etc.—i.e., a basic interpersonal communication skills (BICS) level (Cummins, 2008). Examples of appropriate topics are intercultural communication, the environment, education, medical issues, etc. However, the texts selected should not require specialized background knowledge.

Detail was also provided to participants on the CEFR (Council of Europe, 2001) as a source (especially the writing assessment descriptors) for defining the difference between an "easier" Task A at the B1 level and a more advanced Task B at the B2 level. It was felt that bringing the CEFR into the test design from the beginning would facilitate stakeholder understanding of the test scores and task requirements. It might also be useful to report scores not only as scale scores but in bands which can indicate to test takers their approximate level in terms of some external criterion, and the CEFR offered possibilities here.

The CEFR descriptors from the most relevant scales provided one source of background guidance alongside consideration of other existing scales, from which the TEAP TLU-specific descriptors were developed for the rating scales. This was done with the explicit intention of making close connections with the CEFR levels in the rating scales and test design for the purposes of reporting the results to test takers. Linking the test to a widely used outside criterion was intended to increase transparency and interpretability and give added value to feedback for test users. However, a number of important changes to the CEFR performance descriptors were made, where the scales were either inadequate or not sufficiently comprehensive, well calibrated, or transparent.

A further day was subsequently spent with the team at Eiken on refining the assessment criteria identified in the earlier whole group discussion, editing the rubrics, and improving the draft tasks and criteria, and the iterative process of revision continued thereafter. An important part of this further development of the test specifications was the identification of useful quantitative indices to be used for evaluating contextual parameters relevant to the production of input reading texts for both Task A and Task B. As described in Green, Ünaldi, and Weir (2010), such indices provide an empirical basis for the review of test task content. It was decided from the outset that the level of the input reading texts should be controlled to be at one level below the expected performance level of the task. For example, the input texts for Task A (B1) would be controlled to be accessible at an A2+ level of reading ability, and the input texts for Task B (B2) should be accessible to students with a B1-level reading ability. It was felt that controlling for difficulty in the input reading tests would reduce the potential interference of reading ability in the interpretation of the test results primarily as a performance test of writing ability. The empirical indices suggested by the consultant for this purpose were incorporated into the test task specifications and provided concrete guidelines for item development and review and a transparent means of controlling for reading difficulty in the input texts.

Post-meeting Development

The team at Sophia University and the external consultant as well as Eiken staff, were all involved in the subsequent ongoing revision of the tasks, criteria, and test procedures. In the months that followed the meeting at Sophia, draft tasks and specifications were revisited iteratively and the sample tasks were revised and tried out on opportunity samples of native speakers (NS) and nonnative speakers (NNS), together with the marking schemes. Initial small-scale trialing was conducted after more advanced drafts of test specifications and matching tasks were created. For the writing tasks, this was done with Eiken staff, including 10 NS and 10 NNS speakers, to check if interpretations of the summary tasks would be consistent across these groups and also across ability levels. The criteria steadily evolved and were progressively tightened up through multiple revisions. These mini-trials shed initial light on timing for the tasks, word limits, test-taker understanding of tasks (e.g., clarity of rubrics), applicability of criteria to output, and appropriateness of texts.

On the basis of this series of small mini-trials and iterative consultancy on Skype, the team produced a draft version of the specifications, from which they finalized two prototype test Tasks—A and B—together with appropriate sets of marking criteria for each (covering content, grammar, lexis, cohesion, and coherence). For Task A, cohesion and coherence were treated as one category in the marking scheme, whereas they were separated for Task B. This was agreed during the workshop, as it was felt that, because Task A is designed to be answered in a single paragraph, there was little scope for coherence in terms of organizational structure. Inconsistencies and problems in the scales were identified and eliminated. Wording was continually improved in terms of appropriateness, transparency, brevity, and accessibility to stakeholders.

2. Empirical Studies: Trial Test and Pilot Test

2.1 Scope of the Two Studies

There were two stages in the TEAP developmental pathway:

- **Study 1—Trial:** An initial small-scale trial test with 61 first-year university students in December 2010 and two raters from the Eiken project team. All scripts were double-marked. A subset of scripts was further marked by six raters, three from the Eiken project team (including the two main raters) and three from CRELLA. This subset of scripts provided the basis for in-depth discussion of the rating scales between the two teams.
- **Study 2—Pilot:** A large-scale pilot test with 120 third-year high school students in December 2011 and four raters who were recruited externally. These raters were required to have advanced degrees in TESOL or a relevant field and/or extensive teaching experience in the relevant contexts for the TEAP (Japanese high school and university EFL). As well as teaching experience, all four raters had extensive previous training and experience in the rating of writing-proficiency tests.

2.2 Research Questions to Be Answered by the Investigation

Major attention would be paid to the following questions:

- **RQ1:** How well does the test function in terms of scoring validity after incorporating modifications from the earlier trialing?
- **RQ2:** Is there any evidence from test-takers' output language that validates the descriptors used to define the levels on each rating scale?

Consideration would also be given to establishing the participating students' perceptions of the testing procedures and the participating raters' views of the scoring procedures.

2.3 Study 1—Initial Trial

Data was first collected in a small-scale trial test to examine how well the draft test materials and rating scales operationalized the test construct described in the test specification in terms of contextual, cognitive, and scoring validity. The trial also allowed detailed analysis of how the rating scales functioned in practice. Research was carried out *inter alia* into:

- the adequacy of the timing for both tasks
- the length of text that students produced
- acceptable word limits
- whether the students understood what they were expected to do
- the appropriateness of texts for target-group students

- the usefulness of the marking scheme for operational testing
- the perceptions of the test as a suitable test for incoming university students

Different analyses were carried out on test scores, ratings, and feedback questionnaires from raters and students. All of these sources of empirical validity evidence offered useful information to verify or modify the test specifications, test tasks and rubrics, and rating scale descriptors in preparation for the main Study 2. Study 2 would focus on scoring validity and linguistic features of the test-takers' output language, and would confirm that changes made after the trial test functioned in the ways that the test designers intended.

2.3.1 Participants

As described previously, the students were all first-year university students at Sophia University. They represented a range of different departments, but were all enrolled in the same general foreign-language courses. Initially, it was intended to obtain a small sample representing three broad proficiency groups of elementary, intermediate, and advanced levels. Due to a number of practical constraints, in practice the students came from two broad proficiency levels as determined by placement procedures for the courses they were taking; intermediate (N=30) and advanced (N=31).

2.3.2 Data Collection Procedures

The writing test was administered as a part of the students' regular foreign-language classes. The small-scale, iterative, internal trialing described previously had resulted in an initial estimate of 70 minutes as an appropriate time for students at the relevant level of ability to complete both tasks. Teachers were issued with instructions to allow students 20 minutes for Task A and 50 minutes for Task B. Although the initial test design called for two Task As, for this initial trial it was decided to administer only one of each task type. This was done for practical reasons, as the trial depended on a convenience sample that was obtained by administering the tests during normal class times. It was also unclear whether two Task As were in fact necessary, and it was decided to await the results of the trial before making a final decision on this matter.

The post-test questionnaire for students included questions on whether they felt the time allocation was relevant as well as questions asking about the explicitness and interpretability of the test task instructions, whether they had understood what the tasks required, the appropriateness of the level of the input texts, and the appropriateness of the tasks overall for measuring writing ability. This final topic was felt to be particularly important for evaluating the face validity of the test design for potential test users, as the tasks were innovative for proficiency tests. Each student was issued a test booklet that contained the instructions and test tasks, space for writing responses, and space for planning and editing, if desired (although this was not required). The questionnaire was printed on the back of the test booklet and contained multiple-choice items on the issues described above, as well as space for written responses from students, if they wished to supply any extra comments. All test materials were collected by the classroom teachers and returned to the project team at Eiken for analysis.

2.3.3 Data Analysis

All test scripts were double-marked by two raters who were members of the Eiken project team and had extensive training and experience in rating EFL performance tests. After rating an initial 10 scripts for each task, the two raters met to discuss issues of interpretation of the rating scale descriptors and to standardize their judgments. After agreeing on scores for the initial scripts, each rater marked the remaining scripts individually. Raters allocated scores of 0-3 for each analytic rating scale (0 = below A2, 1 = A2, 2 = B1, 3 = B2). Classical testing analyses were carried out on the rating data and descriptive statistics and score distributions produced for review by the principal consultant and the team at CRELLA. Inter-rater reliability was assessed through indices of agreement, adjacent agreement, and the correlations between scores by the two raters for each of the analytic scales as well as the overall composite scores for each task. Composite scores were derived by summing the results of each rater's separate analytic scale scores to derive a total for Task A and a total score for Task B. A subset of 10 scripts for each task was further independently rated by another project member at Eiken as well as three members of CRELLA. This common subset of scripts marked by six raters provided a basis for an in-depth, post-rating discussion of the rating scales and suggested revisions.

Descriptive statistics were produced for the questionnaire results, and all free responses by test takers were inputted into an Excel spreadsheet for easy review and reference by project team members. All scripts were inputted as text files by a research assistant, and analysis of the vocabulary levels of the scripts using Range vocabulary analysis software were carried out. A number of quantitative measures for analyzing the linguistic features of the scripts were obtained through tools available in Microsoft Word, including readability indices, the number of paragraphs, number of words per script, average number of words per sentence, and average number of sentences per paragraph. An internal report on the results of these different analyses was prepared and distributed to all stakeholders for review, including the project team at Eiken, CRELLA, and the development team at Sophia.

2.3.4 Results of Initial Trial of TEAP Writing Tasks and Marking Scripts

General Findings Based on Results of the Trial

- The general impression, based on the data available, is that the test works as intended.
- The time allocated for each task and the word limits appear appropriate. Students can write to the intended lengths.
- Students are positive about both integrated tasks and see them as appropriate for measuring academic writing ability. The research literature supports this view.
- The majority of those responding to the questionnaire were clear on what was required of them by the test.
- The level of the tasks appears appropriate in terms of both contextual and cognitive parameters.
- Assuming that the participating students in Study 1 are more proficient or as proficient as the target population of the test, we expect that many high school students (at the A2 or B1 benchmark levels recommended for high school graduates by the Japanese Ministry of Education) are likely to struggle with Task B, which is a B2-level task designed to distinguish between more advanced intermediate students and lower-intermediate B1-level students. This is as expected. Nevertheless nearly all were able to attempt the task in a conscientious manner.

It is necessary that the test enables us to distinguish between any students who have performed at a B2 level and those who have not. It is also necessary for Task B to correctly identify those who have performed at a B1 level rather than an A2 level or lower. It appears to provide appropriate information on both distinctions.

Specific Recommendations for Overall Task Specifications

- There only needs to be *one* summary task for Task A. This task is performing well on its own and providing sufficient information to make judgments as to whether someone has performed at a B1 level or not.
- It appears appropriate to keep the original time set for the test (70 minutes total test time for tasks A and B) even though this proved to be more than sufficient with this sample (which may be of a slightly higher average ability than the eventual test population).

Recommendations for Rating Scales

- Remove “orthographic control” as a criterion from the scales. The category does not add relevant information and its omission will facilitate any future transition to computer-based training (CBT) versions.
- Further review of the wording and subcategories within each scoring category is needed to reduce redundancy and tighten descriptors in order to reduce the processing burden on raters.
- For the moment, in Task B, keep “coherence” and “cohesion” as separate scales for feedback purposes, but monitor their separate values when the test goes live.
- For Task A “main ideas,” it is appropriate to maintain the importance of identifying main ideas in order to get a score of 2 (B1 level). Students unable to identify the theme in the input text should get a score of 1 (A2 level).
- For Task B “main ideas,” similarly, it is important to maintain the focus on students having to identify the overlapping solution offered in both input texts. This was an integral part of the original task design. Content is the key criterion in the view of most academics (along with organization).
- For B1, the CEFR states that an appropriate response for processing text uses “original text wording and ordering.” Task A is designed to be a B1-level task. It is appropriate for students at a B1 level who can make a summary using the original wording to receive a score of 2. At the B2 level, they should be encouraged in the rubric to use their own words. The descriptors at the relevant levels of the scale for “lexical range and accuracy” also make these distinctions clear. A student would not be able to achieve a B2 level on this scale without demonstrating sufficient lexical resources of their own. The score scales thus allow for positive feedback targeted at an appropriate level for different test takers: writing to a B1 level as designated in the CEFR, appropriately reusing the wording from the input text can be an appropriate goal for many high school students. At the same time, the limitations of this level of performance are made clear by the descriptors for this scale, and in order to demonstrate an advanced B2 level of performance, a test taker would need to demonstrate more sophisticated lexical resources.
- When assigning an overall level for feedback purposes, it may be appropriate to use a mixed conjunctive-compensatory model. In other words, students cannot be classed as B2 level unless they receive a B2 rating in “main ideas” (or in other important categories, possibly “coherence/cohesion”).
- Writing significantly less or more than the prescribed word limit is likely to result in lower scores in a number of the criteria, so this might be pointed out in the handbook rather than on the exam paper, as we do not want unnecessary word counting while the exam is in pencil-and-paper mode.

We feel that this interactive, *a priori* validation approach—in which results from various data-collection stages were discussed and acted on by the researchers at Eiken, colleagues at Sophia, and the principal consultant and his team in CRELLA—meant that the test versions employed in the main pilot study had already been refined to a high level. The results from second-stage trials for writing described previously indicated that the tests worked well at a practical level, and that the rating scales seemed to be measuring the constructs intended and were being interpreted consistently by raters.

The results from the trial were reviewed iteratively from January to September 2011, and minor changes were made to the writing scales, test content, and test specifications, with ongoing feedback from the CRELLA team. Minor changes were thus made to both rating scales and test content for both tests before the main pilot to reflect the trial results.

2.4 Study 2—Main Pilot

2.4.1 Participants

A total of 120 third-year high school students were recruited through the network of private Catholic schools associated with Sophia University. Students were given book coupons for participating. These students were considered to be representative of the typical test-taker population for Sophia University, which is one of the leading liberal arts universities in Japan. As the initial stage of test development was focused on designing a test appropriate for Sophia University, one of the principal stake holders in the project, this sample of test takers was felt to be appropriate. It should be noted, however, that these test takers may not be representative of the broad range of typical high school students in Japan. In particular, it can be expected that these students were more highly motivated, and certainly had a particular interest in studying English as a foreign language, as evidenced by their willingness to give up their free time to voluntarily participate in a pilot version of a performance test for no significant immediate personal gain. The level of motivation of these students can be further inferred by the participation rate. Of the 120 applicants who applied to take part in the test, 113 attended on the date of the test, a participation rate of 94%. Although 113 students took the speaking pilot test, which was administered on the same day as the writing test, during the writing test one student became slightly ill and was unable to complete the test. Although the student had attempted Task A before leaving the test room, this student's data was removed from the analysis for both tasks. This resulted in a total sample of 112 students for the writing pilot.

2.4.2 Data Collection Procedures

The test was conducted on a Sunday. The TEAP Writing Test was being developed at the same time as a separate project to develop a speaking component for the TEAP testing program, and the final pilot tests for both of these test components were conducted in tandem on the same day. The 120 students who applied to take part were split into four groups of 30 students each. Two groups took both tests (writing and speaking) in the morning, and two groups took both tests (writing and speaking) in the afternoon. For both the morning and afternoon time slots, one group took writing first and the other group took speaking first. In this way, the test sample was completely counter-balanced to avoid order effects, etc. As described above, of the 120 applicants, a total of 113 participants actually took part in the speaking test, and data from 112 test takers was obtained for the writing test. A questionnaire based on the same one used during the trial was administered to all participants after they had completed the writing test.

For the writing test, a total of four raters marked all scripts. As described in Section 2.1, these raters were recruited externally and were required to meet strict conditions for education and experience. All of the raters were trained and experienced in the rating of writing-proficiency tests, including for high-stakes purposes. Rating was conducted over a three-day period. The first half-day consisted of a three-hour training session. In the first 90 minutes, raters were given background information on the TEAP test, the CEFR, and the levels targeted by the TEAP Writing Test. This was followed by review and discussion of the TEAP writing tasks and writing scales. The second 90-minute half of the training session consisted of rating standardized exemplars of performance of the test tasks. The standardized exemplars were selected from the Study 1 trial scripts, which had already been extensively reviewed. A selection of these scripts on which the raters had achieved high levels of agreement were further reviewed by three members of the Eiken project team, to choose scripts on which all three members could agree on the analytic scores for each scale. In this way, solid examples of scripts which scored across the main three score bands—A2, B1, and B2—were able to be selected for use as standardized exemplars. Training with exemplars followed the following pattern:

- Raters read through the script for Standardized Exemplar 1, and the Eiken training staff provided the results and the reasons for the benchmark scores.
- Raters were given the script for Standardized Exemplar 2 and asked to rate the script first. The scores from each rater were elicited and the raters discussed the reasons for their scores before the Eiken training staff provided the Eiken benchmark scores and the rationale for these scores.
- The final two exemplars were rated consecutively by raters, and the scores for each were elicited from each rater with discussion. The benchmark scores were provided by Eiken training staff with the rationale for these scores.

The above procedure was repeated for both Task A and Task B. After training, raters each individually rated all 112 scripts for both tasks. Raters were not permitted to discuss their scores with each other during rating.

2.4.3 Data Analysis

The same set of classical test theory (CTT) analyses on rating data was conducted on the pilot test scores as for the Study 1 trial. Agreement indices were calculated for all pairs of raters and descriptive statistics prepared for the students' questionnaire results. In addition to the CTT analysis of test scores, the rating data was analyzed using multifaceted Rasch analysis with the FACETS program, the details and results of which are described below.

The writing scripts for all students were transferred to text-file format for analysis of linguistic features using the Coh-Metrix program. The results of this analysis are also described in more detail below.

2.4.4 Results of Writing Pilot (Administered December 18, 2011)

Student Scores

We present below a summary of student performance on the two tasks in the TEAP writing pilot. The results in terms of CEFR levels for separate analytic scales and overall CEFR levels for each task were derived from the ratings of all four raters in the pilot. Analyses of rater reliability (see Section 3.1 for details) showed that all of the raters were interpreting the scales in a suitably uniform manner and were performing with adequate levels of consistency, and this supported the use of all four raters' data in the estimation of preliminary CEFR levels for the purposes of this study. Tables 1 and 2 provide an overview of the descriptive statistics for final scores awarded on each of the separate criteria and for overall scores for both tasks. Figures 1 and 2 show the distribution of test takers who have been allocated to each CEFR level across each of the rating criteria used for both Task A and Task B. The majority of the students were, in the main, performing at the B1 level across the criteria, as might have been expected by our knowledge of the test-taker potential at this level in Japan. There were few B2 or below-A2 performances. The higher scores for lexis in Task A are perhaps attributable to the students being able to use words directly from the texts provided in their answers. The shortness of the summary in Task A and in the input passage itself similarly appears to lift the coherence and cohesion scores for many candidates.

Table 1: ***Task A—TEAP Writing Pilot Descriptive Statistics***

	N	Minimum	Maximum	Mean	Std. Deviation
Overall	112	.00	2.69	1.6696	.58938
Main ideas	112	.00	2.50	1.5446	.60517
Coherence and cohesion	112	.00	3.00	1.7277	.67413
Lexical range and accuracy	112	.00	3.00	1.7232	.63703
Grammatical range and accuracy	112	.00	3.00	1.6830	.62600

Table 2: ***Task B—TEAP Writing Pilot Descriptive Statistics***

	N	Minimum	Maximum	Mean	Std. Deviation
Overall	112	.00	2.90	1.6884	.53059
Main ideas	112	.25	2.75	1.6897	.59203
Coherence	112	.25	3.00	1.6563	.63239
Cohesion	112	.25	3.00	1.7701	.57357
Lexical range and accuracy	112	.25	2.75	1.6674	.53628
Grammatical range and accuracy	112	.25	3.00	1.6585	.56850

Figure 1: **Task A—TEAP writing pilot student performance**

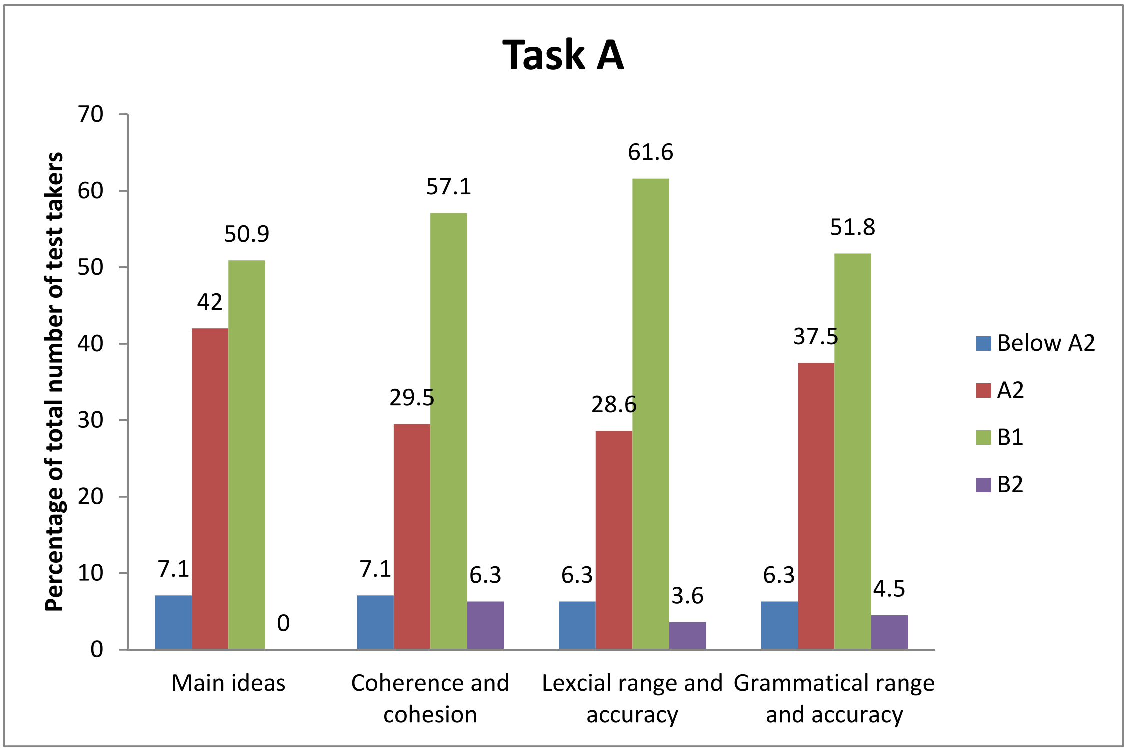
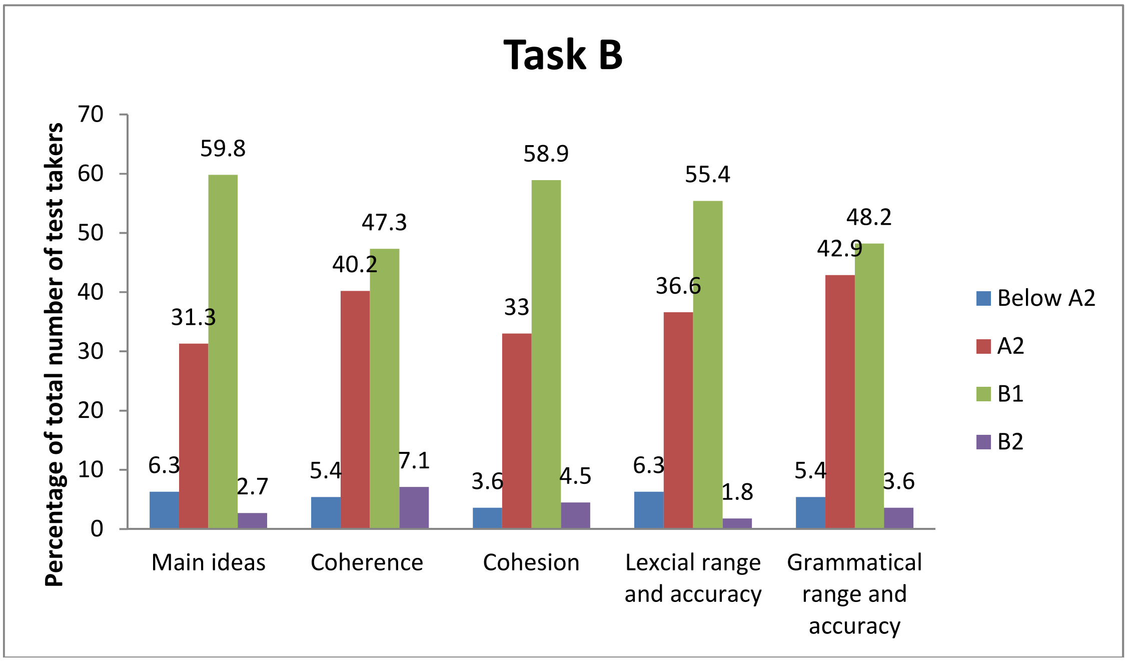


Figure 2: **Task B—TEAP writing pilot student performance**



Task B places more demand on the students in terms of overall "coherence," as the required output is twice the length of Task A. Fewer candidates achieved B1 grades in terms of "lexical range and accuracy" and "grammatical range and accuracy," as they were encouraged to use their own words and information has to be *transformed* rather than *retold*. These trends are shown in the mean scores reported below.

