A Research Report on the Development of the Test of English for Academic Purposes (TEAP) Speaking Test for Japanese University Entrants – Study 1 & Study 2

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

Table of Contents

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

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) speaking 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スピーキ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。

- Four test tasks were designed to reflect language functions considered important in high school and university education in Japan. Examinations of language functions elicited via different parts of the test confirmed that targeted functions were elicited by the relevant parts of the test as intended.

- 日本の高校、大学教育において重要とされる言語機能を反映する４つのテストタスクが考案された。受験者の発話の言語機能の分析により、それぞれのタスクが引き出すべき言語機能を引き出していることが検証された。

- A study carried out into the scoring validity of the rating of the TEAP Speaking 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スピーキングテストの「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。

- Linguistic features of test takers' output were quantified in relation to key assessment features specified in the five draft analytical rating scales. All examined features of test-taker output varied according to the assessed proficiency level, providing evidence that the rating scales are differentiating test takers' performance in a way congruent with the test designers' intention.

- ５つの評価基準に記述された重要な項目において、TEAPスピーキングテストで実際に受験者が使用した言語が、評価官に評価されたレベルと整合性があるかどうか分析された。検証された全ての言語指標において、受験者の使用言語は点数に応じて高度になっていることが立証され、評価表がテストの出題者の意図に沿って、受験者のスピーキング能力を測っていることが証明された。

- Questionnaire surveys and a focus group discussion were carried out to understand the participating students', interlocutors', and raters' perceptions of the test content and the test procedures. In general, students perceived the test content and the test procedures positively. Interlocutors found the interlocutor training session useful, and felt that the task timings, instructions, questions, and general test administration were appropriate. Raters found the training session and rating scales useful and effective, and the training session gave them confidence in using the rating criteria to assess test-taker performance.

- 受験者、面接官、評価官のTEAPスピーキングテストの経験後の意見、感想がアンケートとフォーカスグループディスカッションにより調査された。受験者はテストの内容と実施方法について、面接官は面接の方法のトレーニング、タスクの時間、面接の指示、質問事項、実施方法について、評価官は評価方法のトレーニングや評価基準について、肯定的な意見、感想を述べた。

1. Introduction

This report describes two *a priori* validation studies of the speaking component of the Test of English for Academic Purposes (TEAP), a new test of academic English proficiency for university entrance purposes in Japan. Drawing on Weir's socio-cognitive framework for validating speaking tests (Weir, 2005; further elaborated in Taylor [Ed.], 2011), this project has collected different types of *a priori* validity evidence during the development of the speaking test, which informed test design and contributed to a validity argument prior to the administration of the operational tests.

The two studies presented in this report involve:

- **Study 1:** A small-scale trial test with 23 first-year university students, three interlocutors, and three raters
- **Study 2:** A large-scale pilot test with 120 third-year high school students, five interlocutors, and six raters

Study 1 examined how well the test materials and rating scales operationalized the test construct described in the draft test specifications in terms of certain aspects of *context* validity and *scoring* validity. Different analyses were carried out on linguistic and functional features of test takers' output language; test scores; feedback questionnaires from test takers, interlocutors, and raters; and a post-marking focus group discussion of raters. All of these sources of empirical validity evidence have offered useful information to verify or modify the draft test specifications and rating scale descriptors for Study 2. Study 2 focused mainly on *scoring* validity, to confirm that changes made after the trial test functioned in ways that the test designers intended.

In this section, we will first provide a brief overview of the aims of the TEAP Speaking Test and then describe background information regarding how the draft test specifications, rating scales, and test materials were developed prior to the two studies presented in this report.

1.1 The TEAP Speaking Test

The TEAP test, which includes separate papers on four skills (reading, listening, writing, and speaking)[1], was designed to measure the language ability of Japanese high school students intending to study at Japanese universities. While specifically taking into account the needs of students intending to study at Sophia University, which is a partner in the development of the test, from the outset the test has been intended to have the potential for wider application beyond one institution. A more 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 two million test takers a year, and Sophia University, one of the leading private universities in Japan. Following the involvement of Professor C. J. Weir in the TEAP writing project, Dr. Fumiyo Nakatsuhara at the Centre for Research in English Language Learning and Assessment (CRELLA)

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

at the University of Bedfordshire in the UK was contracted to provide specialist assistance to the TEAP speaking project.

In the first year of the TEAP speaking project (April 2010 to March 2011), the role of Dr. Nakatsuhara as a consultant—drawing on her previous experience in researching speaking assessment in Japan (2009, 2011, forthcoming)—was to provide a literature review on the assessment of speaking for English as a foreign/second language (EFL/ESL) learners and to develop draft test specifications while communicating with the Eiken and Sophia University project teams. In the second year (April 2011 to March 2012), her consultancy involved developing draft rating scales while communicating with the other project team members, providing advice on various aspects of the TEAP Speaking Test (e.g., test tasks, test administration, interlocutor frame, rater and interlocutor training materials and procedures), planning two *a priori* validation studies, and analyzing the data from these studies: Study 1 (trial test) and Study 2 (pilot test). These studies were designed to provide *a priori* validity evidence during the development of the speaking test (see Section 1.2 for details on *a priori* validity). Such evidence was intended to inform test design and validation prior to the introduction of the test on an operational basis.

The TEAP is intended to evaluate the preparedness of high school students to understand and use English when taking part in typical learning activities at Japanese universities. The target language use (TLU) tasks relevant to 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) 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." Like the TEAP Writing Test, the TEAP Speaking Test would thus cover academic contexts relevant to studying at university in the EFL context of Japan. It is related directly to studying and learning, rather than general, everyday activities or interactions that fall in the personal/private domain.

The TEAP is a test of academic English proficiency which it is envisaged will be used for the purpose of university admissions, and, as such, results must be able to discriminate between an appropriate range of student ability levels. At the same time, the 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 examinations. Following the decision made for the TEAP Writing Test through consultation with the main stakeholders in light of guidelines published by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (2002), it was decided that for the TEAP Speaking Test as well, the main area of interest would be whether students attain a level of proficiency relevant to the B1 level of speaking ability defined in the Common European Framework of Reference (CEFR) (Council of Europe, 2001) (see Weir, 2012 for details).

It should be noted here that the CEFR played a central role in the whole TEAP project as a source for identifying criterial features of the different ability levels to be targeted by different test tasks. The CEFR descriptors were also useful starting points for developing the necessarily more specific descriptors needed for use in rating scales. It was felt that bringing the CEFR into the test design from the beginning would facilitate stakeholders' understanding of the test scores and task requirements. It should also be useful to report scores not only as scale scores but in bands which can indicate to test takers their approximate level in terms of some external criterion, and the CEFR offered possibilities here.

Following the decision made for the TEAP Writing Test, it was decided that the TEAP Speaking 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 forward 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 domain, and be able to provide useful feedback for students at this more advanced level of ability.

As mentioned above, a long-term aim of the TEAP is to foster a positive impact on English education in Japan. As described in Sasaki's (2008) summary of the 150-year history of English-language education and assessments in Japan, greater emphasis is now placed on the teaching of speaking skills as *practical communication abilities.* The current course of study for high schools encourages the use of communicative speaking activities in the classroom, a trend also emphasized in the Action Plan to Cultivate Japanese with English Abilities (MEXT, 2003). The new course of study (MEXT, 2008), which will be implemented from 2013, maintains this focus, encouraging the use of integrated tasks for both speaking and writing. To achieve the goal of equipping students with practical communication abilities, some innovations in English education have been made in recent years, such as the inclusion of a listening component in the National Center Test for University Admissions administered by the National Center for University Entrance Examinations (NCUEE) from 2005. Nevertheless, despite these recent innovations, practical information on how to assess students' speaking abilities has not been made sufficiently accessible to classroom teachers. That is, the revised course of study, as with the current version, does not provide guidelines or a rating scale for speaking assessment in high schools, and there is no plan for introducing a speaking component into the National Center Test for University Admissions (personal communication with the chief researcher at the NCUEE, Ishizuka, 2004). Although no reliable figures are available on the number of universities, either public or private, which at present use a speaking test as part of their entrance examinations, such cases are rare and usually restricted to the final screening stage for specific departments, such as Foreign Languages. A significant gap, therefore, still remains between policy goals and changes to actual practice on the ground. As already noted, the TEAP project has from the outset placed importance on creating positive washback, and the TEAP development team strongly hopes that the introduction of a standardized TEAP Speaking Test with transparent test specifications could help to promote the testing of speaking abilities in Japan and provide a transparent model for designing a speaking test suitable for the context in which the TEAP will be used.

1.2 Background to the Studies: Designing the TEAP Speaking Test

MEXT Guidelines

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 schools (MEXT, 2008) regarding the types of compulsory and optional English modules and example language-use situations and example language functions to be focused on in these modules. This review provided valuable information for understanding trends in the Japanese education sector relevant to the TEAP.

Background Review Report

As part of the first preparatory work undertaken prior to drafting test specifications and deciding on speaking test formats, the consultant provided the TEAP development team with a background review report (Nakatsuhara, November 2010; in Japanese). The report included a review of the assessment literature on speaking ability and incorporated the results of a survey of 760 high school students and 172 high school teachers previously undertaken by the consultant (Nakatsuhara, 2010). This was to establish a common understanding among the TEAP development team before a face-to-face meeting to discuss the most appropriate speaking tasks and associated criteria for use in the TEAP Speaking Test. An overview of the contents of the report is presented below.

Chapter 1: Socio-cognitive framework for validating speaking tests
Chapter 2: Test-taker characteristics
- 2.1 Japanese high school students' experiential characteristics
- 2.2 Japanese high school students' psychological and physical/physiological characteristics

Chapter 3: Context validity and cognitive validity
- 3.1 Discourse features and language functions elicited via different speaking test formats
- 3.2 Speaking tasks used for classroom activities and assessments in Japanese high schools
- 3.3 Different types of speaking test tasks

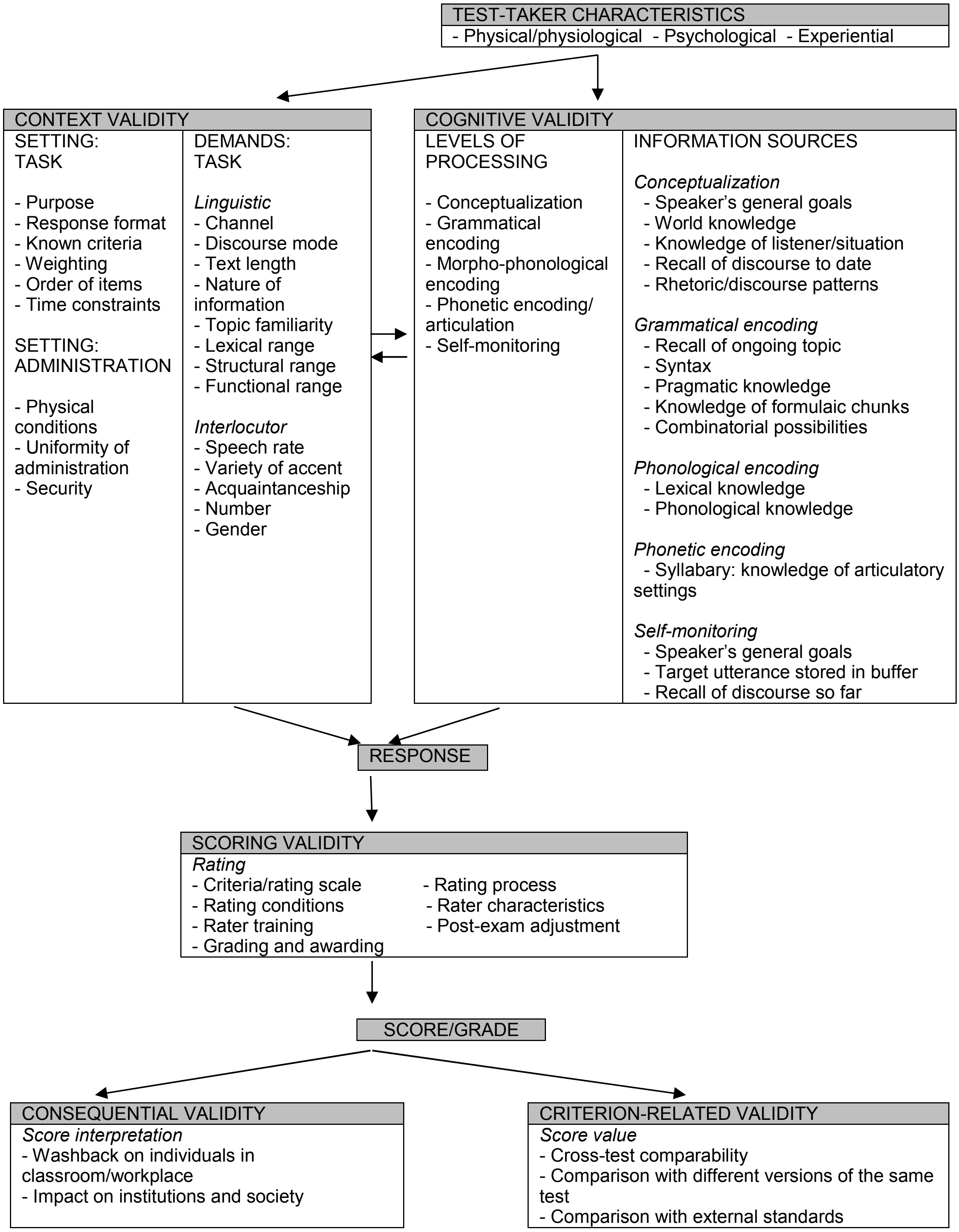
Chapter 4: Scoring validity and criterion-related validity
- 4.1 Comparison of scoring validity in different speaking test formats
- 4.2 Different rating criteria used for the assessment of speaking

Chapter 5: Consequential validity
- 5.1 Conditions for fostering positive washback
- 5.2 Japanese high school teachers' and students' perceptions of the use of speaking tests in the classroom and as part of university entrance examinations

The report illustrated the socio-cognitive framework for validating speaking tests presented in Weir (2005) and further elaborated in Taylor (Ed., 2011). O'Sullivan and Weir (2011, p. 20) describe the framework as "the first systematic attempt to incorporate the social, cognitive and evaluative (scoring) dimensions of language use into test development and validation." Weir (2005) provides versions of the framework adapted for each of the four skills, and the framework for speaking has been applied and refined in Taylor (2011). Taylor (2011, p. 25-28) provides a useful overview of the benefits of using the framework. The framework for speaking, as shown in Figure 1, represents a unified approach to gathering validation evidence for a speaking test. It is particularly valuable that the framework conceptualizes different aspects of validity in terms of temporal sequencing, thus offering test developers a clear plan of what validity evidence should be collected at what stage. The framework comprises *context* validity and *cognitive* validity, which should be established before the test becomes operational (i.e., *a priori* validation), as well as *scoring* validity, *consequential* validity, and *criterion-related* validity, which are usually examined and reported after the test event (i.e., *a posteriori* validation).

Using the socio-cognitive validation framework, the report touched upon different aspects of validity while referring to how they relate to the target Japanese context and what critical questions the TEAP development team should be addressing in applying this framework to the development of the TEAP Speaking Test.

***Figure 1:* The socio-cognitive framework for conceptualizing speaking test validity (Taylor, 2011, p. 28; adapted from Weir, 2005, p. 46)**



TEST-TAKER CHARACTERISTICS

- Physical/physiological - Psychological - Experiential

CONTEXT VALIDITY

SETTING: TASK

- Purpose
- Response format
- Known criteria
- Weighting
- Order of items
- Time constraints

SETTING: ADMINISTRATION

- Physical conditions
- Uniformity of administration
- Security

DEMANDS: TASK

Linguistic

- Channel
- Discourse mode
- Text length
- Nature of information
- Topic familiarity
- Lexical range
- Structural range
- Functional range

Interlocutor

- Speech rate
- Variety of accent
- Acquaintanceship
- Number
- Gender

COGNITIVE VALIDITY

LEVELS OF PROCESSING

- Conceptualization
- Grammatical encoding
- Morpho-phonological encoding
- Phonetic encoding/ articulation
- Self-monitoring

INFORMATION SOURCES

Conceptualization

- Speaker's general goals
- World knowledge
- Knowledge of listener/situation
- Recall of discourse to date
- Rhetoric/discourse patterns

Grammatical encoding

- Recall of ongoing topic
- Syntax
- Pragmatic knowledge
- Knowledge of formulaic chunks
- Combinatorial possibilities

Phonological encoding

- Lexical knowledge
- Phonological knowledge

Phonetic encoding

- Syllabary: knowledge of articulatory settings

Self-monitoring

- Speaker's general goals
- Target utterance stored in buffer
- Recall of discourse so far

RESPONSE

SCORING VALIDITY

Rating

- Criteria/rating scale
- Rating conditions
- Rater training
- Grading and awarding
- Rating process
- Rater characteristics
- Post-exam adjustment

SCORE/GRADE

CONSEQUENTIAL VALIDITY

Score interpretation

- Washback on individuals in classroom/workplace
- Impact on institutions and society

CRITERION-RELATED VALIDITY

Score value

- Cross-test comparability
- Comparison with different versions of the same test
- Comparison with external standards

Early Version of the Draft Test Specifications and Language Function Surveys

Based on the literature review in the report and a number of discussions with the TEAP development team via Skype, the consultant prepared an initial draft of the test specifications that drew on the socio-cognitive framework for validating speaking tests (Weir, 2005; further elaborated in Taylor (Ed., 2011).

During the development of the initial draft, the survey results reported in the background review report (Nakatsuhara, 2010) regarding the language functions that Japanese high school teachers want their students to acquire were found to be very informative and useful in selecting appropriate task types. The survey results were taken from a questionnaire study (Nakatsuhara, 2010) in which 172 Japanese high school teachers were given a list of language functions, and were asked to judge (yes or no) whether they would want their students to master each of the language functions by the end of their high school study. The language function list used was a slightly modified version of O'Sullivan, Weir, and Saville's (2002) function checklist, and the details of the list are provided in Section 3.1.

At the same time, considering the role of the TEAP as a university entrance examination, it was also felt that the test specifications should reflect the language functions that university teachers consider to be important for a student to be successful in first-year undergraduate classes. For this reason, a questionnaire survey was carried out in January 2011 with 24 English teachers at Sophia University who were teaching first-year students at the time of the data collection. Their teaching experience in tertiary education ranged from 3 years to 35 years. Using the same language function checklist described above based on O'Sullivan et al. (2002), the 24 teachers were asked to rate the extent to which they thought each language function would be important for a student to be successful in their first-year undergraduate classes. The rating was undertaken using a four-point scale (4: very important, 3: important, 2: somewhat important, 1: not important). The results for both high school teachers and Sophia University teachers are provided in Appendix 1. Below is a brief summarization of the results of both surveys.

Sophia university teachers (N=24)

- *Informational* and *interactional* functions were in general considered to be more important than *managing interaction* functions.
- All types of *informational* functions were considered to be "very important" or "important," but the following five functions were especially thought to be essential:
  - Justifying opinions
  - Expressing opinions
  - Comparing
  - Elaborating
  - Providing personal information
- Among the range of *interactional* functions, the following four functions were rated higher than others:
  - Agreeing
  - Negotiating meaning
  - Asking for information
  - Asking for opinions
- Regarding the *managing interaction* functions, only one was thought to be especially important:
  - Reciprocating

High school teachers (N=167)

- *Informational* and *interactional* functions were in general considered to be more important than *managing interaction* functions.
- Over 80% of the participating teachers reported that the functions they would want their students to achieve were:
  - Providing personal information, expressing opinions/preferences, justifying opinions (informational functions)
  - Agreeing, disagreeing, asking for information (interactional functions)
- From 60% to 80% of the participating teachers reported that the functions they would want their students to achieve were:
  - Elaborating, staging, describing, summarizing, suggesting (informational functions)
  - Asking for opinions (interactional function)
  - Reciprocating (managing interaction function)
- From 40% to 60% of the participating teachers reported that the functions they would want their students to achieve were:
  - Comparing, speculating (informational functions)
  - Commenting/modifying, persuading, negotiating meaning (interactional functions)
  - Initiating, changing, deciding (managing interaction functions)

Focus Group Meeting for Key Project Staff Convened at Sophia University

With the early version of the draft test specifications and the survey results, a one-day, face-to-face meeting was held at Sophia University in March 2011, including the key project staff members from Eiken and Sophia University and the consultant. The draft specifications included the following points, and each point was extensively discussed until a full consensus was reached:

- Test purposes
- Theoretical framework and empirical support: socio-cognitive framework
- TLU domain
- Administration schedule
- Ability levels targeted
- Actual level of test takers
- Rating criteria
- Interlocutors' and raters' roles
- Interlocutor and rater training / interlocutor frame
- Preparation of the test handbook
- Duration of the test
- Verbal/written prompts for each task
- Contextual factors that need special attention in the development and administration of the test as a whole and of each task
- Cognitive demands that need special attention in the development and administration of the test as a whole and of each task
- Test structure
- Sample tasks
- The CEFR scales and descriptors relevant to each task in relating each task to the CEFR levels

In discussing the above points, a set of contextual parameters to be addressed by the test developers was recurrently referred to, as recommended by Weir (2005), Taylor (Ed., 2011, Chapter 1), and Galaczi and ffrench (2011). They include:

- Setting (task)
    - Purpose
    - Response format
    - Known criteria
    - Weighting
    - Order of items
    - Time constraints
- Setting (administration)
    - Physical conditions
    - Uniformity of administration
    - Security
- Demands (task)
    - Linguistic (input and output)
        - Discourse mode
        - Channel
        - Length
        - Nature of information
        - Content knowledge
        - Lexical
        - Structural
        - Functional
    - Interlocutor
        - Speech rate
        - Variety of accent
        - Acquaintanceship
        - Number
        - Gender

Furthermore, when discussing the types of tasks, it was repeatedly emphasized that consideration should be given to the cognitive demands that each task would make on the test takers. Following Field's (2011) model of grading cognitive demands of speaking tasks, the development team paid attention to cognitive demands in relation to *grammatical encoding* and *conceptualization*.

Grammatical encoding: Field (2011) argues that linguistic content and the degree of cognitive demands in relation to grammatical encoding can be specified in the form of language functions to be performed by test takers. He identified two possible criteria for grading the functions in terms of cognitive demands:

- The semantic complexity of the function to be expressed
- The number of functions elicited by a particular task

The language function survey results were used in conjunction with this cognitive discussion to make an informed decision regarding task formats. Although some members of the project team were initially keen to include paired or group oral formats to elicit richer *informational* and *managing interaction* functions (ffrench, 2003), in light of the survey results, it was agreed that a role-play task, where candidates could demonstrate their ability to ask for information and to ask for opinions, would be more appropriate.

Conceptualization: Field (2011, p. 87-88) suggests that in L2 speaking tests, conceptualization can be considered under two main headings—provision of ideas and integrating utterances into a discourse framework:

- Provision of ideas: The complexity of the ideas which test takers have to express and the extent to which the ideas are supplied to them. The task demands can be increased or decreased by the type of support that may help the test takers to build ideas to express, and the provision of pre-task planning time will also affect the demands.
- Integrating utterances into a discourse framework: The extent to which test takers are assessed on their ability to relate utterances to the wider discourse. Different cognitive challenges could be posed by different interactional patterns elicited by test formats: interviewer-candidate (I-C), candidate-candidate (C-C), solo candidate (C), and three-way exchange (I-C-C).

The provision of pre-task planning time and how specific and complex the task cues should be were discussed for each task, considering the demands made by different interactional patterns.

Here, it is also important to note that, given the central role of the CEFR in the TEAP project, the development team constantly referred to the CEFR scales and descriptors appropriate for each task type when selecting task formats. This was thought to be vital, as linking the test to a widely used outside criterion would increase transparency and interpretability and give added value to feedback for test takers and other stakeholders.

Another significant issue discussed during the meeting was the role of interlocutors and raters. At all stages of the development process, the team explicitly took into account the high-stakes nature of the decisions that will be made based on the test. To maintain fairness and consistency in the testing procedures, it was decided that interlocutors in the TEAP Speaking Test would concentrate only on interviewing the candidates. This would allow them to efficiently manage the various tasks in the test and to focus on applying the task instructions in a consistent way to elicit appropriate samples of speech from candidates. The team agreed that all test performances would be video-recorded and the recorded performances would be assigned marks by raters. The team also considered the possibility of allowing raters to watch videos more than once, when necessary, so as to increase the reliability of the test.

By the end of the one-day discussion, the development team had agreed on a draft test structure, as illustrated in Table 1.

Several features of the test structure are worth mentioning at this point. As can be seen, different tasks were designed to be appropriate for eliciting different levels of performance. The tasks gradually increase in difficulty (in terms of their cognitive demands), beginning with tasks designed to be accessible to A2/B1-level candidates. The final two tasks are aimed at higher proficiency levels, specifically the B2 level, thought to be appropriate for the TLU domain of university undergraduate classes. This structure is consistent with the overall aims of the test to provide useful feedback to students across these ability levels. A2-level candidates may indeed find the B2-level tasks inaccessible, but useful feedback will still be available to these students. This structure also allows a kind of probing and hypothesis-testing approach to evaluating a candidate’s performance through the accumulation of evidence derived from performances across the various tasks.

Table 1: *Test Structure*

Part	Task (Target Level)	Time	Cognitive Demands: *Grammatical Encoding*	Cognitive Demands: *Conceptualization*	Example Topics
1	Interview (A2–lower B1)	2 min.	e.g.: -Providing specific personal information at different temporal frames (present, past, future)	*a) Ideas* Low *b) Discourse framework (I-C)* Low	e.g.: Study, languages, career, high school life, university life
2	Role play (B1)	2 min.	e.g.: -Initiating interaction -Asking for information/opinions -Commenting	*a) Ideas* Low *b) Discourse framework (C-I)* High	e.g.: Interviewing a high school teacher, interviewing a university student who has come back from study abroad
3	Monologue (B1–B2)	2 min. (incl. 30 sec. for prep.)	e.g.: -Agreeing/disagreeing -Justifying opinions -Elaborating	*a) Ideas* Mid – with prep. time *b) Discourse framework (C)* Mid–high	A topic related to the one discussed in Part 2
4	Extended interview (B2)	4 min.	e.g.: -Expressing opinions -Justifying opinions -Comparing -Speculating -Elaborating	*a) Ideas* High *b) Discourse framework (I-C)* High	Two subject areas that are more topical and abstract than those in the previous parts. e.g.: Means of transportation, festivals, health, studying and traveling abroad; education system

The different levels targeted by the tasks operationalize key concepts in the criterial features of each ability level described in the CEFR. For example, Part 1 in particular is restricted to the kind of “familiar matters regularly encountered in work, school, leisure, etc.” contained in the B1 descriptor of the Global Scale (Council of Europe, 2001, p. 24). Parts 3 and 4, on the other hand, aim to operationalize the more complex, abstract elements of language use described in various scales of the CEFR for the B2 level. Parts 2, 3, and 4 are clearly relevant to the TLU domain for the TEAP test described earlier. Part 1, while talking about school and school events, does not quite comply with the TLU definition provided earlier, which explicitly restricted tasks to those relevant to teaching and learning contexts. However, in order to provide useful feedback to test takers at an A2 level of ability, it was felt appropriate to design Part 1 around topics that would “require a simple and direct exchange of information on familiar and routine matters” that characterize the A2 level (Council of Europe, 2001, p. 24).

As described earlier, Part 2 was designed to be a role play in which candidates would specifically have to ask questions. Candidates are provided with a topic card which explains the situation and gives a set list of topics about which the test takers must ask questions (information they must obtain from the interlocutor). The task card instructs candidates that they may ask extra questions. While the task is limited in scope, it was also felt to be a significant step in the context of language testing in Japan to incorporate such a task. While students are accustomed to asking questions in information gap activities, etc., in oral communication classes at school, it was anticipated that students would not be familiar with such a format in the context of a test. As such, this aspect was specifically focused on in interviews and questionnaires with test takers during subsequent data collection in trialing and piloting.

Based on the meeting minutes and a few post-meeting email exchanges, draft test specifications were revised and submitted to the TEAP project teams at Sophia University and Eiken in April 2011.

Rating Scale Development

After the task types and rating categories were agreed upon in the face-to-face meeting, the consultant started drafting rating scales. Given the vital role of the CEFR in designing the TEAP test, the CEFR descriptors from the most relevant scales were used as the criterion benchmarks from which the TEAP, TLU-specific descriptors were developed. This was done with the explicit intention of building the CEFR into the rating scales and test design for the purposes of reporting the results to test takers. Alongside consideration of the CEFR descriptors, other established rating scales such as the Cambridge ESOL Common Scale for Speaking (Galaczi, ffrench, Hubbard, & Green, 2011) and rating scales that were developed for Japanese learners of English—like the Standard Speaking Test (SST) rating scales (ALC, 2006) and Kanda English Proficiency Test (KEPT) rating scales (Kobayashi & Van Moere, 2004)—also informed the rating scale development. Moreover, a number of changes to the CEFR performance descriptors were made where the scales were either inadequate or not sufficiently comprehensive, well calibrated, or transparent.

Once an early version of the rating scales was drafted, the scales were discussed within Eiken, within Sophia University, and within CRELLA separately, and comments and suggestions for changes were shared in several Skype meetings. Based on these discussions, draft rating scales were revised and submitted to the TEAP development team in June 2011.

The draft scales contained five analytical categories (grammatical range and accuracy, lexical range and accuracy, fluency, pronunciation, and interactional effectiveness), each of which had four levels (0 = below A2, 1 = A2, 2 = B1, and 3 = B2). The development team decided to use an analytic scoring approach, considering its advantage in helping raters to focus on the aspects of language to be measured, thus resulting in better scoring validity, and in enabling score reporting for diagnostic purposes (Taylor & Galaczi, 2011). At this stage, the development team was still investigating two different approaches to scoring: part-scoring, in which raters apply each scale to each part of the test, resulting in analytic scale scores for each part, or overall scoring, in which raters assign one mark for each analytic scale based on performance across the whole test. The issue of part-scoring and overall scoring is revisited and discussed in more detail later in this section and in Section 3.4.

In the meantime, project members at Eiken drafted an interlocutor frame and task prompt cards and prepared interlocutor and rater training materials. All the materials were reviewed by the project team and the consultant, and draft versions were finalized through several Skype meetings.

All of the materials were piloted in a mini-trial test carried out by TEAP project members with three first-year university students from Sophia University who were at approximately A2, B1, and B2 levels. The performances were video-recorded by the project team to review all elements of the testing procedures and the potential to apply the rating scales. The three students were interviewed in Japanese by a project member after taking the test. The interviews provided an opportunity to pilot questions that would be used in the questionnaire for test takers in later trials, but also to take the opportunity to talk in more depth about their impressions of the testing procedures.

The mini-trial provided the opportunity to make adjustments to elements of the testing procedures and the wording of task instructions and questions to be asked in the test. Feedback from the students regarding Part 2 was instructive and confirmed that this task type is both relevant to the TLU domain for the TEAP as well as relevant to the actual experience of test takers in real-life language-use situations. Two of the three students mentioned that they had experience in conducting interviews in high school in both English and Japanese, and the third had experience doing this in Japanese. All felt the task was relevant and realistic.

After reviewing the three performances, a decision was made at this stage to employ overall scoring rather than part-scoring for the trial in Study 1. This decision was taken for three reasons. First, the development team felt that the amount of language elicited in some parts, particularly parts 1 and 2, would not be sufficient for independently scoring these sections. Second, for parts 1 and 2, which are designed to be accessible to students at an A2/B1 level, task constraints may create an artificial ceiling effect, in which test takers with potentially higher ability than the task aims to elicit are not able to display their ability. Third, it was felt that the test structure with four different parts, eliciting different types of language functions and gradually increasing in difficulty, lent itself to a system of probing a test taker's ability. This would allow the raters to form a hypothesis on the approximate level of the test taker based on their performance on preceding parts of the test, and then to test that through the accumulation of evidence across all parts of the test. In this view, the different parts of the test all provide necessary evidence to contribute to a final decision on a test taker's performance.

This approach was in fact built into the training procedures for raters and also had an impact on the wording of descriptors for some scales. For example, parts of descriptors for "lexical range and accuracy" are displayed below. It can be seen that it would be possible to form an initial impression of whether a candidate has met the requirements for A2 or B1 based on parts 1 and 2. Indeed these parts of the test would be most appropriate for investigating this level, as the topics are designed to elicit language that is familiar and everyday. If a test taker had comfortably displayed the ability to meet the requirements of B1 by the end of Part 2, the rater would then consider whether the candidate is capable of moving beyond that and examine their performance in Part 3, finally confirming the judgement made there by evaluating the performance on the B2-level task in Part 4. The decision to employ overall scoring was validated by examining feedback from raters after the Study 1 trial, and will be discussed again later.

> B2: Uses a range of vocabulary sufficient to deal with the full range of topics presented in the test.
> B1: Uses a range of vocabulary sufficient to manage most everyday topics.
> A2: Vocabulary is limited to routine, everyday exchanges.

After the mini-trial, some minor modifications were also made to the testing and data-collection procedures prior to the Study 1 trial, such as the use of the timer and video recording equipment and the seating plan for the interlocutor and the test taker.

Thus far, we have described how the draft test specifications, rating scales, and test materials were developed, drawing on the socio-cognitive framework, on the basis of some empirical data (i.e., language function surveys) and through iterative discussions in the TEAP development team at Eiken, Sophia University, and CRELLA.

2. The Two Studies

2.1 Scope of the Two Studies

As mentioned previously, the draft specifications for the TEAP Speaking Test were developed drawing on the updated version of Weir's socio-cognitive framework for validating speaking tests (Taylor, Ed., 2011). The test consists of four parts: interview, role play, monologue, and extended interview. These tasks were selected to offer different levels of cognitive demand in terms of *conceptualization* and *grammatical encoding*, and to elicit the types of language functions that were considered important by educators at Japanese high schools and Sophia University teachers. Tasks were designed to elicit language functions that were congruent with the results of the two surveys as well as those that were considered relevant based on the literature review (see Table 1 for the test structure).

We now move on to reporting the two *a priori* validation studies carried out in July 2011 (Study 1) and December 2011 (Study 2). As mentioned in Section 1.2, establishing validity evidence should start at the before-the-test event stage, and the studies described here did so by collecting data for *context* validity (which also gave some indication of the *cognitive* demands placed on the candidates) and *scoring* validity.

Study 1 aimed to examine how well the draft test materials and rating scales operationalized the test construct in terms of certain aspects of context validity and scoring validity. Based on the findings from Study 1, some modifications were made to the test materials and rating scales. Study 2 investigated how well the test functioned in terms of scoring validity after incorporating the modifications suggested by Study 1.

2.2 Research Questions

Five research questions were investigated through Study 1 and/or Study 2:

- **RQ1:** To what extent does the test elicit intended language functions in each task? (Study 1)
- **RQ2:** Is there any evidence from test takers' output language that validates the descriptors used to define the levels on each rating scale? (Study 1)
- **RQ3:** What are the participating interlocutors' and students' perceptions of the testing procedures? (Studies 1 and 2)
- **RQ4:** What are the participating raters' perceptions of the testing and rating procedures? (Studies 1 and 2)
- **RQ5:** How well does the test function in terms of scoring validity, after incorporating modifications suggested in Study 1? (Study 2)

2.3 Research Design: Study 1

2.3.1 Participants

Study 1 involved 23 university students, three trained interlocutors, and three trained raters. The 23 students were recruited from different English classes at Sophia University, so as to cover a wide range of proficiency levels. They were all first-year students, who had spent only three months at Sophia University at the time of the data collection. This was a convenience sample, with students recruited on campus through various avenues. As such, issues such as gender balance were not able to be taken into account in the research design. The final sample consisted of 15 females and 8 males.

The three interlocutors were recruited from English teachers at Sophia University. One was a bilingual speaker of English and Japanese, and the other two were native speakers of English. All of them were experienced university teachers (with 37 years, 19 years, and 7 years of experience), but their experience in acting as an interlocutor in standardized speaking tests was limited. They all attended an interlocutor training session prior to the test event. It was considered that the profiles of the three interlocutors selected for Study 1 would reflect those of prospective interlocutors in the operational TEAP Speaking Test. Studies 1 and 2, then, both provided the opportunity to investigate important validity aspects beyond the actual task structure and rating scales, such as the efficacy of interlocutor training procedures and the physical aspects of administering the test.

The three raters were selected by Eiken. All raters were experienced teachers at Japanese universities but with different levels of experience as professional raters in standardized speaking tests. One did not have any experience, another had two to three years' experience rating the EIKEN speaking tests, and another had five years' experience rating the International English Language Testing System (IELTS), Business Language Testing Service (BULATS), and EIKEN speaking tests. They all attended a rater training session prior to the test event. The varied level of experience was considered an advantage, as it would provide an opportunity to trial and review training procedures to see whether novice raters and raters trained for another test would rate speech samples in a consistent manner after the training session.

2.3.2 Data Collection

As mentioned in Section 1.2, all the test procedures were piloted with three students prior to Study 1. This mini-trial test was to confirm that the planned testing procedures would work smoothly, and also to help predict what problems, if any, should be anticipated in the Study 1 data collection. Recordings of these students were also used in the interlocutor training session and the rater training session as sample performances. The samples had been assigned ratings by the Eiken project team using the draft scales prior to the training session.

Study 1 was carried out in July 2011. An overview of the test procedures and task instructions was provided in Japanese to the 23 students for perusal while waiting to take the test. The students did not have access to this sheet during the test. The information was provided in Japanese in advance, as the TEAP is a new test and students did not have prior access to information about the test structure. For operational versions of the test, it is envisaged that information on the test structure will be readily available for potential test takers. The speaking test consisted of the four tasks as described in Table 1. Immediately after their participation, they were asked to complete a feedback questionnaire about their test-taking experience.

Prior to the test event, three interlocutors and three raters took part in the interlocutor and rater training sessions, respectively. After these training sessions, they were asked to fill out a feedback questionnaire on the effectiveness of the training sessions.

During the Study 1 data collection, the three trained interlocutors interviewed 7 or 8 students each, but they did not assign marks to students' live performances. After completing their test sessions, they were asked to fill in a feedback questionnaire about different aspects of the interlocutor frame and interviewing procedures.

All performances were video-recorded, and all the 23 recorded performances were rated by the three trained raters using the draft rating scales described earlier.

2.3.3 Data Analysis

As noted earlier, this research drew on Weir's (2005) socio-cognitive framework for validating speaking tests, which was further elaborated in Taylor (Ed., 2011). The socio-cognitive framework was useful for shaping the research design in this study, not only because the test specifications were developed based on the framework but also because the framework includes a list of contextual parameters that could influence task demands in relation to test takers' outputs required to fulfill the task, such as lexical, structural, and functional features. The use of the framework also enabled us to pinpoint which validity aspect each analysis was targeting. Furthermore, since the TEAP Writing Test has also been developed and validated based on the comparable framework for writing (Weir, 2012), the outcome of this research will fit into a wider validity argument for the TEAP test.

The analysis of the data collected in Study 1 was carried out as follows.

Transcribing the Video-Recorded Performance

All video recordings were transcribed using a slightly simplified version of conversation analysis (CA) notation (Atkinson & Heritage, 1984; the transcription notation is provided in Appendix 2). CA transcription is informative and enables us to examine micro-analytic features of interaction between the examiner and the candidate.

The recordings were transcribed by a research assistant who had previous experience in transcribing speaking-test data and who is a native speaker of English but is also familiar with Japanese speakers of English. The consultant carefully checked the first couple of transcripts, and some modifications were suggested before the research assistant commenced the rest of the transcriptions. An interactive, consensus approach was taken to ensuring consistency in transcriptions. Several transcriptions were checked throughout the process by the Eiken project team member overseeing the transcription, and one full transcript was reviewed by the whole project team at Eiken. Any differences in interpretation were resolved through discussion between the research assistant, the consultant, and the project member overseeing transcription. Prior to a second research assistant carrying out segmentation of the transcripts for the linguistic and discourse analysis described below, all transcripts were double-checked by the second research assistant while watching all recorded samples.

Language Function Analysis

The transcripts were first analyzed for the coverage of language functions elicited in each task. O'Sullivan, Weir, and Saville's (2002) observation checklist was slightly modified for use with the given data. The observation checklist consists of an extensive table of *informational* (e.g., expressing opinion, justifying opinion), *interactional* (e.g., asking for information, negotiating meaning), and *managing interaction* functions (e.g., initiating, reciprocating). While the checklist was originally developed for analyzing language functions elicited from candidates in paired speaking tasks of the Cambridge Main Suite examinations, the potential to apply the list to other speaking tests such as the IELTS Speaking Test (Brooks, 2003) and the Graded Examinations in Spoken English (GESE) (Nakatsuhara & Field, 2012; O'Sullivan, Taylor, & Wall, 2011) has been explored. Since the list draws on Bygate's (1987) speaking model, the applicability of the checklist is not limited to any particular types of L2 candidates' speech, and the list was also useful for examining a range of language functions elicited in the TEAP Speaking Test.

This was the same list used for the language function surveys with educators at Japanese high schools and Sophia University which informed our selection of the task formats in the test (described in Section 1.2). Therefore, by using the same checklist in this validation study, we were able to directly compare language functions specified in the test specifications with functions that were actually elicited from target test takers.

Linguistic and Discourse Analysis of Students' Speech Samples

This analysis was aimed at examining whether test takers' output language validates the descriptors used to define the levels on each rating scale. Previous studies have employed this approach to rating scale validation, including Brown (2006a) and Brown, Iwashita, & McNamara (2005). A variety of linguistic measures were selected to reflect the features of performance relevant to the test construct defined within the draft analytical rating scales, so as to investigate whether these measures differ in relation to the proficiency levels of the candidates assessed using the rating scales. The transcripts were coded for these features by a research assistant. As with transcription, an interactive consensus approach to coding was taken. The project member who oversaw the data preparation reviewed several complete transcripts after they had been coded, and any differences in interpretation were resolved through discussion between the research assistant, the consultant, and the project member overseeing the data preparation.

Three trained raters rated the 23 students' video-recorded test sessions, using the draft TEAP Speaking Test rating scales, which consist of the following five categories:

a. Grammatical range and accuracy
b. Lexical range and accuracy
c. Fluency
d. Pronunciation
e. Interactional effectiveness

Since it is crucial that speech samples selected for the analysis are reliable representatives of a particular level for each analytical category, the test scores were first of all analyzed using multifaceted Rasch analysis.

Once the score analysis had confirmed that the rating scores were assigned by the three raters in a satisfactory and consistent way, as judged by Rasch fit indices and other statistical measures (see Section 3.2.1 for details), the video-recorded speech samples and their transcripts were analyzed for the linguistic characteristics illustrated in Table 2. These linguistic features were selected to reflect elements of performance covered in the draft rating scale descriptors, with the exception of the last three measures listed under "Other—The amount of talk."

The linguistic features were analyzed to investigate the extent to which each of these features differs between the adjacent levels of the rating scales. Since not all measures were relevant for all parts of the test, appropriate parts were selected for different analyses. Section 3.2.2 describes how each of the characteristics was selected and measured.

However, it is important to note that there is no assumption that the measurement criteria listed in Table 2 fully cover the five analytical categories of the rating scale. They relate to some representative aspects of the five categories and thus can only be broadly indicative in quantifying each one.

Table 2: ***Linguistic Measures***

Corresponding Rating Category	Focus	Measure	Parts of the Test Applied
a. Grammatical range and accuracy	Complexity	Ratio of subordinate clauses to AS-units**	1, 2, 3, 4
		Number of words per AS-unit	1, 2, 3, 4
	Accuracy	Percentage of error-free AS-units	1, 2, 3, 4
b. Lexical range and accuracy	Range	Lexical frequency coverage (K1+ K2 words)	1, 2, 3, 4
		Academic Word List coverage	1, 2, 3, 4
	Accuracy*	-	-
c. Fluency	Hesitation	Number of unfilled pauses (utterance initial) per 50 words	1, 2, 3, 4
		Total pause time as a percentage of speaking time	3
	Disfluency	Ratio of repair, false starts, and repetition to AS-units	1, 2, 3, 4
	Temporal	Speech rate in Part 3	3
		Articulation rate in Part 3	3
d. Pronunciation	L1 influence	Number of words pronounced with noticeable L1 influence (*katakana*-like) as percentage of total words produced	1, 2, 3, 4
e. Interactional effectiveness	Length of response	Average words per response	1, 4
	Number of extra questions	Number of separate questions asked that were not on required list in Part 2	2
	Back-channeling and comments	Number of instances of back-channeling and comments in Part 2	2
f. Other—The amount of talk	Length of long turn	Total number of words produced in Part 3	3
	Total production	Total amount of production across all parts of the test, measured in words	1, 2, 3, 4
		Total number of AS-units produced across all parts of the test	1, 2, 3, 4

* Lexical accuracy was not measured in this analysis for reasons described in Section 3.2.2b.

** AS-unit = analysis-of-speech unit.

Analysis of Interlocutors' and Students' Feedback Questionnaires (RQ3)

A feedback questionnaire was given to the three interlocutors at two stages: after the initial training session and after the trial test.

Questions asked after the initial interlocutor training session were about:

- Usefulness of the training session
- Clarity of the interlocutor frame
- Clarity of explanations given about the assessment procedures and criteria
- Usefulness of the training video
- Usefulness of the practice test session during training
- Confidence in acting as an interlocutor in the live test sessions, having finished the training
- Any suggestions to improve the training session

Questions after the trial test asked:

- If they found the time for each part appropriate
- If they found the task instructions appropriate
- If they found task cards in parts 2 and 3 appropriate
- If they found the main questions and follow-up questions appropriate in parts 1 and 4
- If they found the interlocutor's responses to the test takers' questions appropriate in Part 2
- If they found keeping time manageable
- If they found the distance between the interlocutor and the test taker appropriate
- If they had to deviate from the interlocutor frame
- Any comments on the test procedures

A feedback questionnaire was also given to the participating 23 students immediately after the trial test sessions. It included questions about:

- Clarity of the task instructions
- Appropriateness of the time allocated for preparation time in the Part 2 role-play task and in the Part 3 monologue task
- Appropriateness of the speaking time allocated across all tasks
- Relevance of the Part 2 and Part 3 tasks to their target language use
- Appropriateness of topics selected in parts 2, 3, and 4
- Comfort of physical testing conditions (distance between the examiner and the candidate; beep sound of the timer to notify the beginning and end of preparation time; video recording equipment)

Interlocutors' and students' responses to these questionnaires were analyzed using descriptive statistics. All question items were accompanied by comment boxes, where the respondents could elaborate on their dichotomous or multiple-choice responses. Comments provided for each question were used to interpret and/or elaborate on the statistical findings.

Analysis of Raters' Feedback Questionnaires and of Raters' Focus Group Discussion Data (RQ4)

A feedback questionnaire was also given to the three raters at two stages: after the initial rater training session and after the trial test.

Questions asked after the rater training session were about:

- Usefulness of the training session
- Usefulness of watching the interlocutor training video and discussing the interlocutor frame before reviewing rating criteria, as background information
- Clarity of the rating criteria
- Usefulness of the standardized exemplars
- Appropriate number of the standardized exemplars to understand how to apply the rating criteria
- Usefulness of rating the standardized exemplars and discussing the raters' scores before looking at the benchmark scores
- Confidence in being able to apply the rating criteria in rating samples of test-taker performance, having finished the training
- Any suggestions to improve the training session

Questions asked after the trial test were concerned with:

- Whether the descriptors for each of the five analytical categories were easy to understand and interpret in linking them to the candidates' performance
- Whether the descriptors for each score point distinguished well between the levels of the scales for each of the five categories
- How distinct each rating scale was from the others
- Whether the descriptors for each score point were appropriate
- Whether the quality of the video recordings was sufficient for rating the speaking samples
- Whether they needed to watch the video samples more than once to rate them
- Whether they thought the test format provided a sufficient quantity of language to rate appropriately
- At what stage of the rating process they finalized their marks for each category
- Describing the process or processes they followed when rating the samples (free responses)
- Any comments on the rating procedure or suggestions for improving the rating scales (free responses)

Responses to these questionnaires were analyzed using descriptive statistics. All question items in the rater questionnaires were also accompanied by free-response space, where their dichotomous or multiple-choice responses could be elaborated on.

Focus Group Discussion

After the three raters had completed the rating of all recorded performances and had filled in a feedback questionnaire, they were invited to return for a focus group feedback session. The session took place a few days after rating was completed to allow for the scores awarded by raters to be analyzed. The focus group discussion was designed to elicit the raters' reasons for choosing the scores they had awarded. It was hoped that data from this process would provide insights into the rating process(es) and help to identify key performance features that might have influenced raters' decisions. A researcher in Eiken acted as a facilitator of the focus group discussion.

The three raters were given a copy of the five analytical rating scales, plus original handwritten score sheets for each rater, with his/her own comments on them.

The facilitator selected two speech samples on which raters generally agreed (Student 2-2 at Level 3 [B2] and Student 3-6 at Level 2 [B1]) and one for whom there was significant disagreement (Student 2-3 at Level 1 [A2] or Level 2 [B1]). As previously discussed, the B1 level was considered to be a benchmark level of performance, as it represented the upper ability level recommended by MEXT as an appropriate goal for high school graduates. The decision was thus made to focus attention on a test taker that appeared to represent a borderline A2/B1-level performance. During the meeting, the raters watched the video for each of the three test takers again, and the video was paused after each task to allow for discussion. The facilitator asked questions related to items in the feedback questionnaire.

2.4 Research Design: Study 2

After incorporating some modifications to the draft task materials and rating scales based on the Study 1 results, Study 2 was carried out in December 2011 to confirm whether or not these changes improved the quality of the test in terms of scoring validity. Another aim of Study 2 was to obtain feedback from all of the participants, including test takers, interlocutors, and raters. This report focuses on the results obtained from the raters to further validate the training and examining procedures.

2.4.1 Participants

A total of 120 third-year high school students were recruited to participate in the Part 2 study. They were recruited from the network of private high schools associated with Sophia University, and were considered to be representative of the typical test-taker population for Sophia University. Gender was not controlled for in this study, as priority was given to recruiting a sufficient number of high school students willing to take part in pilot test procedures, which included both speaking and writing components, as described in the validation study for the TEAP Writing Test (Weir, 2012). The participation rate was extremely high, with only seven absences on the day of testing, resulting in a total sample of 113 students.

Five interlocutors were recruited for Study 2, and students were randomly and evenly distributed across all five interlocutors. Two of the interlocutors had participated in the Study 1 trial. Of the three new interlocutors, one was a bilingual speaker of Japanese and English, and two were native English speakers. Two of the new interlocutors were teaching at Sophia University. The remaining interlocutor was not a university educator but was a trained and experienced rater for the EIKEN speaking tests.

A total of six raters were involved in Study 2. All raters were native speakers of English. As shown in Table 3, they were fairly experienced teachers at Japanese universities and/or junior high and high schools, as well as being experienced raters in standardized speaking tests. Two of the raters had previously participated in rating for the Study 1 trial.

Table 3: ***Rater Profile in Study 2***

Rater	Teaching Experience	Examining Experience at Standardized Speaking Tests
R1	University: $5^{9}/_{12}$ yrs Jr./high school: 7 yrs Private English school: 6 yrs	EIKEN Grade Pre-1: 1 yr EIKEN Grade 2: 3 yrs
R2	University: 14 yrs Jr./high school: 4 yrs	EIKEN: 3 yrs
R3	University: 4 yrs Private English school: 10 yrs	EIKEN Grade 1: $5^{9}/_{12}$ yrs IELTS: $3^{1}/_{12}$ yrs
R4	Jr./high school & private English school: 2½ yrs	EIKEN Grade Pre-1: $1^{1}/_{12}$ yrs EIKEN Grade 2: $3^{1}/_{12}$ yrs
R5	University: 25 yrs	EIKEN Grade 1: 7 yrs EIKEN Grade 2: 3 yrs
R6	University: 6 yrs Business English school: 19 yrs	EIKEN: 3 yrs BULATS: 4 yrs IELTS: 1½ yrs

2.4.2 Data Collection and Analysis

All students were fully informed about the purpose of the pilot test in the process of recruiting participants. Students were given book coupons for participating. As with the Study 1 trial, on the day of the test, prior to entering the test room, students were given an overview of the test procedures and task instructions in Japanese.

The 120 students were split into four groups of 30 students each. These participants took part in this speaking pilot study (Study 2) as well as a writing pilot study reported in Weir (2012). The two data sets were collected from the same students on the same day. Two groups out of the four took both speaking and writing tests in the morning, and the other two groups took both tests in the afternoon. For both the morning and afternoon test sessions, one group took the writing test first and the other group took the speaking test first. This counter-balanced research design was to avoid order effects, if any (see Weir, 2012, for the TEAP Writing Test validation studies).

All speaking test sessions were video-recorded, and the six raters independently rated 60 video-recorded performances each. Test scores were examined using both classical test theory (CTT) analysis and multifaceted Rasch analysis using the FACETS program. The ratings plan created a rating matrix to ensure sufficient overlap (see Table 4) to enable the analysis of the data with the FACETS program. This ensured that individual raters' scores were calibrated to investigate rater severity levels.