As Table 3 shows, the majority of students achieved similar grades on the two tasks. Nine students who achieved a B1 grade on Task A did worse on Task B, but only three students who were awarded a B1 grade on Task B did worse on Task A. This suggests that for most candidates Task B was more difficult than Task A, as intended. Cross-tabulations for individual criteria performances are reported in Appendix A.

Table 3: ***Cross-tabulation of Overall Scores on Tasks A and B***

		Task A Overall			
		Below A2	A2	B1	B2
		Count	Count	Count	Count
Task B Overall	Below A2	7	0	0	0
	A2	1	37	9	0
	B1	0	3	53	1
	B2	0	0	0	1

The following sections describe analyses on the data produced by the pilot and focus on several aspects of scoring validity. These analyses directly address the research questions (RQ1 and RQ2) described earlier in Section 2.2.

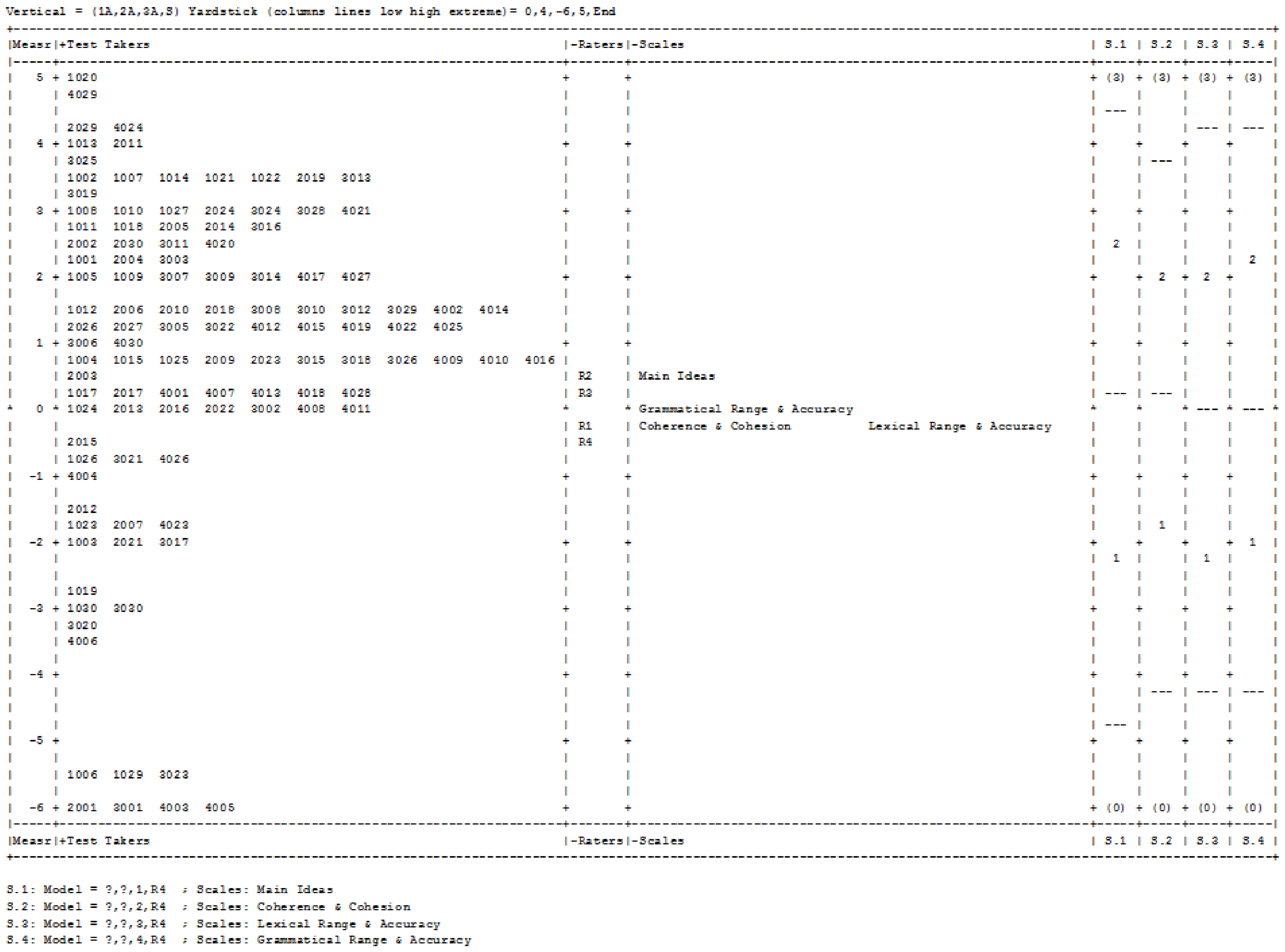
3. Analysis of Rating Performance

3.1 Case Study 1—Facet Analysis of Rating in TEAP Writing Pilot Test

Multifaceted Rasch analysis was carried out using three major facets for the score variance in this study: examinees, raters, and rating categories. As Task A and Task B were separately rated using different rating scales, this section analyses the quality of the ratings for these two tasks separately.

Figures 3 and 4 show an overview of the results of the partial credit analysis for Task A and Task B, respectively, plotting estimates of examinee ability, examiner harshness, and rating scale difficulty. They were all measured by the uniform unit (logits) shown on the left side of the map labeled measure ("measr"), making it possible to directly compare all the facets. The more able examinees are placed towards the top and the less able towards the bottom. The more lenient examiners and the easier rating categories appear towards the bottom, and the harsher examiners and the more difficult rating categories towards the top (e.g., rater 2 [R2] is the harshest examiner in Figure 3). The right-hand column, "scale," refers to the levels of the rating scales.

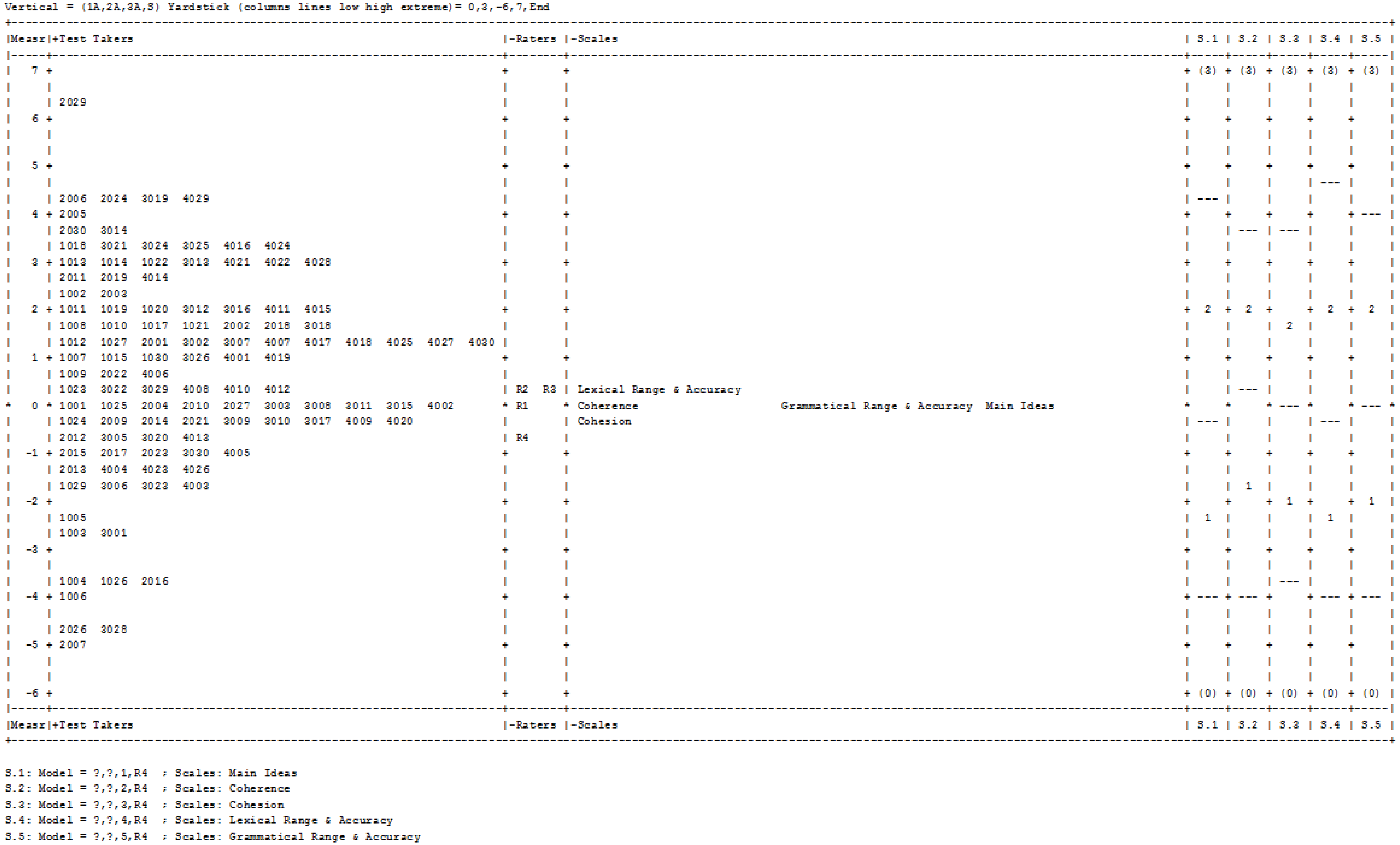
Figure 3: Overall facet map—Task A (partial credit analysis)



```
Vertical = (1A,2A,3A,S) Yardstick (columns lines low high extreme)= 0,4,-6,5,End
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Test Takers                                                      |-Raters|-Scales                                                     | S.1 | S.2 | S.3 | S.4 |
|-----+------------------------------------------------------------------+-------+------------------------------------------------------------+-----+-----+-----+-----|
|   5 + 1020                                                             +       +                                                            + (3) + (3) + (3) + (3) |
|     | 4029                                                             |       |                                                            |     |     |     |     |
|     |                                                                  |       |                                                            | --- |     |     |     |
|     | 2029  4024                                                       |       |                                                            |     |     | --- | --- |
|   4 + 1013  2011                                                       +       +                                                            +     +     +     +     |
|     | 3025                                                             |       |                                                            |     | --- |     |     |
|     | 1002  1007  1014  1021  1022  2019  3013                         |       |                                                            |     |     |     |     |
|     | 3019                                                             |       |                                                            |     |     |     |     |
|   3 + 1008  1010  1027  2024  3024  3028  4021                         +       +                                                            +     +     +     +     |
|     | 1011  1018  2005  2014  3016                                     |       |                                                            |     |     |     |     |
|     | 2002  2030  3011  4020                                           |       |                                                            |  2  |     |     |     |
|     | 1001  2004  3003                                                 |       |                                                            |     |     |     |  2  |
|   2 + 1005  1009  3007  3009  3014  4017  4027                         +       +                                                            +     +  2  +  2  +     |
|     |                                                                  |       |                                                            |     |     |     |     |
|     | 1012  2006  2010  2018  3008  3010  3012  3029  4002  4014       |       |                                                            |     |     |     |     |
|     | 2026  2027  3005  3022  4012  4015  4019  4022  4025             |       |                                                            |     |     |     |     |
|   1 + 3006  4030                                                       +       +                                                            +     +     +     +     |
|     | 1004  1015  1025  2009  2023  3015  3018  3026  4009  4010  4016 |       |                                                            |     |     |     |     |
|     | 2003                                                             | R2    | Main Ideas                                                 |     |     |     |     |
|     | 1017  2017  4001  4007  4013  4018  4028                         | R3    |                                                            | --- | --- |     |     |
*   0 * 1024  2013  2016  2022  3002  4008  4011                         *       * Grammatical Range & Accuracy                               *     *     * --- * --- *
|     |                                                                  | R1    | Coherence & Cohesion          Lexical Range & Accuracy     |     |     |     |     |
|     | 2015                                                             | R4    |                                                            |     |     |     |     |
|     | 1026  3021  4026                                                 |       |                                                            |     |     |     |     |
|  -1 + 4004                                                             +       +                                                            +     +     +     +     |
|     |                                                                  |       |                                                            |     |     |     |     |
|     | 2012                                                             |       |                                                            |     |     |     |     |
|     | 1023  2007  4023                                                 |       |                                                            |     |  1  |     |     |
|  -2 + 1003  2021  3017                                                 +       +                                                            +     +     +     +  1  |
|     |                                                                  |       |                                                            |  1  |     |  1  |     |
|     |                                                                  |       |                                                            |     |     |     |     |
|     | 1019                                                             |       |                                                            |     |     |     |     |
|  -3 + 1030  3030                                                       +       +                                                            +     +     +     +     |
|     | 3020                                                             |       |                                                            |     |     |     |     |
|     | 4006                                                             |       |                                                            |     |     |     |     |
|     |                                                                  |       |                                                            |     |     |     |     |
|  -4 +                                                                  +       +                                                            +     +     +     +     |
|     |                                                                  |       |                                                            |     | --- | --- | --- |
|     |                                                                  |       |                                                            |     |     |     |     |
|     |                                                                  |       |                                                            | --- |     |     |     |
|  -5 +                                                                  +       +                                                            +     +     +     +     |
|     |                                                                  |       |                                                            |     |     |     |     |
|     | 1006  1029  3023                                                 |       |                                                            |     |     |     |     |
|     |                                                                  |       |                                                            |     |     |     |     |
|  -6 + 2001  3001  4003  4005                                           +       +                                                            + (0) + (0) + (0) + (0) |
|-----+------------------------------------------------------------------+-------+------------------------------------------------------------+-----+-----+-----+-----|
|Measr|+Test Takers                                                      |-Raters|-Scales                                                     | S.1 | S.2 | S.3 | S.4 |
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------+

S.1: Model = ?,?,1,R4  ; Scales: Main Ideas
S.2: Model = ?,?,2,R4  ; Scales: Coherence & Cohesion
S.3: Model = ?,?,3,R4  ; Scales: Lexical Range & Accuracy
S.4: Model = ?,?,4,R4  ; Scales: Grammatical Range & Accuracy
```

Figure 4: Overall facet map—Task B (partial credit analysis)



```
Vertical = (1A,2A,3A,S) Yardstick (columns lines low high extreme)= 0,3,-6,7,End
+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Test Takers                                                       |-Raters |-Scales                                                                                | S.1 | S.2 | S.3 | S.4 | S.5 |
|-----+------------------------------------------------------------------+--------+---------------------------------------------------------------------------------------+-----+-----+-----+-----+-----|
|   7 +                                                                  +        +                                                                                       + (3) + (3) + (3) + (3) + (3) |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|     | 2029                                                             |        |                                                                                       |     |     |     |     |     |
|   6 +                                                                  +        +                                                                                       +     +     +     +     +     |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|   5 +                                                                  +        +                                                                                       +     +     +     +     +     |
|     |                                                                  |        |                                                                                       |     |     |     | --- |     |
|     | 2006  2024  3019  4029                                           |        |                                                                                       | --- |     |     |     |     |
|   4 + 2005                                                             +        +                                                                                       +     +     +     +     + --- |
|     | 2030  3014                                                       |        |                                                                                       |     | --- | --- |     |     |
|     | 1018  3021  3024  3025  4016  4024                               |        |                                                                                       |     |     |     |     |     |
|   3 + 1013  1014  1022  3013  4021  4022  4028                         +        +                                                                                       +     +     +     +     +     |
|     | 2011  2019  4014                                                 |        |                                                                                       |     |     |     |     |     |
|     | 1002  2003                                                       |        |                                                                                       |     |     |     |     |     |
|   2 + 1011  1019  1020  3012  3016  4011  4015                         +        +                                                                                       +  2  +  2  +     +  2  +  2  |
|     | 1008  1010  1017  1021  2002  2018  3018                         |        |                                                                                       |     |     |  2  |     |     |
|     | 1012  1027  2001  3002  3007  4007  4017  4018  4025  4027  4030 |        |                                                                                       |     |     |     |     |     |
|   1 + 1007  1015  1030  3026  4001  4019                               +        +                                                                                       +     +     +     +     +     |
|     | 1009  2022  4006                                                 |        |                                                                                       |     |     |     |     |     |
|     | 1023  3022  3029  4008  4010  4012                               | R2  R3 | Lexical Range & Accuracy                                                              |     | --- |     |     |     |
*   0 * 1001  1025  2004  2010  2027  3003  3008  3011  3015  4002       * R1     * Coherence                   Grammatical Range & Accuracy  Main Ideas                  *     *     * --- *     * --- *
|     | 1024  2009  2014  2021  3009  3010  3017  4009  4020             |        | Cohesion                                                                              | --- |     |     | --- |     |
|     | 2012  3005  3020  4013                                           | R4     |                                                                                       |     |     |     |     |     |
|  -1 + 2015  2017  2023  3030  4005                                     +        +                                                                                       +     +     +     +     +     |
|     | 2013  4004  4023  4026                                           |        |                                                                                       |     |     |     |     |     |
|     | 1029  3006  3023  4003                                           |        |                                                                                       |     |  1  |     |     |     |
|  -2 +                                                                  +        +                                                                                       +     +     +  1  +     +  1  |
|     | 1005                                                             |        |                                                                                       |  1  |     |     |  1  |     |
|     | 1003  3001                                                       |        |                                                                                       |     |     |     |     |     |
|  -3 +                                                                  +        +                                                                                       +     +     +     +     +     |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|     | 1004  1026  2016                                                 |        |                                                                                       |     |     | --- |     |     |
|  -4 + 1006                                                             +        +                                                                                       + --- + --- +     + --- + --- |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|     | 2026  3028                                                       |        |                                                                                       |     |     |     |     |     |
|  -5 + 2007                                                             +        +                                                                                       +     +     +     +     +     |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|     |                                                                  |        |                                                                                       |     |     |     |     |     |
|  -6 +                                                                  +        +                                                                                       + (0) + (0) + (0) + (0) + (0) |
|-----+------------------------------------------------------------------+--------+---------------------------------------------------------------------------------------+-----+-----+-----+-----+-----|
|Measr|+Test Takers                                                       |-Raters |-Scales                                                                                | S.1 | S.2 | S.3 | S.4 | S.5 |
+-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+

S.1: Model = ?,?,1,R4  ; Scales: Main Ideas
S.2: Model = ?,?,2,R4  ; Scales: Coherence
S.3: Model = ?,?,3,R4  ; Scales: Cohesion
S.4: Model = ?,?,4,R4  ; Scales: Lexical Range & Accuracy
S.5: Model = ?,?,5,R4  ; Scales: Grammatical Range & Accuracy
```

Examinees in Task A and Task B

The test was able to discriminate well between examinees. As shown in Tables 4 and 5, the fixed (all same) chi-square test was statistically significant in both Task A and Task B analysis (Task A: χ^2 (111) = 1572.7, p<.005; Task B: χ^2 (111) = 2046.9, p<.005). The separation index was 4.06 and 4.19, and the examinees were able to be separated into 5.75 and 5.92 statistically separate strata in Task A and Task B, respectively. In other words, there were strong empirical grounds for using at least the three performance levels employed in the TEAP. The person reliability, analogous to a Cronbach's alpha reliability estimate in a CTT analysis, was .94 in Task A and .95 in Task B. The ability to separate the examinees into statistically distinct strata is important for the TEAP test, since it will be used for entrance purposes to discriminate between students of different ability levels.

For fit analysis, we follow Wright and Linacre's suggestion that infit mean-square values in the range of 0.5 to 1.5 are "productive for measurement" (Wright & Linacre, 1994). Out of the 112 students analyzed, 9 students in Task A and 13 students in Task B were identified as misfitting (see Tables 4 and 5). The percentage of misfitting students in the data sets was 8% and 12%, respectively. These values are much greater than the 2% level that any test development should aim at (McNamara, 1996, p. 178). Perhaps because neither writing nor speaking has been taught or studied as systematically as reading and listening skills in the English education system in Japan, some students had jagged profiles across the five rating criteria. That is, some students might have shown a strong performance in certain categories and still had distinct individual weaknesses in other areas, which was picked up as misfitting in the analysis.

Table 4: ***TEAP Writing Pilot Task A0—Test-Taker Measurement Report Summary and List of Misfitting Test Takers (Partial Credit Analysis)***

```
+-----------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair-M|       Real | Infit     Outfit  |Estim.| Corr. |                    |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Num Test
Takers   |
|------------------------------+-------------+--------------------+------+-------+--------------------|
|   11     16      .7   .69|  -3.61   .73 | 2.02  2.9  2.02  2.8| -.70 |   .45 |  88 4006          |
|   14     16      .9   .88|  -2.75   .79 | 1.98  1.8  2.02  1.8|  .27 |  -.13 |  18 1019          |
|   17     16     1.1  1.06|  -1.74  1.01 | 3.04  2.8  3.00  2.7| -.27 |   .03 |  22 1023          |
|   20     16     1.3  1.24|   -.82   .96 | 3.35  4.3  3.60  4.2|-1.48 |   .19 |  25 1026          |
|   24     16     1.5  1.50|    .17   .61 | 1.59  2.0  1.55  1.8|  .08 |   .33 | 110 4028          |
|   27     16     1.7  1.69|    .87   .60 | 1.53  1.5  1.57  1.6|  .31 |  -.19 |  36 2009          |
|   31     16     1.9  1.94|   1.88   .82 | 2.52  2.7  2.31  2.3| -.08 |  -.09 |   5 1005          |
|   34     16     2.1  2.12|   2.68   .71 | 1.90  1.9  1.90  1.8|  .27 |   .54 |  33 2005          |
|   35     16     2.2  2.18|   2.94   .67 | 1.74  1.7  1.74  1.7|  .30 |   .39 |  50 2024          |
|------------------------------+-------------+--------------------+------+-------+--------------------|
| Total  Total  Obsvd  Fair-M|       Real | Infit     Outfit  |Estim.| Corr. |                    |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Num Test
Takers   |
|------------------------------+-------------+--------------------+------+-------+--------------------|
|   26.7   16.0    1.7  1.67|    .70   .60 |  .99  -.1   .99  -.1|      |   .15 | Mean (Count: 112)  |
|    9.4     .0     .6   .59|   2.71   .25 |  .51  1.3   .52  1.3|      |   .21 | S.D. (Population)  |
|    9.4     .0     .6   .59|   2.72   .26 |  .51  1.4   .53  1.3|      |   .21 | S.D. (Sample)      |
+-----------------------------------------------------------------------------------------------------+
   With extremes, Real, Populn: RMSE .65  Adj (True) S.D. 2.63  Separation 4.06  Strata 5.75
Reliability .94
   With extremes, Real, Sample: RMSE .65  Adj (True) S.D. 2.65  Separation 4.08  Strata 5.77
Reliability .94
Without extremes, Real, Populn: RMSE .56  Adj (True) S.D. 2.04  Separation 3.66  Strata 5.22
Reliability .93
Without extremes, Real, Sample: RMSE .56  Adj (True) S.D. 2.05  Separation 3.68  Strata 5.24
Reliability .93
With extremes, Real, Fixed (all same) chi-square: 1572.7  d.f.: 111  significance (probability): .00
With extremes, Real,  Random (normal) chi-square: 96.0  d.f.: 110  significance (probability): .83
-----------------------------------------------------------------------------------------------------
```

Table 5: ***TEAP Writing Pilot Task B—Test-Taker Measurement Report Summary and List of Misfitting Test Takers (Partial Credit Analysis)***

```
+-----------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair-M|       Real | Infit     Outfit   |Estim.| Corr. |              |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Num Test
Takers    |
|-------------------------------+--------------+---------------------+------+-------+---------------------|
|   12     20      .6   .60|  -3.60  .56 | 1.69  2.7  1.69  2.7| -.50 |  -.02 |   4 1004          |
|   17     20      .9   .85|  -2.65  .87 | 3.77  4.9  4.03  5.1|-1.55 |   .33 |   3 1003          |
|   21     20    1.1  1.05|  -1.82  .64 | 1.98  2.1  1.98  2.0|  .31 |   .40 |  85 4003          |
|   22     20    1.1  1.10|  -1.61  .75 | 2.73  3.3  2.75  3.2| -.29 |   .56 |  27 1029          |
|   33     20    1.6  1.65|    .45  .60 | 1.94  2.8  2.09  3.1| -.39 |   .43 |  22 1023          |
|   36     20    1.8  1.80|   1.04  .68 | 2.22  2.7  2.19  2.6| -.02 |  -.29 |   7 1007          |
|   38     20    1.9  1.90|   1.47  .68 | 2.06  2.2  2.09  2.1|  .24 |  -.38 |  26 1027          |
|   39     20    2.0  1.95|   1.70  .62 | 1.65  1.4  1.70  1.5|  .52 |  -.07 |  16 1017          |
|   41     20    2.0  2.04|   2.16  .67 | 1.92  1.9  1.79  1.6|  .43 |   .28 |  11 1011          |
|   42     20    2.1  2.09|   2.39  .83 | 3.02  3.5  2.91  3.2| -.46 |  -.24 |   2 1002          |
|   43     20    2.2  2.14|   2.62  .62 | 1.72  1.7  1.99  2.1|  .33 |  -.16 |  96 4014          |
|   44     20    2.2  2.19|   2.84  .63 | 1.85  2.1  1.74  1.7|  .26 |  -.17 |  21 1022          |
|   46     20    2.3  2.29|   3.26  .67 | 2.19  3.3  2.18  3.0| -.48 |   .23 |  78 3025          |
|-------------------------------+--------------+---------------------+------+-------+---------------------|
| Total  Total  Obsvd  Fair-M|       Real | Infit     Outfit   |Estim.| Corr. |              |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Num Test
Takers    |
|-------------------------------+--------------+---------------------+------+-------+---------------------|
|   33.8   20.0    1.7  1.69|    .71  .49 | 1.00  -.1  1.00  -.2|      |  .12 | Mean (Count: 112)  |
|   10.6     .0     .5   .53|   2.14  .09 |  .55  1.4   .57  1.4|      |  .22 | S.D. (Population)  |
|   10.6     .0     .5   .53|   2.15  .09 |  .55  1.4   .58  1.4|      |  .22 | S.D. (Sample)      |
+-----------------------------------------------------------------------------------------------------------+
Real, Populn: RMSE .50  Adj (True) S.D. 2.09  Separation 4.19  Strata 5.92  Reliability .95
Real, Sample: RMSE .50  Adj (True) S.D. 2.10  Separation 4.21  Strata 5.95  Reliability .95
Real, Fixed (all same) chi-square: 2046.9  d.f.: 111  significance (probability): .00
Real,  Random (normal) chi-square: 105.9  d.f.: 110  significance (probability): .59
-------------------------------------------------------------------------------------------------------------
```

Raters and Rating Criteria in Task A

In relation to these main pilot test results, the most important consideration is that of whether raters are able to make effective use of the scale categories when scoring the test takers' scripts. It is to be expected that some raters will be harsher than others, but as long as they behave consistently, such differences can be taken into account in scoring the performances. We used the FACETS program to examine to what extent raters were performing consistently and rating criteria were functioning systematically.