Table 4: ***Study 2 Rating Matrix***

Speech Sample Groups	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
Group 1 (30 students)	✓	✓				
Group 2 (30 students)			✓	✓		
Group 3 (30 students)					✓	✓
Group 4 (30 students)	✓	✓	✓	✓	✓	✓

As in Study 1, all six raters filled in a feedback questionnaire on their rating experience immediately after they had completed the rating of all speech samples assigned to them. The questionnaire was the same as the one used in Study 1 except for one minor change (see Section 4.2). Responses to the questionnaire were analyzed using descriptive statistics, and comments provided in the free-response space for each question were utilized to interpret and/or elaborate on the statistical findings. Any differences from Study 1 results were highlighted.

3. Results and Discussion: Study 1

3.1 Language Functions (RQ1)

This section reports language functions elicited via each of the four tasks and examines whether language functions that the test designers intended to elicit were actually produced by test takers.

As mentioned in Section 2.3.3, a slightly modified version of O'Sullivan, Weir, & Saville's (2002) observation checklist was used for this analysis[2]. The function list, as shown in Table 5, consists of an extensive table of *informational* (e.g., expressing opinions, justifying opinions), *interactional* (e.g., asking for information, negotiating meaning), and *managing interaction* functions (e.g., initiating, reciprocating). The number of turns in which each language function was produced by a test taker was counted for each task. A turn was defined generally as a continuous stream of speech by the test taker bounded by speech by the examiner[3]. When one turn fulfilled more than one function, all functions were coded for that turn.

This investigation also provides empirical evidence in relation to the cognitive validity of the test. The draft test specifications were built based on Field's (2011) model of grading cognitive demands in spoken test tasks in terms of grammatical encoding and conceptualization. As illustrated in Section 1.2, task demands for grammatical encoding can partly be assessed by the types and combination of language function required to complete the tasks. This analysis can therefore also provide some indication of the cognitive validity of the test.

Language functions that were designed to be elicited across the four parts of the test are:

1. **Part 1 (interview):** Providing specific personal information in different temporal frames (present, past, future)
2. **Part 2 (role play):** Initiating interaction, asking for information, asking for opinions, commenting
3. **Part 3 (monologue):** Agreeing, disagreeing, justifying opinions
4. **Part 4 (extended interview):** Expressing opinions, justifying opinions, comparing, speculating, elaborating

Table 5 shows the average number of turns in which each function was produced per participant across the four parts of the test. From this table, functions with an average realization rate of 0.7 turns or above per test taker are in bold. This seems like an appropriate threshold for identifying the main functions elicited in the test, as it was thought to be reasonable to say that those functions produced on average by 70% or more of the participants reflect the task characteristics rather than just being produced by pure chance.

[2] Modifications were (1) combining *expressing opinions* and *expressing preferences* as one category, as they were difficult to differentiate, (2) adding *commenting, greeting,* and *thanking* functions that occurred frequently in the given dataset, and (3) excluding *persuading,* which was unlikely to occur due to the task formats.

[3] Instances of non-significant back-channeling by the examiner that did not interrupt the examinee were not considered to have interrupted a turn.

Table 5: ***Language Functions Elicited Across Four Parts of the Test***

	Part 1		Part 2		Part 3		Part 4	
	Mean	**St. Dv.**	**Mean**	**St. Dv.**	**Mean**	**St. Dv.**	**Mean**	**St. Dv.**
Giving personal info. (present)	**1.70**	1.02	0.00	0.00	0.00	0.00	**1.04**	1.07
Giving personal info. (past)	**2.30**	1.11	0.04	0.21	0.17	0.39	0.39	0.84
Giving personal info. (future)	**1.74**	0.75	0.13	0.34	0.00	0.00	0.04	0.21
Expressing opinions/preferences	**4.52**	2.00	0.00	0.00	0.04	0.21	**5.09**	1.93
Elaborating	**2.52**	2.59	0.00	0.00	**0.74**	1.10	**2.17**	2.04
Justifying opinions	0.65	0.83	0.00	0.00	**1.74**	1.14	**3.35**	1.07
Comparing	0.00	0.00	0.00	0.00	0.00	0.00	**2.30**	1.36
Speculating	0.00	0.00	0.00	0.00	0.00	0.00	**0.87**	1.18
Staging	0.00	0.00	0.22	0.42	0.04	0.21	0.00	0.00
Describing a sequence of events	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Suggesting	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.21
Agreeing	0.00	0.00	0.00	0.00	**0.96**	0.56	0.13	0.34
Disagreeing	0.00	0.00	0.00	0.00	0.17	0.39	0.09	0.29
Modifying	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Asking for opinions	0.00	0.00	**1.87**	0.87	0.00	0.00	0.00	0.00
Asking for info.	0.22	0.52	**3.87**	1.58	0.00	0.00	0.04	0.21
Commenting	0.00	0.00	**1.70**	2.62	0.00	0.00	0.09	0.29
Asking for permission	0.04	0.21	**0.83**	0.65	0.00	0.00	0.00	0.00
Greeting	**2.04**	0.77	**0.83**	0.49	0.00	0.00	0.39	0.58
Thanking	**0.78**	0.52	**1.30**	0.97	**1.09**	0.90	**2.48**	0.99
Negotiating meaning - checking understanding	0.43	0.59	0.39	0.72	0.09	0.29	**0.74**	1.63
- indicating understanding	0.04	0.21	**1.00**	1.13	0.04	0.21	0.17	0.49
- asking for clarification	0.17	0.58	0.09	0.42	0.00	0.00	**0.74**	1.10
- correcting others	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
- responding to a clarification request	0.00	0.00	0.17	0.39	0.04	0.21	0.04	0.21
Initiating	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	0.00
Changing	0.00	0.00	0.04	0.21	0.00	0.00	0.00	0.00
Reciprocating	0.22	0.52	**2.22**	2.35	0.04	0.21	0.00	0.00

We will now examine whether the main functions observed in the data are congruent with the goals of each part. In doing so, we will also relate those functions to the three main types of language functions (informational, interactional, and managing interaction) in order to obtain an overall picture across all four parts. This is useful for understanding that there are clear differences between the four parts in their capability of eliciting different types of function. How each of the main functions was realized will also be exemplified, as we should also ensure that ways in which these functions were elicited were in line with the test designers' intentions.

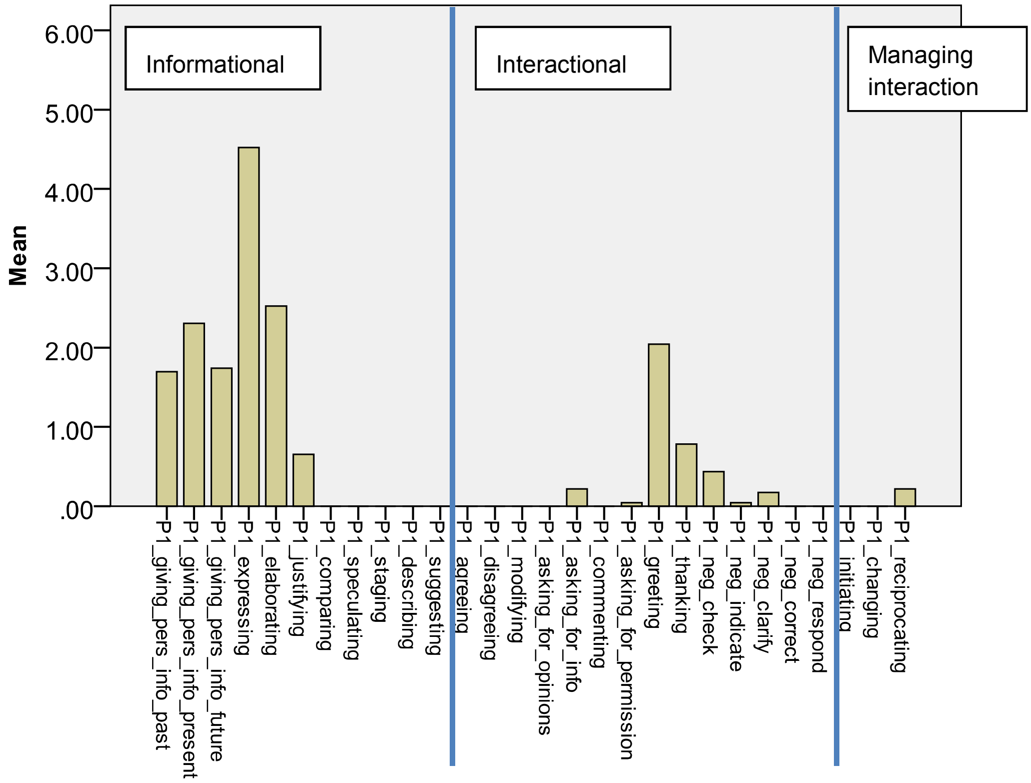
Part 1 (Interview)

Figure 2 illustrates the language functions elicited in Part 1. The X-axis shows language functions investigated and the Y-axis indicates the average number of turns per test taker in which each function was produced.

As shown in Figure 2, Part 1 (interview) of the test mainly elicited *informational* functions such as giving personal information in different temporal frames, expressing opinions/preferences, and elaborating, as well as some *interactional* functions like greeting and thanking.

As intended in the test specifications, personal information in different temporal frames was extensively elicited; 1.70 turns on average for present information; 2.30 turns for past information, and 1.74 turns for future information. Examples (1), (2), and (3) show how these functions were realized, demonstrating that these functions were elicited by those questions designed to target these functions.

Figure 2: **Language functions elicited in Part 1 (interview)**



Example (1) Giving personal information (present)
*1 **E:** What do you like to do in your free time?*
*2→ **S1-1:** (1.4) I like: $reading?$ ahm ¥ reading books:? Or ah $kum-$ comic books ((laughs)) [((laughs))*
*3 **E:** [((laughs))*
*4→ **S1-1:** Yeah. And:: (1.0) often I listen: ah music hhh. (1.1) mmmmm (.) .hhh And: (2.2) my (0.6)*
5→ best like thinking is, sleeping.=
*6 **E:** =[Mmmmm*
*7→ **S1-1:** [((laughs)) .hhh or: (0.5) watching a TV. ((waves left hand))*

Example (2) Giving personal information (past)
*1 **E:** I see. Next. I'm sure there were many events at your high school. Which event did you enjoy the most?*
*2→ **S1-4:** (1.2) Um (0.8) The cultural festival, because um I (.) I was a member of (0.5) drama club, and (1.2) ahm*
3 I: (0.6) I <played a role> (1.1) the main character $at the-$ (0.5) at the stage (0.5) for (.) the culture festival

Example (3) Giving personal information (future)
*1 **E:** What kind of job, (0.9) would you like to have (.) in the future?*
*2 **S1-1:** .Hhh (2.5) I::: (2.0) I (.) hhh want to be a teacher. (0.7) Yeah. And (0.9) $I: teach: I- I, I will? (.) I, want ¥*
3 to teach (.) the (theology).

However, although the results were satisfactory overall, examining how the functions were realized can provide useful hints for more effective phrasing when developing question alternatives for future test forms. In this case, it could be useful to rephrase the question "What do you like to do in your free time?" to "What do you usually do in your free time?" to elicit present presentation. This is because, as shown in Example (1), while S1-1 described what she often does in line 4, she first talked about a favorite way of spending her free time in lines 2, 5, and 7, due to the way the question was asked. Some candidates, in fact, expressed favorite ways of spending free time only (which will fall into the function category of "expressing opinions/preferences"), without mentioning if they are really spending their free time in that way. As a result, the participants produced on average 4.52 turns that expressed opinions/preferences, and these cases were often followed by cases of elaborating (2.52 turns on average), as illustrated in line 2 in Example (4) and line 3 in Example (5), respectively.

Example (4) Expressing opinions/preferences

1 ***E:*** *Ah OK. Ah, first, I'd like to learn a bit about you. All right? So, what do you like to do in your free time.*
2→ ***S3-5:*** *Ah, (0.4) mm I like ¥ to: (0.9) read (.) a: book? (0.7) Ah but especially I (0.4) like to read essay.*
3 *About ah (0.8) uh: (1.4) many many ah (everyday) life (2.4) of the (1.2) many people.*

Example (5) Elaborating

1 ***S2-2:*** *Um: I like (.) to:: listen to my music.*
2 ***E:*** *Really can you tell me a little bit more.*
3→ ***S2-2:*** *OK, (.) my favorite artist is Lady Gaga. At first, I didn't like her, because (.) her (0.7) her clothes*
4 *are so:: (0.9) ((laughs)), but when I heard (.) her song, [the (1.1) =%I don't know ¥ how to say%*
5 *$the (0.5) lysic?$ No, (1.7) =*

Additionally, since Part 1 is the beginning of the interaction, this part also elicited language to greet (2.04 turns on average) and to thank (0.78 turns on average), as shown in lines 1, 3, and 9 in Example (6) below.

Example (6) Greeting / thanking

1→ ***S1-6:*** *Good afternoon.*
2 ***E:*** *Please have a seat.*
3→ ***S1-6:*** *Thank you. ((sits down))*
4 ***E:*** *Ready?*
5 ***S1-6:*** *(0.6) Hh- oh I'm OK.*
6 ***E:*** *Good. Me too. ((laughs)) My name is XXX. May I ask your name?*
7 ***S1-6:*** *(0.5) My name is XXXX.*
8 ***E:*** *Nice to meet you, XXXX.=*
9→ ***S1-6:*** *Nice to meet you too.*

Part 2 (Role Play)

In contrast, language functions elicited in Part 2 (role play), as presented in Figure 3, were characterized more as *interactional*, such as asking for opinions, asking for information, commenting, asking for permission, greeting, thanking, and negotiating meaning (indicating understanding). The elicitation of language functions to *manage interaction*, like initiating interaction and reciprocating, was also noticeable.

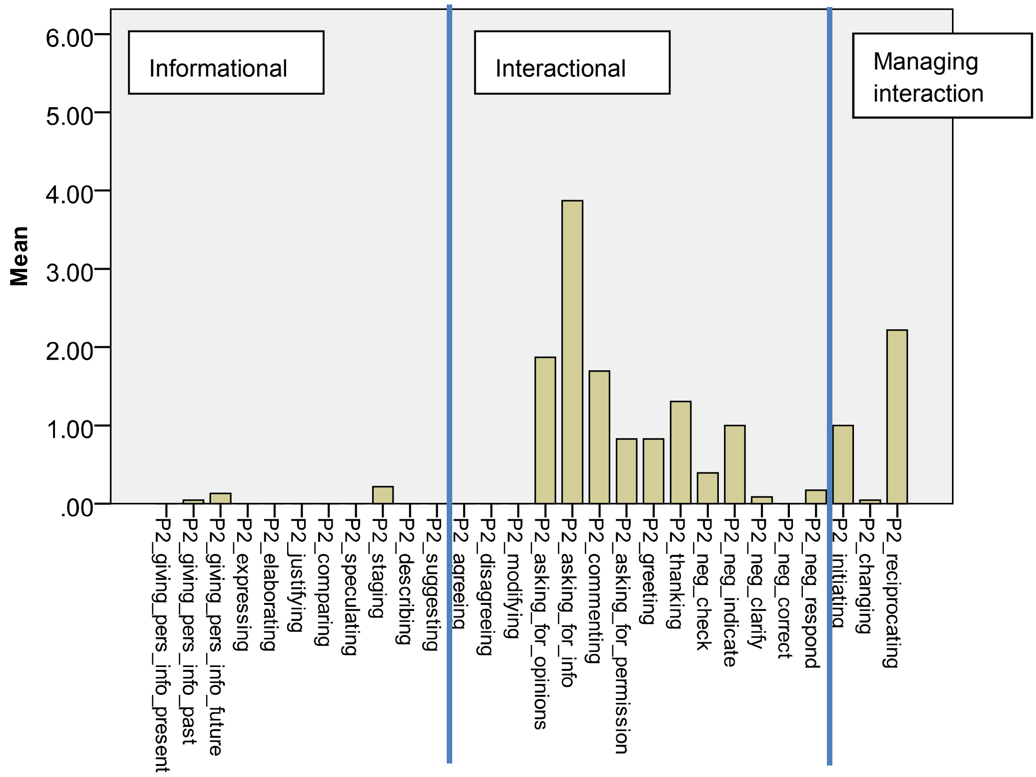
As the Part 2 task requires candidates to initiate an interview interaction with the interviewer, all candidates initiated an interaction without fail (1.00 turns on average), often by greeting (0.87 turns on average) and/or by asking for permission (0.83 turns on average), as shown in Example (7).

Example (7) Initiating / greeting / asking for permission

S1-7: *$H-$ hello, (0.6) may I ask (.) you some questions?*

Test takers also produced 1.87 turns on average which included instances of asking for opinions and 3.87 turns which included instances of asking for information, which are of course part of the task requirements and so numerous instances of these functions in this part were anticipated. A few examples are given in Example (8).

Figure 3: **Language functions elicited in Part 2 (role play)**



Example (8) Asking for opinions / asking for information
S2-8: *Ah: OK. (1.1) Oh, (1.0) mm (0.6) ah, (0.7) $Why do you wan-$ (0.7) why did you want ¥ to be (.) a teacher?*

S3-1: *Do you have any advice for future high schools (0.5) teachers.*

S2-7: *Yeah and (0.4) What subject do you teach?*

After the interviewer had answered these questions posed by the candidates, they sometimes commented on the interviewer's responses before moving on to the next question (1.70 turns on average), as illustrated in line 4 in Example (9) as well as line 5 of Example (10). They also indicated their understanding of what the interviewer said (1.00 turns on average), as in line 3 in Example (10).

Example (9) Commenting
1 S2-1: *Ah. (0.8) And uh (1.1) do you have (.) problem in the class?*
2 E: *Yes, students get sleepy [in the afternoon. ((laughs))*
3 S2-1: *[Ah.*
4→ S2-1: *$I-$ I always (0.4) sleep in the afternoon*

Example (10) Negotiating meaning (indicate understanding)
1 S3-5: *And (0.8) what subject (0.5) do you teaches?*
2 E: *Ah I teach English writing.*
3→ S3-5: *Oh, English writing.*
4 E: *Mm hm? =*
5 S3-5: *= It is very (.) hard. ((laughter))*

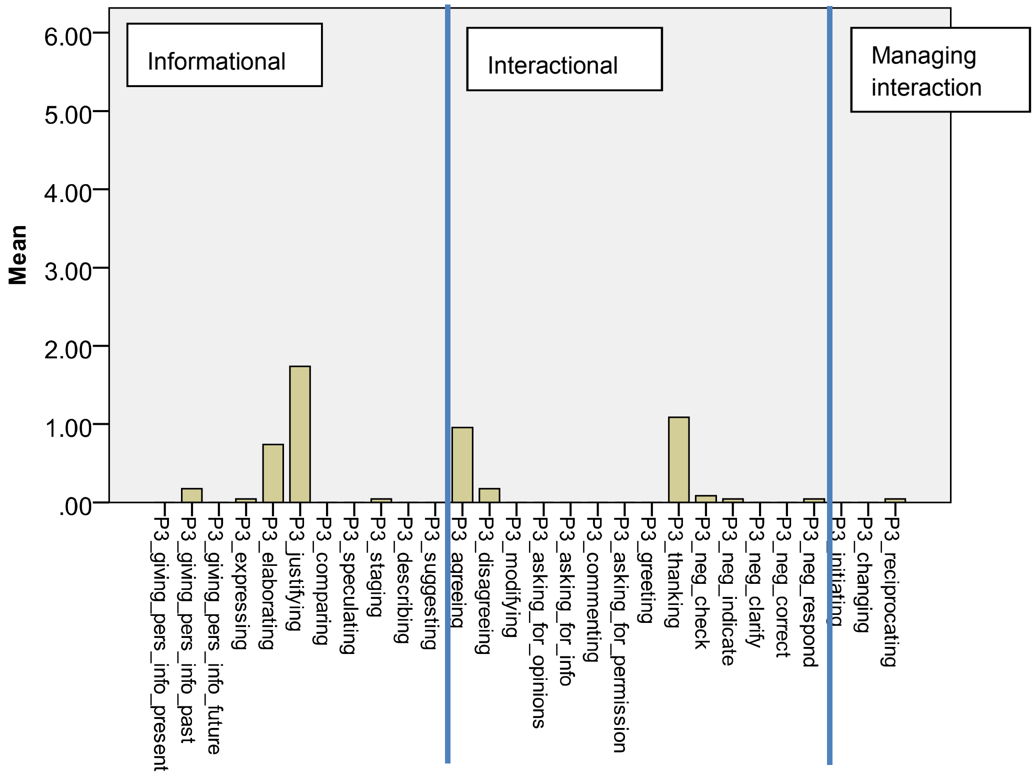
During or at the end of the role play, the participants also thanked the interviewer, who in fact was acting as an interviewee in this task (1.30 turns on average), and many of the candidates asked one or more extra questions to keep the conversation going. This is recognized as an attempt to share the responsibility for developing the interaction; in other words, "reciprocating" (2.22 turns on average). Example (11) shows a typical instance where the thanking and reciprocating functions were realized.

Example (11) Thanking / reciprocating
S2-4: *(1.1) Thank you. (0.6) {ah, ja} (.) ah: (.) do: you like ¥ to teach English?*

Part 3 (Monologue)

Part 3 (monologue) of the test, as illustrated in Figure 4, elicited a limited number of language functions. However, language to agree/disagree and to justify opinions expected from the task requirement was successfully observed.

Figure 4: **Language functions elicited in Part 3 (monologue)**



Example (12) illustrates a typical monologue produced by the participating candidates. They usually state their position first, whether they agree or disagree with the given statement (on average 0.96 turns included instances of agreeing, against 0.17 turns for disagreeing), as in line 1. This is usually followed by justifying opinions, as in lines 4 and 5 (1.76 turns on average), and the justification is often elaborated on, as in lines 7–10 and lines 14–17 (0.74 turns on average). This part usually finishes with the candidates thanking the interviewer when they return a task card, as in line 19 (1.09 turns on average).

Example (12): Typical monologue in Part 3

*1→ **S3-5:** Ahm, (0.4) I agree with this statement of (ahm) good ¥ to teach English in Japanese (.)*
2 elementary $student-$ schools,
*3 **E:** Mm hm,*
*4→ **S3-5:** ¥ Ah because ah: (.) many: countries like Korea and China started (0.8) ah: $studying (.) i-$*
5 ah teach and studying English, ah:, for (.) very very young,
*6 **E:** Mm hm,*
*7→ **S3-5:** (Years). Ah so (0.4) Japan $sh-$ (1.0) should $t-$ teach English ah: (0.4) too. Ahm (0.9) and (.)*
8 most of Japanese (0.4) can't speak English in ah (0.7) ¥ when start (0.7) working, and (.)
9 because of the (1.7) ah the (1.2) teach English (.) from (0.7) junior high school ¥ students is (0.4)
10 very late,
*11 **E:** Mm hm,*
*12 **S3-5:** And: (0.6)*
*13 **E:** Mm hm,*
*14→ **S3-5:** (1.9) Late for: (1.3) $to s-$ (0.5) to working. And (2.3) $(it's calculate as)$ (2.8) mm (0.9) mm:*
15 (0.5) Japanese many workers starts (0.8) nn ¥ studying English ah (0.6) six years later than (.)
16 many other country students, and (0.4) this is, ah, bad things $for:$ (1.8) about (.) English (.)
17 $to s-$ (.) about speaking or writing.
*18 **E:** Mm hm? OK, well, thank you very much. May I have the card back please.*
*19→ **S3-5:** ((gives card back)) Thank you.*

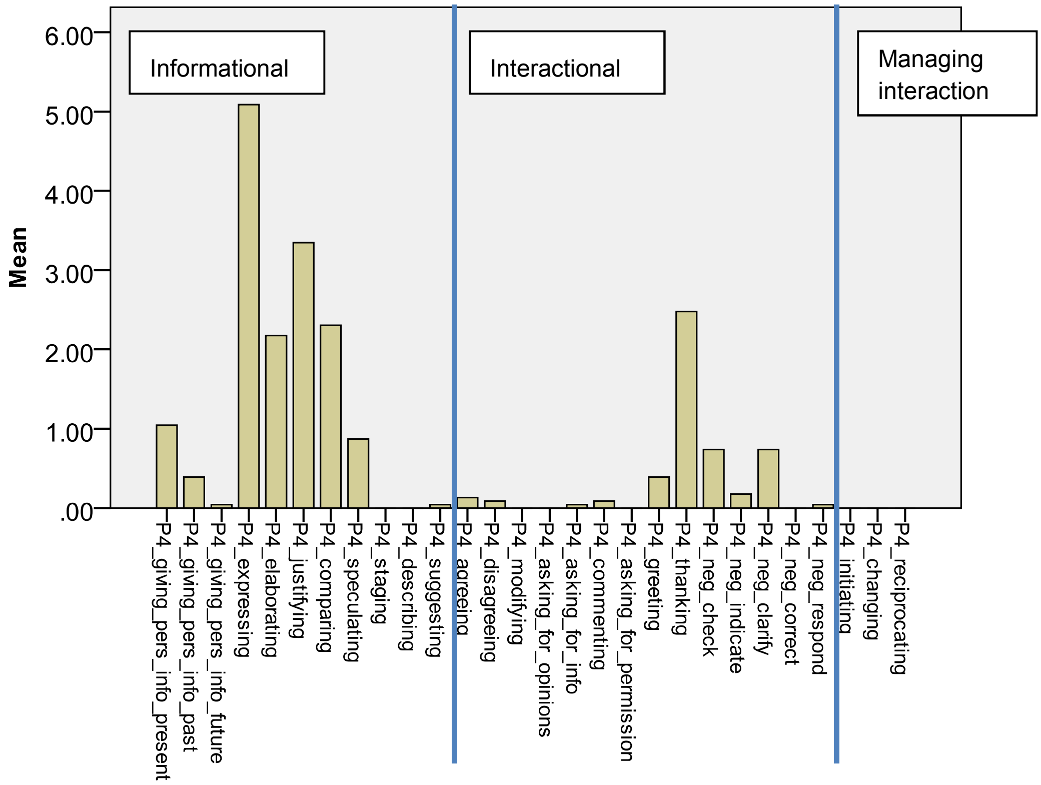
While the functions elicited in this part were satisfactory, it was noted that the amount of back-channeling provided by the interviewer varied from one session to another. This could have been influenced by how each candidate talked (e.g., the number and length of pauses during the discourse and the intonation with which each utterance ends). However, since the number of verbalized response tokens could influence raters' impressions of candidates' fluency when they rate video-recorded performances (Nakatsuhara, 2008), it is worth considering limiting the interviewer's contribution in this task to non-verbalized response tokens (such as nodding, smiling), unless the interlocutor's help is necessary. In fact, interlocutors were trained to follow a set script in the interlocutor frame. For Part 3, interlocutors were instructed only to listen and not to offer verbal feedback. Even in the interactive parts of the test, such as Part 4, the list of potential follow-up questions is tightly controlled. The results of this detailed analysis of the transcripts have thus provided important information to take into account in future training procedures to ensure fairness and consistency in testing procedures.

Part 4 (Extended Interview)

Figure 5 shows that Part 4 of the test elicited a number of *informational* functions, such as giving personal information (present), expressing opinions/preferences, elaborating, justifying opinions, comparing, and speculating. Test takers also negotiated meaning (checking understanding/asking for clarification) and thanked the interviewer, both of which are *interactional* functions.

Part 4 was the most effective in eliciting language to express opinions/preferences (5.09 turns on average), which is usually followed by justification(s) (3.35 turns on average). When they provided reason(s) for their opinions/preferences, they often speculated about a possible consequence (0.87 turns on average), as illustrated in Example (13), gave personal information about their present activities (1.04 turns on average), as in Example (14), made comparisons (2.30 turns on average), as in Example (15), and elaborated on their prior talk (2.17 turns on average), as in Example (16).

Figure 5: **Language functions elicited in Part 4 (extended interview)**



Example (13): Expressing opinions/preferences; speculating; justifying opinions

*1 **E:** =Now, I'd like to have us- to talk about the Internet and the media. (So), start? Should parents limit*
2 children's use of the Internet?
*3→ **S2-2:** (1.3) Uhm: yes, I think so. Because (0.9) $t- ahm: (0.4) if (0.4) if-$ if children (0.4) use their (.) so much*
4 time on the Internet, ¥ maybe they will not communicate to their parents, or their sisters or brothers
5 or their friends, ¥ so maybe it's

Example (14): Giving personal info. (present); justifying opinions

*1 **E:** I see. OK, let's go to the next question. Do you think social media such as Facebook and Twitter are*
2 changing the way people interact?
*3 **S1-4:** (1.5) ahm: (4.2) yes. ((laughs)) yes I think so (0.8) $because$ (0.8) ah (2.2) mmm (5.8)*
4→ well (2.1) so (1.7) me and my friends? (0.5) are (1.5) very involved hh $in [there –$ in this (0.8)
*5 **E:** [((laughs))*
*6 **S1-4:** mm social $madi-$ media? (1.1) and (4.1) at first, (0.7) $it (.) was (.) only::$ (.) it was (1.7)*
7 um only a social media? (0.9) mmm ¥ $to:$ (.) to kill time? [((laughs))
*8 **E:** [((laughs))*
*9 **S1-4:** hh but (0.9) ahm (1.6) but now, (0.8) mmm (1.4) $I- I feel:$ I feel ¥ that (0.7) mmm (0.9)*
10 $we: are-$ we are (0.5) very (1.1) involved ((laughs)) .hhh $in it in,$ in them (1.2) hmm (0.7) ((laughs))

Example (15): Comparing

*1 **E:** Thank you, I see. And, do you think reading newspapers is better than watching news on TV.*
*2→ **S2-4:** (0.4) Ah: (1.0) .hhh mmm, I: think (4.2) yes. Eh:, because (.) newspaper have (.) a lot of (2.0)*
3 {eh toh} (0.4) information than what (0.4) TV. (1.0) Eh (2.0) in case of TV, (0.4) {toh} (1.1) they have
4 a limit of time. (1.0) so: (1.2) they: limited (0.8) {toh} to (3.8) eh: ((laughs)) (.)
5 $con- converse.$ ¥ [converse the information.

Example (16): Elaborating
1 E: *Should parents limit children’s use of the Internet?*
2 S3-3: *(2.7) I think ¥ it depends on $their age$ the children’s age?*
3 E: *Mm?*
4 S3-3: *(0.6) (Because) (0.6) if (0.6) the children is already in high school, ¥ then (1.3) they (.) may (0.5)*
5 *manage their (0.5) selves. (0.9) But if they are (elementary) school,*
6 E: *Mm hm?*
7 S3-3: *(2.2) ¥ I think the parents (.) should limit the (.) use of Internet.*
8 E: *Ah OK. [All right=*
9 S3-3: *[$Because-$*
10 S3-3: *=Because sometimes they don’t know the (0.7) (payments) [of the (.) Internet,*
11 E: *[Mm,*
12→ S3-3: *(0.6) and if (0.4) they made a mistake and just $click one,$ one click, (1.2) ¥ I think ¥ (0.5)*
13 *it $m-$ might cause the payment (0.8) [ah I think ¥ you have to limit.*

Test takers negotiated meaning, especially when a new topic was introduced, by asking for clarification (0.74 turns on average) and by checking understanding (0.74 turns on average), as shown in lines 2 and 4 of Example (17), respectively.

Example (17): Negotiating meaning (checking understanding) (asking for clarification)
1 E: *Yeah. Um, should parents limit children’s use of the Internet.*
2→S2-1: *(2.0) I’m sorry, I couldn’t catch.*
3 E: *Should parents limit children’s use of the Internet?*
4→ S2-1: *(1.0) Parents limit?*
5 E: *Limit.*

Since Part 4 is the last task of the test, test takers often thanked the interviewer more than once (2.48 turns on average), as shown in Example (18). It should be noted, however, that two out of the three interviewers in this study reformulated the way the interviewer should terminate the test interaction, as in this example, although the interlocutor frame specifies how to end the test as “OK, thank you very much. This is the end of the test.”

Example (18): Thanking
1 E: *Ahm, XXX, this brings us to the end. [Thank you very much.*
2 S2-7: *[Yes.*
3→ S2-7: *Ah, thank you very much. ((gets up))*
4 E: *Thank you.*
5→ S3-7: *Thank you:*

In this section, we have discussed types of language function elicited via each part of the test, and how each function was actually realized. As summarized in Table 6, the data confirms that the types of function observed in each part are congruent with the goals of each part, fully covering the functions described in the draft test specifications. It was also encouraging to find evidence that targeted language functions were not only elicited but also elicited in ways that the test designers intended. This indicates that the intended constructs of the four tasks are appropriately operationalized.

Three minor modifications could be suggested based on the analysis:
1. Part 1: Pay attention to phrasing of questions to elicit candidates’ personal information in the present tense.
2. Part 3: Limit the interviewer’s contribution to only non-verbalized response tokens (such as nodding, smiling).
3. Part 4: Standardize the interviewer’s utterance to terminate the interaction at the end of Part 4.

Table 6: ***Comparison of Language Functions Targeted and Major Language Functions Elicited (Over 0.7)***

	Part 1		Part 2		Part 3		Part 4	
	Targeted	Elicited	Targeted	Elicited	Targeted	Elicited	Targeted	Elicited
Giving personal info. (present)	✓	✓						✓
Giving personal info. (past)	✓	✓						
Giving personal info. (future)	✓	✓						
Expressing opinions/preferences		✓					✓	✓
Elaborating		✓			✓	✓	✓	✓
Justifying opinions					✓	✓	✓	✓
Comparing							✓	✓
Speculating							✓	✓
Staging								
Describing a sequence of events								
Suggesting								
Agreeing					✓	✓		
Disagreeing					(✓)			
Modifying								
Asking for opinions			✓	✓				
Asking for info.			✓	✓				
Commenting			✓	✓				
Asking for permission				✓				
Greeting		✓		✓				
Thanking		✓		✓		✓		✓
Negotiating meaning - checking understanding								✓
- indicating understanding				✓				
- asking for clarification								✓
- correcting others								
- responding to a clarification request								
Initiating			✓	✓				
Changing								
Reciprocating				✓				

3.2 Linguistic and Discourse Features of Test Takers' Output in Relation to the Proficiency Levels of the Candidates (RQ2)

This section reports on an investigation of test takers' output language, in which the relationship between quantifiable linguistic and discourse features and test scores is examined to validate the descriptors used to define the levels on each rating scale.

However, before reporting the analysis of linguistic and discourse features, this section first of all reports how well the Study 1 ratings worked with the first draft of the speaking test rating scales and with the three newly trained raters for these scales. The purposes of the Study 1 score analysis are twofold:

1. To examine whether the Study 1 ratings were carried out satisfactorily enough to be relied on as a basis for test takers' proficiency levels, to which the linguistic features of their speech samples will be related in Section 3.2.2
2. To offer some scoring information to be combined later on with information obtained from the rater feedback questionnaire and the post-rating discussion, which will be reported on in Section 3.4

Since 3 out of the 23 students' speech samples were used as standardized exemplars, the remaining 20 students' scores were analyzed here. We should bear in mind that with such a small number of students and raters, we cannot draw any firm conclusions. Nevertheless, the results should still be useful to offer some indication about the draft rating scales and the efficacy of the training procedures.

3.2.1 Score Analysis

Multifaceted Rasch analysis was carried out using three major facets for the Study 1 score variance: examinees, raters, and rating categories. Figure 6 shows the overview of the results of the rating scale model analysis, 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 “measr” (measure), making it possible to directly compare all the facets. The more able examinees are placed towards the top (e.g., S1-2 is the most able) and the less able towards the bottom (e.g., S3-7 is the least able). 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 1 is the harshest examiner). The right-hand column, “scale,” refers to the levels of the rating scales.

Figure 6: **Overall facet map (Study 1)—Rating scale model analysis**

```
+-------------------------------------------------------------------------------------------------------+
|Measr|+Examinees           |-Raters|-Rating Categories                                             |Scale|
|-----+--------------------+-------+--------------------------------------------------------------+-----|
|   8 +                    +       +                                                              + (3) |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |     |
|     | 1-2                |       |                                                              |     |
|   7 +                    +       +                                                              +     |
|     | 2-2                |       |                                                              |     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |     |
|   6 +                    +       +                                                              +     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              | --- |
|   5 + 1-4  2-7  2-8  3-2 +       +                                                              +     |
|     |                    |       |                                                              |     |
|     | 3-5                |       |                                                              |     |
|     |                    |       |                                                              |     |
|   4 +                    +       +                                                              +     |
|     | 3-4                |       |                                                              |     |
|     | 1-3  1-6           |       |                                                              |     |
|     |                    |       |                                                              |     |
|   3 + 3-6                +       +                                                              +     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |  2  |
|     |                    |       |                                                              |     |
|   2 +                    +       +                                                              +     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |     |
|   1 +                    +       +                                                              +     |
|     |                    |       | Grammatical range and accuracy                               |     |
|     | 2-1                | R1    | Lexical range and accuracy                                   |     |
|     | 2-3                |       |                                                              |     |
*   0 * 1-7                * R3    * Fluency                                                      *     *
|     | 1-5  2-4           |       |                                                              | --- |
|     |                    | R2    |                                                              |     |
|     | 2-5                |       | Interactional effectiveness   Pronunciation                  |     |
|  -1 + 2-6                +       +                                                              +     |
|     |                    |       |                                                              |     |
|     |                    |       |                                                              |     |
|  -2 + 2-9                +       +                                                              +     |
|     |                    |       |                                                              |     |
|     | 3-7                |       |                                                              |     |
|     |                    |       |                                                              |  1  |
|  -3 +                    +       +                                                              + (0) |
|-----+--------------------+-------+--------------------------------------------------------------+-----|
|Measr|+Examinees           |-Raters|-Rating Categories                                             |Scale|
+-------------------------------------------------------------------------------------------------------+
```

Results for each of the three facets (i.e., examinees, raters, and rating categories) are briefly described in the following section. Apart from ensuring that the scoring system has functioned with an acceptable level of consistency and describing some notable findings on the relative difficulty of the five rating categories, we will not further interpret or elaborate on scoring results in Study 1. Scoring results in Study 2, based on a larger data set (see Section 4.1), will provide a fuller interpretation of the scoring results.

Examinees

The test was able to discriminate well between examinees. The fixed (all same) chi-square test was statistically significant ($\chi^2(19) = 448.4$, $p<.005$). The separation index was 4.54, and the examinees were able to be separated into 6.38 statistically separate strata. The person reliability, analogous to a Cronbach's alpha reliability estimate in a CTT analysis, was .95.

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 20 students, only 1 student was identified as misfitting (S3-7: infit mean square = 2.08; outfit mean square = 2.21). The percentage of misfitting students in the Study 1 data set was 5%, which seems acceptable for this small data set, although it is a little greater than the 2% that any test development should aim at (McNamara, 1996, p. 176).

Raters

As shown in Table 7, all the three raters showed quite good fit, indicating that they performed with a satisfactory degree of consistency.

Table 7: ***Study 1 Rater Measurement Report***

Rater	Fair Average	Measure	Real S.E.	Infit MnSq	Outfit MnSq
Rater 2	2.04	-.55	.22	.90	.86
Rater 3	1.98	.00	.24	1.17	1.13
Rater 1	1.91	.55	.23	.85	.80

While the analysis showed that the differences in the raters' degree of severity were statistically significant ($X^2(2) = 11.9$, $p<.005$), the severity range was very small. The difference between the harshest rater (i.e., Rater 1) and the most lenient rater (i.e., Rater 2) was only 0.13 of a band.

In terms of exact agreement of raw scores, raters showed exact agreement in 57.7% of the total cases (i.e., 173 agreements out of the total of 300 inter-rater agreement opportunities), and adjacent agreement (exact + plus/minus 1) was 100%. However, given that the scale has only four steps, we should hope for greater exact agreement.

Rating Categories

As illustrated in Table 8, none of the rating criteria was misfitting. This is an encouraging result, as this indicates that the assumption of uni-dimensionality holds for this data (Bonk & Ockey, 2003). In other words, the separate analytic rating scales seem to be contributing to a common construct of "speaking ability." This is important for the TEAP Speaking Test, which aims to provide a composite score by summing scores across the separate analytic scales.

Table 8: ***Study 1 Rating Category Measurement Report***

Rating Category	Fair Average	Measure	Real S.E.	Infit MnSq	Outfit MnSq
Pronunciation	2.07	-.74	.31	1.19	1.10
Interactional Effectiveness	2.07	-.74	.30	1.06	1.02
Fluency	1.97	.08	.29	.86	.84
Lexical range and accuracy	1.91	.57	.29	1.01	1.04
Grammatical range and accuracy	1.87	.83	.30	.75	.67

The analysis showed that the five rating categories exhibited different degrees of difficulty and these differences were statistically significant ($X^2(4) = 23.5$, $p<.005$). Judging from Table 8 and Figure 6, we can identify three groups of scale difficulty—with "grammatical range and accuracy" and "lexical range and accuracy" being the most difficult, followed by "fluency." "Pronunciation" and "interactional effectiveness" were the easiest categories.

Additional analysis for the rating categories was carried out using the partial credit model. Probability curves, shown in Figures 7-1 to 7-5, could be an indicator of whether the wording of each scale step (and also the training procedures for raters) worked as intended. For all rating categories, the rating scale steps progressed in the order as designed, with each step being progressively more difficult than the lower step on the scale. These figures also confirmed the above result of the rating scale model analysis, indicating that "pronunciation" and "interactional effectiveness" tended to be easier than the other categories (e.g., students at logit 0.0 are most likely to score 2 in the "pronunciation" and "interactional effectiveness" categories, although their chance to score 2 and 1 for the other categories is around 50% each).

Only one misfit value was found for all of the scale steps across all the rating categories. Pronunciation for a score point of 1 showed a misfitting outfit mean square value. To investigate this, all unexpected responses (with residuals greater than 2.0) were examined, and it seems that Rater 3 was awarding 1's for pronunciation when 2's were expected. This issue will be revisited in Section 3.4.2, where the post-rater discussion is reported.

To summarize the score analysis presented in this section, the scoring system generally worked well, demonstrating that all raters behaved with an acceptable level of consistency and, for all rating categories, the rating scale steps progressed in the expected order, indicating that the scales were interpreted by the raters, broadly, in the way intended. The information from the score analysis also provided insights into the interpretability of the rating scales, and in one instance this contributed to some minor adjustments in wording to the descriptors on one scale. This issue is discussed further in Section 3.4.2.

Identifying Test Takers at Three Levels of Proficiency

Having confirmed that the Study 1 ratings seemed to be of satisfactory quality, the 23 test takers were categorized into three proficiency levels for each rating scale category ("pronunciation," etc.): Level 1 (A2), Level 2 (B1), and Level 3 (B2). The above analysis demonstrated that all raters showed an acceptable level of consistency, and the adjacent agreement for all rating categories was 100%. As all three raters examined all speech samples in Study 1, each test taker's proficiency level on each analytical category was determined based on the mode across the three raters.

Figure 7-1: **Pronunciation scale**

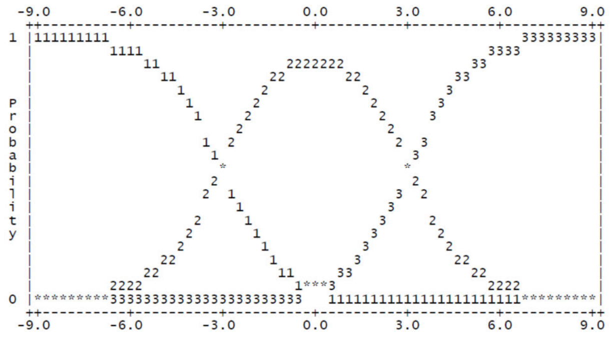


Figure 7-2: **Grammatical range and accuracy scale**

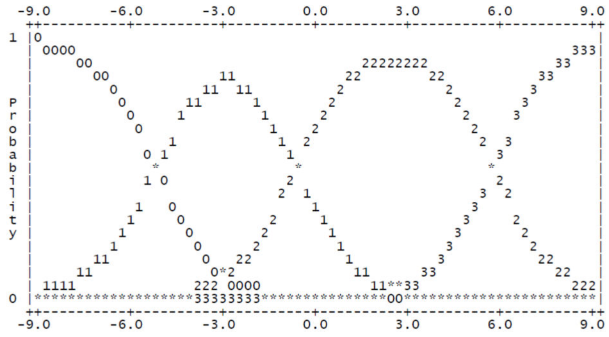


Figure 7-3: **Lexical range and accuracy scale**

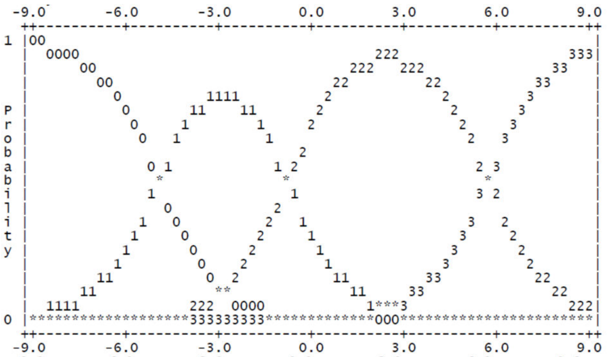


Figure 7-4: **Fluency scale**

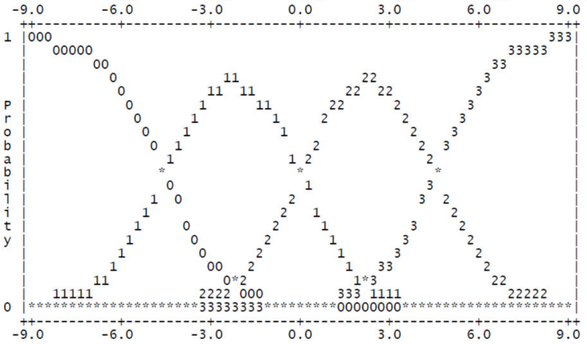


Figure 7-5: **Interactional effectiveness scale**

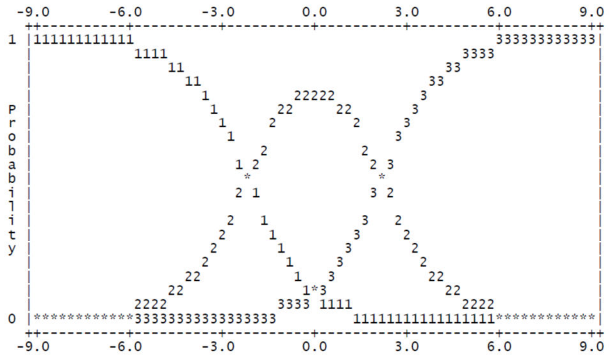


Table 9 shows the mode scores for each student on each rating scale category. It also shows the number of test takers who fell into Level 1, Level 2, and Level 3 for the five analytical categories. It was found that test takers were relatively well spread across levels for "fluency" and "interactional effectiveness." However, there were only a limited number of test takers in Level 1 on the "pronunciation" scale, in Level 3 on the "grammatical range and accuracy" scale, and on the "lexical range and accuracy" scale. As there was only one candidate who scored 1 for "pronunciation," Level 1 and Level 2 students were combined for this scale for the linguistic analysis. It should be noted therefore that the speech samples for these levels were not representative enough to make any firm conclusions. The results, however, are still useful for indicating possible trends. This is consistent with the purpose of the Study 1 trial, which was to acquire *a priori* validity evidence early on in the development process that might be useful in spotting problems early and making necessary adjustments before the larger-scale pilot test.

Overall scores, shown in Table 9, were simply computed by calculating the mean of all analytical scores and rounding up for scores at or above 0.5 of a level, and rounding down for scores below. It should be remembered here that these procedures were undertaken only in the context of collecting indicative *a priori* validity evidence from a rating plan that involved using three raters to rate all samples. The final form of feedback and the details of operational rating plans for the TEAP Speaking Test will, of course, be developed following the completion of all piloting, and will necessarily take into account a number of other factors impacting on scoring validity and test validity, including issues of practicality for test administration.

Table 9: ***Identifying Test Takers at Three Levels of Proficiency***

Examinee ID	Pronunciation	Grammatical Range and Accuracy	Lexical Range and Accuracy	Fluency	Interactional Effectiveness	Overall
S1-1	2	1	1	1	2	1
S1-2	3	3	3	3	3	3
S1-3	2	2	2	2	2	2
S1-4	3	2	2	2	2	2
S1-5	2	1	1	1	2	1
S1-6	2	2	2	2	3	2
S1-7	2	1	2	2	1	1
S2-1	2	2	1	2	2	2
S2-2	3	3	2	3	3	3
S2-3	2	2	2	1	1	2
S2-4	1	1	2	2	2	2
S2-5	2	1	1	1	2	1
S2-6	2	1	1	1	1	1
S2-7	3	2	2	3	3	3
S2-8	3	2	2	2	3	2
S2-9	2	1	1	1	1	1
S3-1	2	2	2	2	2	2
S3-2	2	2	2	3	3	2
S3-3	3	3	3	3	3	3
S3-4	2	2	2	2	2	2
S3-5	3	2	2	2	3	2
S3-6	2	2	2	2	2	2
S3-7	2	1	1	1	1	1
Level 1 (A2)	**1 (4.3%)**	**8 (34.8%)**	**7 (30.4%)**	**7 (30.4%)**	**5 (21.7%)**	**7 (30.4%)**
Level 2 (B1)	**15 (65.2%)**	**12 (52.2%)**	**14 (60.9%)**	**11 (47.8%)**	**10 (43.5%)**	**12 (52.2%)**
Level 3 (B2)	**7 (30.4%)**	**3 (13.0%)**	**2 (8.7%)**	**5 (21.7%)**	**8 (34.8%)**	**4 (17.4%)**

3.2.2 Linguistic and Discourse Analysis

Following Brown's (2006a) methodology, we first listed focused assessment areas for each rating category, and then reviewed the literature to identify linguistic features that could quantify the focused areas. It should be noted that the selected linguistic features merely relate to representative aspects of some of the focused areas. Due to practical constraints and a lack of objective measures appropriate for some key features, we were not always able to cover all key areas. However, it was hoped that the measures selected for each rating category would be broadly indicative in quantifying some related features for each rating category.

a. Grammatical Range and Accuracy

Focused assessment areas within the draft scale for grammatical range and accuracy describe a range and complexity of grammatical structures, frequency of grammatical errors, and their impact on the communication. To quantify test takers' output in relation to these two syntactic aspects, two measures of syntactic complexity and one measure of syntactic accuracy were selected.

The two measures for syntactic complexity were (1) the ratio of subordinate clauses to AS-units (analysis-of-speech unit) and (2) the number of words per AS-unit. They are both commonly used methods to examine the syntactic complexity of L2 learners' discourse (e.g., Brown, 2006a; Elder & Iwashita, 2005; Inoue, 2010; Nakatsuhara & Field, 2012; Tavakoli & Foster, 2008; Wigglesworth & Elder, 2010). An AS-unit is defined as "a single speaker's unit consisting of an independent clause, or sub-clausal unit, together with any subordinate clauses(s) associated with either" (Foster, Tonkyn, & Wigglesworth, 2000, p. 365), and the unit is made suitable for quantifying spoken language. There are specific rules about how we should take into account the fragmental nature of spoken language, and according to the rules, short response tokens such as "OK" and "all right" are counted as one unit.

For the former measure, we first cleaned the data set by removing all instances of repair, repetition, and false starts, and then manually counted the number of AS-units (Foster et al., 2000) and the number of subordinate clauses per each part session. The latter measure was added, because, in interactive oral discourse as in the current data set, the use of subordinate clauses is often rather limited compared with monologue oral production. It was therefore considered that the two measures would give supplementary information on the degree of syntactic complexity in the data set. Before measuring the number of words per AS-unit, we removed filled pauses, Japanese words, and all instances of repair, repetition, and false starts from the data set.

Syntactic accuracy was measured by the percentage of error-free AS-units. Variations of this measure have also been used by different researchers, like the percentage of error-free clauses (e.g., Tavakoli & Foster, 2008), error-free T-units (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008) and error-free utterances (e.g., Brown, 2006a). However, given the relevance of AS-units in segmenting interactional spoken language, we decided to use the percentage of error-free AS-units in this study. The data set was first cleaned of all instances of repair, false starts, and repetition, and errors were manually coded on the transcripts. Errors included syntactic and morphological errors such as tense markings and plural forms, word order, article usage, pronoun usage, and preposition usage. They did not include incorrect lexical choice.

To investigate the extent to which the syntactic complexity and accuracy of test-taker output differ in relation to their test scores on the "grammatical range and accuracy" scale, three measures of syntactic complexity and accuracy were compared between Level 1, Level 2, and Level 3 candidates. As mentioned above, since the sample size in each level is small, only descriptive statistics are carried out.

The results of these analyses are shown in Table 10 and Figures 8-1 to 8-3.