Table 6 shows that on Task A, all four raters were consistent in applying the scales. All of the raters made use of every scale category on every criterion and, following Wright and Linacre's suggestion that infit mean-square values in the range of 0.5 to 1.5 are "productive for measurement," no rater exhibited misfit (an indication of inconsistent scoring behavior) or overfit (over-predictability of rater behavior) to the model. However, the four raters involved differed in their levels of harshness/leniency, and these differences were statistically significant (p<.01). Rater 2 was the harshest, awarding scores that were, on average, a quarter of a band lower than those awarded by rater 4, the most lenient. It should be noted that the magnitude of the differences, with a range spanning just over 1 logit, is in fact quite good in terms of studies which have used FACETS to analyze the efficacy of rater training, with the spread of severity measures in this sample being close to those obtained in post-training analyses; for example, Elder, Knoch, Barkhuizen, & von Randow (2005); Knoch, Read, & von Randow (2007); and Weigle (1998). O'Sullivan (2002) suggests that when the range of severity of raters is much smaller than the range of ability measures, we can interpret this as an indication that differences in rater severity do not have much practical impact on scores. In this case, the range of ability measures is more than 10 times that of the rater severity measures.

Table 10 shows the levels of agreement between raters. Levels of agreement were high, with all four raters in agreement on 58% of the scores awarded. Average adjacent agreement (the percentage of cases of exact agreement in addition to cases in which raters differed by plus or minus one score band) is almost 100% for both writing tasks. Obviously, the steps are broad (only 4 points on the scale).

Of the four rating criteria in Task A, "main ideas" produced the lowest and "lexical range and accuracy" the highest scores for these test takers. There were no misfitting or overfitting scale criteria (see Table 7). The raw data shows that raters 1, 2, and 4 were harshest on "main ideas" and most generous either on "lexical range and accuracy" (rater 1) or "grammatical range and accuracy" (raters 2 and 4). Conversely, rater 3 awarded his or her highest scores for "main ideas" and lowest for "grammatical range and accuracy." This suggests that rater 3 has interpreted the scale criteria somewhat differently to the other raters, but, as the fit statistics (infit and outfit) displayed in Table 7 show, this has not seriously impacted the measurement qualities of the scale.

Table 6: ***TEAP Writing Pilot Task A—Rater Measurement Report (Partial Credit Analysis)***

```
+--------------------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair-M|       Real | Infit     Outfit  |Estim.| Corr. | Exact Agree. |                |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Obs %  Exp % | N Raters         |
|------------------------------+--------------+---------------------+------+-------+--------------+---------------------|
|  801    448     1.9   1.86|   -.51   .10 | 1.04   .5  1.02   .3|  .97 |   .51 |  57.0   55.5 | 4 R4            |
|  770    448     1.8   1.78|   -.21   .10 | 1.00   .0  1.02   .2|  .99 |   .53 |  56.7   56.5 | 1 R1            |
|  732    448     1.7   1.69|    .15   .11 | 1.15  2.1  1.13  1.7|  .85 |   .51 |  55.0   56.5 | 3 R3            |
|  689    448     1.6   1.58|    .57   .10 |  .80 -3.2   .77 -3.4| 1.22 |   .54 |  58.6   55.2 | 2 R2            |
|------------------------------+--------------+---------------------+------+-------+--------------+---------------------|
|  748.0  448.0    1.7   1.73|    .00   .10 | 1.00  -.1   .99  -.3|      |   .52 |              | Mean (Count: 4)    |
|   41.9     .0     .1    .11|    .40   .00 |  .13  2.0   .13  1.9|      |   .01 |              | S.D. (Population)  |
|   48.4     .0     .1    .12|    .46   .00 |  .15  2.3   .15  2.2|      |   .01 |              | S.D. (Sample)      |
+--------------------------------------------------------------------------------------------------------------------+
Real, Populn: RMSE .10  Adj (True) S.D. .39  Separation 3.88  Strata 5.51  Reliability (not inter-rater) .94
Real, Sample: RMSE .10  Adj (True) S.D. .45  Separation 4.52  Strata 6.36  Reliability (not inter-rater) .95
Real, Fixed (all same) chi-square: 66.0  d.f.: 3  significance (probability): .00
Real,  Random (normal) chi-square: 2.9  d.f.: 2  significance (probability): .23
Inter-Rater agreement opportunities: 2592  Exact agreements: 1473 = 56.8%  Expected: 1450.2 = 55.9%
----------------------------------------------------------------------------------------------------------------------
```

Table 7: ***TEAP Writing Pilot Task A—Scale Measurement Report (Partial Credit Analysis)***

```
+---------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair-M|        Real | Infit      Outfit   |Estim.| Corr. |                          |
|  Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | N Scales
|
|-------------------------------+--------------+----------------------+------+-------+---------------------------------|
|   774     448      1.8   1.77|   -.22   .09 |  .95  -.7   .94  -.8| 1.06 |    .53 | 2 Coherence & Cohesion         |
|   772     448      1.8   1.82|   -.14   .10 |  .87 -1.9   .85 -2.1| 1.13 |    .55 | 3 Lexical Range & Accuracy      |
|   754     448      1.7   1.76|   -.04   .10 |  .96  -.5   .96  -.5| 1.04 |    .53 | 4 Grammatical Range & Accuracy
|
|   692     448      1.6   1.58|    .39   .11 | 1.21  2.9  1.19  2.5|  .78 |    .47 | 1 Main Ideas                  |
|-------------------------------+--------------+----------------------+------+-------+---------------------------------|
|   748.0   448.0     1.7   1.73|    .00   .10 | 1.00  -.1   .99  -.2|      |   .52 | Mean (Count: 4)               |
|    33.3      .0      .1    .09|    .24   .01 |  .13  1.8   .13  1.7|      |   .03 | S.D. (Population)            |
|    38.4      .0      .1    .11|    .27   .01 |  .15  2.1   .15  2.0|      |   .04 | S.D. (Sample)               |
+---------------------------------------------------------------------------------------------------------------------+
Real, Populn: RMSE .10  Adj (True) S.D. .21  Separation 2.10  Strata 3.14  Reliability .82
Real, Sample: RMSE .10  Adj (True) S.D. .25  Separation 2.50  Strata 3.66  Reliability .86
Real, Fixed (all same) chi-square: 19.9  d.f.: 3  significance (probability): .00
Real,  Random (normal) chi-square: 2.6  d.f.: 2  significance (probability): .27
-----------------------------------------------------------------------------------------------------------------------
```

Raters and Rating Criteria in Task B

On Task B, the raters showed differences in terms of harshness, and these differences were once again statistically significant (p<.01) (see Table 8). Rater 2 was once again the harshest and rater 4 the most lenient, with a difference of around a quarter of a band between them. Again, there were no misfitting raters, suggesting, as we would hope, that these raters were generally able to employ the scales consistently in scoring these test takers. Levels of agreement were again high on Task B, with all four raters in agreement on 57% of the scores awarded (Table 10).

However, the raw data shows that on Task B, rater 3 did not award any band 3 scores for "lexical range and accuracy" (although he or she did award band 3 to at least nine test takers on other criteria). On the other hand, rater 4 only awarded the lowest score of 0 to one test taker. This suggests that some raters may need more encouragement to award scores at the extreme points of the scale on one or more criteria. Reluctance to use the top and bottom of the rating scale is well recognized in the literature on raters and rating behavior, and thus needs to be made a priority in rater training.

Table 8: ***TEAP Writing Pilot Task B—Rater Measurement Report (Partial Credit Analysis)***

```
+-------------------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair-M|      Real | Infit     Outfit  |Estim.| Corr. | Exact Agree. |              |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Obs %  Exp % |
N Raters        |
|------------------------------+--------------+--------------------+------+-------+-------------+--------------------|
| 1019    560     1.8   1.85|  -.54   .09 | 1.21  3.4  1.22  3.1|  .77 |   .42 |  53.3   54.9 | 4 R4           |
|  967    560     1.7   1.76|  -.16   .09 |  .87 -2.3   .84 -2.5| 1.16 |   .52 |  60.4   56.1 | 1 R1           |
|  907    560     1.6   1.65|   .28   .09 | 1.02   .2  1.03   .5|  .98 |   .47 |  57.0   55.9 | 3 R3           |
|  889    560     1.6   1.61|   .41   .09 |  .90 -1.7   .90 -1.6| 1.10 |   .54 |  58.6   55.5 | 2 R2           |
|------------------------------+--------------+--------------------+------+-------+-------------+--------------------|
|  945.5  560.0    1.7   1.72|   .00   .09 | 1.00  -.1  1.00  -.1|      |   .49 |             | Mean (Count: 4)   |
|   51.3     .0     .1    .09|   .37   .00 |  .13  2.3   .14  2.2|      |   .05 |             | S.D. (Population) |
|   59.3     .0     .1    .11|   .43   .00 |  .15  2.6   .17  2.6|      |   .06 |             | S.D. (Sample)     |
+-------------------------------------------------------------------------------------------------------------------+
Real, Populn: RMSE .09  Adj (True) S.D. .36  Separation 4.13  Strata 5.85  Reliability (not inter-rater) .94
Real, Sample: RMSE .09  Adj (True) S.D. .42  Separation 4.81  Strata 6.75  Reliability (not inter-rater) .96
Real, Fixed (all same) chi-square: 69.1  d.f.: 3  significance (probability): .00
Real,  Random (normal) chi-square: 2.9  d.f.: 2  significance (probability): .23
Inter-Rater agreement opportunities: 3360  Exact agreements: 1926 = 57.3%  Expected: 1868.1 = 55.6%
---------------------------------------------------------------------------------------------------------------------
```

Of the five criteria used in scoring Task B (see Table 9), "cohesion" produced the highest scores: three of the four raters awarded their highest scores on this criterion (the exception being rater 3). Raters 1 and 2 awarded their lowest scores for any criterion on "coherence," while rater 3 was again harshest on "grammatical range and accuracy" and rater 4 gave his or her lowest scores on "main ideas." As with Task A, there were no misfitting or overfitting scale criteria. As with Task A, rater 3 appears to have interpreted the scale criteria a little differently to the other raters, but without seriously compromising the quality of the measurement.

The result that all rating criteria in Task A and Task B showed good fit values is encouraging, as this indicates that the assumption of unidimensionality holds for this data (Bonk & Ockey, 2003). In other words, the separate analytic rating scales seem to be independent but contributing to a common construct of “writing ability” measured by Task A and Task B of the test. This is important for the TEAP Writing Test, as it aims to provide a composite score as one source of feedback by summing scores across the separate analytic scales in Task A and Task B.

Table 9: ***TEAP Writing Pilot Task A—Scale Measurement Report (Partial Credit Analysis)***

```
+--------------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair-M|       Real | Infit     Outfit  |Estim.| Corr. |                        |
| Score  Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | N Scales
|
|------------------------------+--------------+---------------------+------+-------+-------------------------------|
|  793    448     1.8  1.81|  -.27  .10 | 1.07  1.0  1.06   .8|  .93 |   .48 | 3 Cohesion                |
|  742    448     1.7  1.63|  -.01  .09 | 1.00   .0   .99   .0| 1.02 |   .50 | 2 Coherence               |
|  743    448     1.7  1.67|   .04  .10 | 1.08  1.1  1.08  1.1|  .92 |   .47 | 5 Grammatical Range &
Accuracy |
|  757    448     1.7  1.76|   .07  .10 |  .89 -1.7   .89 -1.4| 1.10 |   .51 | 1 Main Ideas              |
|  747    448     1.7  1.74|   .17  .10 |  .94  -.9   .96  -.4| 1.06 |   .50 | 4 Lexical Range & Accuracy    |
|------------------------------+--------------+---------------------+------+-------+-------------------------------|
|  756.4  448.0    1.7  1.72|   .00   .10 |  .99  -.1  1.00   .0|      |  .49 | Mean (Count: 5)           |
|   19.1     .0     .0   .06|   .15   .00 |  .07  1.1   .07   .9|      |  .02 | S.D. (Population)         |
|   21.3     .0     .0   .07|   .16   .00 |  .08  1.2   .07  1.1|      |  .02 | S.D. (Sample)             |
+--------------------------------------------------------------------------------------------------------------+
Real, Populn: RMSE .10  Adj (True) S.D. .11  Separation 1.13  Strata 1.84  Reliability .56
Real, Sample: RMSE .10  Adj (True) S.D. .13  Separation 1.36  Strata 2.15  Reliability .65
Real, Fixed (all same) chi-square: 11.3  d.f.: 4  significance (probability): .02
Real,  Random (normal) chi-square: 3.0  d.f.: 3  significance (probability): .40
--------------------------------------------------------------------------------------------------------------
```

Table 10: ***Inter-rater Agreement in Task A and Task B***

Writing Task A							
Exact		**Plus/Minus 1**		**Adjacent Agreement**		**Differ by 2 or more categories**	
MEAN	58%	MEAN	39%	MEAN	97%	MEAN	3%
MAX	76%	MAX	51%	MAX	100%	MAX	6%
MIN	47%	MIN	23%	MIN	94%	MIN	0%

Writing Task B							
Exact		**Plus/Minus 1**		**Adjacent Agreement**		**Differ by 2 or more categories**	
MEAN	57%	MEAN	40%	MEAN	97%	MEAN	3%
MAX	77%	MAX	58%	MAX	100%	MAX	10%
MIN	39%	MIN	21%	MIN	90%	MIN	0%

Inter-rater Agreement

Having established the acceptable level of marker reliability in the rating of the test tasks, we then investigated whether there was any external empirical evidence to support the classification of the scripts into the criterion levels mapped out in the marking scheme. We took the 112 Task A and 112 Task B candidate scripts from the pilot study, typed them up into electronic files, and submitted them to Coh-Metrix analysis to examine whether the scripts of students adjudged to be at the B1 level were in fact significantly different in terms of a variety of lexical, syntactic, and cohesion indices in Coh-Metrix than student scripts at the A2 level. This case study is reported next.

There were two reasons for focusing on criterial differences between the B1 and A2 levels only for the purposes of this study. Firstly, the B1 and A2 levels are of course relevant to the benchmark levels of English proficiency for high school graduates recommended by MEXT. Secondly, as reported earlier in Section 2.4.4, there were in practice very few performances rated at a B2 level.

3.2 Coh-Metrix Analysis of Scripts at A2 and B1 Levels on Tasks A and B

The literature on the potential sources of text complexity is extensive, and so this brief survey is limited to those studies that considered a broad range of parameters of relevance to this case study. Firstly, a number of descriptive typologies of text characteristics have been designed for use in test development and validation studies (see, for example, Freedle & Kostin, 1993; Freedle, 1997; Bachman, Davidson, Ryan, & Choi, 1995; Fortus, Coriat, & Fund, 1998; Enright et al., 2000; Masi, 2002; Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu, 2004; and Khalifa & Weir, 2009). Bachman et al.'s 1995 test comparability study identified textual properties such as the nature of the text, length, vocabulary, grammar, cohesion, distribution of new information, type of information, topic, genre, rhetorical organization, and illocutionary acts. Freedle and Kostin (1993; see also Freedle, 1997) took into consideration referentials, rhetorical organizers, fronted structures, vocabulary, concreteness/abstractness, subject matter, coherence, and length of various segments such as word, sentence, paragraph, and passage as text-related variables. Fortus, Coriat, and Fund (1998) investigated length, number of negations, number of referential markers, vocabulary, grammatical complexity, abstractness, topic, and rhetorical structure as textual variables contributing to the level of difficulty of reading-comprehension items. Enright et al. (2000) identified three groups of salient textual features: grammatical/discourse features, pragmatic/rhetorical features, and linguistic variables. Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu (2004) included text source, authenticity, discourse type, domain, topic, nature of content, text length, vocabulary, and grammar as relevant features for text analysis. Khalifa and Weir (2009) examined the contextual features proposed in the research literature and established a subset which enabled Cambridge ESOL to make criterial distinctions between levels of proficiency in its reading examinations. There appears to be a measure of consensus in the subjective judgments of these different authors on the features to be addressed when considering text complexity.

Until relatively recently, we lacked the quantitative tools necessary to compare automatically and accurately the various contextual characteristics of the range of written texts at different levels of ability (Biber, Conrad, Reppen, Byrd, Helt, Clark, et al., 2004). However, recent advances in automated textual analysis and computational linguistics have now made it feasible to provide more quantitative approaches focusing analytically on a wide range of individual characteristics (Crossley, Louwerse, McCarthy, & McNamara, 2007; Crossley & McNamara, 2008; Graesser, McNamara, Louwerse, & Cai, 2004; Graesser, McNamara, & Kulikowich, 2011; Green, Ünaldi, & Weir, 2010; Green, 2011; Weir, Bax, Chan, Field, Green, & Taylor, 2012; and Wu, 2012). New technologies offer examination boards the potential of a more systematic, efficient way of describing a number of the contextual parameters in texts (see Green, Ünaldi, & Weir, 2010).

Whilst most of the research studies in the area have been carried out on reading texts in tests (i.e., test input), there seemed, *mutatis mutandis,* to be no real barrier to employing the same sorts of analysis on the writing scripts produced in first- and second-language examinations (i.e., test output). Many automatic rating systems for scoring second-language writing performance are premised on similar automated algorithms relating to measures of lexical complexity, structural complexity, and text-level representation (Shermis & Berstein, 2003).

Graesser, McNamara, and Kulikowich (2011, p. 223) make a strong case for using a system called Coh-Metrix:

> Recent advances in numerous disciplines have made it possible to computationally investigate various measures of text and language comprehension that supersede surface components of language and instead explore deeper, more global attributes of language. They have allowed the analysis of many deep-level factors of textual coherence and processing to be automated, permitting more accurate and detailed analyses of language to take place. A synthesis of the advances in these areas has been achieved in Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis.

They argue that such “automated analysis is unquestionably more reliable and objective than approaches that involve humans annotating and rating texts by hand” (Graesser, McNamara, & Kulikowich, 2011, pp. 223-234).

Our study was informed by a survey of the Coh-Metrix-related literature and of other literature on text difficulty to determine which indices had been found useful in the past for establishing the relative complexity of written texts at different levels, such as school readers across the US grade range (Crossley, Greenfield, & McNamara, 2008; Graesser, McNamara, & Kulikowich, 2011).

It took into account the recent work done by Wu (2012) in her doctoral study which compared Cambridge Main Suite and GEPT Taiwan examinations in ESOL at the B1 and B2 levels in terms of contextual and cognitive parameters, the work by Green, Ünaldi, and Weir (2010) in comparing IELTS and undergraduate texts at British universities, and the investigation by Green (2011) of texts targeted at different levels of the CEFR (Council of Europe, 2001).

In addition to using Coh-Metrix to analyze the linguistic features of reading material (Crossley, Louwerse, McCarthy, & McNamara, 2007; Crossley & McNamara, 2008), the same researchers have relatively successfully applied Coh-Metrix to students’ writing performance in order to explore various features of the written output produced by both L1 writers (Crossley & McNamara, 2010a; McNamara, Crossley, & McCarthy, 2010) and L2 writers (Crossley & McNamara, 2010b). There is also research comparing writing performance across L1 and L2 writers (Crossley & McNamara, 2009, 2011). Some of this research has compared high- and low-scoring writing performance for L2 writers (Crossley & McNamara, 2010b) and L1 writers (McNamara, Crossley, & McCarthy, 2010). Of particular relevance to this project is the fact that several of the L2-focused studies were undertaken with a population not dissimilar to the target population for the TEAP Writing Test. For example, Crossley and McNamara (2010b, p. 17) report on the essay performances of graduating high school students in Hong Kong, in which they observed that high-scoring L2 writers produced more linguistically sophisticated texts that “use less frequent, less familiar words, while also deploying a more diverse range of words.” Interestingly, the Coh-Metrix analyses seemed to suggest that while the performance of high-level L2 writers was marked by syntactic complexity, lexical diversity, and word frequency, it proved harder to identify cohesion and coherence as a distinguishing measure, at least based on those indices available using Coh-Metrix. In relation to L1 writers, McNamara, Crossley, and McCarthy (2010, p. 76) similarly concluded, “some of the textual features of good student writing may not be the same as those features that are considered to be facilitative for reading (i.e., cohesion indices).”

Crossley and McNamara (2009, pp. 132-133) called for more research to be conducted in this area, especially studies linking computational analyses with human judgments of writing quality:

> Future studies should attempt to evaluate texts using independent measures such as human rating to examine variance between L1 and L2 writings. … Specifically, we see the need for studies examining different language populations, studies that consider how well computational tools predict human judgments of text proficiency, and the possibility of using computational tools such as Coh-Metrix to evaluate L2 essays.

This brief literature review outlined above provided some encouragement and justification for applying Coh-Metrix to the TEAP writing scripts, thus responding to calls for further research.

Given the large, and sometimes seemingly overlapping, number of indices (54) offered in the publicly available online version of Coh-Metrix, Weir, Bax, Chan, Field, Green, and Taylor (2012) first eliminated 13 indices which did not exhibit any significant differences across three levels of reading proficiency in the well-established Cambridge ESOL examinations—FCE (B2), CAE (C1), and CPE (C2). In our study of the TEAP writing scripts, 6 of these 13 indices did not exhibit any significant differences across tasks A and B at the B1 and A2 levels and were thus eliminated from further investigation in this study (index identification numbers below are those used in the Coh-Metrix output file):

- 13 Incidence of negative temporal connectives
- 24 Number of conditional expressions, incidence score
- 36 Average sentences per paragraph
- 49 Concreteness mean for content words
- 50 Incidence of positive logical connectives
- 52 Ratio of intentional particles to intentional content

In the TEAP writing study, an additional set of 13 indicators exhibited no significant differences across the two proficiency levels A2 and B1, and so they too were eliminated from further investigation, as listed below:

- 9 Positive additive connectives
- 10 Incidence of positive temporal connectives
- 11 Positive causal connectives
- 15 All connectives
- 18 Adjacent anaphor reference
- 21 Anaphor reference all
- 23 Pronoun ratio
- 27 LSA sentence adjacent
- 53 Intentional content
- 55 Sentence syntax similarity adjacent
- 56 Syntactic similarity all
- 57 Syntactic similarity all within paragraphs
- 60 Minimum concreteness content words

Twenty-three further indices were eliminated after iterative focus-group discussion of the results obtained. The list of potential complexity indices was pruned where it was felt they:

- Overlapped, even if a significant difference between texts was also present in the associated indices: e.g., 20 Stem overlap overall (vs. 17 Adjacent stem overlap), 19 Argument overlap all distances (vs. 16 Argument overlap adjacent).
- Did not tell us anything useful: e.g., 22 Noun phrase incidence, 34 Number of paragraphs, 33 Number of sentences, 14 Incidence of negative causal connections.
- Showed unsatisfactory results on box plots: i.e., significant but only very small, real difference between the two adjacent levels of proficiency or anomalous results (e.g., A2 more complex than B1):
    - 7 Incidence of causal verbs, links, and particles
    - 12 Incidence of negative additive connectives
    - 25 Number of negations incidence score
    - 26 Logical operators
    - 28 LSA all sentences combination mean
    - 29 LSA paragraph to paragraph mean
    - 30 Personal pronoun incidence score
    - 31 Noun hypernym
    - 32 Mean hypernym value of verbs
    - 42 Higher-level constituents
    - 44 Type token ratio
    - 45 Raw frequency content words
    - 47 Min raw freq content words
    - 51 Incidence of negative logical connectives
    - 54 Temporal cohesion

- 58 Content word overlap
- 59 Mean of location and motion ratio scores

We finally decided on a total of 12 key indices that were clearly useful in establishing differences in the text complexity of written scripts rated at the A2 and B1 levels in our study (see Table 11). The descriptive data for these indices can be found in Appendices A and B.