Table 10: *Syntactic Complexity and Accuracy*

Focus	Measure	Parts Applied	Level	N	Min	Max	Mean	Std. Dv.
Complexity	Ratio of subordinate clauses to AS-unit	1, 2, 3, 4	**Level 1 (A2)**	8	0.09	0.52	**0.26**	0.15
			Level 2 (B1)	12	0.13	0.59	**0.36**	0.13
			Level 3 (B2)	3	0.37	0.51	**0.45**	0.07
	Number of words per AS-unit	1, 2, 3, 4	**Level 1 (A2)**	8	3.29	6.09	**4.84**	0.93
			Level 2 (B1)	12	4.83	8.63	**6.64**	1.10
			Level 3 (B2)	3	6.07	8.02	**7.22**	1.02
Accuracy	Percentage of error-free AS-units	1, 2, 3, 4	**Level 1 (A2)**	8	0.45	0.71	**0.58**	0.09
			Level 2 (B1)	12	0.37	0.84	**0.54**	0.13
			Level 3 (B2)	3	0.59	0.75	**0.67**	0.08

Figure 8-1: **Number of subordinate clauses to AS-units (Parts 1–4)**

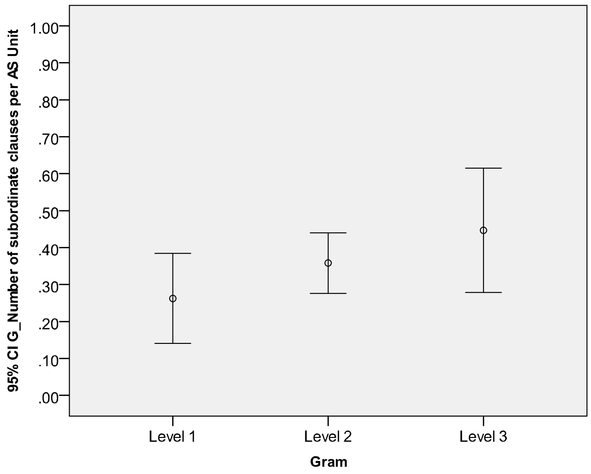


Figure 8-2: **Number of words to AS-units (Parts 1–4)**

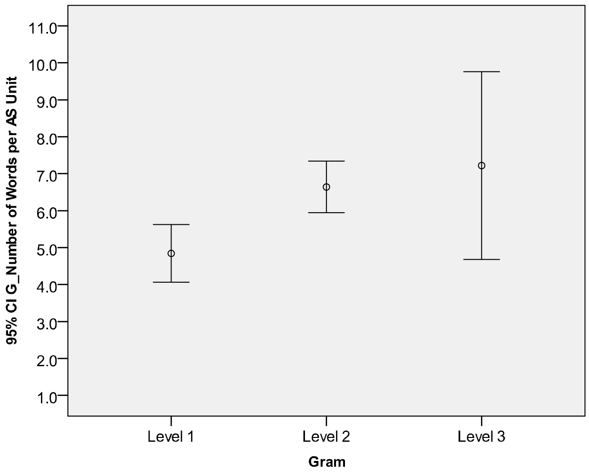
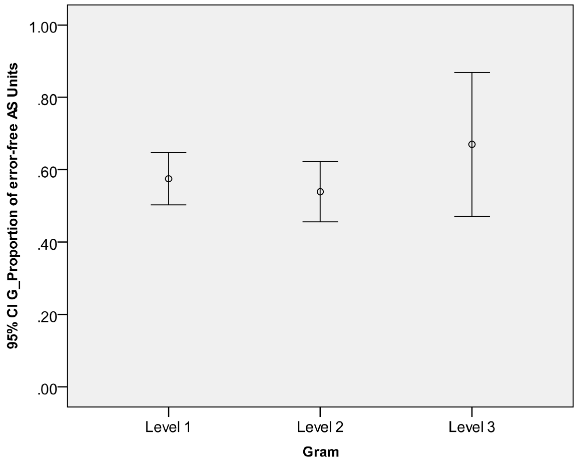


Figure 8-3: **Percentage of error-free AS-units (Parts 1–4)**



As expected from the rating results, the greatest use of subordinate clauses was observed in Level 3 candidates' output (on average 45% of the total AS-units), followed by Level 2 candidates (36%) and Level 1 candidates (26%). The number of words per AS-unit was also in the expected order: 7.22 words by Level 3 candidates, 6.64 words by Level 2 students, and 4.84 words by Level 1 students. It was noted that the number of words per AS-unit was in general remarkably small, but examinations of transcripts identified that this was due to their extensive use of repair, false starts, and repetition, which were all excluded from the analysis here (see Section 3.2.2c. for repair fluency), and due to the frequent use of short response tokens, which were counted as one AS-unit.

The average percentage of error-free AS-units between Level 3 and Level 2 students was also in accordance with the expected order—67% of the AS-units produced by Level 3 students were error-free, while 54% of them by Level 2 students were error-free. However, Level 1 students produced slightly greater accuracy than Level 2 students, with 58% of error-free AS-units. While this outcome may be surprising, this is in fact congruent with the literature that B1-level learners (i.e., Level 2 students in this study) tend to attempt more complex utterances at the expense of accuracy (Hawkins & Filipović, 2012).

b. Lexical Range and Accuracy

Within the draft scale for lexical range and accuracy, key features described are a range of vocabulary sufficient to deal with the full range of topics presented in the test as well as frequency and the impact of incorrect word choice.

An analysis of lexical frequency was carried out using the Range program based on the General Service List (GSL) (West, 1953) and the Academic Word List (AWL) (Coxhead, 2000). This follows the approach taken to investigate frequency profiles in test-taker output across different score bands of the IELTS test by Read and Nation (2006). In fact, two versions of the Range program are now available, one which uses the GSL/AWL lists and another which employs a 14-level list derived from the spoken corpora of the British National Corpus (BNC) developed by Nation (2006). The BNC version would, of course, provide a more finely turned vocabulary profile across more varied levels. However, it was decided to run the vocabulary analysis using the GSL version of the Range program, as this provided access to the AWL. The percentage of words in learner output covered by the AWL has been demonstrated to be a reliable indicator of the academic nature of written texts (Green, Unaldi, & Weir, 2010). As the TEAP is a test of academic English proficiency, this measure was deemed to be potentially relevant. However, as Nation (2006) points out in discussing the development of the 14-level BNC word lists, the one drawback of those lists is that the AWL words are scattered across different levels, and so this useful indicator is hidden.

After cleaning the transcripts by removing language used for filled pauses, non-word back-channeling signals (e.g., "uh huh"), false starts, repair, and repetition, the coverage of the most frequent 2,000 items in the General Service List[4] and the coverage of the words on the Academic Word List were examined on each participant's transcript on this whole part of the test. The analysis was undertaken on word types rather than word tokens, as the analysis aimed at finding out the lexical variation of the students across the three levels.

[4] The results presented here, however, should be interpreted with some caution, as off-list words in the given data also included words for new computer communication tools such as Facebook, which are familiar to the target students.

Results are presented in Table 11 and Figures 9-1 and 9-2. The most frequent 2,000 items covered, on average, 87.89% of the students' output language at Level 1, 86.32% at Level 2, and 85.71% at Level 3, indicating some progression in the use of less frequent words across the three levels. Similarly, Level 3 students used more academic words (4.69%) than Level 2 students (3.10%) and Level 1 students (2.61%). It is worth noting that, for the use of academic words, the difference between levels 3 and 2 was more salient than that between levels 2 and 1. This is congruent with the test designers' intention that Level 3 is the B2 level, at which learners start being more capable of coping with academic English.

Table 11: ***Lexical Frequency Coverage and Academic Word Coverage (By Word Type) Across the Three Proficiency Levels (%)***

Focus	Measure	Parts Applied	Level	N	Min	Max	Mean	Std. Dv.
Lexical range (type)	K1+K2 (%)	1, 2, 3, 4	**Level 1 (A2)**	7	83.54	94.26	**87.89**	3.60
			Level 2 (B1)	14	78.20	91.78	**86.32**	3.23
			Level 3 (B2)	2	80.46	90.96	**85.71**	7.42
	AWL (%)	1, 2, 3, 4	**Level 1 (A2)**	7	0.64	6.10	**2.61**	1.99
			Level 2 (B1)	14	1.69	4.59	**3.10**	1.04
			Level 3 (B2)	2	4.30	5.08	**4.69**	0.55

Figure 9-1: **Coverage of K1+K2 word types (Parts 1–4)**

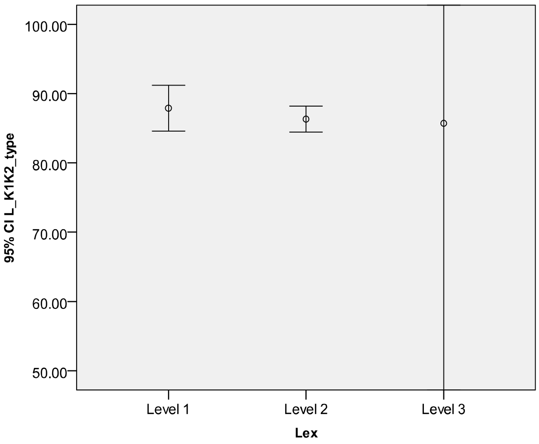
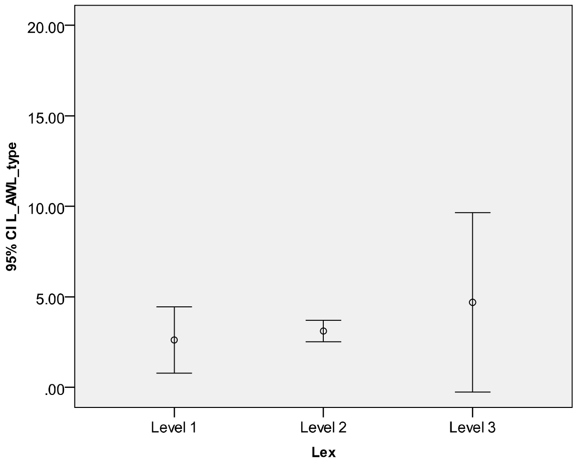


Figure 9-2: **Coverage of academic word types (Parts 1–4)**



Unlike the lexical range analysis, the examination of lexical accuracy turned out to be problematic. No measure was found in the literature for objectively judging lexical accuracy or appropriateness. It was attempted to identify incorrect words and possible target words that students wanted to produce. However, unlike morphological errors included in the grammatical accuracy analysis in Section 3.2.2a., it was hard to reliably identify word-choice errors. In most cases, although we were able to tell that a lack of vocabulary somewhat impeded communication at a discourse level, it was not possible to pinpoint which word was actually chosen wrongly. For instance, in the following example, where S1-1 tried to give a reason why she disagrees with the idea of teaching English in Japanese elementary schools, the word "patient" might have been incorrectly selected, but there is not enough context to judge it reliably, especially because she failed to make herself understood.

***S1-1:** (2.2) First: (.) $I:$ (1.8) I: disagree this: (.) statement (.) because $I: ehm I?$ (1.0) I was patient ¥ for uhm study English, and (1.1) $I'm very$ (1.4) I'm very very study hard English, and (1.8) $I think the Japanese: elementary school?*

Therefore, due to the difficulty in reliably quantifying lexical accuracy, we had to exclude the accuracy examination from the lexical analysis here.

c. Fluency

Key assessment features specified within the draft scale for fluency describe hesitation, disfluency features such as reformulation, and speech rate.

After reviewing the literature on measuring fluency (e.g., Brown, 2006a; Elder & Iwashita, 2005; Foster & Skehan, 1996; Inoue, 2010; Iwashita et al., 2008; Kormos & Dénes, 2004; Nakatsuhara & Field, 2012; Tavakoli & Foster, 2008; Wigglesworth & Elder, 2010), it was decided to use two measures for hesitation, one measure for disfluency, and two measures for speed fluency.

The two measures for hesitation were (1) the number of unfilled pauses per 50 words in all four parts and (2) total pause time as a percentage of speaking time in Part 3. The former was measured by the number of pauses of 0.3 seconds or longer that occurred after an examinee had begun speaking, divided by the number of words and multiplied by 50. The latter was calculated by the total length of pauses of 0.3 seconds or greater as a percentage of total speaking time for Part 3. Different researchers use different length of pauses to be counted; Iwashita et al. (2008) and Foster and Skehan (1996) counted pauses of 1 second or more, while Kormos and Denes (2004) counted pauses of more than 0.2 seconds and Tavakoli and Foster (2008) pauses of 0.4 seconds or more. For the present analysis, we decided to set our cutoff point as 0.3 seconds, as repeated listening of the recordings by the research group confirmed that intra-utterance pauses of 0.3 seconds and above were recognizable to listeners, while pauses shorter than 0.3 seconds did not seem to affect listeners' perception about test takers' pausing behavior.

Disfluency was measured by the total number of words coded as instances of repair, false starts, or repetition divided by the number of AS-units across the four parts. To do so, these disfluency features were firstly coded on the transcripts manually, as in the excerpts below between the two $ signs (i.e., "$I like:$," "$I was belonged to-$").

Example (1) Giving personal information (present)
***S1-1:** And hhh (1.6) $I like:$ (3.5) .hhh {iya-} (1.0) um, I like ¥ decorate (.) ((gestures with hands)) the classroom?*
***S2-2:** $I was belonged to-$ I belonged to: (0.5) choir- %you know choir?%*

Previous studies have used different formulations for disfluency analysis, such as the number of disfluency features per 100 words (e.g. Brown, 2006a) and the number of disfluency features per 60 seconds (Iwashita et al., 2008; Kormos & Denes, 2004) and the total number of disfluency features as they are (e.g., Wigglesworth, 1997). However, the present study used the ratio to AS-units, as it was considered to represent more accurately the extent to which repair (dis)fluency would affect the message conveyed by the candidates.

Speech rate and articulation rate were calculated as indicators of speed fluency. Speech rate was computed by calculating the total number of syllables divided by total speaking time including pauses, whereas articulation rate was calculated by counting the total number of syllables divided by total duration of pure speech time. These rates were measured only for Part 3 (monologue). This is because in interactional parts of the test (parts 1, 2, and 4), it is not possible or even desirable to determine the ownership of unfilled pauses between turns; that is, both conversants (i.e., interviewer and candidate) are responsible for such pauses unless the previous speaker nominates the next speaker (for example, by explicit questioning). Following Kormos and Denes (2004) and Inoue (2010), we pruned filled pauses, uncompleted single words, and non-words (including Japanese words) from the syllable count, and the total number of syllables was divided by the total speaking time with and without pauses, respectively, for Part 3, measured in seconds.

Table 12 and Figures 10-1 to 10-5 show the results.

The means of the three groups on all five fluency measures varied in accordance with the rating scores that candidates obtained. In terms of hesitation, the number of unfilled pauses (0.3 seconds and above) by Level 3 students was on average 10.89 instances, while Level 2 students had 17.44 instances and Level 1 students 22.44 instances. Total pause time as a percentage of speaking time in Part 3 monologue was 24% by Level 3 students, 36% by Level 2 students and 45% by Level 1 students.

The ratio of disfluency features such as repair, false starts and repetition to AS units clearly increased as the fluency scores decreased; Level 3 candidates showed on average one disfluency feature in four out of five AS-units (0.8), while Level 2 candidates had 1.27 features per AS-unit and Level 1 candidates had 1.43 features per AS-unit.

As for speed fluency, across the three proficiency levels, speech rate and articulation rate in Part 3 changed in the expected direction. Level 3 students on average uttered 1.68 syllables and articulated 3.02 syllables per second, Level 2 students uttered 1.71 syllables and articulated 2.88 syllables, and Level 1 students uttered 1.67 syllables and articulated 2.32 syllables.

Table 12: ***Fluency Measures Across the Three Proficiency Levels***

Focus	Measure	Parts Applied	Level	N	Min	Max	Mean	Std. Dv.
Hesitation	Number of unfilled pauses per 50 words	1, 2, 3, 4	**Level 1 (A2)**	7	13.95	32.32	**22.44**	5.71
			Level 2 (B1)	11	9.91	24.83	**17.44**	4.44
			Level 3 (B2)	5	6.69	18.02	**10.89**	4.23
	Total pause time as a percentage of speaking time	3	**Level 1 (A2)**	7	0.30	0.73	**0.45**	0.16
			Level 2 (B1)	11	0.20	0.60	**0.36**	0.13
			Level 3 (B2)	5	0.11	0.36	**0.24**	0.09
Disfluency	Ratio of repair, false starts, and repetition to AS-units	1, 2, 3, 4	**Level 1 (A2)**	7	0.86	2.19	**1.43**	0.53
			Level 2 (B1)	11	0.61	2.22	**1.27**	0.52
			Level 3 (B2)	5	0.53	1.08	**0.80**	0.19
Temporal	Speech rate	3	**Level 1 (A2)**	7	1.17	2.08	**1.68**	0.36
			Level 2 (B1)	11	0.42	2.73	**1.71**	0.65
			Level 3 (B2)	5	1.56	2.34	**1.96**	0.33
	Articulation rate	3	**Level 1 (A2)**	7	1.18	3.13	**2.32**	0.62
			Level 2 (B1)	11	2.15	3.23	**2.88**	0.35
			Level 3 (B2)	5	2.63	3.32	**3.02**	0.27

Figure 10-1: **Number of unfilled pauses (utterance initial) per 50 words (Parts 1–4)**

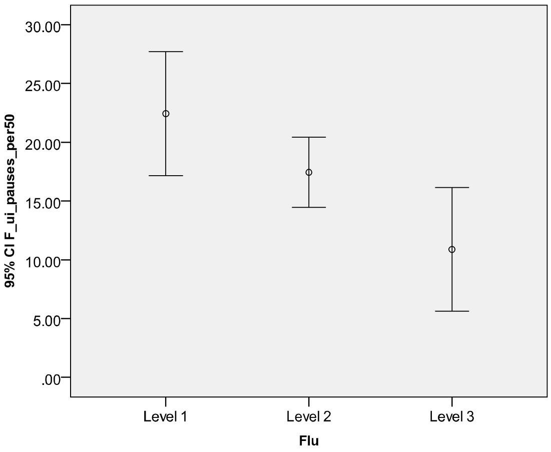


Figure 10-2: **Total pause time as a percentage of speaking time (Part 3)**

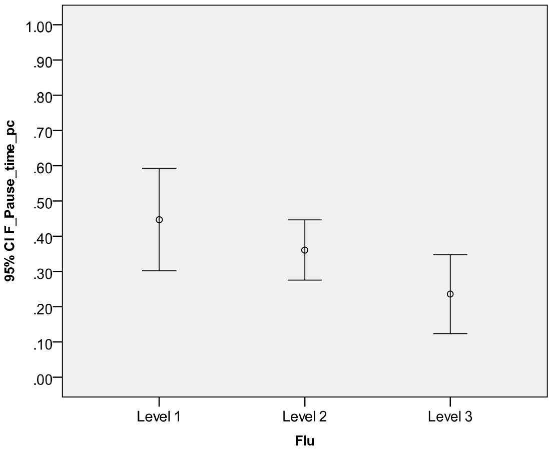


Figure 10-3: **Ratio of repair, false starts, and repetition to AS-units (Parts 1–4)**

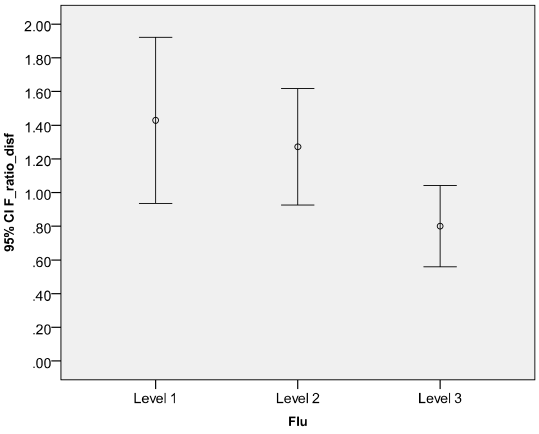


Figure 10-4: **Speech rate (Part 3)**

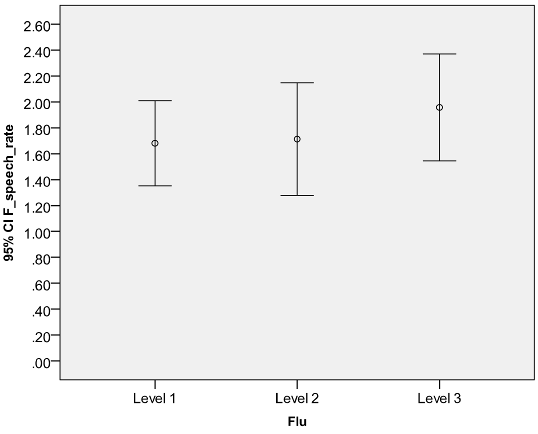
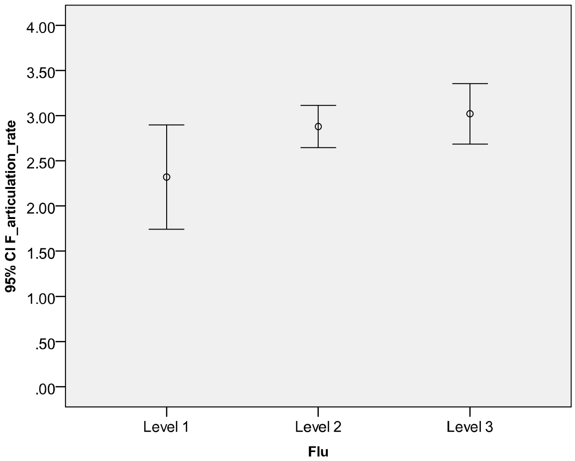


Figure 10-5: **Articulation rate (Part 3)**



d. Pronunciation

Key assessment areas specified in the draft "pronunciation" scale are intelligibility, prosodic features such as intonation, rhythm, word/sentence stress, assimilation/elision, and L1 influence.

Pronunciation was the hardest category to quantify. Iwashita et al. (2008) employed measures of phonology using specialists' judgements on different phonological features. Brown (2006a) dropped phonology from her analysis because of the difficulty of measurement. Post (2011), in her study focusing on L2 pronunciation features, used acoustic analysis software to analyze pronunciation features. The method used by Post is the most accurate way of quantifying pronunciation features, but it involves incredibly labor-intensive work by segmenting each phoneme and judging its discrepancy from, for example, received pronunciation (RP). Iwashita et al.'s method would be less labor-intensive, but it was not feasible for this study either. It still requires specialists to listen repeatedly to audio recordings of test takers' oral productions and to rate the appropriateness of each phonological feature while coding them on transcripts.

Instead, for the purpose of this study, we decided to measure only the quantity of L1-influenced (Japanese *katakana*-like) words, by counting the number of L1-influenced words as a percentage of total words produced. Words spoken with noticeable *katakana*-like pronunciation such as inserting extra vowels (e.g., [dogʊ] for [dog]), all syllables evenly stressed without using [ə], or L1-influenced consonants (e.g., [ɹ] for [l], [s] for [θ]) were coded on the transcripts. Examples include:

S1-1: $And::$ (1.8) ((laughs)) (.09) and: what's (.) your ah *problem* [pʊɹobʊɹemʊ] in class:.
S2-4: (.) Ah, I: enjoyed (1.5) *club* [kɹɑbʊ].

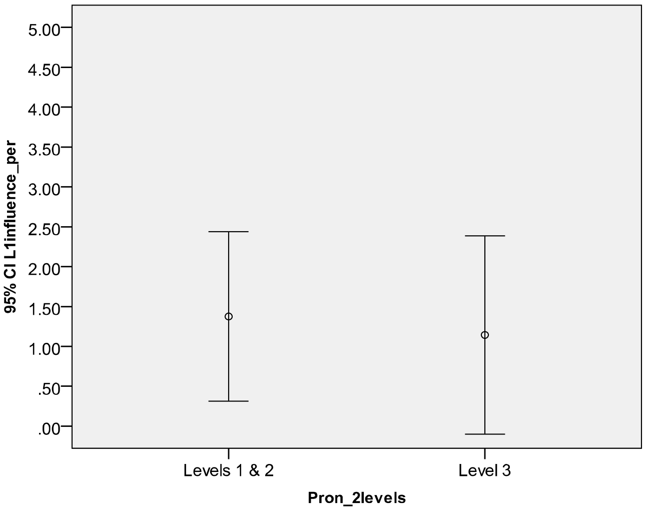
Although this covers only one of the key assessment aspects for the pronunciation scale, it was the only feasible analysis in this study. It was hoped that this would give us a rough indication for one of the pronunciation aspects.

Table 13 and Figure 11 show the results. As mentioned earlier, since there was only one student who scored a 1 in pronunciation, the analysis here combines levels 1 and 2 students together as one category and compares this category to Level 3 students. Level 3 students showed less L1 influence than level 1 and 2 students. While the percentage of words pronounced with L1 influence by Level 3 students was 1.14%, that by level 1 and 2 students was 1.38%.

Table 13: ***Pronunciation Measures Across the Three Proficiency Levels (%)***

Focus	Measure	Parts Applied	Level	N	Min	Max	Mean	Std. Dv.
L1 influence	Percentage of words pronounced with L1 influence	1, 2, 3, 4	**Level 1 + 2 (A2 & B1)**	1 + 15	0.00	8.00	**1.38**	2.00
			Level 3 (B2)	7	0.00	4.00	**1.14**	1.35

Figure 11: **Percentage of words pronounced with L1 influence (Parts 1–4)**



e. Interactional Effectiveness

Key assessment areas for the draft "interactional effectiveness" scale are effectiveness in participating in the interaction both actively and receptively (e.g., expanding and developing ideas and showing understanding of what the interlocutor said) and sensitivity to the turn-taking system in Part 2 of the test.

While there is no straightforward way of measuring interactional effectiveness by quantification, three measures were used to gain some indication for this scale. For the part 1 and 4 interviews, the length of responses was measured by average words per response. For the Part 2 role play, where candidates are required to ask the interviewer questions and to maintain the interaction, the number of extra questions and the number of instances of back-channeling and comments were counted.

As illustrated in Table 14 and Figures 12-1 to 12-3, a much larger difference in all three measures was observed between levels 3 and 2 than between levels 2 and 1. While the average number of words per response in parts 1 and 4 for level 1 and 2 students was 21.58 words and 21.81 words, respectively, the average number of words for Level 3 students was 34.90. Level 1 and 2 students on average asked 1.40 and 1.50 respective extra questions in Part 2, while Level 3 students asked on average 2.38 questions. Similarly, the numbers of back-channeling and comments in Part 2 by level 1 and 2 students were 6.80 and 6.70 times, respectively, while Level 3 students provided 9.63 response tokens, whether they were back-channeling or comments.

Table 14: *Interactional Effectiveness Measures Across the Three Proficiency Groups*

Focus	Measure	Parts Applied	Level	N	Min	Max	Mean	Std. Dv.
Length of response	Average words per response	1, 4	**Level 1 (A2)**	5	8.83	34.50	**21.58**	11.31
			Level 2 (B1)	10	6.93	43.56	**21.81**	11.18
			Level 3 (B2)	8	15.15	67.13	**34.90**	16.72
Number of extra questions	Number of separate questions asked that were not on required list in Part 2	2	**Level 1 (A2)**	5	1.00	2.00	**1.40**	0.55
			Level 2 (B1)	10	0.00	3.00	**1.50**	0.97
			Level 3 (B2)	8	1.00	6.00	**2.38**	1.69
Back-channeling and comments	Number of instances of back-channeling and comments in Part 2	2	**Level 1 (A2)**	5	4.00	13.00	**6.80**	3.83
			Level 2 (B1)	10	3.00	11.00	**6.70**	2.63
			Level 3 (B2)	8	2.00	20.00	**9.63**	6.26

Figure 12-1: **Average words per response (Parts 1 and 4)**

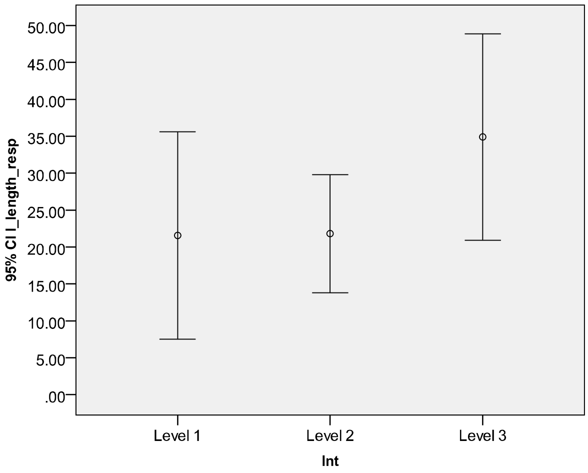


Figure 12-2: **Number of separate questions asked that were not on required list in the role-play task (Part 2)**

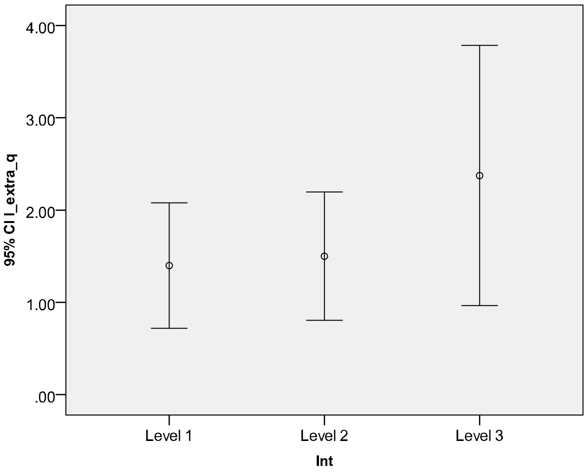
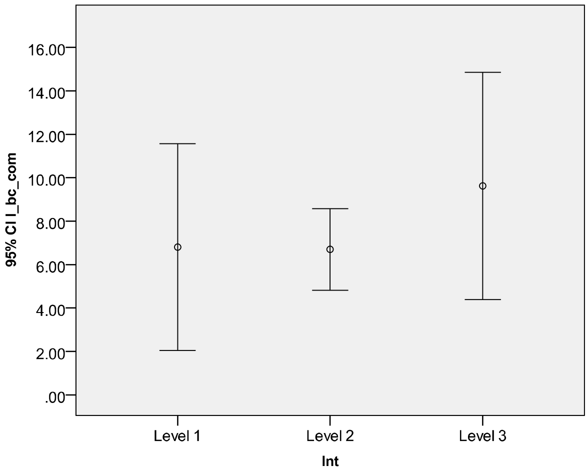


Figure 12-3: **Number of instances of back-channeling and comments in the role-play task (Part 2)**



f. Other—The Amount of Talk

Additionally, the quantity of production was also measured in three ways. These measures are not related to any of the rating scales. It was, however, thought to be important to confirm that the test was capable of eliciting a relatively similar amount of talk; in other words, equally ratable speech samples, from the three proficiency groups. For this analysis, overall scores were used to divide test takers into the three proficiency groups. For these production measures, filled pauses, Japanese words, and all instances of repair, false starts, and repetition were excluded from the analysis.

Results are presented in Table 15 and Figures 13-1 to 13-3. The results confirm that the test did not seem to show a great discrepancy among the three groups in terms of the effectiveness in eliciting contributions from students. As expected from the results of articulation rate and speech rate analyses, the average total number of words produced across all parts of the test by Level 3 students was the largest (458.75 words), followed by Level 2 students (434.08 words) and Level 1 students (417.71 words). However, the total number of AS-units in the whole test and the total number of words produced in Part 3 did not follow the same order. In general, it seems that the difference in the amount of speech produced by the three proficiency groups was not so great as to affect the ratability of speech samples. Nevertheless, it is worth noting that there were large individual differences within levels.

Table 15: ***Quantification of Production***

Focus	Measure	Parts Applied	Level	N	Min	Max	Mean	Std. Dv.
Length of long turn	Total number of words produced in Part 3	3	Level 1 (A2)	7	37.00	104.00	**74.14**	25.52
			Level 2 (B1)	12	17.00	178.00	**78.75**	42.67
			Level 3 (B2)	4	51.00	92.00	**71.50**	18.45
Total production	Total amount of production across all parts of the test, measured in words	1, 2, 3, 4	Level 1 (A2)	7	242.00	602.00	**417.71**	131.68
			Level 2 (B1)	12	164.00	890.00	**434.08**	192.99
			Level 3 (B2)	4	332.00	656.00	**458.75**	138.77
	Total number of AS-units produced across all parts of the test	1, 2, 3, 4	Level 1 (A2)	7	47.00	115.00	**72.71**	23.18
			Level 2 (B1)	12	35.00	111.00	**67.92**	19.52
			Level 3 (B2)	4	62.00	101.00	**77.50**	16.70

Figure 13-1: **Total number of words produced in the monologue task (Part 3)**

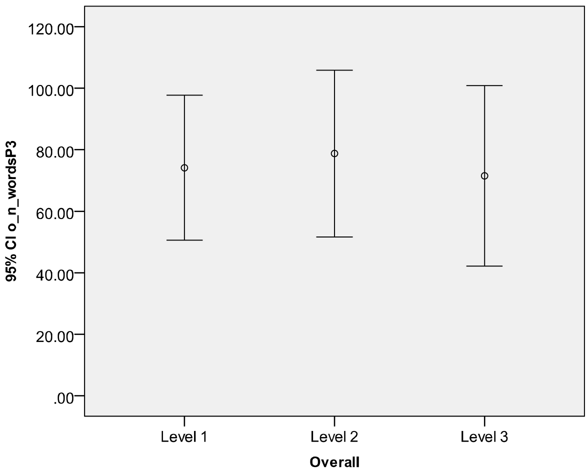


Figure 13-2: **Total amount of production across all parts of the test, measured in words (Parts 1–4)**

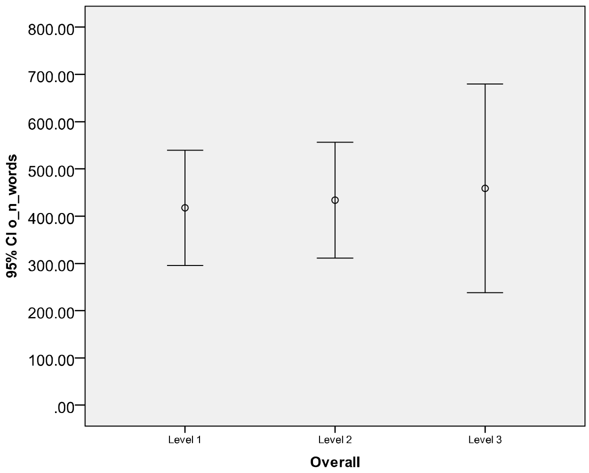
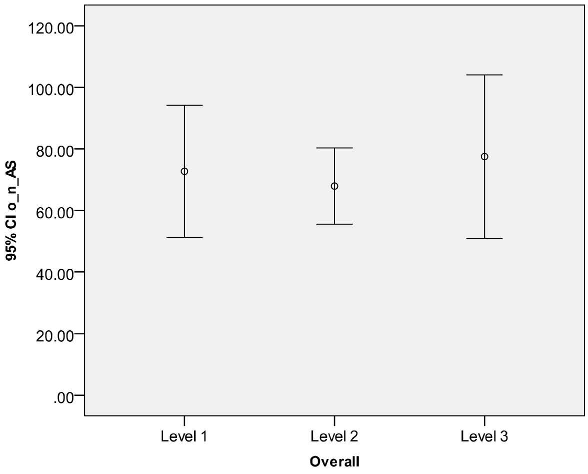


Figure 13-3: **Total number of AS-units produced across all parts of the test (Parts 1–4)**



3.2.3 Summary

Section 3.2.2 has quantified the linguistic and discourse features of test-taker output that are related to key assessment features specified in the draft analytical rating scales: “grammatical range and accuracy,” “lexical range and accuracy,” “fluency,” “pronunciation,” and “interactional effectiveness.” In general, all examined features of test-taker output varied according to the assessed proficiency level (Level 1, Level 2, and Level 3). All measures broadly exhibited changes in the expected direction across the three levels, providing the evidence that the rating scales are differentiating test takers’ performance in a way congruent with the test designers’ intention.

However, it was also worth noting that, for a few measures, the difference between two adjacent levels was not as expected. One of these examples is grammatical accuracy, suggesting a possible trade-off between grammatical accuracy and complexity for Level 2 students. Furthermore, for some scales, the differences between levels were greater at one boundary than the other; for example, a larger difference was observed between levels 3 and 2 than between levels 2 and 1 for the interactional effectiveness category. This result is in accordance with previous research (e.g., Brown, 2006a; Pollitt & Murray, 1996), indicating that specific aspects of performance are probably more relevant to differentiate particular levels. This finding is worth following up to better understand the nature of test-taker performance in the TEAP test. We should also bear in mind that it is necessary to replicate this study with a larger data set, as the small sample size of Study 1 did not allow us to use any inferential statistics. As can also be noted from examination of the error bars, many of the features examined, while demonstrating trends for the mean number of features in the expected direction across levels, demonstrated considerable overlap. Such results are also consistent with previous research in which small sample sizes make it difficult to achieve statistically significant results (Brown, 2006a). Once again, however, we would stress that the purpose of Study 1 was to acquire *a priori* validity evidence early on to confirm if the test developers intentions were being realized, and, if not, what changes might be required. Such information was intended to inform refinement of the specifications and scales before larger-scale piloting.

In addition to the linguistic measures related to the five rating categories, this section also analyzed overall production quantity across the three proficiency groups. The results confirmed that the test was capable of eliciting equally ratable samples from different proficiency groups.

Following the function and linguistic analyses, the next two sections (Sections 3.3 and 3.4) will examine the participating interlocutors', students', and raters' perceptions of the testing procedures.

3.3 Interlocutors' and Students' Feedback (RQ3)

3.3.1 Interlocutor Questionnaire

As mentioned in Section 2.3.2, after participating in an interlocutor training session, three interlocutors involved in the trial test filled out a short feedback questionnaire, using five-point Likert scales (1: strongly disagree, 2: disagree, 3: neither disagree or agree, 4: agree, 5: strongly agree).

Table 16 shows that the interlocutors generally found the training session useful, agreeing or strongly agreeing with all statements about different aspects of the training session. For their confidence in acting as an interlocutor after this training session (Q6), Interlocutor 1 rated 3 (neither agree nor disagree) and commented "need more practice," although he commented that "[he] thought it was a good session, nothing to add" for the final free response question (Q7). Interlocutor 2 suggested that "[trainers should] send materials in advance for people who want to read them" (Q7). Therefore, while the current training session does not seem to require major changes, it is worth considering sending an interlocutor handbook from which interlocutors can gain some prior knowledge before the actual face-to-face session, and including a few more practice sessions during the training session to ensure that all interlocutors will have full confidence in interviewing candidates after the training session.

Table 16: ***Interlocutor Questionnaire After the Training Session***

		Mean
Q1	I found the training session useful.	4.67
Q2	The interlocutor frame was clear.	4.33
Q3	The assessment procedures and criteria were clearly explained.	4.50 (1 missing)
Q4	The training video was helpful.	4.33
Q5	The practice test session during training was useful.	5.00
Q6	Having finished the training, I am confident in acting as an interlocutor in the live test sessions.	4.00
Q7	Do you have any suggestions to improve the training session?	-

The interlocutors were also asked to fill out a feedback questionnaire after Study 1 test sessions. A list of questions and their responses are summarized in Table 17. For all questions, there was free comment space for the respondents to provide comments or suggestions.

All three interlocutors in general felt the task timings, instructions, questions, and general test administration were appropriate. There were some suggestions for improvement and comments on each question, such as:

1. The question sequence should be more natural in the Part 1 interview ("Perhaps indicate the number of questions per task or find some way to make sequence more natural." [Q1.2, Interlocutor 2]).
2. The Part 2 role-play instructions should be clearer ("Some students didn't seem to understand that they were supposed to ask about the points mentioned." [Q2.3, Interlocutor 1]).
3. Part 4 follow-up questions were useful ("[Follow-up questions were] important for lower-level students." [Q4.3, Interlocutor 2]).
4. The use of a timer needs practicing ("[Throughout the test, keeping time was manageable], although I kept forgetting to start the clock." [Q5.1, Interlocutor 3]).
5. It is inevitable to deviate from the interlocutor frame in minor ways ("Not sure [how I deviated] but there were numerous times [when I needed to deviate.]" [Q5.3, Interlocutor 2]; "When students had difficulties, I used non-verbal body language and gestures." [Q5.3, Interlocutor 2]).

Table 17: ***Interlocutor Feedback Questionnaire After Study 1 Test Sessions***

	Questions	Response Options: Frequency
Part 1: Interview		
Q1.1	I found the time for the interview was	Too short: 0 Appropriate: 3 Too long: 0
Q1.2	I found the task instructions	Were appropriate: 2 Need changing: 1
Q1.3	I found the questions	Were appropriate: 3 Need changing: 0
Q1.4	I found the follow-up questions	Were appropriate: 3 Need changing: 0
Part 2: Role play		
Q2.1	I found the time for the role play was	Too short: 0 Appropriate: 3 Too long: 0
Q2.2	I found the task instructions	Were appropriate: 3 Need changing: 0
Q2.3	I found the task card	Was appropriate: 3 Needs changing: 0
Q2.4	I found the interlocutor's responses to the test taker's questions	Were appropriate: 3 Need changing: 0
Part 3: Monologue		
Q3.1	I found the time for the monologue was	Too short: 0 Appropriate: 3 Too long: 0
Q3.2	I found the instructions	Were appropriate: 3 Need changing: 0
Q3.3	I found the task card	Was appropriate: 3 Needs changing: 0
Part 4: Extended interview		
Q4.1	I found the time for the extended interview was	Too short: 0 Appropriate: 3 Too long: 0
Q4.2	I found the instructions	Were appropriate: 3 Need changing: 0
Q4.3	I found the questions	Were appropriate: 3 Need changing: 0
Q4.4	I found the follow-up questions	Were appropriate: 3 Need changing: 0
General test administration		
Q5.1	I found keeping time was	Easy: 0 Manageable: 2 Too difficult: 0 (missing: 1)
Q5.2	I found the distance between the interlocutor and the test taker was	Too close: 0 Appropriate: 2 Too far: 0 (missing: 1)
Q5.3	Did you have to deviate from the interlocutor frame: If so, how?	Yes: 2 No: 0 (missing: 1)
Q5.4	If you have any other comments on the test procedures, please write them down.	-

3.3.2 Student Questionnaire

As well as the interlocutors, test takers were also asked to fill out a feedback questionnaire on their test-taking experience immediately after they had completed their speaking test session. Table 18 shows the results. The test takers' questionnaire and the test takers' comments written in the spaces provided on the questionnaire were in Japanese. For the purposes of reporting results, the questions and comments cited below have been translated into English.

Table 18: ***Student Feedback Questionnaire Results***

Questions		Response Options: Frequency (%)
About the interviewer's English in general		
Q1	Did you find any parts of the interviewer's instructions unclear?	Yes: 1 (4.3%) No: 22 (95.7%)
Q2	Did you find any parts of the interviewer's questions unclear?	Yes: 6 (26.1%) No: 17 (73.9%)
About Part 2		
Q3	How did you perceive the length of preparation time given in Part 2?	Too short: 0 (0.0%) Appropriate: 22 (95.7%) Too long: 1 (4.3%)
Q4	Did you find it comfortable to attempt the Part 2 task that required you to ask questions?	Comfortable to attempt: 17 (73.9%) Uncomfortable to attempt: 6 (26.1%)
Q5	Do you think that the Part 2 task where you asked your teacher questions was relevant to your real-life English-use situation?	Yes: 22 (95.7%) No: 0 (0.0%) (missing value: 1 [4.3%])
Q6	Have you ever interviewed anyone either in Japanese or in English?	Yes: 14 (60.9%) No: 9 (39.1%)
About Part 3		
Q7	How did you perceive the length of preparation time given in Part 3?	Too short: 4 (17.4%) Appropriate: 17 (73.9%) Too long: 1 (4.3%)
Q8	Have you ever given a speech (short or long) in English?	Yes: 14 (60.9%) No: 9 (39.1%)
Q9	Do you think the Part 3 topic (English education at elementary schools) is relevant for third-year high school students?	Yes: 18 (78.3%) No: 5 (21.7%)
About Part 4		
Q10	Do you think the Part 4 topic (Internet and media) is relevant for third-year high school students?	Yes: 21 (91.3%) No: 2 (8.7%)
About test administration conditions		
Q11	What did you think about the distance between you and the interviewer?	Too close: 0 (0.0%) Appropriate: 22 (95.7%) Too far: 1 (4.3%)
Q12	In Part 2 and Part 3, you heard a beep sound of the timer to notify the beginning and the end of preparation time. Which do you prefer, having the beep sound or not?	Without beeps: 9 (39.1%) With beeps: 14 (60.9%)
Q13	As a whole, what did you think about the time allocated to the test taker's speaking time?	Too short: 3 (13.0%) Appropriate: 19 (82.6%) Too long: 1 (4.3%)
Q14	Did video recording distract your attention during the speaking test?	Yes: 0 (0.0%) No: 23 (100.0%)

It is encouraging that 95.7% of the students found all test instructions clear (Q1) and that 73.9% of them found all interviewer's questions clear (Q2). It seems that some lack of clarity about the interviewer's questions reported by six students was related to a lack of vocabulary (S1-1: "there was a word whose meaning I didn't know"), rather than ambiguity posed by vague questions. It is also worth noting that five out of those six students were Level 1 (A2) students, based on the scores awarded in this speaking test.

The length of preparation time in Part 2 was perceived as appropriate by 95.7% of the students (Q3). Since the Part 2 role-play task is a new test task type in the Japanese context, Q4 explored whether the participating students felt comfortable in attempting this task, which they had probably not experienced thus far. Of all the students, 73.9% found it comfortable to attempt the Part 2 task (Q4), giving positive comments such as "this task is innovative in giving test takers an initiative and I enjoyed it" (S1-6), while some students wanted to have more freedom in thinking about their own questions (S2-3: "I would have felt more comfortable if I had been asking my own questions only, rather than following a given list of questions"). Of all the students, 95.7% thought that the task reflected their real-life language-use situations (Q5). As many as 18 students elaborated on the usefulness of the Part 2 task for their English-use situations; for example, "in English classes, we often have to ask questions to native-speaker teachers in English, and it is in fact very important to be able to ask questions in English" (S1-4); "this is a good task because I sometimes hesitate to ask a question in English to native-speaker teachers, even when I have a question" (S3-1); and "whether to teachers or others, we have to ask questions in everyday life" (S2-5). Of all the students, 60.9% also had experience in interviewing either in Japanese or English (Q6).

The length of preparation time in Part 3 was recognized as appropriate by 73.9% of the students (Q7), and the task also seemed to be relatively familiar to them, as 60.9% had previously given a speech in English (Q8). Of all the students, 78.3% thought the topic (English education at elementary schools) was relevant for third-year high school students (Q9). Positive comments given on this question included "this topic has been often discussed in news" (S1-4); "after studying English for six years, everybody should have some opinions about this topic" (S1-5); and "this topic does not require any technical vocabulary" (S1-6), while some students thought "this topic might be too easy" (S2-2).

The Part 4 topic was perceived as relevant for third-year high school students by 91.3% of the participants (Q10). Many of them made comments like "Internet is part of our life" (S2-2) and "it is in fact an important issue to think about before starting university study" (S1-5). However, a concern was also raised that "not everybody uses or knows about Facebook and Twitter" (S3-7).

The physical distance between the interviewer and the candidate was perceived as appropriate by 95.7% of the students (Q11), and the candidate's speaking time overall was also perceived as appropriate by 82.6% of the students (Q13). It seems that students had split opinions about the beep sound of the timer. Of all the students, 60.9% thought that having the sound was good, as "it makes it very clear that I have used up 30 seconds" (S3-5). On the other hand, 39.1% of them disagreed that the sound was good, because "the timer sound made me nervous, and instead the interviewer can just let the candidate know that the time is over" (S2-3). Finally, it was very encouraging to find that none of the students thought that video recording distracted their attention during the speaking test (Q14).

3.4 Raters' Feedback (RQ4)

Like the interlocutors, three raters were also asked to complete feedback questionnaires after a rater training session and after rating Study 1 performances.

3.4.1 Rater Questionnaire

Rater Training Feedback Questionnaire

Table 19 summarizes the results of the rater feedback questionnaire after rater training.

Table 19: ***Rater Training Questionnaire***

	Questions	Mean*
Q1	I found the training session useful.	4.33
Q2	Watching the interlocutor training video and discussing the interlocutor frame before reviewing the rating criteria was useful as background information.	4.00
Q3	The rating criteria were clearly explained.	4.67
Q4	The standardized exemplars were good examples of the scoring categories for the different criteria.	4.00
Q5	The number of standardized exemplars (3) was sufficient to help me understand how to apply the rating criteria.	3.50 (1 missing)
Q6	Rating the standardized exemplars (2 & 3) and discussing the raters' scores before looking at the benchmark scores was useful practice.	4.33
Q7	Having finished the training, I am confident that I will be able to apply the rating criteria in rating samples of test-taker performance.	4.00
Q8	Do you have any suggestions to improve the training session	-

*1: strongly disagree – 5: strongly agree

In general, it seems the three raters found the training session useful and effective, and the session gave them confidence in using the rating criteria to assess test-taker performance. Their free responses were also mostly positive ones, but a few aspects to be improved were also identified as to:

- When the interlocutor training video should be shown ("In future, you may want to try a quick introduction of the scales before viewing the video. That may be more helpful." [Q2, Rater 2])
- The number of the standardized exemplars ("Good range. As always, more would have been better. I felt I was really getting it after explanation of the final exemplar." [Q4 & Q5, Rater 1])
- When to show the benchmark scores ("But I think it would be helpful to show benchmarks for the initial exemplar." [Q6, Rater 1])

Rating Feedback Questionnaire

The results of the rater feedback after the Study 1 test sessions are summarized in Table 20. Unlike their very positive feedback on the rater training session, it seems that they had some difficulties when it came to actually applying the scales to test-taker performance. For instance, Rater 1 reported that "for both the [grammatical and lexical] range categories, I often felt that the descriptors for two levels could describe the same speaker." The interactional effectiveness category was perceived as the most difficult to use, receiving comments such as "too long and confusing" (Rater 3) and "[fluency and interactional effectiveness categories] are very hard to rate, as they are the result of a holistic impression."

It was interesting to note that, in accordance with previous research (e.g., Brown, 2006b), raters in the current study also felt that pronunciation was distinct from the rest of the rating scales, while they felt that all the other categories had either some or significant overlap with each other. Rater 2, for instance, commented "All of these influence our perceptions of a person's ability in a foreign language. Hard to tease apart and look at separately."

Table 20: ***Rating Feedback Questionnaire***

	Rating Questionnaire	Responses
Q1	**The descriptors are easy to understand and interpret, when applying them to students' performance** (1: strongly disagree – 5: strongly agree)	**Mean**
Q1-1	Pronunciation	3.33
Q1-2	Grammatical range and accuracy	3.00
Q1-3	Lexical range and accuracy	3.00
Q1-4	Fluency	3.33
Q1-5	Interactional effectiveness	2.33
Q2	**The descriptors for each score point distinguish well between each of the levels of the scales** (1: strongly disagree – 5: strongly agree)	
Q2-1	Pronunciation	3.33
Q2-2	Grammatical range and accuracy	3.00
Q2-3	Lexical range and accuracy	3.33
Q2-4	Fluency	3.33
Q2-5	Interactional effectiveness	2.33
Q3	**How distinct are the scoring scales? Use the following three-point scale to describe the amount of overlap between different pairs of scales.**	**Frequency**
Q3-1	Pronunciation & grammatical range and accuracy	Very distinct: 3 Some overlap: 0 Significant overlap: 0
Q3-2	Pronunciation & lexical range and accuracy	Very distinct: 3 Some overlap: 0 Significant overlap: 0
Q3-3	Pronunciation & fluency	Very distinct: 2 Some overlap: 1 Significant overlap: 0
Q3-4	Pronunciation & interactional effectiveness	Very distinct: 3 Some overlap: 0 Significant overlap: 0
Q3-5	Grammatical range and accuracy & lexical range and accuracy	Very distinct: 0 Some overlap: 1 Significant overlap: 2
Q3-6	Grammatical range and accuracy & fluency	Very distinct: 0 Some overlap: 1 Significant overlap: 2
Q3-7	Grammatical range and accuracy & interactional effectiveness	Very distinct: 0 Some overlap: 1 Significant overlap: 2
Q3-8	Lexical range and accuracy & fluency	Very distinct: 0 Some overlap: 0 Significant overlap: 3
Q3-9	Lexical range and accuracy & interactional effectiveness	Very distinct: 0 Some overlap: 1 Significant overlap: 2
Q3-10	Fluency & interactional effectiveness	Very distinct: 0 Some overlap: 1 Significant overlap: 2

Table 20 (Cont'd): *Rating Feedback Questionnaire*

Rating Questionnaire		Responses
Q4	**The descriptors for each score point are:**	Too short: 0 Appropriate: 2 Too long: 1
Q5	**Was the quality of the videos sufficient for rating the speaking samples?**	Yes: 3 No: 0
Q6	**Did you need to watch the video samples more than once to rate them?**	Yes: 2 No: 1
Q7	**Does the test format provide a sufficient quantity of language to rate appropriately?**	Yes: 2 No: 1
Q8	**Does the format provide a sufficient sample of language to distinguish between the intended levels?**	Yes: 3 No: 0
Q9	**At what stage of the rating process did you finalize your mark for each category? (Free response)**	-
Q10	**Please describe the process or processes you followed when rating the samples?**	-
Q11	**If you have any other comments on the rating procedure or suggestions for improving the rating scales, please write them below. (Free response)**	-

Regarding the number of descriptors for each score point (Q4), raters' perceptions varied; Rater 1 thought "maybe too many overall," while Rater 3 thought "compared to the CEFR descriptors, in most cases they are shorter—more succinct." Two out of the three raters needed to watch the video samples more than once to rate them, but "not in their entirety" (Rater 1), just to check part of the performance (Q6). Rater 1 did "not feel that the role play was effective" in providing a sufficient quantity of language to rate appropriately (Q7). However, overall it was encouraging that all three raters felt the format provided a sufficient sample of language to distinguish between the intended levels (Q8). The stage of the rating process when the three raters finalized their mark for each category (Q9) was different; their responses were "after the interview was completed" (Rater 1); "at the end of the sample usually, sometimes pronunciation earlier" (Rater 2); and "depended on the student" (Rater 3). The process(es) that the three raters followed also varied. Raters 1 and 2 had similar processes: "I changed my system as I went through videos; I started off taking lots of notes and working with three levels by writing them and then circling as each threshold was made or failed" (Rater 1) and "hypothesis tested for each category during each part of the test; made final decision at end" (Rater 2). By contrast, Rater 3 had a fixed order in rating: "Pronunciation, lexis, grammar, fluency, and finally interactional effectiveness."