Table 11: ***Twelve Indices Where There Was a Significant and Meaningful Difference Between B1 and A2 Candidates on Task A or B***

Indices with significant and meaningful difference	
Cohm 8 Causal cohesion	Cohesion
Cohm 16 Argument overlap, adjacent, unweighted	Cohesion
Cohm 17 Adjacent stem overlap	Cohesion
Cohm 35 Number of words	Lexical
Cohm 37 Average words per sentence	Syntactic
Cohm 38 Average syllables per word	Lexical
Cohm 39 Flesch Reading Ease	Lexical
Cohm 40 Flesch-Kincaid	Lexical
Cohm 41 Mean number of modifiers per noun-phrase	Syntactic
Cohm 43 Mean number of words before main verb of main clause in sentences	Syntactic
Cohm 46 Log frequency of content words	Lexical
Cohm 48 Log min. freq of content words	Lexical

A number of general caveats need to be expressed, arising from the experience of making these comparisons, before discussing the findings in detail.

Few reported studies have attempted to use Coh-Metrix to establish differences between students' exam writing scripts at different proficiency levels (based on examiner ratings), and so we must remain cautious about this innovative method of enquiry for validating marking scales (Roscoe, Crossley, Weston, & McNamara, 2011; Crossley & McNamara, 2010b, 2011; McNamara, Crossley, & McCarthy, 2010). Additionally, limitations are imposed by the available algorithms. Only 54 of around 600 algorithms resulting from weighted statistical models combining extracted linguistic properties are publicly available from Coh-Metrix.

Furthermore, the rating of written scripts necessarily takes account of a number of features which Coh-Metrix cannot account for:

(a) the extent and degree of lexical and syntactic errors in the scripts,

(b) the relevance and adequacy of the scripts in relation to the demands regarding task content set out in the test rubrics,

(c) the appropriateness of the use of lexis or grammar,

(d) the overall effect of an exam script on the reader,

(e) the overall appropriateness of the script in terms of intended genre and rhetorical task for the discourse community concerned,

(f) the writer's style or other expressive differences, and

(g) coherence beyond that activated by cohesive markers (McNamara, Crossley, & McCarthy, 2010, pp. 73-75); i.e., clear structuring so that main ideas and arguments are clearly presented, staying on topic.

It is important to note that the TEAP scripts in this study were assigned to overall CEFR levels B1 and A2 through rating procedures which included a number of the dimensions described in (a)–(g) above. The grades assigned were not based solely on the occurrence or non-occurrence of the various linguistic features (lexical, syntactic, or cohesive) that Coh-Metrix is designed to analyze and for which it can produce data. This means that the assignment of exam essays to different grades may not necessarily reflect a linear relationship in terms of a number of the Coh-Metrix lexical, syntactic, and cohesion indices that one might expect to find; for example, in studies which have looked at the difficulty level of textbooks at various stages in the school system. Solid performance on some of the TEAP assessment criteria may have lifted a student into the B1 rather than the A2 classification, as distinguished from more complex use of lexis, syntax, or cohesion. Conversely, poor performance on a number of the TEAP assessment criteria may have placed another student in the A2 category, despite a superior script in terms of a number of the Coh-Metrix text complexity indices.

Another factor to be kept in mind is that Task A was designed as an *easier* task at the B1 level, and it was felt appropriate that students should be allowed to use wording/syntax from the text at this level rather than their own words. As a result, we might not expect to find any real differences in the performances in Task A in respect of lexical and syntactic indices in Coh-Metrix. The few differences that occur in Task A largely relate to those indices which Coh-Metrix uses to determine the cohesion/coherence of text (see Appendix B, Task A indices for details). In Task B, there are clearer differences with respect to the lexical and syntactic indices, as candidates have to write far more and are encouraged and, in the case of reporting nonverbal information, forced to use their own words and syntactic knowledge.

Given that the ability to organize written work at the discourse level is normally expected only at the B2 level (Shaw & Weir, 2007, Chapter 3), it is perhaps not surprising that only a few cohesion/coherence indices indicate significant differences between A2 and B1 scripts. Cohesion and coherence may not be sufficiently well grounded in these ability levels yet, and the shortness of the piece of writing may exacerbate this situation (see details in Appendix B of where significant differences occurred between the two proficiency levels in terms of the Coh-Metrix indices).

There may be a more general condition at work, however. McNamara, Crossley, and McCarthy (2010, p. 57) found that:

> Using 26 validated indices of cohesion from Coh-Metrix, none showed differences between high and low proficiency essays and no indices of cohesion correlated with essay ratings. These results indicate that the textual features that characterise good student writing are not aligned with those features that facilitate reading comprehension. Rather, essays judged to be of higher quality were more likely to contain linguistic features associated with text difficulty and sophisticated language.

This point is taken up again in Section 3.2.3.

Finally, in Appendix D, we report the results of a multiple regression run on the Coh-Metrix indices that exhibited statistically significant and meaningful differences between levels A2 and B1. The results show that although the indices identified clear differences (see above and Appendices B and C), the predictive power of these indices was relatively limited even for Task B ($R^2 = 0.371$), albeit with only two levels to regress against. This should sound a note of caution for anyone contemplating using such measures as part of an automated scoring package. The criteria (a)–(g) we listed previously still require application by humans. Coh-Metrix offers complementary evidence that there are critical differences in texts produced by students in the TEAP writing A2 and B1 categories, but we should be cautious in expecting that it can do any more than that.

Nevertheless, our study does show clear differences between the two levels A2 and B1 in terms of a range of lexical, syntactic, and to a lesser extent cohesive indices, and it would be foolish to ignore such evidence of the scoring validity of the TEAP Writing Test. We finish this report by describing the 12 parameters where there were significant and meaningful differences between the A2 and B1 bands on the TEAP pilot writing scripts. The discussion that follows offers some theoretical and empirical underpinning for each of the 12 selected parameters, drawing on the research literature, especially in the fields of psycholinguistics and L1/L2 literacy. This provides some additional justification for the validity of the indices and for the application of computational tools, such as Coh-Metrix, to analyze text.

3.2.1 Lexical Complexity

The lexical indices that showed significant differences in scripts at the A2 and/or B1 levels are presented in Table 12 (see Appendices B and C for details).

Table 12: ***Lexical Indices Showing Significant Differences in Scripts at A2 and/or B1 Levels***

Coh-Metrix	**Task**	**CEFR**	**Min**	**Max**	**Mean**	**Std. Dv.**	**Mann-Whitney Test**
Cohm 35 Number of words	Task A	A2 (40)	35	154	82.35	24.49	z=-0.7; p=0.457
		B1 (62)	51	135	83.98	18.60	
	Task B	A2 (47)	41	253	171.96	46.34	z=-3.1; p<0.01
		B1 (57)	130	290	200.63	33.60	
Cohm 38 Average syllables per word	Task A	A2 (40)	1.21	1.51	1.33	0.07	z=0; p=0.992
		B1 (62)	1.10	1.61	1.33	0.08	
	Task B	A2 (47)	1.36	1.63	1.47	0.06	z=-2.8; p=0.01
		B1 (57)	1.37	1.64	1.50	0.06	
Cohm 39 Flesch Reading Ease	Task A	A2 (40)	34.10	93.73	79.77	9.80	z=-0.9; p=0.377
		B1 (62)	55.83	100	79.73	6.72	
	Task B	A2 (47)	53.47	79.49	69.33	5.22	z=-4.4; p<0.01
		B1 (57)	54.22	74.75	64.56	5.05	
Cohm 40 Flesch-Kincaid	Task A	A2 (40)	2.65	12	5.35	1.84	z=-1.6; p=0.099
		B1 (62)	2.34	12	5.75	1.54	
	Task B	A2 (47)	4.58	9.51	6.84	1.074	z=-4.6; p<0.01
		B1 (57)	5.70	12	8.02	1.223	
Cohm 46 Celex, logarithm, frequency of content words	Task A	A2 (40)	2.17	2.84	2.53	0.13	z=-0.9; p=0.369
		B1 (62)	2.31	2.96	2.56	0.14	
	Task B	A2 (47)	2.29	2.81	2.58	0.10	z=-3.8; p<0.01
		B1 (57)	2.39	2.67	2.52	0.07	
Cohm 48 Celex, logarithm, min. raw frequency of content words	Task A	A2 (40)	1.03	1.84	1.47	0.22	z=-0.8; p=0.425
		B1 (62)	1.09	1.91	1.44	0.19	
	Task B	A2 (47)	1.10	2.09	1.56	0.20	z=-4; p<0.01
		B1 (57)	1.00	1.69	1.40	0.17	

There is some overlap between these indices, but as well as being part of the Coh-Metrix range of indices, the first four are also available in the Microsoft Word program and are therefore directly accessible with no effort to test developers. So, although word length is included in the formula for Flesch Reading Ease Score and the latter is used to calculate Flesch-Kincaid Grade average, we decided to keep these in, as readers and researchers can readily make use of them.

Number of Words

What is perhaps more important than simple text length is the density and complexity of idea units within the text (Bachman, 1990), but Coh-Metrix cannot analyze this; only human raters can. That said, in a testing context, the longer the written output, the more complex the text is likely to be, in that it is likely to contain a greater number of idea units. The difference in average length between A2 and B1 scripts on Task B is noticeable.

Average Syllables per Word

Readers take longer to process a multisyllabic word than a monosyllabic one, allowing for frequency effects (Rayner & Pollatsek, 1989). The demands of decoding a text at the lexical level are thus better measured by counting syllables than by counting whole words. B1-level candidates use slightly longer words in Task B.

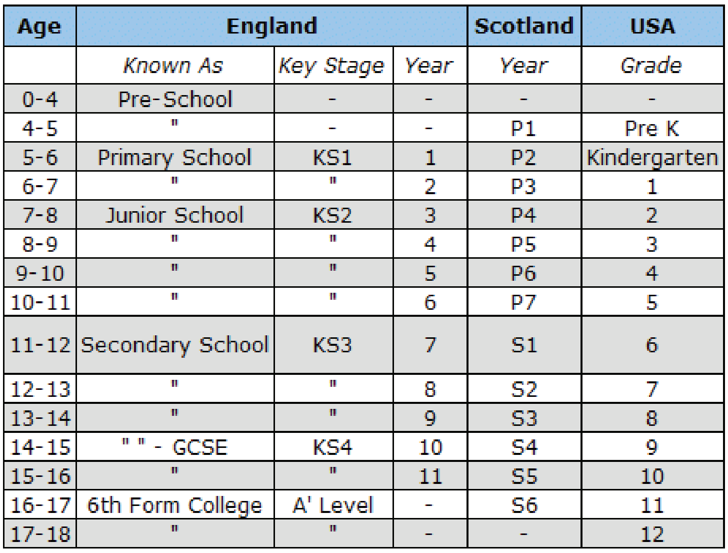
Flesch Reading Ease Score and Flesch-Kincaid Grade Estimates

In test development, readability formulas combining relatively crude syntactic and lexical features such as word and sentence length (including the Flesch Reading Ease Score and Flesch-Kincaid Grade Level) are often used as convenient indicators of text complexity. Such indices have been widely criticized as inadequate to reveal textual complexity (see, for example, Gervasi & Ambriola, 2002; Masi, 2002) and as being inappropriate for L2 readers (Brown, 1997), but nevertheless they still constitute important indices in more recent and detailed analyses of textual complexity.

These readability statistics ("Cohm 39 Flesch Reading Ease" and "Cohm 40 Flesch-Kincaid Grade Level"), widely used in test development, are also available through Microsoft Word: both measures being based on the relative numbers of syllables, words, and sentences found in a text. Flesch Reading Ease scores range from 0 to 100, with lower scores reflecting more challenging texts. A score below 50 is said to require college-level reading skills. The Flesch-Kincaid Grade Level is based on the US school system, with 12 representing the final year of high school and 13 to 16 the college level.

The more common Flesch-Kincaid Grade Level formula converts the Reading Ease Score to a US grade-school level. For example, a score of 5.0 means that a fifth-grader; i.e., a Year 6, average 10-year-old can understand the document. Table 13 shows a conversion table for British school years to US school grades. The scores end at university entrance level.

Table 13: ***Conversion Table for USA, England, and Scotland School Years***



Age	England			Scotland	USA
	Known As	*Key Stage*	*Year*	*Year*	*Grade*
0-4	Pre-School	-	-	-	-
4-5	"	-	-	P1	Pre K
5-6	Primary School	KS1	1	P2	Kindergarten
6-7	"	"	2	P3	1
7-8	Junior School	KS2	3	P4	2
8-9	"	"	4	P5	3
9-10	"	"	5	P6	4
10-11	"	"	6	P7	5
11-12	Secondary School	KS3	7	S1	6
12-13	"	"	8	S2	7
13-14	"	"	9	S3	8
14-15	" " - GCSE	KS4	10	S4	9
15-16	"	"	11	S5	10
16-17	6th Form College	A' Level	-	S6	11
17-18	"	"	-	-	12

Source: http://learning.covcollege.ac.uk/content/Jorum/CEH_Jorum/page_41.htm

Graesser, McNamara, Louwerse, and Cai (2004, p. 394) argue that readability formulas have widespread use even though they rely exclusively on word length and sentence length:

> ... Certainly these features have considerable validity as indices of text difficulty. However, such shallow aspects alone explain only a part of text comprehension, and ignore many language and discourse features that are theoretically influential at estimating comprehension difficulty… Texts are no doubt more difficult to read when they contain longer words and lengthier sentences. Longer words tend to be less frequent in the language, as we know from Zipf's (1949) law, and infrequent words take more time to access and interpret during reading (Just & Carpenter, 1980). We do not deny that the word- and sentence-length parameters in these readability formulas have some approximate degree of validity.

We need to bear in mind that Coh-Metrix is stated to work best on scripts equal to or greater than 200 words; since the length of the TEAP scripts is likely to be below 200 words (about 70 words for Task A and about 200 words for Task B), for indices such as type-token ratio, we might not be getting a totally accurate picture in this data. The literature also suggests that a text should generally have more than 200 words before the Flesch Reading Ease and Flesch-Kincaid Grade Level scores can successfully be applied; therefore, the data for Task B is likely to be the more reliable. The B1 texts in Task B are clearly more complex.

Word Frequency

Crossley, Greenfield, and McNamara (2008, pp. 482, 488) point out that:

> Frequency effects have been shown to facilitate decoding, with frequent words being processed more quickly and understood better than infrequent ones (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). Rapid or automatic decoding is a strong predictor of L2 reading performance (Koda, 2005). Texts which assist such decoding (e.g., by containing a greater proportion of high-frequency words) can thus be regarded as easier to process…

Crossley et al (2008, p. 483) notes:

> Coh-Metrix calculates word frequency information through CELEX frequency scores. The CELEX database (Baayen, Piepenbrock, & Gulikers, 1993) consists of frequencies taken from the early 1991 version of the COBUILD corpus, a 17.9-million-word corpus. For this study CELEX logarithm mean for content words was selected as the lexical-level variable. This measure was selected because frequency effects have been shown to facilitate decoding. Frequent words are processed more quickly and understood better than infrequent ones (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). Rapid or automatic decoding is a strong predictor of L2 reading performance (Koda, 2005). Texts which assist such decoding (e.g., by containing a greater proportion of high-frequency words) can thus be regarded as easier to process.

In our study, two CELEX frequency scores were selected:

(1) Log frequency of content words: FRQCLacw (Index 46)

This is the log frequency of all content words in the text. Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content.

Log frequency can be used as a predictor of text difficulty. Weir, Bax, Chan, Field, Green, and Taylor (2012) report a well-established frequency effect in reading results in slower decoding times for less frequent words (Garman, 1985).

(2) Log min. raw frequency of content words: FRQCLmcs (Index 48)

This initially computes the lowest log frequency score among all of the content words in each sentence. A mean of these minimum log frequency scores is then computed. The logarithm is to base 10. Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content. The word with the lowest log frequency score is the most rare word in the sentence (scores range from 0 to 6).

McNamara, Crossley, and McCarthy (2010) similarly found that word frequency (as measured by CELEX, logarithm for all words) was one of the three most predictive indices of essay quality.

B1-level students are using less-frequent words more often than A2-level students are.

3.2.2 Syntactic Complexity

Crossley, Greenfield, and McNamara (2008, p. 482) observe:

> …[in careful reading] a reading text is processed linearly, with the reader decoding it word by word; but, as he or she reads, the reader also has to assemble decoded items into a larger scale syntactic structure (Just & Carpenter, 1987; Rayner & Pollatsek, 1994). Clearly, the cognitive demands imposed by this operation vary considerably according to how complex the structure is (Perfetti, Landi, & Oakhill, 2005).

Texts with less complex grammar tend on the whole to be easier than texts with more complex grammar. Berman (1984) investigated how opacity and heaviness of sentence structures could result in increased difficulty in processing. A considerable number of indices have been suggested in the literature for the estimation of grammatical complexity (see Wolfe-Quintero, Inagaki, & Kim,1998; Ortega, 2003). Based on the earlier review process, we employed a range of the quantitative measures available through Coh-Metrix.

The Coh-Metrix indices that showed differences between A2 and B1 scripts are shown in Table 14 (see Appendices B and C for details).

Table 14: ***Coh-Metrix Indices Showing Differences Between A2 and B1 Scripts***

	Task	**CEFR**	**Min**	**Max**	**Mean**	**Std. dv.**	**Mann-Whitney test**
Cohm 37 Words per sentence	Task A	A2 (40)	7.33	32	13.94	3.80.	z=-2.54; p=0.01
		B1 (62)	9.71	32	14.57	3.83	
	Task B	A2 (47)	8.61	21.29	13.09	2.33	z=-3.7; p<0.01
		B1 (57)	10.36	25.71	15.17	3.04	
Cohm 41 Modifiers per noun phrase	Task A	A2 (40)	0.23	1	0.57	0.17	z=-1.6; p=0.121
		B1 (62)	0.17	1.04	0.52	0.15	
	Task B	A2 (47)	0.33	1	0.61	0.16	z=-2.2; p=0.03
		B1 (57)	0.32	0.94	0.66	0.11	
Cohm 43 Words before main verb	Task A	A2 (40)	1.2	6.25	3.61	1.35	z=-1.4; p=0.15
		B1 (62)	1.83	6	3.22	1.01	
	Task B	A2 (47)	1.36	7.2	3.26	1.09	z=-2.2; p=0.03
		B1 (57)	1.65	5.43	3.66	0.93	

Average Sentence Length

Weir, Bax, Chan, Field, Green, and Taylor (2012) claim that this index would appear to be a rough measure of both the syntactic complexity and the lexical density of a sentence. Clearly, the number of words in a sentence must often correlate loosely with the sentence's complexity in terms of number of clauses. Alternatively, or in addition, a longer sentence might contain longer and more complex phrases—i.e., might be denser in lexical terms. This measure partly relates to processing at the level of structure building (Gernsbacher, 1990) in that, the more complex the sentence, the more elaborate the structure that has to be assembled. If one assumes that longer sentences might also result from longer and more densely packed clauses, then the measure is also an indicator of difficulty of parsing. In parsing, a reader has to hold a series of words in the mind until such time as he/she reaches the end of a clause and can trace a syntactic pattern in the string (Rayner & Pollatsek, 1989). The longer the clause, the more words the reader has to hold in the mind. Lewis, Vasishth, and van Dyke (2006) suggest that processing items towards the end of longer sentences will be harder, since they usually have to be integrated with items that have occurred earlier on in the sentence. Graesser, Karnavat, Daniel, Cooper, Whitten, and Louwerse (2001) also suggest that longer sentences tend to place more demands on working memory and are therefore more difficult.

B1-level students in Task B are clearly writing longer sentences, on average more than two words extra.

Number of Modifiers Per Noun Phrase

The mean number of modifiers per noun phrase is an index of the complexity of referencing expressions. Barker (1998) argues that noun phrases carry much of the information in a text and computerized systems that attempt to acquire knowledge from text must first decompose complex noun phrases to get access to that information.

Graesser, McNamara, Louwerse, and Cai (2004) suggest that sentences with difficult syntactic composition have a higher ratio of constituents per noun phrase than do sentences with simple syntax. The presence of modifiers in the form of adjectives or prepositional phrases extends the length of a subject noun phrase, and thus delays the point at which the verb is reached. However, the same argument would clearly not apply in the case of an object noun phrase in a subject-verb-object (SVO) sentence. Weir, Bax, Chan, Field, Green, and Taylor (2012) feel that a more satisfying explanation relates to the burden upon parsing: the inclusion of modifiers increases the length and complexity of the string of words that a reader has to hold in the mind, while imposing a syntactic pattern upon it.

B1-level students in Task B use a slightly higher average number of modifiers per noun phrase than A2-level students.

Mean Number of Words Before the Main Verb

Weir, Bax, Chan, Field, Green, and Taylor (2012) maintain that the justification given in the Coh-Metrix specifications that "Sentences that have many words before the main verb are taxing on working memory" is not a convincing one; there, the authors refer to working memory as a very general notion and do not specify at all how it operates in this case. The best explanation would seem to be a syntactic one associated with parsing. Critical to the parsing of a clause is the verb, which not only provides a predicator for the event being described but also signals the likely syntactic structure of the whole sentence through its valency (Trueswell, Tanenhaus, & Kello, 1993). The presence of modifiers in the form of adjectives or prepositional phrases extends the length of a subject noun phrase, and thus delays the point at which the verb is reached. An alternative explanation relates to parsing. The words that occur before the verb are the first in a sentence to be analyzed, and the longer the subject noun phrase is, the greater the burden imposed at this early stage upon working memory.

McNamara, Crossley, and McCarthy (2010) found that the number of words before the main verb was one of the three most predictive indices of essay quality. They found that, of all the indicators of syntactic complexity, the mean number of words before the main verb displayed the largest effect size in the discriminant analysis. Crossley and McNamara (2011, p. 186) report that the number of modifiers in a noun phrase and the number of words before the main verb were the two indices which were "predictive of higher essay quality." In the case of both indices (ibid), "an increase in syntactic complexity equated to an increase in human ratings of essay quality."

B1-level students in Task B use a slightly higher average number of words before the main verb than A2-level students.

3.2.3 Cohesion and Coherence

Cohesion

For the purposes of this study, we adopt the Graesser, McNamara, Louwerse, and Cai (2004, p. 193) definition of cohesion as a property of a text that involves:

> …explicit features, words, phrases or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas and in connecting ideas to higher level global units (e.g., topics and themes)…

McNamara, Graesser, and Louwerse (in press) argue that:

> Cohesion arises from a variety of sources, including explicit argument overlap and causal relationships, and can operate between sentences, groups of sentences, paragraphs, and chapters…

These cohesive devices cue the reader on how to form a coherent representation. The coherence relations are constructed in the mind of the reader and depend on the skills and knowledge that the reader brings to the situation. In other words, coherence is a psychological construct, whereas cohesion is a textual construct.

While Alderson (2000) notes that an absence of cohesive devices does not seriously damage comprehension when the topic is relatively familiar to readers, it has been argued that explicit cohesive devices help in establishing textual coherence (Goldman & Rakestraw, 2000). There is a common assumption that the essays of better students will be more cohesive than those of lower-scoring students; however, little empirical evidence is available in the research literature (see a summary of evidence in Crossley & McNamara, 2010a) to support this view. In fact, there is growing evidence that in the case of overt cohesive ties this may not be the case, with more *advanced* writers at the C1 and C2 levels (Kennedy & Thorpe, 2007).

McNamara et al (2010, p. 73), using 26 validated indices of cohesion from Coh-Metrix, found:

> …no indication that higher-scored essays were more cohesive. There were no cohesion indices that showed differences between high- and low-proficiency essays.

Their findings are consonant with the earlier studies of Neuner (1987) and Kennedy and Thorpe (2007), which suggested the use of cohesive markers was not necessarily linked to the quality of writing for advanced-level students.

This is replicated to a certain extent in the TEAP pilot study, and in a number of cases there were no significant differences in the use of various cohesive devices between high- and low-scored essays or the difference was marginal and/or in the wrong direction (see Appendix B). However, there were a few notable exceptions to this (see below) in two areas in Task A regarding *argument overlap* and *stem overlap*, where the incidence for B1-level students was greater than that at A2. Additionally, in the case of the *causal cohesion* index, there are significant differences in both tasks A and B between A2- and B1-level students. In the pilot test, there was opportunity for the expression of *causality* in both tasks, and in both Task A and Task B the better, B1-level students were more cohesive in terms of this index than were A2-level students (see Table 15). As McNamara et al (2010, p. 76) noted, judgments of quality of writing may be task-dependent, and we would add that this is true for Coh-Metrix indices as well. Where causal connections are required by the task, as in the TEAP, the indices relating to these may figure prominently.

The differences in student samples between the TEAP study and McNamara et al's study (2010) with regard to these three cohesion indices may also explain the differences to be found in the TEAP data. In our study, we are dealing with quite low-level Japanese EFL students writing in a limited time frame, whereas McNamara et al (2010) were using native speakers writing untimed argumentative essays in a freshman

English program in the United States. It does seem to be the case that *higher*-level students employ less-overt markers of cohesion to promote coherence in the mind of the reader, relying instead on more sophisticated use of language. This finding is suggested by Kennedy and Thorpe both in their review of the literature on cohesion and coherence in student writing (2007, pp. 318-323) and in their own empirical work on IELTS. It is perhaps at the *lower* end of the ability spectrum (A2–B1 in our sample) that students employ overt cohesive devices more often and their presence or absence is more relevant for making grade distinctions. As Kennedy and Thorpe found (2007, p. 317), "...lower levels appear to need overt lexico-grammatical markers to structure their argument... Level 8 writers [IELTS band score] have other means at their disposal."

Cohesion indices that showed significant differences between bands A2 and B1 on Task A in three cases and additionally in Task B in the case of causal cohesion are shown in Table 15. (See Appendices B and C for descriptive statistics and box plots on all cohesion indices.)

Table 15: ***Cohesion Indices Showing Significant Differences Between Bands A2 and B1***

	Task	CEFR	Min	Max	Mean	Std. Dv.	Mann-Whitney Test
Cohm 8 Causal cohesion	Task A	A2 (40)	0	7	0.87	1.15	z=-2.4; p=0.018
		B1 (62)	0.2	5	1.10	0.96	
	Task B	A2 (47)	0.14	2	0.78	0.40	z=-2.4; p=0.02
		B1 (57)	0.2	3	0.97	0.50	
Cohm 16 Argument Overlap, adjacent, unweighted	Task A	A2 (40)	0	1	0.41	0.25	z=-2.3; p=0.019
		B1 (62)	0	1	0.53	0.28	
	Task B	A2 (47)	0.25	1	0.61	0.17	z=-0.4; p=0.73
		B1 (57)	0.23	1	0.59	0.16	
Cohm 17 Adjacent stem overlap	Task A	A2 (40)	0	1	0.38	0.26	z=-2; p=0.05
		B1 (62)	0	1	0.48	0.28	
	Task B	A2 (47)	0.19	1	0.56	0.17	z=-0.4; p=0.72
		B1 (57)	0.17	1	0.56	0.19	

Causal Cohesion

According to the Coh-Metrix website,

> Causal cohesion reflects the extent to which sentences are related by causal cohesion relations. Causal cohesion relations are appropriate when the text refers to events and actions that are related causally, as in the case of science texts with causal mechanisms and stories with an action plot. Causality is not relevant, for example, in texts that describe static scenes and texts that convey abstract logical arguments.
>
> Coh-Metrix needs to first estimate how much of the text refers to events and actions that may be part of causal content. This is accomplished by counting the number of main verbs that are causal, based on WordNet… The higher the incidence of causal verbs in a text, the more the text is assumed to convey causal content.

> Having *causal verbs* in a text does not insure that the reader can connect these events and actions with causal relations. …causal cohesion relations are signaled by *causal particles*… Some causal particles are conjunctions, transitional adverbs, and other forms of connectives, such as *since, so that, because,* and *consequently*. These particles are used to indicate some causal relationship between clauses that refer to events and actions. Other causal particles consist of a small number of verbs that explicitly assert there is a causal relationship between constituents, without specifying the nature of the causal content: *cause, enable, make*.
>
> **Causal cohesion: CAUSC (8)**
>
> This is a ratio of causal particles (P) to causal verbs (V). The denominator is incremented by the value of 1 to handle the rare case when there are 0 causal verbs in the text. Cohesion suffers when the text has many causal verbs (signifying events and actions) but few causal particles that signal how the events and actions are connected.

Lexical Cohesion
Crossley and McNamara (2011, p. 175) describe how "Argument overlap measures how often two sentences have nouns with common stems (including pronouns), while stem overlap measures how often a noun in one sentence shares a common stem with other word types in another sentence (not including pronouns)... Lexical overlap has been shown to aid in text comprehension…"

Adjacent stem overlap: According to Coh-Metrix,

> **Adjacent stem overlap: CREFS1u**
>
> This is the proportion of adjacent sentences that share one or more word stems.
>
> Example: The division of cells with a membrane-bound nucleus and organelles (eucaryotic cells) involves two distinct but overlapping stages, mitosis and cytokinesis. Mitosis occurs to replicate the cell's genetic material in the nucleus, whereas cytokinesis occurs to divide the gel-like liquid surrounding the cell's nucleus, called cytoplasm.
>
> In this example, the word *division* has a stem overlap with *divide*.

Argument overlap: When a noun, pronoun, or noun phrase in one sentence is a co-referent of a noun, pronoun, or noun phrase in another sentence.

The cognitive demands of storing information while reading are considerable. Weir, Bax, Chan, Field, Green, and Taylor (2012) describe how, in addition to (a) holding the surface language of the current sentence in the mind until it can be syntactically parsed and (b) carrying forward a discourse representation of the text so far, a reader also has to carry forward an awareness of what constitutes the current topic or focus (Sanford & Garrod, 1981). Items that have been mentioned in the immediately preceding text are said to be foregrounded and thus more easily matched to subsequent anaphors or incorporated into inferential processes. The value of the argument overlap measure is presumably that it indicates the extent to which the same entity is foregrounded in successive sentences, thus simplifying the process of identifying and carrying forward the topic focus.

Although the essays of B1-level candidates exhibit a higher degree of cohesion than do A2-level candidates in terms of the two lexical cohesion indices—Coh-Metrix 16 and 17 in Task A—this is not repeated in the higher-level Task B. Coh-Metrix Index 8 "causal cohesion," however, shows significant differences in both Tasks A and B. We perhaps need to provide some more specific guidance to raters concerning what cohesive indices are important in the TEAP tasks at the A2 and B1 levels and what specific features they should attend to when rating this category. This could help distinguish better between the levels. As we can see from Appendix E, cohesion appears as a strong predictor for overall performance on both tasks according to the data in our pilot study, so its place in our rating scales for A2- and B1-level students appears justified.

4. General Discussion and Recommendations

4.1 Conclusions

Rigorous and iterative development and trialing procedures produced a pilot test which demonstrated acceptable context and cognitive validity for use as an EAP writing test for students wishing to enter Japanese universities.

The study carried out into the scoring validity of the rating of the TEAP Writing Test indicated acceptable levels of intra- and inter-marker reliability and demonstrated that receiving institutions could depend on the consistency of the results obtained on the test. It should be noted that the final form of feedback to be used in operational versions of the test is still being finalized.

The study carried out on the contextual complexity parameters (lexical, grammatical, and cohesive) of scripts allocated to different bands on the TEAP rating scales indicated that there were significant differences between the scripts in adjacent band levels A2 and B1, particularly in respect of a number of indicators of lexical and syntactic complexity. These findings accord with the literature on writing that suggests more skilled writers are better able to access and use less familiar words as well as more complex syntax in writing essays; i.e., their use of language is more sophisticated.

This data provides complementary evidence that the rating scales are working as intended but also demonstrates that these Coh-Metrix indices cannot be used in place of marker-driven rating scales to place students accurately on proficiency levels.

4.2 Recommendations for Ongoing and Future Research

1. Ways might be researched of enhancing the cognitive validity of the TEAP tasks even further. There is a strong case for including a space on the test paper for use in planning the essay and proactively encouraging candidates to do this. We might encourage students to spend any surplus time on planning, monitoring, and editing, which constitute the vital washback elements of a direct writing task. It would be useful to carry advice on the advantages of planning, monitoring, and revision on the exam paper itself, as individual scores are likely to improve as a result. I would favor the criteria of assessment being reprinted on the paper to facilitate these critical cognitive processes and achieve beneficial washback on the teaching and learning that precedes them.

   Research might be carried out in the future to see if the provision of additional dedicated time for planning before students write, as well as afterwards for monitoring and revision, impacts positively on students' scores, as seems likely from the limited research done on this to date.

2. Further research is required as the TEAP is operationalized, and in particular the usefulness of each criterion should be iteratively reviewed. The marks awarded for lexis in Task A should be monitored. As most of the words used come from the text, consideration might be given in time to reducing the weighting of this criterion if students seem to benefit unduly from this in comparison with other criteria. It might also be useful to investigate whether two separate criteria are needed for cohesion and coherence in Task B or whether these are best amalgamated into one scale, as in Task A. We should provide more specific guidance to raters concerning what cohesive indices are important in the TEAP tasks at the A2 and B1 levels and what specific features they should attend to when rating this category. Our Coh-Metrix analysis suggests that in the examples of cohesion provided for raters in the current TEAP scales it would also be useful to make reference to causality and argument as well as to connectives and anaphoric reference.

3. Given the integrated reading-into-writing nature of Task B, it might be interesting to undertake some close linguistic analysis of high-scoring writing scripts to explore in greater depth precisely how skilled L2 writers generate a new text-level representation that successfully integrates and transforms content from multiple sources.

4. A further core component of any ongoing test-monitoring program concerns the issue of rater attitudes and behavior. If human raters are to be used (as opposed to an automatic scoring system), then careful attention will need to be paid to the ways in which raters are recruited, trained, standardized, and monitored within the wider quality-assurance system for the operational test. In addition to statistical monitoring of rater behavior in terms of raters' application of the assessment criteria and rating descriptors, it might be instructive to engage in a survey of rater attitudes to these elements, together with some (even small-scale) studies of rater behavior using qualitative methodologies, such as focus group or verbal protocol analysis.

5. The procedures employed in this pilot involved rating with four raters in order to facilitate in-depth analyses of the rating scales and rater behavior. The final operational rating plan will be decided based both on the results of this report and further analyses (such as generalizability analysis), but as with any testing program will also need to consider aspects of scoring validity in conjunction with all of the other major elements of test validity, including of course practicality and efficiency. These elements need to be evaluated together to determine how all of them contribute to an overall evaluation of validity or usefulness for the intended uses and interpretations of the test (Bachman & Palmer, 1990; Weir, 2005).

References

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). Specification for item development and classification within the CEF: The Dutch CEFR construct project. *Workshop on Research into and with the CEFR,* University of Amsterdam, February 2004.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study. *Studies in Language Testing 1.* Cambridge: UCLES / Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Barker, K. (1998). A trainable bracketer for noun modifiers. *Proceedings of the Twelfth Canadian Conference on Artificial Intelligence (LNAI 1418)* (pp. 196-210), Vancouver.

Bereiter, C., & Scardamalia, M. (1987a). *The psychology of written composition.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Bereiter, C., & Scardamalia, M. (1987b). Knowledge telling and knowledge transforming in written composition. S. Rosenberg (Ed.), *Advances in applied psycholinguistics, vol. 2: Reading, writing and language learning.* Cambridge: Cambridge University Press.

Berman, R. (1984). Syntactic components of the foreign language reading process. J. C. Alderson & A. Urquhart (Eds.), *Reading in a foreign language*, 139-159. London: Longman.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al. (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus (ETS TOEFL Monograph Series, MS-25)*. Princeton, NJ: Educational Testing Service.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20* (1), 89-110.

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students*. Princeton, NJ: Educational Testing Service.

Britt, M. A., & Sommers, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology, 25*(4), 313-339.

Brown, J. D. (1997). An EFL readability index*. University of Hawaii Working Papers in English as a Second Language, 15*(2), 85-119.

Carrell, P. L. (1982). Cohesion is not coherence**.** *TESOL Quarterly, 16*(4), 479-488.

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

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475-493.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91*(2), 15-30.

Crossley, S. A., & McNamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching, 41*(3), 409-429.

Crossley S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18,* 119-135.

Crossley, S. A., & McNamara, D. S. (2010a). Cohesion, coherence and expert evaluations of writing proficiency. R. Catrambone & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989), Cognitive Science Society, Austin, Texas.

Crossley, S. A., & McNamara, D. S. (2010b). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 18,* 1-21.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *Inter. J. Continuing Engineering Education and Lifelong Learning, 21,* 2/3, 170-191.

Cummins, J. (2008). BICS and CALP: Empirical and Theoretical Status of the Distinction. B. Street & N. H. Hornberger (Eds.), *Encyclopedia of language and education, vol. 2: Literacy* (2nd ed., pp. 71-83). New York: Springer Science + Business Media LLC.

Dunlea, J. (2010). Designing a research agenda to justify the uses and interpretations of the EIKEN tests. *Proceedings of the 12th Academic Forum on English Language Testing in Asia.* The Language Training and Testing Center, Taipei, Taiwan.

Dunlea, J., & Figueras, N. (2012). Replicating Results from a CEFR Test Comparison Project Across Continents. D. Tsagari & I. Csepes (Eds.), *Collaboration in language testing and assessment,* 31-45. New York: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175-196.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL2000 Reading Framework: A working paper*. TOEFL Monograph Series 17. Princeton, NJ: ETS.

Field, J. (2004). *Psycholinguistics. The key concepts*. London: Routledge.

Fortus, R., Coriat, R., & Fund, S. (1998). Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test. A. In Kunnan, A. (Ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Research Colloquium,* 61-87. Long Beach, Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.

Freedle, R. (1997). The relevance of multiple-choice reading test data in studying expository passage comprehension: The saga of a 15-year effort towards an experimental/correlational merger. *Discourse Processes, 23,* 399-440.

Freedle, R., & Kostin, I. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. TOEFL Research Reports, No. RR-93-44.* Princeton, NJ: Educational Testing Service.

Gernsbacher, M.A. (1990). *Language comprehension as structure building.* Mahwah, NJ: Erlbaum.

Gervasi, V., & Ambriola, V. (2002). Quantitative assessment of textual complexity. L. Merlini Barbesi (Ed.), *Complexity in language and text,* 197-228. Pisa: PLUS-University of Pisa.

Goldman, S. R. (1997). Learning from text: Reflections on the past and suggestions for the future. *Discourse Processes, 23,* 357-398.

Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Ferris & D. M. Bloome (Eds.), *Use of intertextuality in classroom and educational research,* 317-351. Greenwich, CT: Information Age Press.

Goldman, S., & Rakestraw, J. (2000). Structural aspects of constructing meaning from text. M. Kamil, P. Rosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research,* 311-335. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Graesser, A. C., Karnavat, A. B., Daniel, F. K., Cooper, E., Whitten, S. N., & Louwerse, M. (2001). A computer tool to improve questionnaire design. *Statistical Policy Working Paper 33*, *Federal Committee on Statistical Methodology,* 36-48. Washington, DC: Bureau of Labor Statistics.

Graesser, A. C., McNamara, D.S., & Kulikowich J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher, Vol. 40,* No. 5, 223-234.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36,* 193-202.

Green, A. (2011). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge: Cambridge University Press.

Green, A., Ünaldi, A., & Weir, C. (2008). The cognitive processes of second language academic readers. *LLAS Pedagogic Research Fund Project report*. Retrieved December 5, 2005 from http://www.llas.ac.uk/projects

Green, A., Ünaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts for testing reading for academic purposes. *Language Testing, 27*(3), 1-21.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). A study of writing tasks assigned in academic degree program. *TOEFL Research Reports*, *RR-95-44*. Princeton: Educational Testing Service.

Hartmann, D. K. (1995). Eight readers reading: The intertextual links of proficient readers reading multiple passages*. Reading Research Quarterly, 30,* 520-561.

Horowitz, D. M. (1986a). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20,* 445-460.

Horowitz, D. M. (1986b). Essay examination prompts and the teaching of academic teaching. *English for Specific Purposes, 5,* 107-120.

Hyland, K. (2002). *Teaching and researching writing, applied linguistics in action series*. London, UK: Longman.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 4,* 329-354.

Kennedy, C., & Thorp, D. (2007). A corpus-based investigation of linguistic responses to an IELTS academic writing task. L. Taylor (Ed.), *IELTS collected papers; research in speaking and writing assessment,* 316-377*.* Cambridge: Cambridge University Press.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading. Studies in Language Testing, 29.* Cambridge: Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Lacroix, N. (1999). Macrostructure construction and organisation in the processing of multiple text passages. *Instructional Science, 27,* 221-233.

Lewis, L. R., Vasishth, S., & van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science, October*, 447-454.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Longman.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

McNamara, D. S., Graesser, A. C, & Louwerse, M. M. (in press). Sources of text difficulty: Across the ages and genres. J. P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. Lanham, MD: R&L Education.

Masi, S. (2002). The literature on complexity. L. Merlini Barbesi (Ed.), *Complexity in language and text,* 197-228. Pisa: PLUS-University of Pisa.

MEXT (2002). *Japanese government policies in education, culture, sports, science, and technology*. Retrieved from http://www.mext.go.jp/b_menu/hakusho/html/hpac200201/hpac200201_2_015.html

MEXT (2003). *Action plan for cultivating Japanese with English abilities*. Retrieved from http://warp.ndl.go.jp/info:ndljp/pid/286794/www.mext.go.jp/b_menu/houdou/15/03/03033101/001.pdf

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes*, *4*, 43-66.

Neuner, J. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English, 21*(1), 92-105.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

O'Sullivan, B. (2002). Investigating variability in a test of second language writing ability. *Research Notes, 7,* 14-17.

Perfetti, C. A. (1997). Sentences, individual differences and multiple texts: Three issues in text comprehension. *Discourse Processes, 23,* 337-355.

Perfetti, C. A., Rouet, J-F., & Britt, M. A. (1999). Toward a theory of document representation. H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading*, 99-122. London: Lawrence Erlbaum Associates Inc.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13,* 111-129.

Plakans, L. (2009). Discourse synthesis in integrated second language writing asessment. *Language Testing, 26,* 561-585.

Plakans, L. (2010). Independent vs. intergrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185-194.

Rastle, K. (2007). Visual word recognition. M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics,* 71-87. Oxford: Oxford University Press.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.

Roscoe R. D., Crossley S., Weston, J., & McNamara, D. (2011). *Automated assessment of paragraph quality: Introduction, body, and conclusion paragraphs, Florida Artificial Intelligence Research Society Conference, Twenty-Fourth International FLAIRS Conference.*

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations of comprehension beyond the sentence.* Chichester: John Wiley.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. S. Rosenberg (Ed.), *Advances in applied psycholinguistics, vol. 2: Reading, writing and language learning.* Cambridge: Cambridge University Press.

Segev-Miller, R. (2007). Cognitive processes in discourse synthesis: The case of intertextual processing strategies. G. Rijlaarsdam, M. Torrance, L. Van Waes, & D. Galbraith (Eds.), *Writing and cognition: Research and applications*, 231-250. Amsterdam: Elsevier.

Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing. Studies in Language Testing, 26*. Cambridge: Cambridge University Press and Cambridge ESOL.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Spivey, N. N. (1991). Transforming texts: Constructive processes in reading and writing. *Written Communication*, *7*(2), 256-287.

Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly 24*(1), 7-26.

Stahl, S. A., Hynd, C. R., Britton, B. K., McNish, M. M., & Bosquet, D. (1996). What happens when students read multiple source documents in history? *Reading Research Quarterly 31*(4), 430-456.

Stromso, H. I., & Braten, I. (2002). Norwegian law students' use of multiple sources while reading expository texts. *Reading Research Quarterly 37*(2), 208-227.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*(3), 528-553.

Van Dijk, T. A., & Kintsch, W. (1983). The notion of macrostructure. T. A. van Dijk & W. Kintsch (Eds.), *Strategies of discourse comprehension*, 189-223. New York: Academic Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263-287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1983). The Associated Examining Board's Test of English for Academic Purposes: An exercise in content validation. A. Hughes & D. Porter (Eds.), *Current developments in language testing*, 147-153. London: Academic Press.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach.* Basingstoke: Palgrave Macmillan.

Weir, C. J., Bax, S., Chan, S., Field, J., Green, A., & Taylor, L. (2012). *Report to Cambridge ESOL UK on the contextual parameters of CAE examinations.*

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawai'i Press.

Wright, B., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370. Retrieved March 27, 2012 from http://www.rasch.org

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.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing*, *25*(4), 521-551.

Appendices

Appendix A: Cross-tabulations of Relative Candidate Performance on Tasks A and B

		Task A Overall			
		Below A2	A2	B1	B2
		Count	Count	Count	Count
Task B Overall	Below A2	7	0	0	0
	A2	1	37	9	0
	B1	0	3	53	1
	B2	0	0	0	1

		Task A Ideas			
		Below A2	A2	B1	B2
		Count	Count	Count	Count
Task B Ideas	Below A2	7	0	0	0
	A2	1	27	7	0
	B1	0	19	48	0
	B2	0	1	2	0

			Task A Coherence and Cohesion			
			Below A2	A2	B1	B2
Task B			Count	Count	Count	Count
Below A2 Cohesion	Coherence	Below A2	4	0	0	0
		A2	0	0	0	0
		B1	0	0	0	0
		B2	0	0	0	0
A2 Cohesion	Coherence	Below A2	2	0	0	0
		A2	2	14	11	0
		B1	0	4	3	1
		B2	0	0	0	0
B1 Cohesion	Coherence	Below A2	0	0	0	0
		A2	0	9	9	0
		B1	0	6	33	4
		B2	0	0	4	1
B2 Cohesion	Coherence	Below A2	0	0	0	0
		A2	0	0	0	0
		B1	0	0	2	0
		B2	0	0	2	1

		Task A Lexical			
		Below A2	A2	B1	B2
		Count	Count	Count	Count
Task B Lexical	Below A2	7	0	0	0
	A2	0	27	14	0
	B1	0	5	54	3
	B2	0	0	1	1

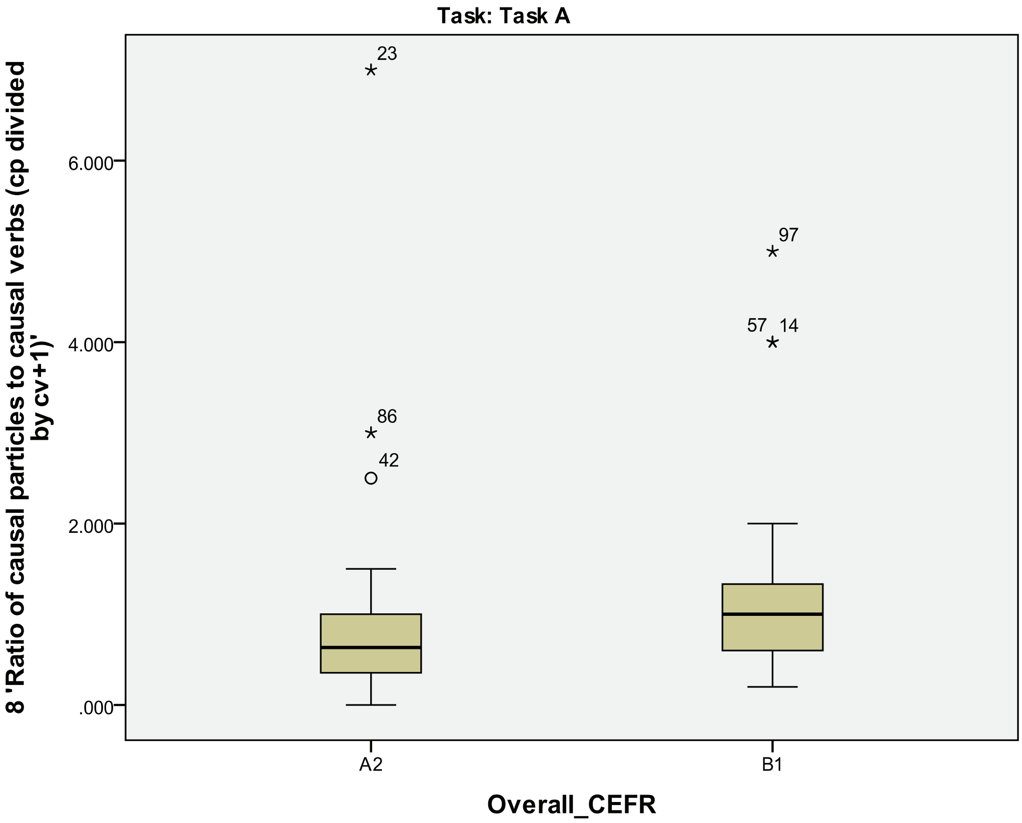
		Task A Grammatical			
		Below A2	A2	B1	B2
		Count	Count	Count	Count
Task B Grammatical	Below A2	6	0	0	0
	A2	1	29	16	2
	B1	0	13	39	2
	B2	0	0	3	1

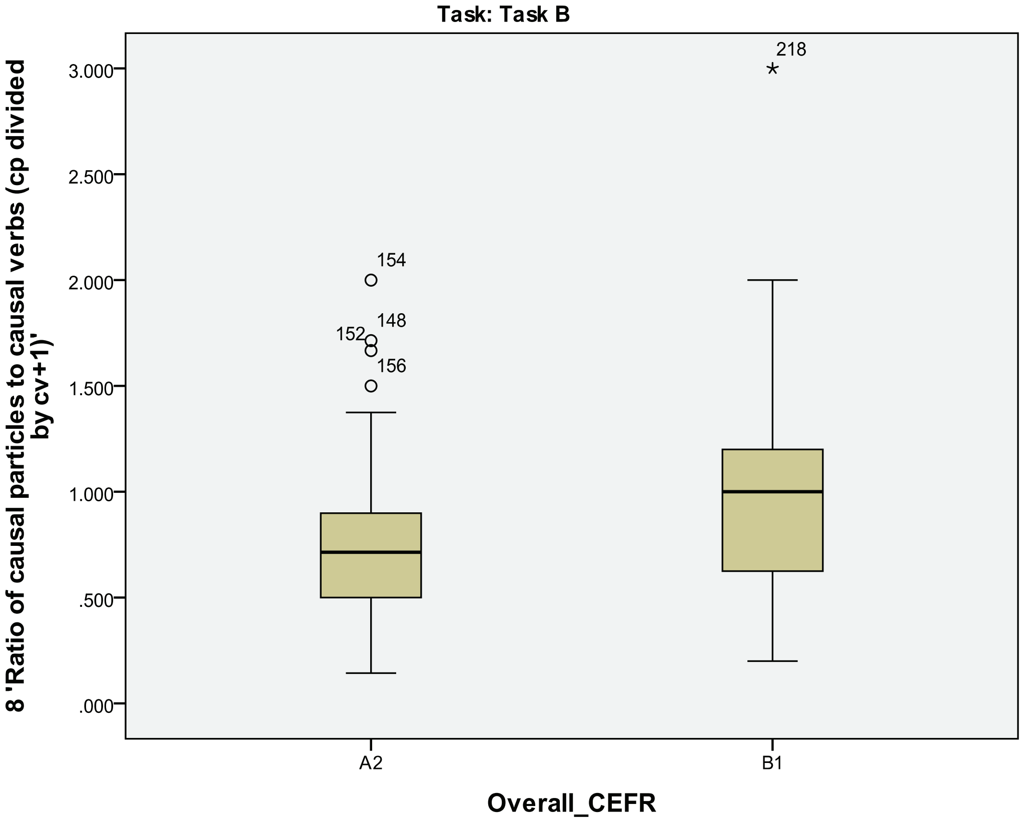
Appendix B: Coh-Metrix Indices Where There Was a Significant Difference in Indices Between A2 and B1 Candidates