3.4.2 Focus Group Discussion

As noted previously, the three raters participated in a post-rating focus group session to discuss reasons for choosing the scores they gave. This was to identify key performance features that might have influenced raters' decisions and to examine whether these salient features to the raters were congruent with the key assessment areas for each rating category to which raters should be paying attention. A researcher in Eiken acted as a facilitator in the focus group discussion.

As mentioned in Section 2.3.3, three speech samples were selected for the discussion; two speech samples on which raters generally agreed (Student 2-2 at Level 3 [B2] and Student 3-6 at Level 2 [B1]) and one for whom there was significant disagreement (Student 2-3 at Level 1 [A2] or Level 2 [B1]). During the meeting, the raters watched the video for each of the three test takers again and

the video was paused after each task to allow for discussion. The facilitator asked questions related to items in the feedback questionnaire.
All raters, as described in Section 2.3.1, were experienced teachers at Japanese universities but with different levels of experience as a professional rater in standardized speaking tests. One did not have any experience (Rater 1), another had two to three years' experience of rating the EIKEN speaking tests (Rater 2), and the other had five years' experience in IELTS, BULATS, and EIKEN speaking tests (Rater 3). Furthermore, all the three raters had previously participated in extensive training in the CEFR for a separate standard-setting project relating the EIKEN speaking tests to the CEFR, enabling them to discuss the TEAP Speaking Test scales with the related CEFR levels in mind.

To describe briefly the three raters' characteristics identified during the training and discussion sessions:

- **Rater 1:** At the beginning of the training session, Rater 1 described himself as tending to prefer a holistic rather than analytic scoring model. However, he quickly adapted to the analytic TEAP scale and, according to the results, was no more prone to assigning flat profiles than any other rater. At the end of the discussion, he expressed the view that the analytic scales helped him to identify different features of performance that would not have been as evident based on a holistic scale.
- **Rater 2:** During training, Rater 2 demonstrated a commitment to following the descriptors exactly as they were written and not allowing personal preferences or past experience to color his interpretation.
- **Rater 3:** During training and discussion sessions, Rater 3 expressed a strong preference for holistic scoring. Although this tended not to significantly affect his rating patterns overall, it would serve as an explanation for why his scores for pronunciation occasionally misfit the model. This was further borne out in post-rating discussion, as described below.

The Study 1 score analysis presented in Section 3.2.1 demonstrated that all raters behaved with an acceptable level of consistency and, for all rating categories, the rating scale steps progressed in the expected order, indicating that the scales were interpreted by the raters, broadly, in the way intended. However, it was also noted that "pronunciation" and "interactional effectiveness" tended to be easier than the other categories. The results also showed that there were some unexpected ratings on the pronunciation scale by Rater 3, who was awarding 1's for pronunciation when 2's were expected. The post-rating discussion was instructive for interpreting these results and informative to decide whether any modifications would be necessary for the "pronunciation" and "interactional effectiveness" scales.

In the post-rating discussion, raters rewatched and discussed one of the unexpected examples of test takers to which Rater 3 had given a 1 for pronunciation. During the discussion, Rater 3 mentioned that after listening to the reasons Rater 1 and Rater 2 gave for awarding a 2 for pronunciation to this test taker, he felt that on reflection the test taker in question was in fact "intelligible" and so would not deserve a score of 1 according to the wording of the descriptor.

The issue of *intelligibility* is central to the construct of pronunciation in these scales, and so careful attention will need to be paid to explaining and illustrating this concept for raters in the future. This case may also give some hints on how to avoid problems in the future by giving clearer instructions on the purpose of analytic rating scales. In fact, Rater 3 several times mentioned during the discussion that he preferred to form a holistic impression by letting the test takers' production "wash over him" without immediately referring to individual scales. It is possible that Rater 3 formed a holistic, general impression of overall effectiveness of the test taker without always attending to the specific wording of the rating scale for pronunciation. Thus, Rater 3's assigning of lower-than-expected pronunciation scores for some test takers was likely the result of an overall impression influencing his individual analytic scores. The problem with unexpected scores for pronunciation might be mitigated somewhat in the future by anticipating this kind of behavior in some raters and giving due attention to pointing out the importance of the key concepts to be attended to in the separate analytic scales.

"Pronunciation" and "interactional effectiveness" tended to be easier than the other scales. Given this result, the development team considered if the interpretation of the scales was appropriate, or whether raters were misinterpreting the scales (or being misled by inappropriate wording) and were awarding inappropriately high scores to performances.

It was decided that changes to wording in the pronunciation scale to increase difficulty did not seem to be necessary for the reasons described below. As mentioned above, the issue of *interpretability* is central to the construct of pronunciation as defined in this rating scale. A lot of the discussion with raters, both during training and post-rating, concerned the interpretation of "impeding communication." Raters felt that they were perhaps able to understand the test takers because of their experience in Japan and familiarity with Japanese test takers. They were concerned about whether this would also apply to "unsympathetic" or "naïve" listeners (they thought it would not). The TLU domain for the TEAP has been defined as the EFL context of Japan. Students will be interacting in this context with instructors who are obviously familiar with their students and, as such, familiar with pronunciation features that are typical of Japanese EFL learners. Given the target context, it was felt that it was appropriate to judge students' pronunciation on the basis of raters' own ease of understanding that pronunciation. In fact, the majority of test takers in this sample were able to be understood in terms of their pronunciation. If intelligibility only in terms of individual sounds and stress patterns is taken into account and not the grammatical accuracy or the coherence of the message (which is what is intended for this analytic scale), then these test takers are generally intelligible and so would receive a 2 for pronunciation (which is relevant to the B1-level description of pronunciation in the CEFR).

The case of interactional effectiveness was a little different from that of pronunciation. Part of the reason that interactional effectiveness became, in effect, easier seemed to stem from the interaction between the descriptors and the actual task for Part 2. Raters pointed out that the rather simplistic nature of the questions test takers must ask in the Part 2 role play, when combined with the original wording for a 2-point performance on the interactional effectiveness scale, would make it very unlikely that even A2-level test takers would receive less than 2 (supposedly a B1-level performance) for this particular category. It was felt that this was resulting in inappropriately high scores being given to performances which would not otherwise merit the B1 level. We did not feel that revisions to the actual task, which is meant to be accessible to the majority of test takers, were warranted. On reflection, we have identified some modifications to the wording of the descriptors for this category, which may in fact make the interpretation clearer, and also make it slightly more difficult to receive a 2. We felt this change would be consistent with our interpretation of what a B1-level learner should be able to produce, particularly given the controlled nature of the task in Part 2.

Based on the rater discussion, we therefore decided to modify the wording of the interactional effectiveness scale, but to keep the wording of the pronunciation scale as it was. Bonk and Ockey (2003) and Nakatsuhara (2009) suggest that differences in scale difficulty do not automatically mean that changes need to be made. This is particularly true when the scales are meant to have a criterion-referenced focus. In this case, the scales are built on and meant to be referenced to the CEFR levels. In the CEFR, pronunciation that does not impede communication is unambiguously B1, which is a 2 on our scale. Therefore, it was decided that the pronunciation scale did not need major revision to make it more difficult, and that the pronunciation construct as defined by the CEFR levels is indeed appropriate for this context. However, this may have implications for providing a composite score. The issue of effective weighting of the scales will need to be investigated further.

The rater discussion was also useful to confirm our decision of not employing part-scoring. As mentioned earlier, during the development of the draft specifications, the test development team initially considered using part-scoring, in which raters mark individual tasks separately rather than assigning marks for test takers' performance across the test as a whole. On the one hand, the part-scoring system has the advantage of focusing attention on the features of performance relevant to each task. As shown in the language function analysis in Section 3.1, the four parts of the test were designed to target different language functions, and learners therefore might display differential abilities according to the nature of each task. Furthermore, part-scoring was also initially considered to be feasible in the case of the TEAP as the TEAP raters would be assigning marks to video-recorded performances, thus enabling them to focus only on rating. They can also watch the recorded performance more than once, if necessary.

Nevertheless, the part-scoring system in fact attracted our concerns in the mini-trial test stage in Study 1, and a tentative decision was made prior to Study 1 that raters assign their marks for performance across the test as a whole. The reasons for this decision were outlined earlier. The function analysis indeed confirmed the impression that some parts are highly controlled and unlikely to elicit a sufficient sample for independent scoring (e.g., only questioning and commenting in Part 2). The decision to take the limitations on elicited speech samples into account when considering whether to use part-scoring or overall scoring is in accordance with one of the caveats expressed by Taylor and Galaczi (2011, p. 187) regarding this issue.

This observation in the mini-trial stage was further confirmed by the rater discussion in Study 1 here. Although raters in Study 1 were not asked to rate each part separately, when the discussion facilitator asked about the raters' rating process(es) in each part, such as "anything in there (Part 2) that might have modified the initial impression?," the unfeasibility of using parts 1, 2, and 3 to constitute ratable speech samples independently was referred to; for example, "not enough language production in this part to move someone's score up" (Rater 1). At the same time, the potential for each part to contribute different kinds of information in the gradual accumulation of evidence over the whole test to justify a final level distinction was also confirmed. An example is provided below, in which the raters are discussing the usefulness of the gradual increase in difficulty and difference in focus across the different parts of the test in probing the upper limits of candidates' performance.

> R1*) As soon as she entered this section she seemed to hit a wall.*
> R2*) I think the test is successful at doing that.*
> R3*) Part 3 and 4—you could actually see it.*
>
> *R1) At the end of Part 1, I had the idea that she'd moved out of 1, gone into safe 2 territory; felt that it'd be unlikely to move beyond a 2.*
> ***J) How about the other categories beyond pronunciation?***
> R1*) Across the board—confirming my hypothesis.*

3.5 Summary of Study 1 and Modifications Made to the Test Materials and Rating Scales

Study 1 investigated various aspects of context validity and scoring validity of the TEAP Speaking Test so as to collect information on the extent to which the test materials and rating scales operationalized the test construct described in the draft test specifications. The investigations included the analysis of functional and linguistic features of test takers' output language, test scores, feedback questionnaires from test takers, interlocutors, and raters, and a post-marking focus group discussion of raters. Based on the findings, several modifications were suggested for different aspects of the test. The main findings and suggested modifications in each analysis are summarized below, before moving on to reporting on Study 2.

a) Language Functions Across the Four Parts of the Test

The analysis confirmed that the types of function observed in each part were congruent with the goals of each part, fully covering the functions described in the draft test specifications. It was also encouraging to find evidence that targeted language functions were not only elicited but also elicited in ways that test designers intended. A few minor suggestions were made to rephrase one of the Part 1 questions and to standardize interviewer behavior in parts 3 and 4.

b) Linguistic Features Across the Three Proficiency Levels of Candidates

Linguistic features of test-taker output were quantified in relation to key assessment features specified in the five draft analytical rating scales. In general, all examined features of test-taker outputs varied according to the assessed proficiency level (Level 1, Level 2, and Level 3), providing evidence that the rating scales are differentiating test takers' performance in a way congruent with the test designers' intention.

While no modification was suggested at this point, it was pointed out that further research with a larger sample size is necessary to offer a better understanding of some results that require elaboration (e.g., a possible trade-off between grammatical accuracy and complexity, a possibility of specific aspects of performance being more relevant to differentiate particular levels).

c) Interlocutor Training/Post-interviewing Questionnaires

All three interlocutors in general found the interlocutor training session useful, and felt that the task timings, instructions, questions, and general test administration were appropriate. Some suggestions were made about the clarity of the test instructions, the need for more practice with the use of a timer, and guidelines for what interlocutors should do when they feel a need for deviating from the interlocutor frame in minor ways.

d) Student Feedback Questionnaire

Students in general perceived the test content and the test procedures positively. The Part 2 role-play task where they were required to ask the interlocutor a series of questions was especially received positively, which confirmed the use of this innovative task in the Japanese context. The use of the beep sound of a timer seems to require monitoring. Some students felt that the increased formality caused by the sound made them nervous, although the majority preferred to have the sound. It was very encouraging to find that none of the students thought that video recording distracted their attention during the speaking test.

e) Test Scores, Rater Training/Post-marking Questionnaires, and Post-marking Rater Discussion

In general, the three raters found the training session useful and effective, and the session gave them confidence in using the rating criteria to assess test-taker performance. Some modifications were suggested regarding the provision of a rater handbook prior to the training event and the need for more standardized exemplars to practice during the training session.

Analysis of the test scores demonstrated that the rating of the video-recorded performance was carried out with a satisfactory level of consistency, and the rating scale seemed to function as the test designers intended. The post-rating questionnaire and focus group discussion were very useful in interpreting some aspects of the score analysis. Some adjustments to the wording of descriptors in the interactional effectiveness scale were suggested. It was also noted that, in future rater training sessions, raters need to be explicitly instructed that an overall impression should not influence their individual analytic scores, especially on the pronunciation scale.

All of these sources of empirical validity evidence described above offered useful information to verify or modify the draft test specifications, test materials, and rating scale descriptors to be used in Study 2. All the modifications suggested were further discussed by the project team, and revised rating scales and revised test materials were prepared for Study 2.

The rest of the report will describe Study 2, which focused mainly on scoring validity, to confirm that changes made after the trial test functioned in ways that the test designers intended.

4. Results and Discussion: Study 2

4.1 Test Scores (RQ5)

As in Study 1, multifaceted Rasch analysis was carried out using three major facets for the score variance in this study: examinees, raters, and rating categories. The scores for the 113 students who participated in the pilot were subject to the analysis here. The partial credit model was used for the analysis.

Figure 14 shows an overview of the results of the partial credit analysis, plotting estimates of examinee ability, examiner harshness, and rating scale difficulty (see Section 3.2.1 for an interpretation of the overall facets map).

Figure 14: Overall facet map (Study 2)—Partial credit model analysis

```
Vertical = (1A,2A,3A,S) Yardstick (column lines low high extreme)= 0,2,-10,9,End

+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Test takers                                                            |-Raters     |-Rating categories                                          | S.1 | S.2 | S.3 | S.4 | S.5 |
|-----+------------------------------------------------------------------------+------------+------------------------------------------------------------+-----+-----+-----+-----+-----|
|   9 + 2029                                                                   +            +                                                            + (3) + (3) + (3) + (3) + (3) |
|     | 1008  2024                                                             |            |                                                            |     |     |     |     |     |
|   8 + 1012                                                                   +            +                                                            +     +     +     +     +     |
|     |                                                                        |            |                                                            |     |     |     |     |     |
|   7 + 1001  1002                                                             +            +                                                            + --- +     +     +     +     |
|     | 1022                                                                   |            |                                                            |     | --- | --- |     |     |
|   6 + 3014                                                                   +            +                                                            +     +     +     + --- +     |
|     |                                                                        |            |                                                            |     |     |     |     | --- |
|   5 + 2019  4016  4017                                                       +            +                                                            +     +     +     +     +     |
|     | 3025                                                                   |            |                                                            |     |     |     |     |     |
|   4 + 4019                                                                   +            +                                                            +  2  +     +     +     +     |
|     | 4022  4029                                                             |            |                                                            |     |     |     |     |     |
|   3 + 2005  2017  2018  3011  3024  4018  4021  4024                         +            +                                                            +     +  2  +  2  +  2  +     |
|     | 4009                                                                   |            |                                                            |     |     |     |     |  2  |
|   2 + 1007  2011  3007  3013  3022  3026  3029  4020  4030                   + R1         +                                                            +     +     +     +     +     |
|     | 1015  1019  1023  1025  1026  2002  2007  2030  3005  4007  4015  4028 |            |                                                            |     |     |     |     |     |
|   1 + 1003  1014  1017  2012  3016  4014                                     +            + Fluency                                                    +     +     +     +     +     |
|     | 1021  1024  2022  4005                                                 |            | Grammatical Range and Accuracy  Lexical Range and Accuracy | --- |     |     |     |     |
*   0 * 1013  1018  1030  2014  4006  4010  4012  4013  4026                   * R3  R4  R6 *                                                            *     *     * --- * --- * --- *
|     | 2003  3002  4027                                                       | R2         | Interactional Effectiveness                                |     | --- |     |     |     |
|  -1 + 1006  1009  1027  2006  2016  2026  3003  3008  3028  4001  4023  4025 +            + Pronunciation                                              +     +     +     +     +     |
|     | 1020  2004  2010  2015  2020  2023                                     |            |                                                            |     |     |     |     |     |
|  -2 + 1005  1010  1011  1029  2001  2013  2021  3015  4008  4011             + R5         +                                                            +     +     +     +     +     |
|     | 3010  3019                                                             |            |                                                            |     |     |     |     |     |
|  -3 + 1004  3001  3006  3009  3017  3018  3023                               +            +                                                            +     +  1  +  1  +  1  +  1  |
|     |                                                                        |            |                                                            |  1  |     |     |     |     |
|  -4 + 3021  4003  4004                                                       +            +                                                            +     +     +     +     +     |
|     | 2027                                                                   |            |                                                            |     |     |     |     |     |
|  -5 + 4002                                                                   +            +                                                            +     +     +     +     +     |
|     | 3012                                                                   |            |                                                            |     |     |     |     | --- |
|  -6 + 3030                                                                   +            +                                                            +     + --- +     + --- +     |
|     | 2009                                                                   |            |                                                            |     |     | --- |     |     |
|  -7 +                                                                        +            +                                                            +     +     +     +     +     |
|     |                                                                        |            |                                                            | --- |     |     |     |     |
|  -8 +                                                                        +            +                                                            +     +     +     +     +     |
|     |                                                                        |            |                                                            |     |     |     |     |     |
|  -9 +                                                                        +            +                                                            +     +     +     +     +     |
|     | 3020                                                                   |            |                                                            |     |     |     |     |     |
| -10 +                                                                        +            +                                                            + (0) + (0) + (0) + (0) + (0) |
|-----+------------------------------------------------------------------------+------------+------------------------------------------------------------+-----+-----+-----+-----+-----|
|Measr|+Test takers                                                            |-Raters     |-Rating categories                                          | S.1 | S.2 | S.3 | S.4 | S.5 |
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
S.1: Model = ?,?,1,R4  ; Rating categories: Pronunciation
S.2: Model = ?,?,2,R4  ; Rating categories: Grammatical Range and Accuracy
S.3: Model = ?,?,3,R4  ; Rating categories: Lexical Range and Accuracy
S.4: Model = ?,?,4,R4  ; Rating categories: Fluency
S.5: Model = ?,?,5,R4  ; Rating categories: Interactional Effectiveness
```

Examinees

The test was able to discriminate well between examinees. The fixed (all same) chi-square test was statistically significant (χ^2 (112) = 2094.9, p<.005). The separation index was 4.00, and the examinees were able to be separated into 5.67 statistically separate strata. The person reliability, analogous to Cronbach's alpha in a CTT analysis, was .94. 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, as in the Study 1 analysis, we follow Wright and Linacre's (1994) suggestion that infit mean square values in the range of 0.5 to 1.5 are "productive for measurement." Out of the 113 students analyzed, 3 students were identified as misfitting (S3012: infit mean square = 2.39, outfit mean square = 3.00; S2020: infit mean square = 2.15, outfit mean square = 2.57; S3011: infit mean square = 2.36, outfit mean square = 1.87). The percentage of misfitting students in the data set was 2.7%. We in fact anticipated having more misfitting examinees, who have jagged profiles across the five rating criteria, in the recognition that speaking has not been taught or studied as systematically as the other three skill areas in the English education system in Japan. These three misfitting students indeed showed a strong performance in certain categories and still had distinct individual weaknesses in other areas, which was picked up as "misfitting" in the analysis. Nevertheless, it was encouraging to find the ratio of misfitting students in Study 2 was only 2.7%, almost satisfying McNamara's (1996, p. 178) expectation that any test development should aim at having misfitting students at or below 2%.

Raters

As shown in Table 21, all the six raters showed quite good fit, indicating that they performed with a satisfactory degree of consistency.

Table 21: *Study 2 Rater Measurement Report*

Rater	Fair Average	Measure	Real S.E.	Infit MnSq	Outfit MnSq
Rater 5	1.93	-2.04	0.15	0.95	1.07
Rater 2	1.66	-0.35	0.14	0.86	0.82
Rater 3	1.61	-0.15	0.15	0.69	0.60
Rater 6	1.54	0.15	0.15	0.96	0.88
Rater 4	1.52	0.22	0.17	1.27	1.29
Rater 1	1.12	2.17	0.15	1.12	1.11

In terms of exact agreement of raw scores, raters showed exact agreement in 59.7% of the total cases (i.e., 1,587 agreements out of the total 2,660 inter-rater agreement opportunities), which was slightly better than the Study 1 result (57.7%). While we could hope for better exact agreement, particularly using the broad CEFR levels as the basis for our rating scale, the figure is still respectable in terms of what is often seen in the literature. Adjacent agreement was in fact 100%, which would normally be taken to be an excellent result, but of course the same caveat regarding the broad steps of the scale as made in regards to exact agreement applies to the interpretation of adjacent agreement.

The analysis showed that the six raters differed in terms of severity, and these differences were statistically significant ($X^2(5)$ = 398.2, p<.005). The severity range was rather small for four out of the six raters: Rater 2, Rater 3, Rater 6, and Rater 4. However, the range was widened by the harshest and most lenient raters, who deviated a little from the rest of the raters. The difference between the harshest rater (Rater 1) and the most lenient rater (Rater 2) was 0.81 of a band. We suggest that these two raters should be retrained, so that their severity levels will be closer to the other raters.

Rating Categories

As illustrated in Table 22, none of the rating criteria was misfitting. This is an encouraging result, as it indicates that the assumption of unidimensionality holds for this data (Bonk & Ockey, 2003). As mentioned in the Study 1 score analysis, this means that the separate analytic rating scales seem to be contributing to a common construct of "speaking ability." This is vital for the TEAP Speaking Test, which aims to provide a composite score by summing scores across the separate analytic scales.

Table 22: ***Study 2 Rating Category Measurement Report***

Rating Category	Fair Average	Measure	Real S.E.	Infit MnSq	Outfit MnSq
Pronunciation	1.74	-1.09	0.16	1.25	1.33
Interactional effectiveness	1.72	-0.44	0.13	1.04	1.01
Lexical range and accuracy	1.50	0.29	0.14	0.78	0.74
Grammatical range and accuracy	1.57	0.34	0.14	0.94	0.90
Fluency	1.32	0.90	0.13	0.86	0.83

The analysis showed that the five rating categories exhibited different degrees of difficulty, and these differences were also statistically significant ($X^2(4) = 114.8$, $p<.005$). "Fluency" was the most difficult category, followed by "grammatical range and accuracy," "lexical range and accuracy," and "interactional effectiveness." "Pronunciation" was the easiest category.

Here, we should note that, based on the Study 1 score analysis and post-marking rater discussion, the "interactional effectiveness" scale was amended to make each step on the scale slightly more demanding. It was decided not to change the "pronunciation" scale. The results in Study 2 described above demonstrated that the amendments seemed to have worked as intended. The modifications made on the "interactional effectiveness" scale successfully made the scale more difficult, and "pronunciation," as expected, remained less demanding, with the lowest average difficulty measure on the common logit scale out of the five rating scales.

Probability curves for each category were also examined, as to whether the wording of each scale steps (and also the training procedures for raters) worked as intended. As shown in Figures 15-1 to 15-5, for all rating categories, the rating scale steps progressed in the order as designed, with each step being progressively more difficult than the lower step on the scale.

Figure 15-1: **Pronunciation scale**

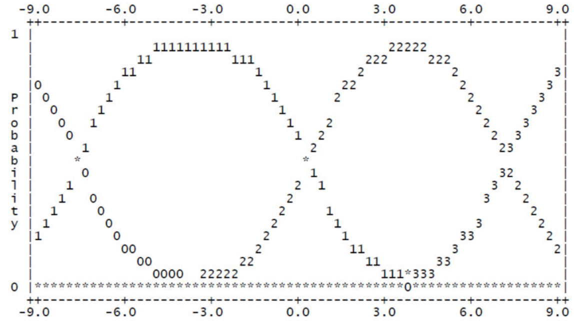


Figure 15-2: **Grammatical range and accuracy scale**

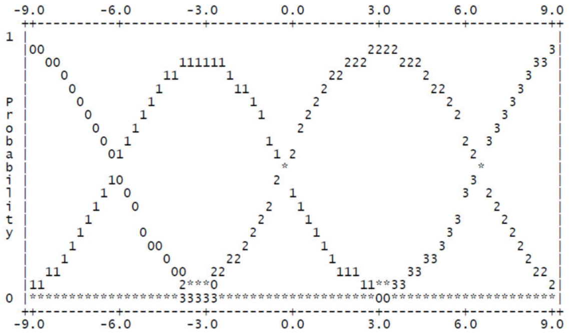


Figure 15-3: **Lexical range and accuracy scale**

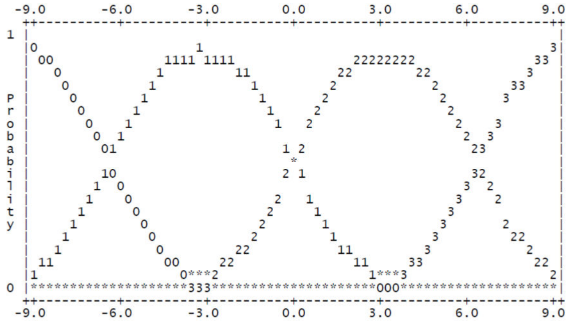


Figure 15-4: **Fluency scale**

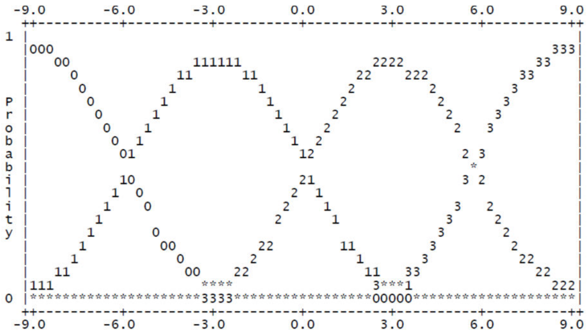
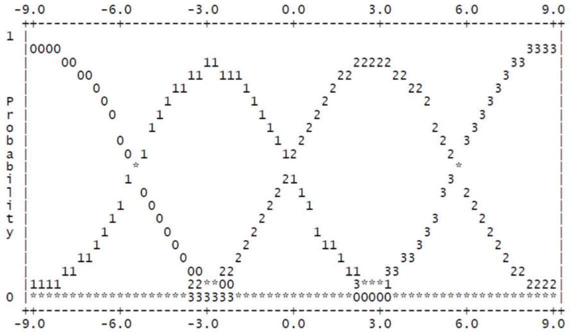


Figure 15-5: **Interactional effectiveness scale**



4.2 Raters' Feedback (RQ4)

Rater Training Feedback Questionnaire

Results of the rater training feedback questionnaire also suggest that some modifications made to the training procedures after Study 1 worked as the test designers intended. Table 23 summarizes the results of feedback given by the six raters involved in Study 2. Compared to the outcomes of the rater training feedback questionnaire for Study 1 reported in Section 3.4.1, raters in Study 2 perceived the training session even more positively.