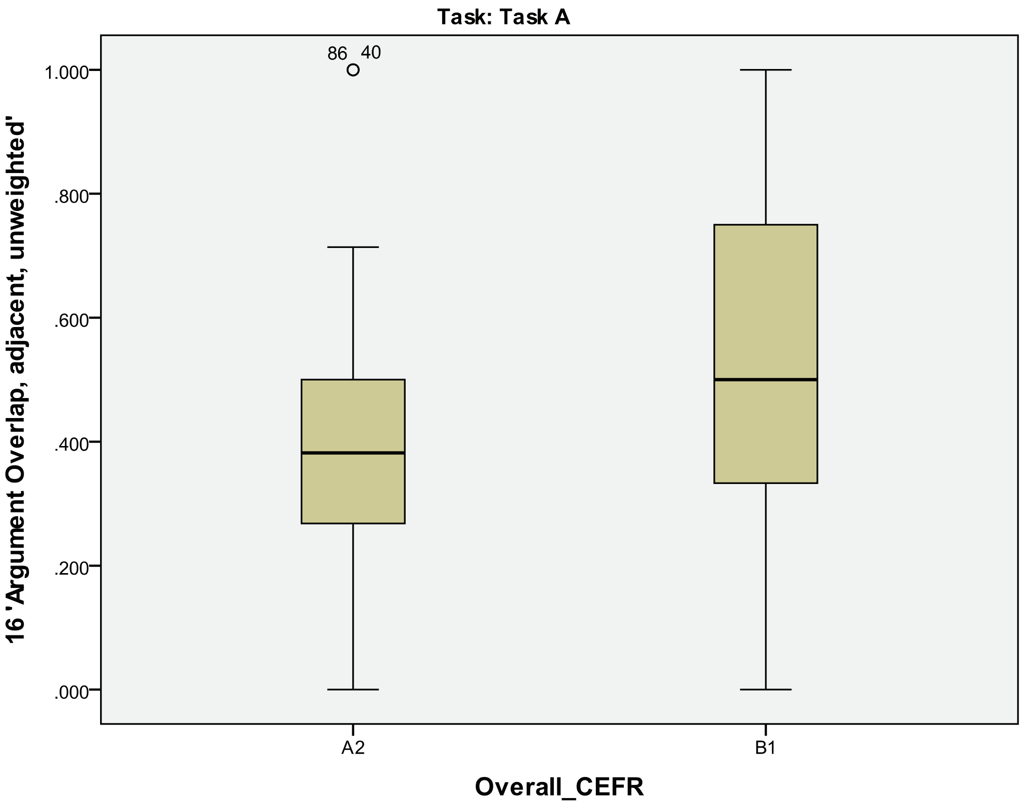
	Coh-Metrix Indices	Groups	Overall Task Grade	Fair A								
	Overall CEFR Level		Mann-Whitney Test		A2 (N=40)				B1 (N=62)			
	TASK A		z	Asymp. Sig. (2- t ailed)	Minimum	Maximum	Mean	Std. Deviation	Minimum	Maximum	Mean	Std. Deviation
7	Causal content	cohesion	-0.8	0.439	28.571	131.148	79.968	25.41225	15.152	157.303	75.797	27.5038
8	Causal cohesion	cohesion	-2.4	0.018	0	7	0.8748	1.145074	0.2	5	1.1028	0.95796
12	Neg. additive connectives	syntactic	-1	0.305	0	42.254	19.806	10.70186	0	43.478	17.578	12.6134
16	Adjacent argument overlap	cohesion	-2.3	0.019	0	1	0.4085	0.25152	0	1	0.53	0.28162
17	Adjacent stem overlap	cohesion	-2	0.05	0	1	0.3799	0.257219	0	1	0.4775	0.27766
19	Argument overlap	cohesion	-2.9	0.004	0	1	0.4084	0.241716	0	1	0.5383	0.24119
20	Stem overlap	cohesion	-3.1	0.002	0	1	0.3638	0.248028	0	1	0.4844	0.24879
25	Negations	syntactic	-0.8	0.419	0	58.824	17.107	12.41464	0	39.474	14.906	10.6912
26	Logic operators	cohesion	-0.2	0.867	10.417	92.437	52.284	20.26157	12.195	79.646	50.995	18.0937
35	No. of words	lexical	-0.7	0.457	35	154	82.35	24.48815	51	135	83.984	18.5945
37	Words per sentence	syntactic	-2.3	0.022	7.333	32	13.94	3.80	9.714	32	14.573	3.83185
38	Syllables per word	lexical	-0	0.992	1.214	1.506	1.3312	0.070941	1.098	1.614	1.3274	0.07591
39	Flesch Reading Ease	lexical	-0.9	0.377	34.095	93.726	79.777	9.800267	55.827	100	79.728	6.71657
40	Flesch-Kincaid	lexical	-1.6	0.099	2.65	12	5.3495	1.835164	2.339	12	5.7507	1.54532
41	Modifiers per NP	syntactic	-1.6	0.121	0.231	1	0.567	0.170003	0.167	1.04	0.5199	0.14814
42	Higher-level constituents	syntactic	-2.3	0.02	0.662	0.897	0.7922	0.046609	0.711	0.902	0.8107	0.03992
43	Words before main verb	syntactic	-1.4	0.15	1.2	6.25	3.6124	1.354462	1.833	6	3.2159	1.008
44	Type-token ratio	lexical	-2.5	0.012	0.667	0.953	0.8182	0.068321	0.578	0.938	0.7825	0.06796
46	Log freq. content words	lexical	-0.9	0.369	2.17	2.84	2.5341	0.130335	2.305	2.964	2.5643	0.13548
47	Min. raw freq. content words	lexical	-0.7	0.472	13.333	192.125	66.586	44.99825	13.8	244	60.241	47.0764
48	Log min. freq. content words	lexical	-0.8	0.425	1.033	1.844	1.4651	0.215099	1.088	1.912	1.4424	0.18755
51	Neg. logical connectives	syntactic	-0.8	0.4	0	42.254	19.806	10.70186	0	43.478	18.112	12.6882

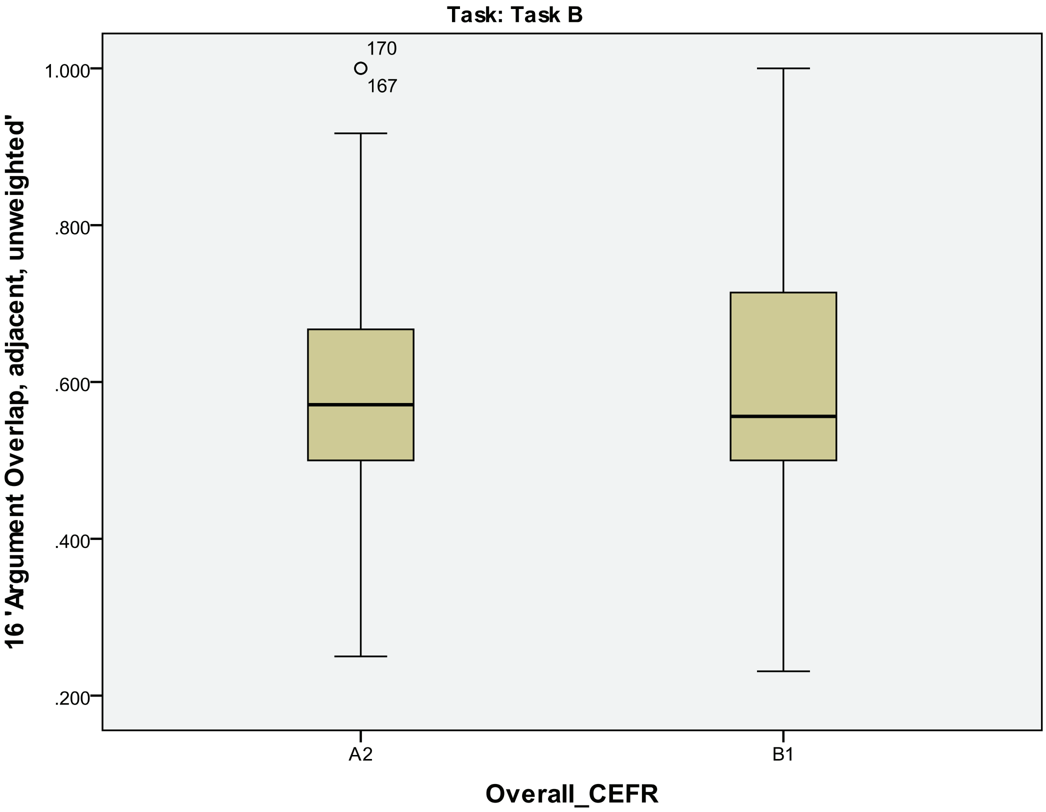
	Coh-Metrix Indices	Groups	Overall	Fair B								
	Overall CEFR Level		Mann-Whitney Tes		A2 (N=47)				B1 (N=57)			
	TASK B		z	Asymp. Sig. (2-tailed)	Minimum	Maximum	Mean	Std. Deviation	Minimum	Maximum	Mean	Std. Deviation
7	Causal content	cohesion	-2.7	0.01	42.254	137.615	79.75649	21.09637	32.68	101.604	67.79225	15.63572
8	Causal cohesion	cohesion	-2.4	0.02	0.143	2	0.772	0.395781	0.2	3	0.96912	0.503669
12	Neg. additive connectives	syntactic	-2.7	0.01	0	28.708	7.31806	6.89688	0	15.306	3.93347	4.390642
16	Adjacent argument overlap	cohesion	-0.4	0.73	0.25	1	0.6063	0.170743	0.231	1	0.58889	0.159346
17	Adjacent stem overlap	cohesion	-0.4	0.72	0.188	1	0.56421	0.166244	0.167	1	0.55558	0.194782
19	Argument overlap	cohesion	-0.4	0.69	0.304	1	0.53621	0.155471	0.248	1	0.52153	0.153283
20	Stem overlap	cohesion	-0.5	0.64	0.235	1	0.4907	0.158052	0.2	1	0.50832	0.16921
25	Negations	syntactic	-3.2	0	0	44.444	12.66847	10.03458	0	25.51	7.19307	7.694361
26	Logic operators	cohesion	-2.8	0.01	24.038	100	47.32613	16.71765	4.505	76.531	37.89228	15.72475
35	No. of words	lexical	-3.1	0	41	253	171.9575	46.34369	130	290	200.6316	33.60359
37	Words per sentence	syntactic	-3.7	0	8.636	21.286	13.09111	2.334421	10.385	25.714	15.16872	3.042582
38	Syllables per word	lexical	-2.8	0.01	1.36	1.634	1.46826	0.058648	1.371	1.644	1.49981	0.055723
39	Flesch Reading Ease	lexical	-4.4	0	53.466	79.488	69.33321	5.218228	54.22	74.747	64.55511	5.052494
40	Flesch-Kincaid	lexical	-4.6	0	4.58	9.506	6.84089	1.070985	5.698	12	8.0234	1.226809
41	Modifiers per NP	syntactic	-2.2	0.03	0.333	1	0.61002	0.160114	0.322	0.935	0.66079	0.113698
42	Higher-level constituents	syntactic	-2	0.04	0.7	0.851	0.78687	0.037509	0.715	0.873	0.77444	0.028295
43	Words before main verb	syntactic	-2.2	0.03	1.364	7.2	3.26245	1.089354	1.647	5.429	3.65709	0.931758
44	Type-token ratio	lexical	-0.1	0.91	0.481	0.831	0.67791	0.079456	0.536	0.797	0.67902	0.053913
46	Log freq. content words	lexical	-3.8	0	2.287	2.807	2.58113	0.095141	2.388	2.667	2.51584	0.068402
47	Min. raw freq. content words	lexical	-3.1	0	19.714	309.368	87.19006	60.70589	18.636	2475.125	102.152	321.5398
48	Log min. freq. content words	lexical	-4	0	1.097	2.085	1.55989	0.204514	1.003	1.685	1.40125	0.167105
51	Neg. logical connectives	syntactic	-2.5	0.01	0	28.708	7.53649	6.916124	0	15.306	4.36933	4.335334

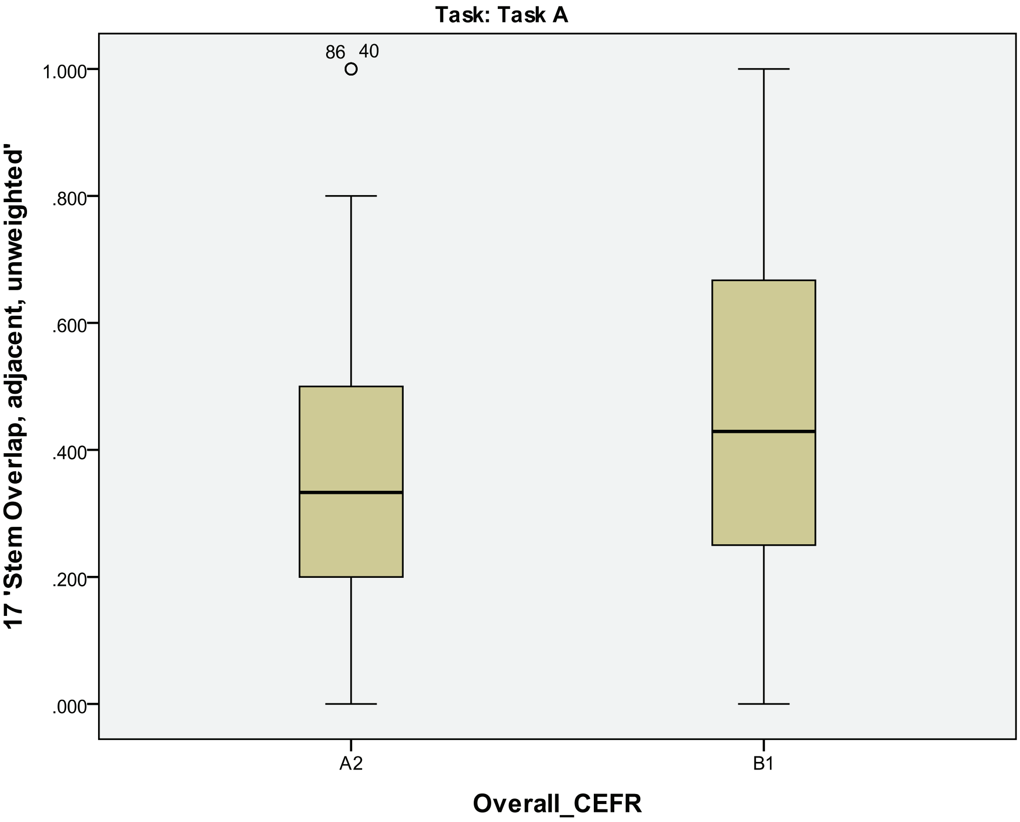
Appendix C: Box Plots of Complexity Indices Where There Was a Significant and Meaningful Difference Between A2 and B1 Performances

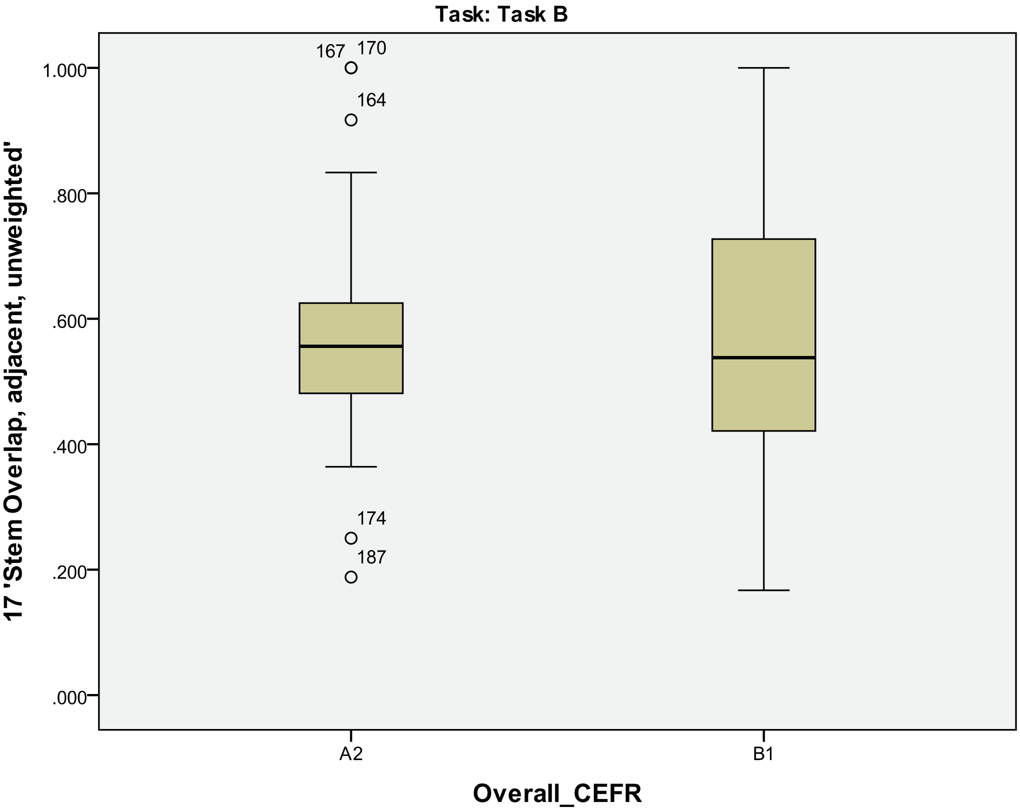


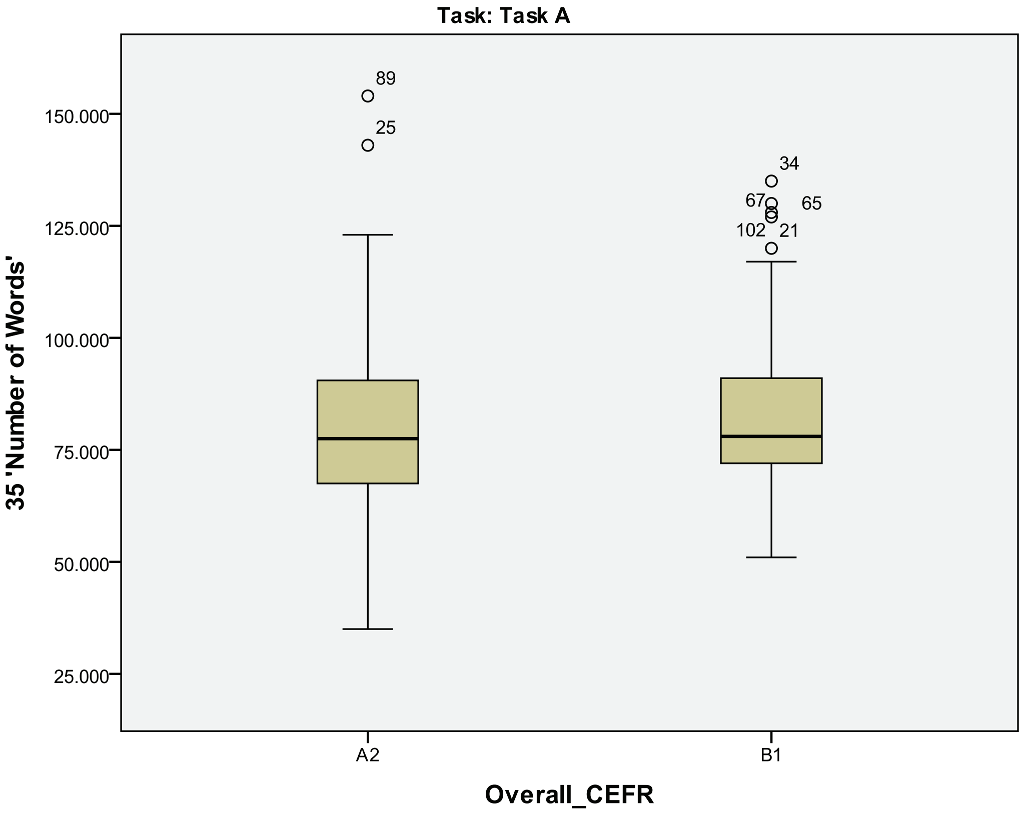


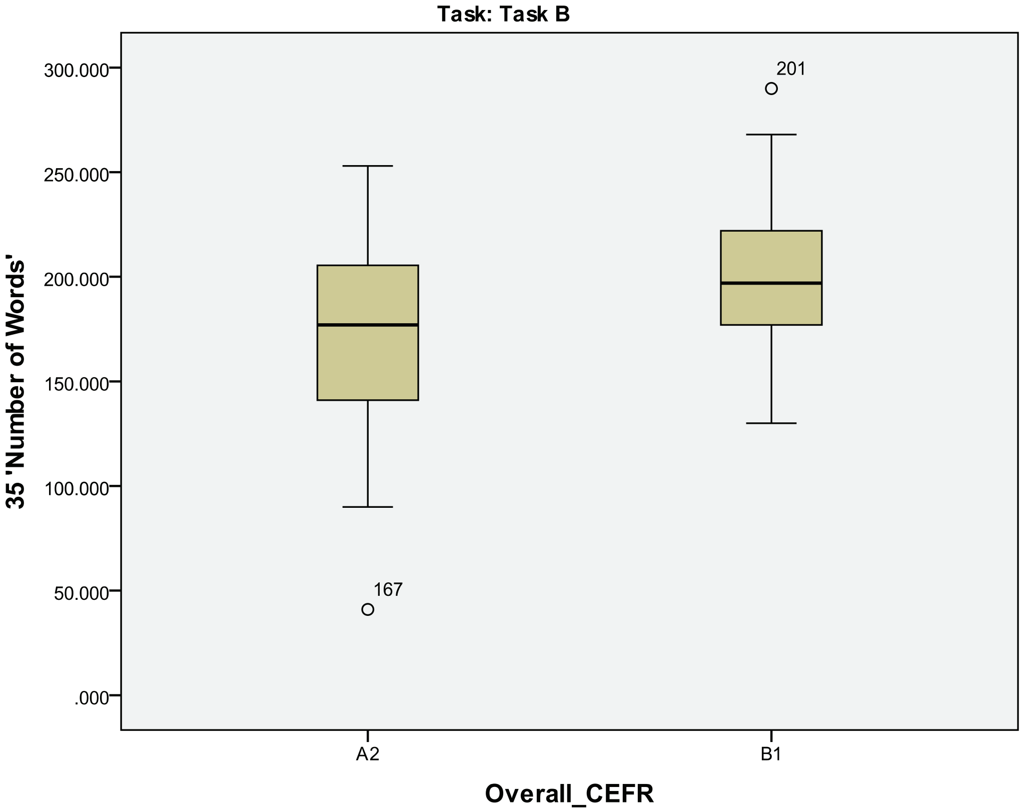

Task: Task A
86
40
1.000
.800
.600
.400
.200
.000
16 'Argument Overlap, adjacent, unweighted'
A2
B1
Overall_CEFR

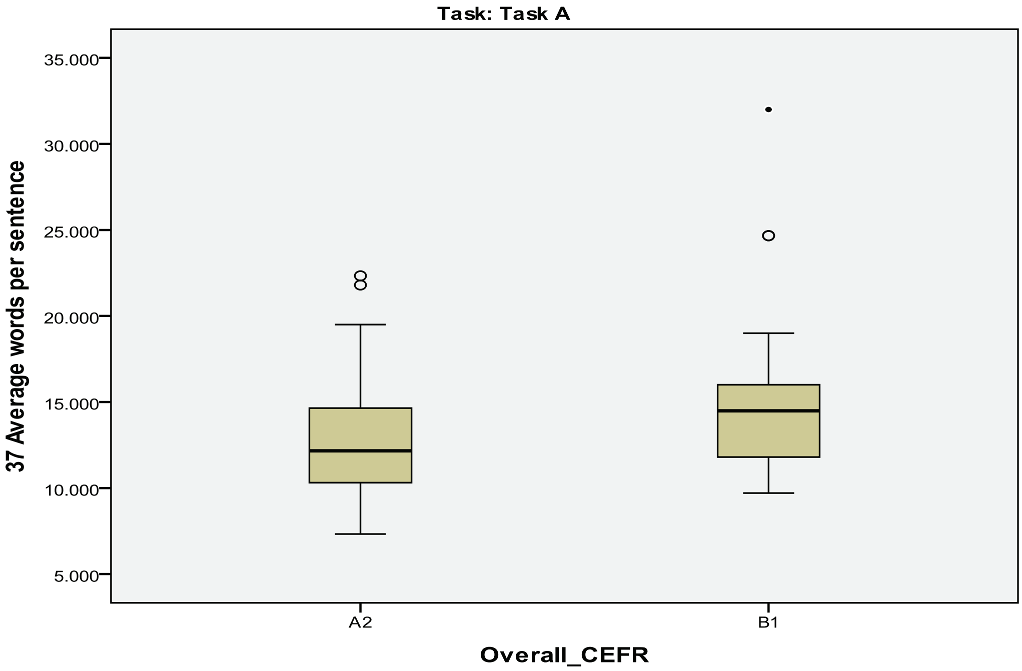

Task: Task B
170
167
1.000
.800
.600
.400
.200
16 'Argument Overlap, adjacent, unweighted'
A2
B1
Overall_CEFR

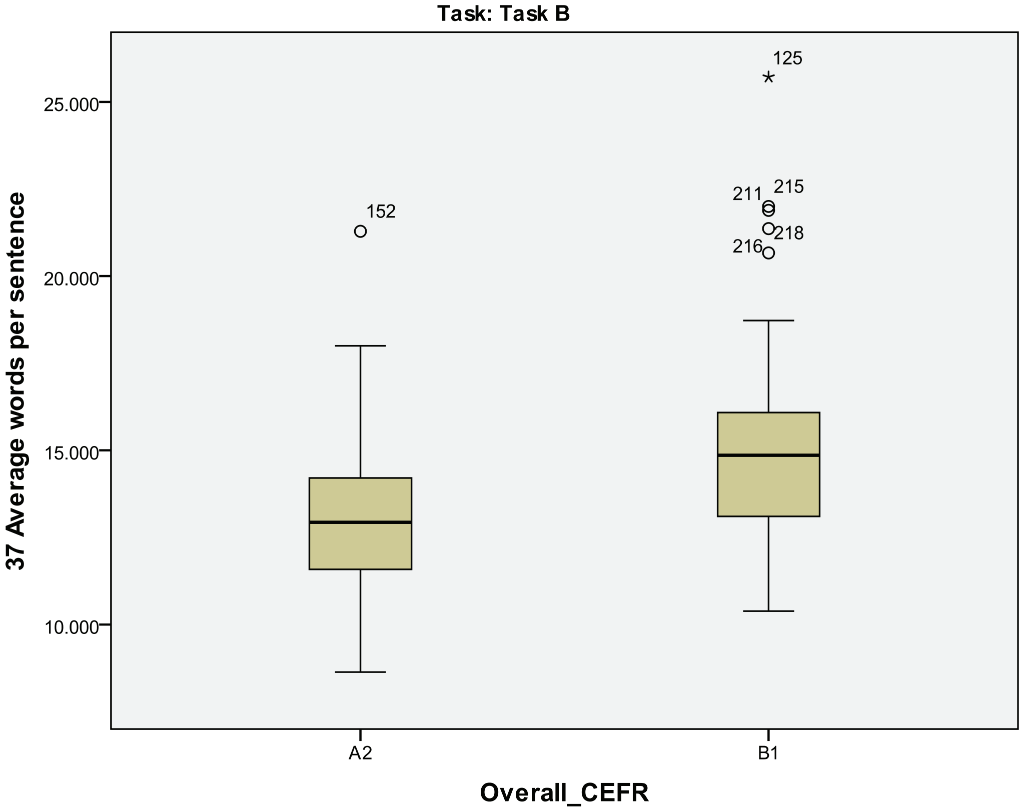

Task: Task A
17 'Stem Overlap, adjacent, unweighted'
1.000
.800
.600
.400
.200
.000
86
40
A2
B1
Overall_CEFR

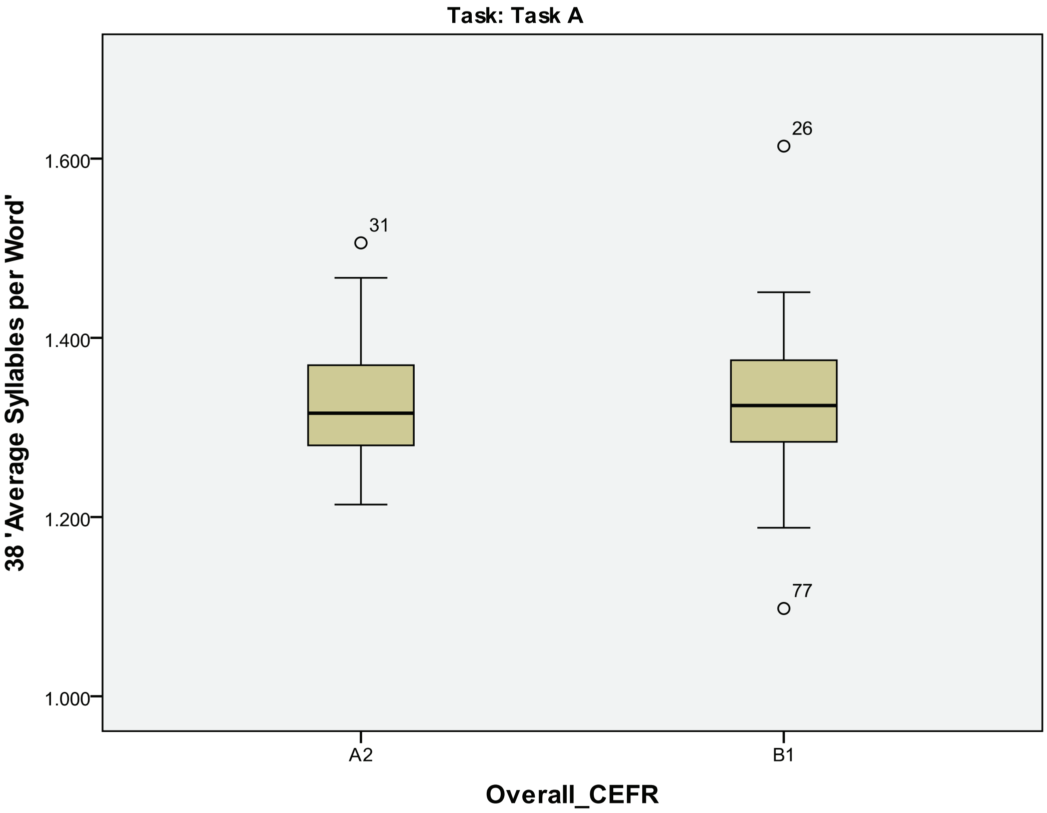

Task: Task B
17 'Stem Overlap, adjacent, unweighted'
1.000
.800
.600
.400
.200
.000
167
170
164
174
187
A2
B1
Overall_CEFR

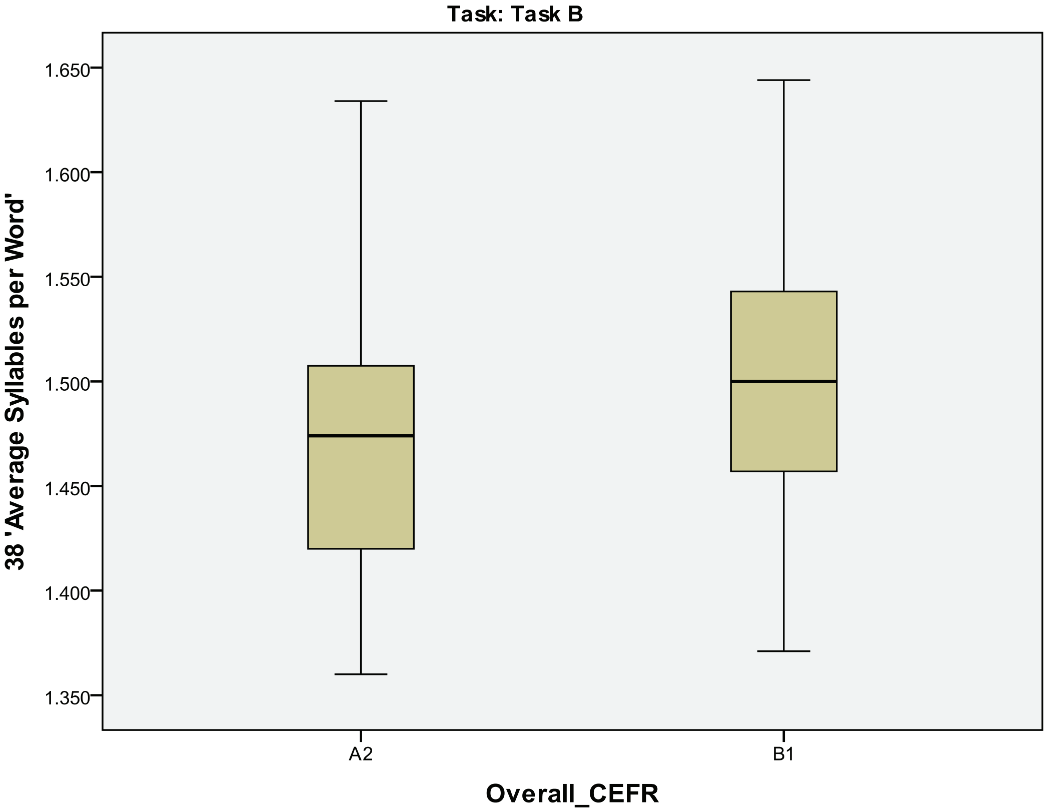

Task: Task A
35 'Number of Words'
150.000
125.000
100.000
75.000
50.000
25.000
89
25
34
67
65
102
21
A2
B1
Overall_CEFR

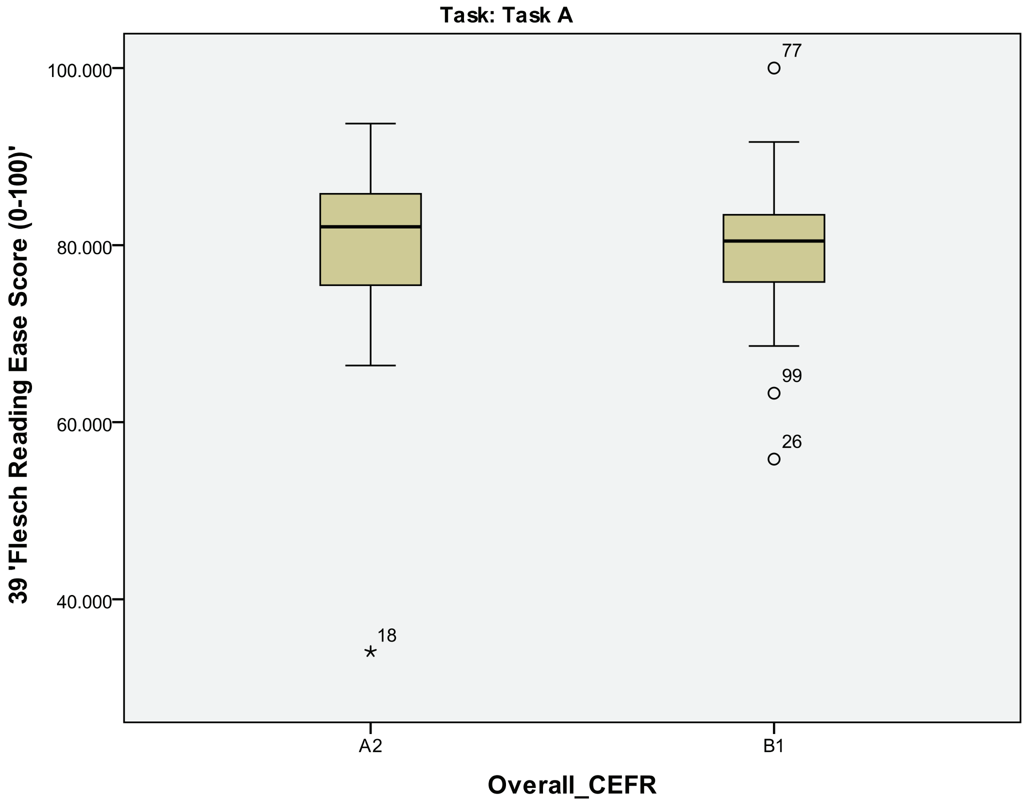

Task: Task B
35 'Number of Words'
300.000
250.000
200.000
150.000
100.000
50.000
.000
201
167
A2
B1
Overall_CEFR

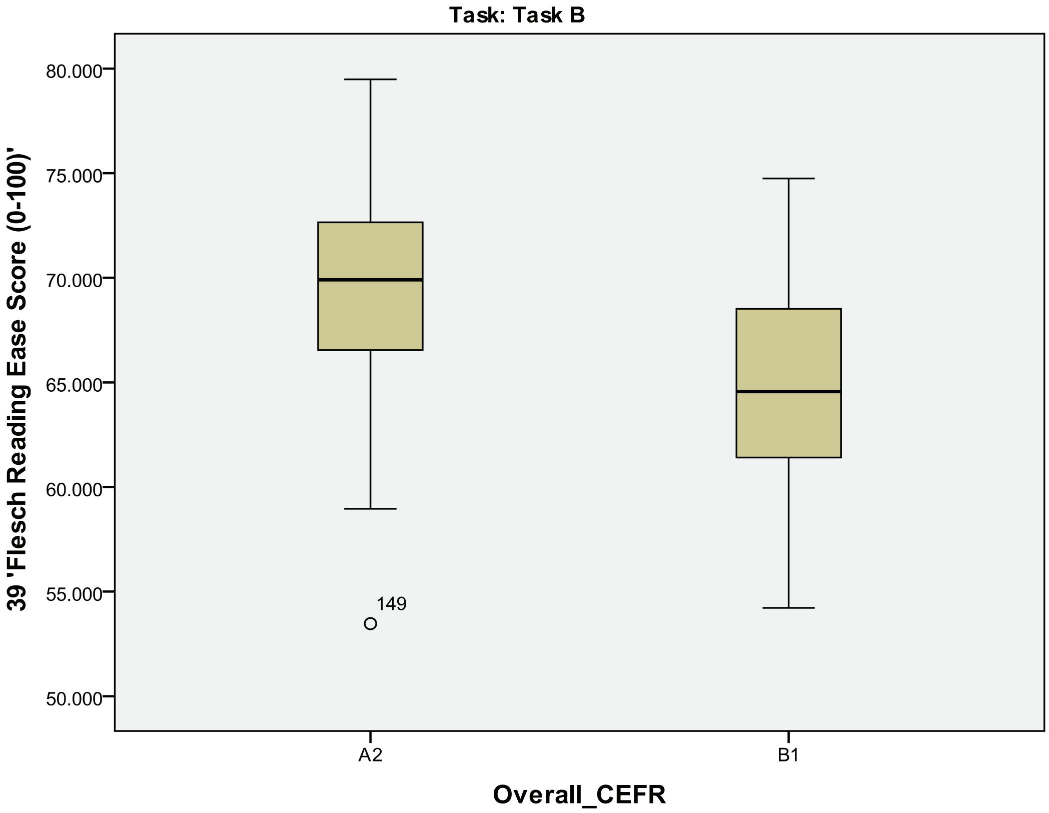

Task: Task A
37 Average words per sentence
35.000
30.000
25.000
20.000
15.000
10.000
5.000
A2
B1
Overall_CEFR

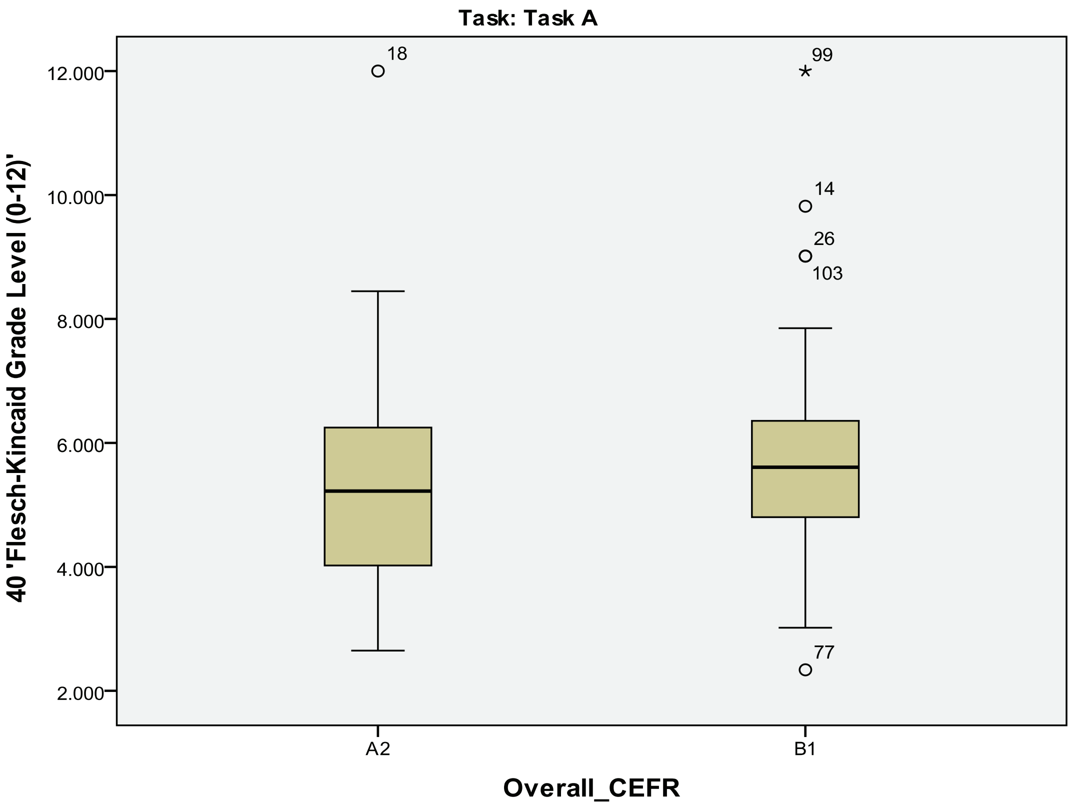

Task: Task B
37 Average words per sentence
25.000
20.000
15.000
10.000
125
152
211
215
218
216
A2
B1
Overall_CEFR

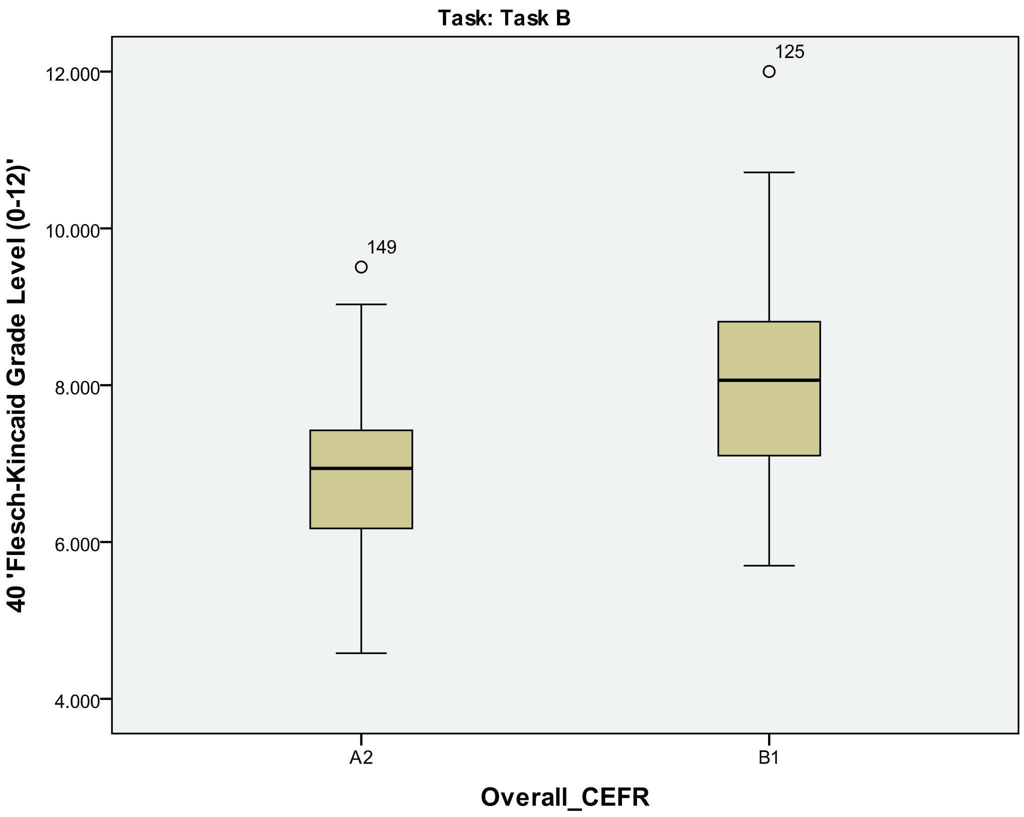

Task: Task A
38 'Average Syllables per Word'
1.600
1.400
1.200
1.000
31
26
77
A2
B1
Overall_CEFR

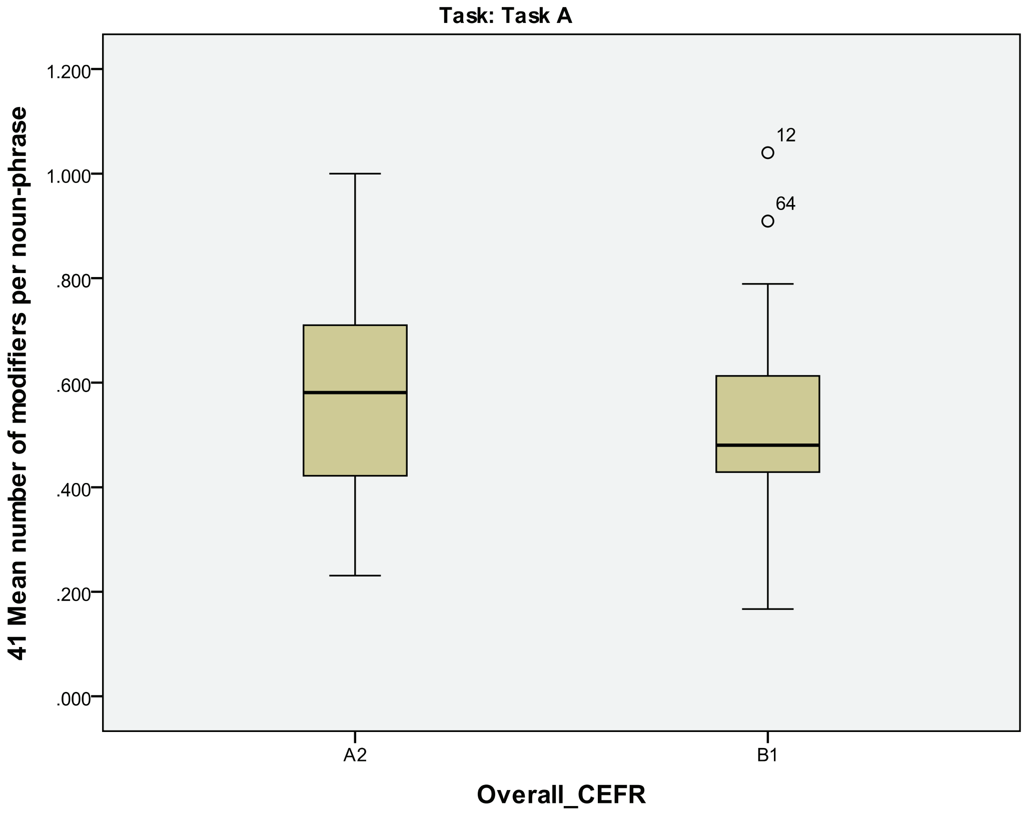

Task: Task B
38 'Average Syllables per Word'
1.650
1.600
1.550
1.500
1.450
1.400
1.350
A2
B1
Overall_CEFR

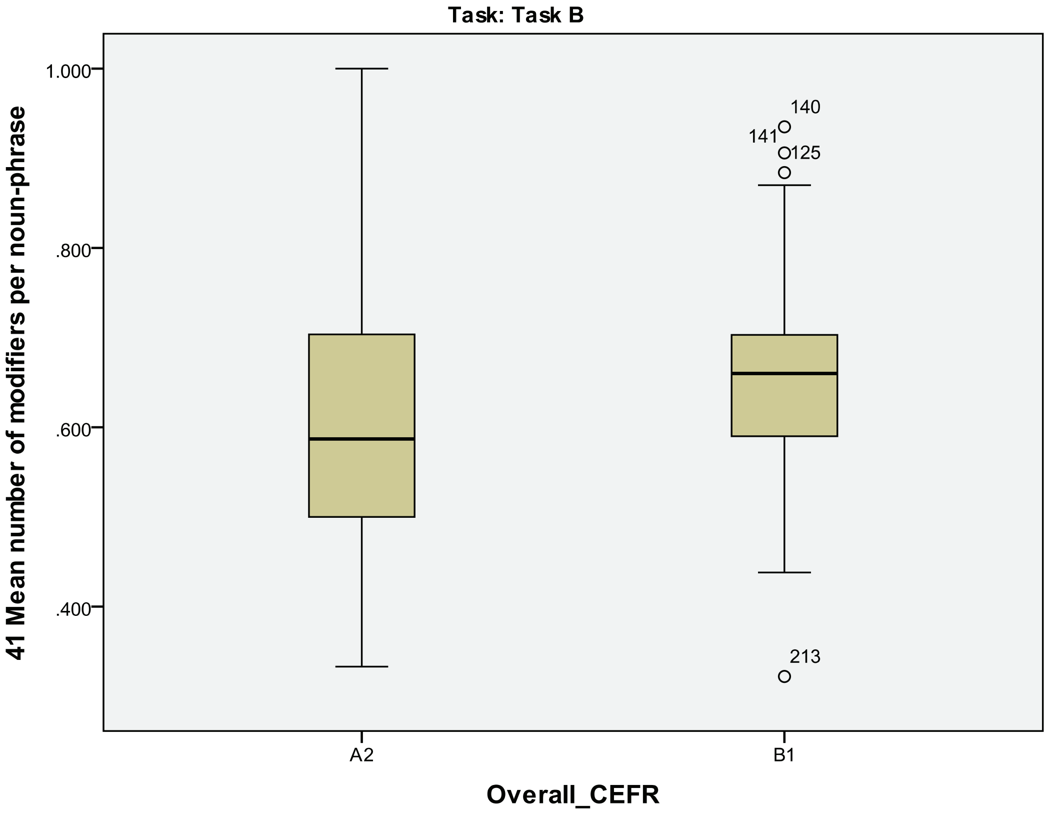

Task: Task A
39 'Flesch Reading Ease Score (0-100)'
100.000
80.000
60.000
40.000
77
99
26
18
A2
B1
Overall_CEFR

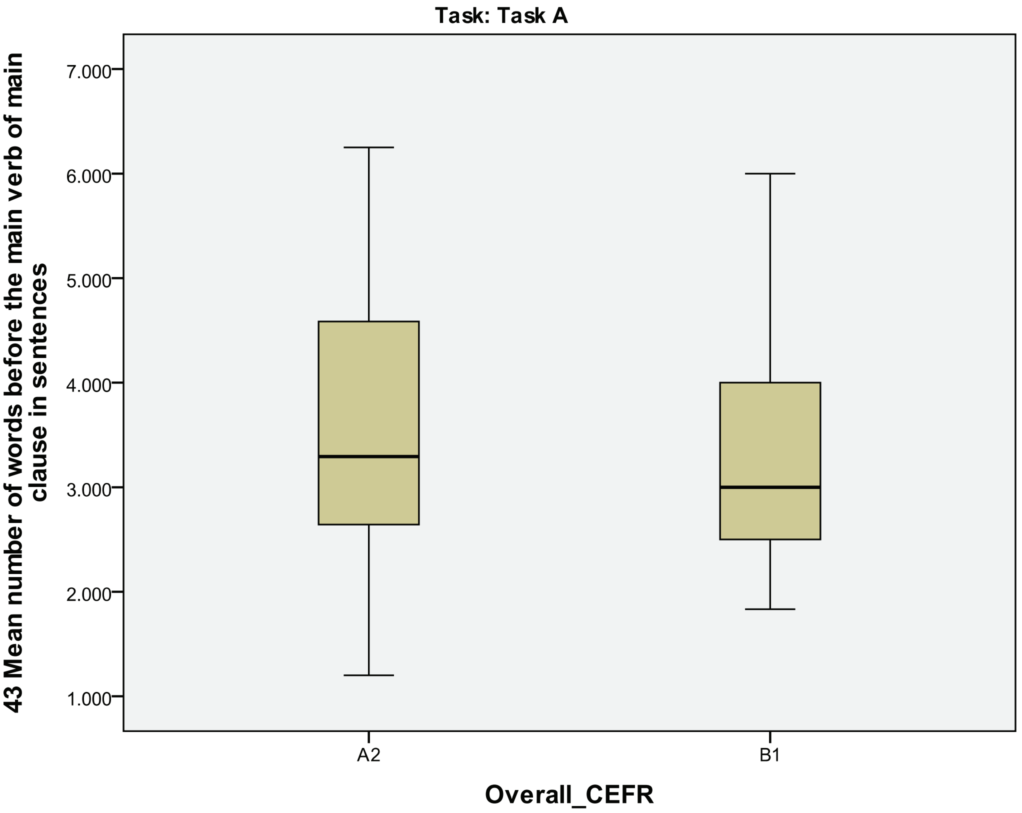

Task: Task B
39 'Flesch Reading Ease Score (0-100)'
80.000
75.000
70.000
65.000
60.000
55.000
50.000
149
A2
B1
Overall_CEFR

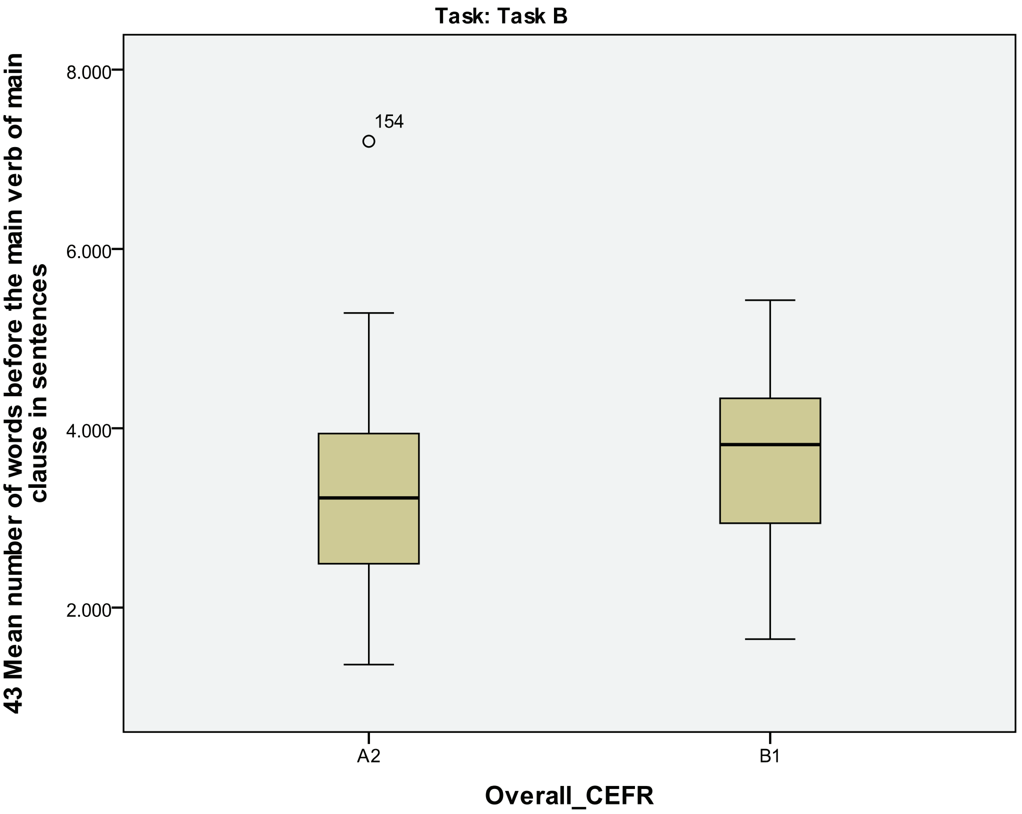

Task: Task A
40 'Flesch-Kincaid Grade Level (0-12)'
12.000
10.000
8.000
6.000
4.000
2.000
18
99
14
26
103
77
A2
B1
Overall_CEFR

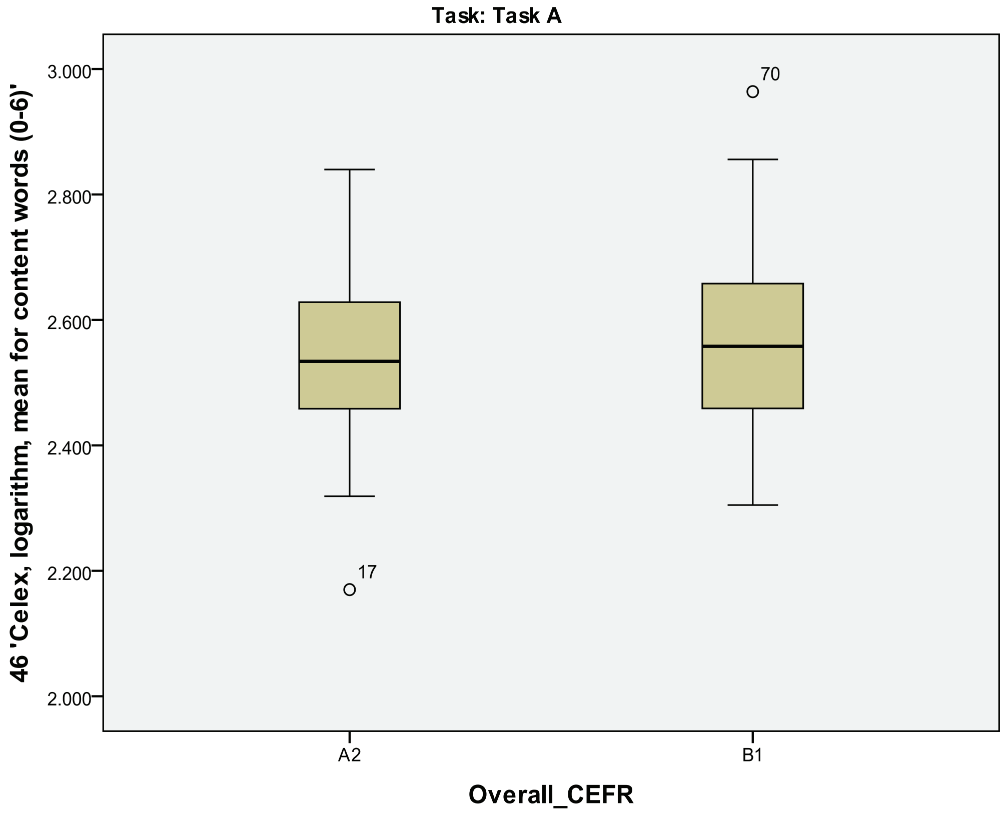

Task: Task B
40 'Flesch-Kincaid Grade Level (0-12)'
12.000
10.000
8.000
6.000
4.000
125
149
A2
B1
Overall_CEFR

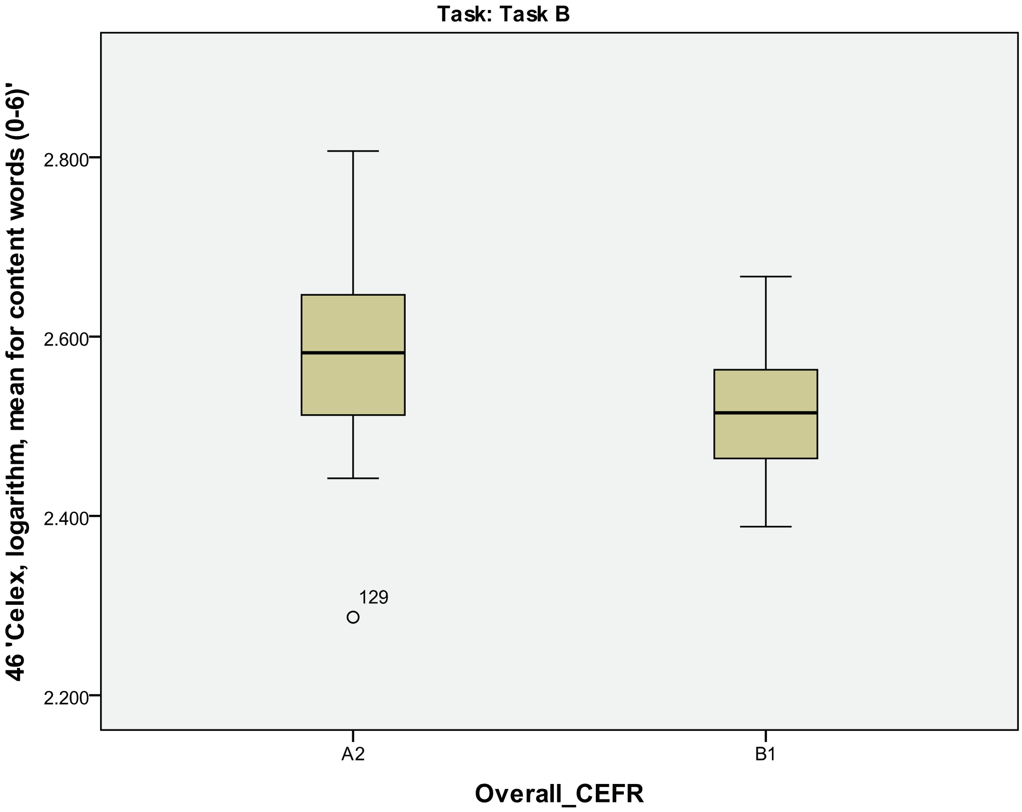

Task: Task A
41 Mean number of modifiers per noun-phrase
1.200
1.000
.800
.600
.400
.200
.000
12
64
A2
B1
Overall_CEFR

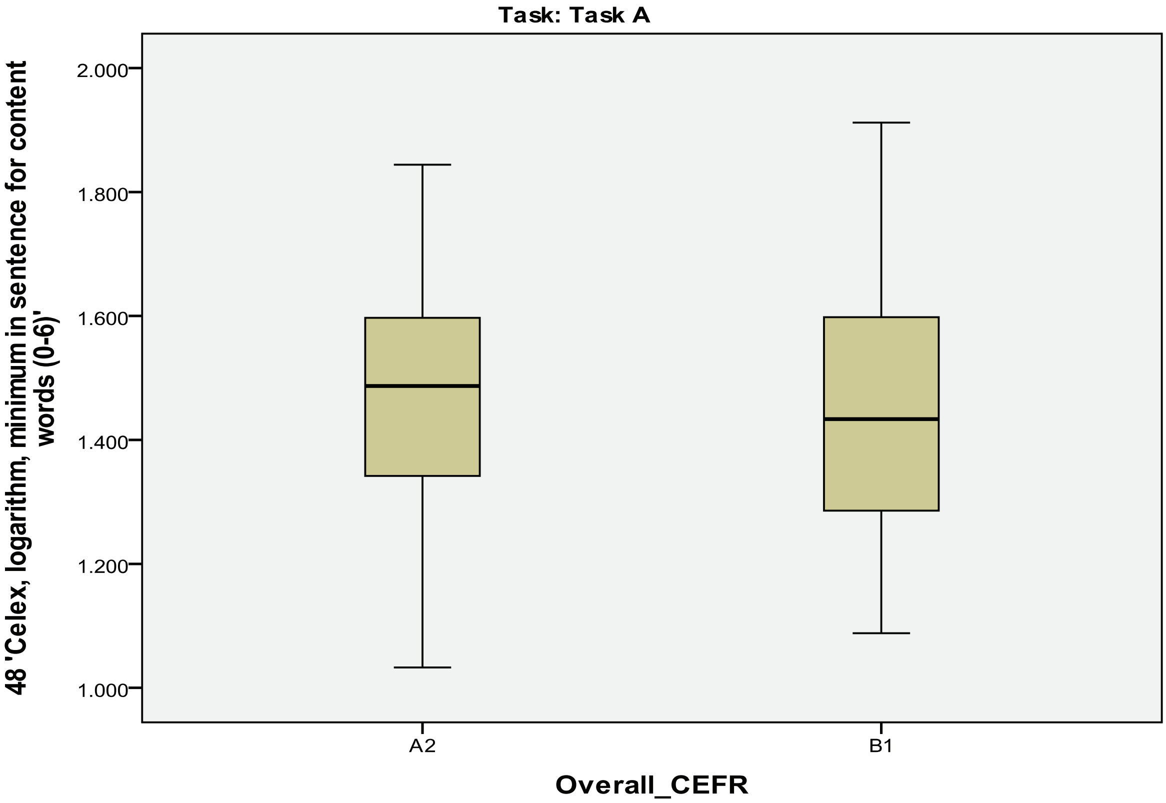

Task: Task B
41 Mean number of modifiers per noun-phrase
1.000
.800
.600
.400
140
141
125
213
A2
B1
Overall_CEFR

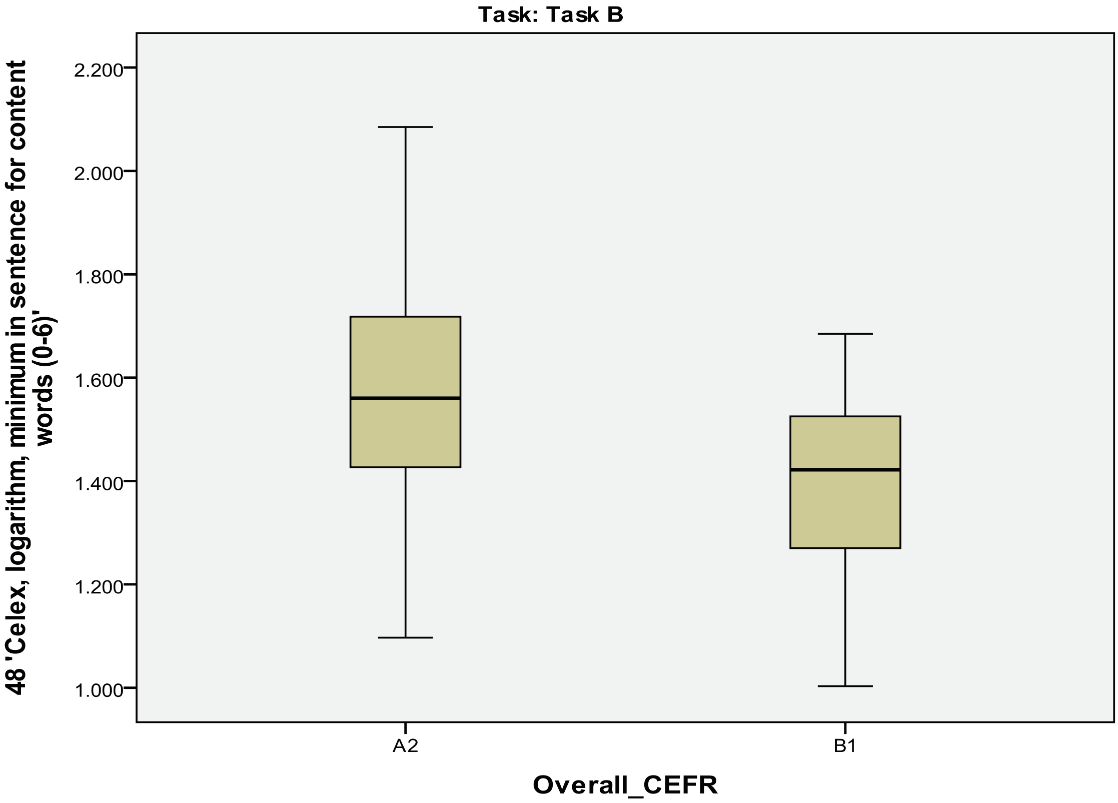

Task: Task A
43 Mean number of words before the main verb of main clause in sentences
7.000
6.000
5.000
4.000
3.000
2.000
1.000
A2
B1
Overall_CEFR


Task: Task B
43 Mean number of words before the main verb of main clause in sentences
8.000
6.000
4.000
2.000
154
A2
B1
Overall_CEFR


Task: Task A
46 'Celex, logarithm, mean for content words (0-6)'
3.000
2.800
2.600
2.400
2.200
2.000
70
17
A2
B1
Overall_CEFR


Task: Task B
46 'Celex, logarithm, mean for content words (0-6)'
2.800
2.600
2.400
2.200
129
A2
B1
Overall_CEFR


Task: Task A
48 'Celex, logarithm, minimum in sentence for content words (0-6)'
2.000
1.800
1.600
1.400
1.200
1.000
A2
B1
Overall_CEFR


Task: Task B
48 'Celex, logarithm, minimum in sentence for content words (0-6)'
2.200
2.000
1.800
1.600
1.400
1.200
1.000
A2
B1
Overall_CEFR

Appendix D: Multiple Regressions of Significant Coh-Metrix Indices Against Band A2 and B1 Scores

Task A

Model Summary[b]

Model	R	R²	Adjusted R²	Std. Error of the Estimate
1	.394[a]	.155	.041	.480

a. Predictors: (Constant), 46 'Celex, logarithm, mean for content words (0-6)', 35 'Number of words', 16 'Argument overlap, adjacent, unweighted,' 37 'Average words per sentence,' 41 'Mean number of modifiers per noun phrase,' 8 'Ratio of causal particles to causal verbs (cp divided by cv+1),' 43 'Mean number of words before the main verb of main clause in sentences,' 38 'Average syllables per word,' 48 'Celex, logarithm, minimum in sentence for content words (0-6),' 17 'Stem overlap, adjacent, unweighted,' 40 'Flesch-Kincaid Grade Level (0-12),' 39 'Flesch Reading Ease Score (0-100)'

b. Task = Task A

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	86.692	112.431		.771	.443
	8 'Ratio of causal particles to causal verbs (cp divided by cv+1)'	-.021	.054	-.045	-.394	.695
	16 'Argument overlap, adjacent, unweighted'	.520	.514	.292	1.011	.315
	17 'Stem overlap, adjacent, unweighted'	-.285	.487	-.158	-.586	.559
	35 'Number of words'	.002	.002	.078	.732	.466
	37 'Average words per sentence'	-.456	.553	-5.870	-.824	.412
	38 'Average syllables per word'	-35.888	45.939	-5.389	-.781	.437
	39 'Flesch Reading Ease Score (0-100)'	-.414	.546	-6.763	-.758	.450
	40 'Flesch-Kincaid Grade Level (0-12)'	.167	.079	.569	2.115	.037
	41 'Mean number of modifiers per noun phrase'	-.132	.341	-.042	-.386	.700
	43 'Mean number of words before the main verb of main clause in sentences'	-.066	.045	-.158	-1.469	.145
	48 'Celex, logarithm, minimum in sentence for content words (0-6)'	-.228	.335	-.092	-.680	.499
	46 'Celex, logarithm, mean for content words (0-6)'	.635	.462	.173	1.376	.172

a. Task = Task A

b. Dependent Variable: Overall CEFR

Task B

Model Summary[b]				
Model	R	R²	Adjusted R²	Std. Error of the Estimate
1	.609[a]	.371	.304	.417

a. Predictors: (Constant), 46 'Celex, logarithm, mean for content words (0-6),' 17 'Stem overlap, adjacent, unweighted,' 41 'Mean number of modifiers per noun phrase,' 8 'Ratio of causal particles to causal verbs (cp divided by cv+1),' 35 'Number of words,' 38 'Average syllables per word,' 43 'Mean number of words before the main verb of main clause in sentences,' 48 'Celex, logarithm, minimum in sentence for content words (0-6),' 37 'Average words per sentence,' 16 'Argument overlap, adjacent, unweighted'

b. Task = Task B

Coefficients[a,b]						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	.341	2.152		.158	.874
	8 'Ratio of causal particles to causal verbs (cp divided by cv+1)'	.121	.110	.113	1.097	.275
	16 'Argument overlap, adjacent, unweighted'	.081	.527	.026	.153	.879
	17 'Stem overlap, adjacent, unweighted'	-.427	.482	-.155	-.886	.378
	35 'Number of words'	.003	.001	.212	2.347	.021
	37 'Average words per sentence'	.037	.021	.214	1.771	.080
	38 'Average syllables per word'	1.934	.781	.228	2.475	.015
	41 'Mean number of modifiers per noun phrase'	.482	.318	.133	1.516	.133
	43 'Mean number of words before the main verb of main clause in sentences'	.001	.045	.002	.027	.979
	48 'Celex, logarithm, minimum in sentence for content words (0-6)'	-.309	.276	-.124	-1.118	.267
	46 'Celex, logarithm, mean for content words (0-6)'	-.955	.595	-.167	-1.604	.112

a. Task = Task B

b. Dependent Variable: Overall CEFR

Excluded Variables[b,c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	39 'Flesch Reading Ease Score (0-100)'	-2144.774[a]	-1.209	.230	-.125	2.138E-9
	40 'Flesch-Kincaid Grade Level (0-12)'	96.280[a]	.725	.470	.075	3.849E-7

a. Predictors in the model: (Constant), 46 'Celex, logarithm, mean for content words (0-6),' 17 'Stem overlap, adjacent, unweighted,' 41 'Mean number of modifiers per noun phrase,' 8 'Ratio of causal particles to causal verbs (cp divided by cv+1),' 35 'Number of words,' 38 'Average syllables per word,' 43 'Mean number of words before the main verb of main clause in sentences,' 48 'Celex, logarithm, minimum in sentence for content words (0-6),'
37 'Average words per sentence,' 16 'Argument overlap, adjacent, unweighted'

b. Task = Task B

c. Dependent Variable: Overall CEFR

Appendix E: Contribution of Individual Criteria to Total Scores on Task A and B

TASK A

Model Summary

Model	R	R²	Adjusted R²	Std. Error of the Estimate
1	.945[a]	.893	.889	.219

a. Predictors: (Constant), Task A Grammatical, Task A Ideas, Task A Coherence and Cohesion, Task A Lexical

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.125	.060		-2.089	.039
	Task A Ideas	.225	.047	.214	4.834	.000
	Task A Coherence and Cohesion	.187	.047	.203	3.995	.000
	Task A Lexical	.267	.055	.268	4.888	.000
	Task A Grammatical	.376	.051	.390	7.410	.000

a. Dependent Variable: Task A Overall

TASK B

Model Summary

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	.941[a]	.885	.880	.21808

a. Predictors: (Constant), Task B Grammatical, Task B Coherence, Task B Cohesion, Task B Ideas, Task B Lexical

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.157	.061		-2.556	.012
	Task B Ideas	.220	.061	.228	3.600	.000
	Task B Coherence	.118	.045	.132	2.586	.011
	Task B Cohesion	.231	.054	.231	4.320	.000
	Task B Lexical	.263	.063	.269	4.193	.000
	Task B Grammatical	.205	.051	.214	4.041	.000

a. Dependent Variable: Task B Overall