Table 23: ***Rater Training Questionnaire***

	Questions	Mean*
Q1	I found the training session useful.	4.83
Q2	Watching the interlocutor training video and discussing the interlocutor frame before reviewing the rating criteria was useful as background information.	4.67
Q3	The rating criteria were clearly explained.	4.67
Q4	The standardized exemplars were good examples of the scoring categories for the different criteria.	4.00
Q5	The number of standardized exemplars (3) was sufficient to help me understand how to apply the rating criteria.	4.33
Q6	Rating the standardized exemplars (2 & 3) and discussing the raters' scores before looking at the benchmark scores was useful practice.	4.83
Q7	Having finished the training, I am confident that I will be able to apply the rating criteria in rating samples of test-taker performance.	4.50
Q8	Do you have any suggestions to improve the training session?	-

*1: strongly disagree – 5: strongly agree

Rating Feedback Questionnaire

The rating feedback questionnaire after the Study 2 ratings was almost identical to the one we used in Study 1, except for Q3. Instead of asking about the amount of overlap between different pairs of scales, in Study 2 we simply asked how distinct each of the five rating scales was from the others.

As shown in Table 24, the revised rating scales seemed to work better in general. The six raters felt it was easier to understand and interpret the descriptors in all five categories (Q1 and Q2), although Rater 1 mentions difficulty distinguishing between levels 1 and 2 for "grammatical range and accuracy," in particular in interpreting the term "reasonably accurate," and Rater 4 mentions problems judging "occasionally vs. frequent" and "incorrect word choice" (raters 1 and 4). As for the overlap between different scales (Q3), raters in general felt the five scales were distinct, even if there was some overlap (e.g., "I feel the grammar and lexis categories, while distinct, in practice are closely linked" [Rater 5]).

The recording quality and speech samples were perceived to be adequate enough, and the raters did not usually have to listen to the recordings twice, as they "rarely" had difficulty in hearing the recorded material (Rater 1) (Q4-7). The rating processes the raters reported (Q10) were interesting. While different raters reported different processes, in general "pronunciation" was the first category they rated that "could be figured out by the end of Part 2" (Q9 and Q10; raters 1 and 4). This is consistent with the Study 1 results.

Table 24: *Rating Feedback Questionnaire*

	Rating Questionnaire	Responses
Q1	**The descriptors are easy to understand and interpret** (1: strongly disagree – 5: strongly agree)	**Mean**
Q1-1	Pronunciation	3.83
Q1-2	Grammatical range and accuracy	4.17
Q1-3	Lexical range and accuracy	3.67
Q1-4	Fluency	4.00
Q1-5	Interactional effectiveness	3.83
Q2	**The descriptors for each score point distinguish well between each of the levels of the scales** (1: strongly disagree – 5: strongly agree)	
Q2-1	Pronunciation	4.00
Q2-2	Grammatical range and accuracy	3.83
Q2-3	Lexical range and accuracy	3.67
Q2-4	Fluency	3.67
Q2-5	Interactional effectiveness	3.83
Q3	**How distinct are the scoring scales? Use the following three-point scale to describe how distinct each of the five scoring scales is from the others.**	**Frequency**
Q3-1	Pronunciation	Very distinct: 4 Some overlap: 2 Significant overlap: 0
Q3-2	Lexical range and accuracy	Very distinct: 2 Some overlap: 3 Significant overlap: 0 (Missing: 1)
Q3-3	Grammatical range and accuracy	Very distinct: 2 Some overlap: 3 Significant overlap: 0 (Missing: 1)
Q3-4	Fluency	Very distinct: 1 Some overlap: 4 Significant overlap: 1
Q3-5	Interactional effectiveness	Very distinct: 4 Some overlap: 1 Significant overlap: 1
Q4	**The descriptors for each score point are:**	Too short: 3 Appropriate: 3 Too long: 0
Q5	**Was the quality of the videos sufficient for rating the speaking samples?**	Yes: 4 No: 2
Q6	**Did you need to watch the video samples more than once to rate them?**	Yes: 4 No: 2
Q7	**Does the test format provide a sufficient quantity of language to rate appropriately?**	Yes: 4 No: 2
Q8	**Does the format provide a sufficient sample of language to distinguish between the intended levels?**	Yes: 3 No: 2 (missing: 1)
Q9	**At what stage of the rating process did you finalize your mark for each category? (Free response)**	-
Q10	**Please describe the process or processes you followed when rating the samples? (Free response)**	-
Q11	**If you have any other comments on the rating procedure or suggestions for improving the rating scales, please write them below. (Free response)**	-

Detailed and useful suggestions for improving the examining procedures were obtained for Q11 by raters 1, 4, and 5, including:

- Part 4 questions should be rephrased so as to make the interviewer–candidate discourse more abstract and objective. ("For Part 4, I suggest eliminating 'ask you some questions' and changing to 'I'd like to discuss some different topics.' Eliminate 'Do you think' and use 'Compare watching TV news and reading newspaper' or 'How social media such as Facebook and Twitter changing the way people communicate'? I think the way questions were posed in second person prompted interviewers to rely too much on personal experience rather than the abstract." [Rater 1])
- More interlocutor standardization is necessary. ("The examiners spoke at very different speeds and gave [or not] help on occasion. Occasionally, examinees asked about words or instructions [with differing responses from the examiners]." [Rater 1]; "Interviewers behaved very differently. They need more guidance and also should watch each other's interviews." [Rater 5])
- Guidance on the interlocutor response to the initial test-taker question in Part 2 is necessary. ("In Part 2, the interviewer's response to the first question 'May I ask you some questions?' is not listed. Variations included: 'Yes,' 'Sure,' 'Certainly,'." [Rater 1])
- The delivery system of test-taker video performance should allow a fast-forwarding function ("Allowing for fast-forwarding through 30-second pauses would not only make grading faster, it makes it easier to keep previous sections in mind while scoring." [Rater 4])
- Guidance on how interlocutors and raters deal with candidates' misunderstanding of the task requirements should be standardized. ("We need guidance on what to do when candidates misunderstand a task, and more generally [whether] it might to be good to rate comprehension." [Rater 5])

These points raised by raters should be discussed by the project team and decisions should be taken as to whether these recommendations should be incorporated into the testing and rating materials and procedures prior to the operational use of the test. Nonetheless, it was very encouraging to find that both the score results and rater questionnaire results in Study 2 demonstrated that the amendments to the scales and test materials made based on Study 1 worked just as the project team intended.

5. Final Remarks

Following a brief overview of the aims of the TEAP Speaking Test and background information about the development of draft test specifications, rating scales, and test materials, this report has described two *a priori* validation studies.

Study 1 examined how well the test materials and rating scales operationalized the test construct described in the draft test specifications in relation to certain aspects of *context* validity and *scoring* validity. The analysis included linguistic and functional features of test takers' output language; test scores; feedback questionnaires from test takers, interlocutors, and raters; and a post-marking focus group discussion of raters. All of these sources of empirical validity evidence offered valuable information to verify the draft rating scales and test materials, but some modifications to the rating descriptors and test materials and procedures were also suggested. These modifications were discussed by the project team and incorporated in the Study 2 test. Study 2 focused mainly on *scoring* validity. It was very encouraging that the Study 2 results demonstrated that the changes made after Study 1 functioned in ways that the test designers intended, and provided further validity evidence for the test.

This report has shown how the TEAP Speaking Test development project took an iterative data-gathering approach following the principles of *a priori* test validation emphasized by Weir (2005). While on-going validation studies are as important as *a priori* validation, on the basis of the two *a priori* validation studies which provided empirical support for the validity of the TEAP Speaking Test, the development team feels confident that the TEAP Speaking Test is operationalizing the test construct that it was designed to measure.

It is hoped that this project offers a model for collecting different types of *a priori* validity evidence during the development stage of a speaking test, to inform test design and contribute to a validity argument prior to the administration of an operational test.

Acknowledgements

I gratefully acknowledge the TEAP project team, who supported and made this validation research possible. Special thanks go to Jamie Dunlea, who oversaw and contributed to all aspects of the project data collection and analysis; to Lydia Takemoto, who provided invaluable research assistance in coding the transcripts for a number of linguistic properties; and Katherine Heins, who transcribed the video data from the trial.

References

ALC (2006). *Standard speaking tests.* Retrieved on February 8, 2007 from http://www.alc.co.jp/edusys/sst/english.html

Atkinson, J. M., & Heritage, J. (1984). *Structures of social action.* Cambridge, New York: Cambridge University Press.

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

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.

Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. *Cambridge ESOL Research Notes, 11*, 20-21.

Brown, A. (2006a). Candidate discourse in the revised IELTS Speaking Test. P. McGovern & S. Walsh (Eds.), *IELTS Research Report, Vol. 6,* 71-89. Canberra: British Council & IDP Australia.

Brown, A. (2006b). An examination of the rating process in the revised IELTS Speaking Test. P. McGovern & S. Walsh (Eds.), *IELTS Research Report, Vol. 6,* 41-69. Canberra: British Council & IDP Australia.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purpose speaking tasks, TOEFL Monograph Series, MS-29.* ETS.

Bygate, M. (1987). *Speaking.* Oxford: Oxford University Press.

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

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

Elder, C., & Iwashita, N. (2005). Planning for test performance. R. Ellis (Ed.), *Planning and task performance in a second language,* 219-237. Amsterdam: John Benjamins.

ffrench, A. (2003). The change process at the paper level, Paper 5, Speaking. C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913-2002,* 367-471. Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking,* 65-111. Cambridge: Cambridge University Press.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning, *Studies in Second Language Acquisition, 18* (3), 299-324.

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

Galaczi, E. D., & ffrench, A. (2011). Context validity. L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking,* 112-170. Cambridge: Cambridge University Press.

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice, 18* (3), 217-237.

Green, A., Unaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing 27*(2), 191-211..

Hawkins, J., & Filipović, L. (2012). *Criterial features in L2 English.* Cambridge: Cambridge University Press.

Hutchby, I., & Wooffitt, R. (1998). *Conversation analysis.* Cambridge: Cambridge University Press.

Inoue, C. (2010). *Investigating the sensitivity of the measures of fluency, accuracy, complexity and idea units with a narrative task.* Paper presented at the *Lancaster University Postgraduate Conference in Linguistics & Language Teaching, 4.* Retrieved July 1, 2011 from http://www.ling.lancs.ac.uk/pgconference/v04.htm

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29* (1), 24-29.

Kobayashi, M., & Van Moere, A. (2004). *Group oral testing: Does amount of output affect scores?* Paper presented at *Language Testing Forum,* Lancaster University, November 28, 2004.

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

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

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

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

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

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *English Language Teaching Journal 62*(3), 266-275

Nakatsuhara, F. (2009). *Conversational styles in group oral tests: How is the conversation co-constructed?* Unpublished PhD thesis, University of Essex.

Nakatsuhara, F. (2010). *A background review report: The development of the Test of English for Academic Purposes (TEAP) speaking paper for Japanese University entrants.*

Nakatsuhara, F. (2011). Effects of the number of participants on group oral test performance. *Language Testing, 28* (4), 483-508.

Nakatsuhara, F. (In press). *The co-construction of conversation in group oral tests.* Peter Lang.

Nakatsuhara, F., & Field, J. (2012). *A study of examiner interventions in relation to the listening demands they make on candidates in the GESE exams.*

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

O'Sullivan, B., Taylor, C., & Wall, D. (2011). *Establishing evidence of construct: A case study.* Paper presented at the *8th annual EALTA conference,* Siena, Italy.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. B. O'Sullivan (Ed.), *Language testing: Theories and practices,* 13-32. Basingstoke: Palgrave.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19* (1), 33-56.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. M. Milanovic and N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium. Studies in Language Testing, 3,* 74-91. Cambridge: Cambridge University Press.

Post, B. (2011). *Using acoustic analysis software to analyse L2 pronunciation features.* Paper presented at the *3rd BAAL TEA SIG annual conference,* University of Warwick, November 18, 2011.

Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS Speaking Test. P. McGovern & S. Walsh (Eds.), *IELTS Research Reports, Vol. 6,* 207-231. Canberra: IELTS Australia.

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

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 58*(2), 439-473.

Taylor, L. (Ed.) (2011). *Examining speaking: Research and practice in assessing second language speaking. Studies in Language Testing, 30.* Cambridge: UCLES/Cambridge University Press.

Taylor, L., & Galaczi, E. D. (2011). Scoring validity. L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking. Studies in Language Testing, 30,* 171-233. Cambridge: UCLES/Cambridge University Press.

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

Weir, C. J. (2012). *Research Report on the Development of the Test of English for Academic Purposes (TEAP) Writing Test for Japanese University Entrants*. Internal research report, Eiken Foundation of Japan.

West, M. (1953). *A general service list of English words.* London: Longman, Green and Co.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*, 1-24.

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

Appendix 1: Language Function Survey Results

Table 25: *Questionnaire Results from 24 English Teachers at Sophia University—Important Language Functions for a Student to Be Successful in a First-Year Undergraduate Class*

Informational Functions				Mean (SD)
#		Gloss: Does a Student…	For Example:	
1	**Providing personal information**	Give information on present/past/future circumstances?	“I live in Saitama.” “I’ve been / I went to… before / last week.” “I’m going / I’ll go to… next week.”	3.35 (1.191)
2	**Expressing opinions/ preference**	Express opinions/ preference?	“I don’t like English food.” “I think…”	3.74 (.619)
3	**Elaborating**	Elaborate on, or modify, one’s own opinion?	“I mean…” “They could also reduce class size, or…”	3.43 (.843)
4	**Justifying opinions**	Express reasons for assertion s/he has made?	“It’s because…” / “Because…” “It’s prettier and cheaper…”	3.74 (.541)
5	**Comparing**	Compare things/ people/events?	“I think X is more useful.”	3.52 (.665)
6	**Speculating**	Speculate?	“She must have paid a fortune for that.” “If we buy this one, we can use it for our school trip.”	2.87 (1.058)
7	**Staging**	Separate out or interpret the parts or sequences of an issue?	“So, first I’ll talk about…” “But first, we have to…” “Now, we must choose…”	3.22 (.998)
8	**Describing**	Describe something/ someone or a sequence of events?	“She is nice and funny.” “I went to buy a ticket and found that the ticket office had already closed.”	3.13 (1.140)
9	**Summarizing**	Summarize what s/he has said?	“So, I think we would choose…” “So you think…” “So we have decided/chosen…”	2.95 (.899)
10	**Suggesting**	Suggest a particular idea?	“What about…” “We could (do)…” “Why don’t we (do)…” / “How about (doing)…?”	3.30 (.876)
11	**Expressing preferences**	Express preferences?	“I think this one would be best.” “I’d rather have a small one.” “I prefer/like this one better.”	3.35 (.832)
Interactional Functions				
12	**Agreeing**	Agree with an assertion made by another speaker?	Can be *marked*: “Yes, I agree.” Can be *unmarked*: “That’s true.”	3.70 (.470)
13	**Disagreeing**	Disagree with what another speaker says?	Can be *marked*: “I don’t think that’s right.” Can be *unmarked*: “Well, that depends on your point of view, but I rather think…”	3.35 (.885)
14	**Modifying/ commenting**	Modify arguments/ comments made by other speaker?	A: “I think intelligence is important for a teacher.” B: “And additionally, if the teacher has a sense of humor, it’s nice, definitely.”	2.74 (1.054)
15	**Asking for opinions**	Ask for opinions?	“What do you think?” “And you?” “Well?”	3.57 (.945)
16	**Persuading**	Attempt to persuade another person?	Can be *cued*: “Don’t you think so? “ Can be *uncued*: “Yes, but you can’t spend it all!”	2.52 (1.123)
17	**Asking for information**	Ask for information?	“What about you? What are your favorite films?” “What are your hobbies?” “Do you know…?”	3.65 (.714)
18	**Conversational repair**	Repair breakdowns in interactions?	Can be “other repair” – breakdown during other speaker’s turn: “I’m sorry I thought you meant…” Can be “self repair” – breakdown during the own turn: “What I wanted to say was…” These repairs may be initiated by the person who is speaking or by the other person and can be verbal or non-verbal.	2.70 (1.020)
19	**Negotiating meaning**	Check understanding?	“OK?” / “Is that clear?” / “So, do I have to…?”	3.57 (.788)
		Indicate understanding of point made by partner?	Can be *verbal*: “Yes, I know what you mean.” “OK, yes.” / Can be *non-verbal*: head nod	3.61 (.839)
		Establish common ground/ purpose or strategy?	“Shall we talk about all of them first before deciding?” “So, we both like this one…”	2.35 (1.152)
		Ask for clarification when an utterance is misheard/ misinterpreted?	“Can you repeat that please?” “What exactly do you mean by wealthy?”	3.61 (.941)

Table 25 (Cont'd): ***Questionnaire Results from 24 English Teachers at Sophia University—Important Language Functions for a Student to Be Successful in a First-Year Undergraduate Class***

Interactional Functions				
19	**Negotiating meaning**	Correct an utterance made by other speaker which is perceived to be incorrect/inaccurate?	"No, we've already decided not to take that one." "You mean…" (a lexical or grammatical correction)	2.70 (1.063)
		Respond to requests for clarification?	Can be *cued*: "What I mean is…" Can be *uncued*: "The blue one."	3.27 (1.077)
Managing Interaction Functions				
20	**Initiating**	Start any interactions?	"Right, so we have to choose the best; what do you think of the blue one?"	2.96 (1.022)
21	**Changing topics**	Take the opportunity to change the topic?	"Yes, that would be the best. So what about the worst?" "I don't like going to a gym, but I like to go for a walk. Last weekend…"	2.30 (1.146)
22	**Reciprocating**	Share the responsibility for developing the interaction?	"What do you think we should do?" "What do you think?" "Have you ever tried to do it?" May simply be "yes," head nod, "uh huh," "mm hmm" to encourage other speaker to continue.	3.22 (.951)
23	**Deciding**	Come to a decision?	"So, we have decided…" "So, let's choose / we've chosen…"	3.00 (1.087)

Figure 16-1: **Informational functions**

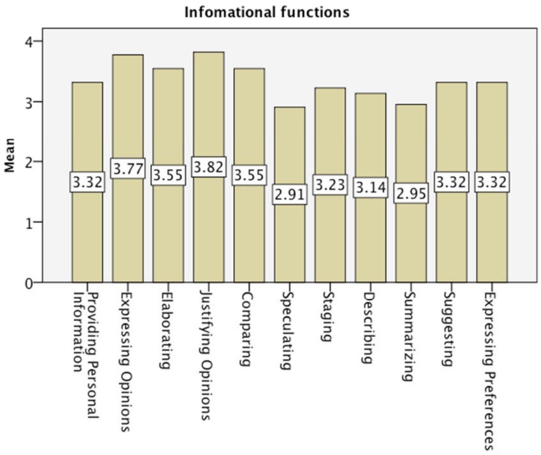
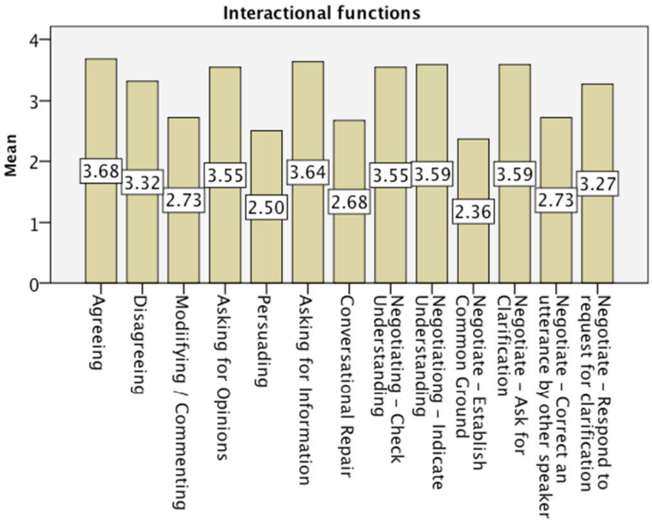


Figure 16-2: **Interactional functions**

Figure 16-3: **Managing interaction functions**


Interactional functions
Mean
0
1
2
3
4
3.68
3.32
2.73
3.55
2.50
3.64
2.68
3.55
3.59
2.36
3.59
2.73
3.27
Agreeing
Disagreeing
Modiifying / Commenting
Asking for Opinions
Persuading
Asking for Information
Conversational Repair
Negotiating – Check Understanding
Negotiationg – Indicate Understanding
Negotiate – Establish Common Ground
Negotiate – Ask for Clarification
Negotiate – Correct an utterance by other speaker
Negotiate – Respond to request for clarification

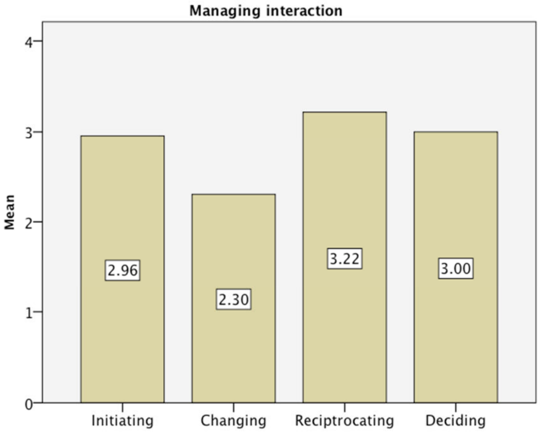

Managing interaction
Mean
0
1
2
3
4
2.96
2.30
3.22
3.00
Initiating
Changing
Reciptrocating
Deciding

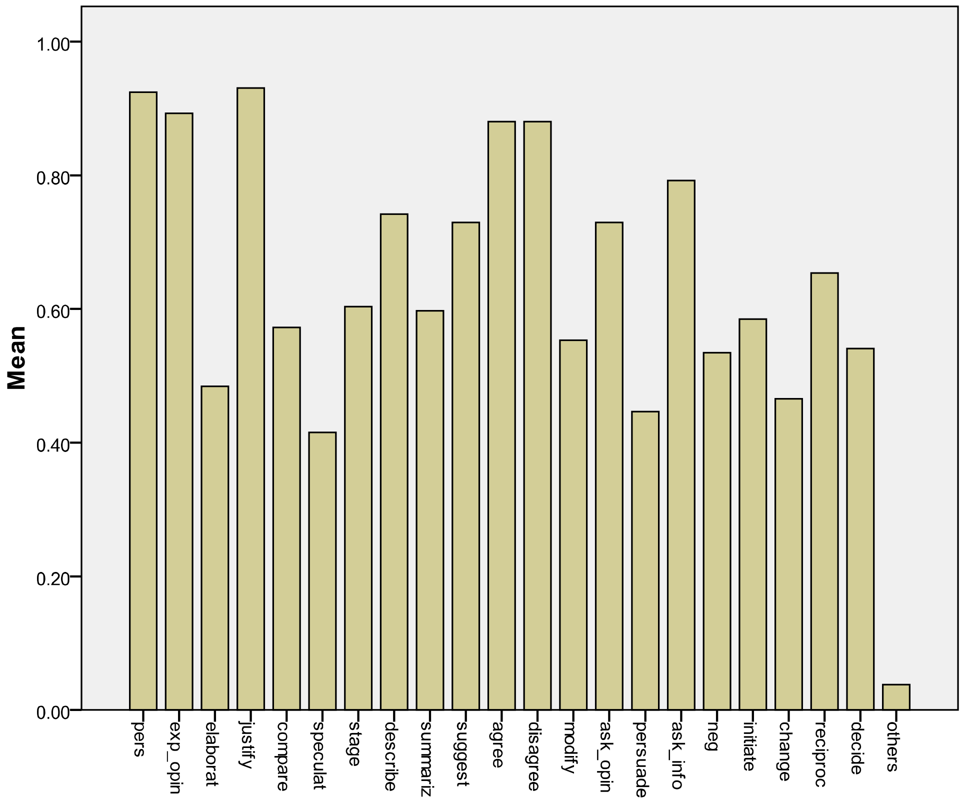
Table 26: ***Questionnaire Results from 172 High School Teachers: Language Functions That Japanese High School Teachers Want Their Students to Acquire (N: Valid = 164, Missing = 8)***

Informational Functions				N=164	
		Gloss: Does a Test Taker…	**For Example:**	**Mean**	**SD**
1	**Providing personal information**	Give information on present/past/future circumstances?	"I live in Saitama." "I've been / I went to… before / last week." "I'm going / I'll go to… next week."	.93	.26
2	**Expressing opinions/ preference**	Express opinions/ preference?	"I don't like English food." "I think…"	.90	.31
3	**Elaborating**	Elaborate on, or modify, one's own opinion?	"I mean…" "They could also reduce class size, or…"	.48	.50
4	**Justifying opinions**	Express reasons for assertion s/he has made?	"It's because…" / "Because…" "It's prettier and cheaper…"	.93	.25
5	**Comparing**	Compare things/ people/events?	"I think X is more useful."	.57	.50
6	**Speculating**	Speculate?	"She must have paid a fortune for that." "If we buy this one, we can use it for our school trip."	.41	.49
7	**Staging**	Separate out or interpret the parts of an issue?	"So, first I'll talk about…" "But first, we have to…" "Now, we must choose…"	.60	.49
8	**Describing**	Describe something/someone or a sequence of events?	"She is nice and funny." "I went to buy a ticket and found that the ticket office had already closed."	.74	.44
9	**Summarizing**	Summarize what s/he has said?	"So, I think we would choose…" "So you think…" "So we have decided/chosen…"	.60	.49
10	**Suggesting**	Suggest a particular idea?	"What about…" "We could (do)…" / "Why don't we (do)…" "How about (doing)…?"	.73	.44
Interactional Functions					
11	**Agreeing**	Agree with an assertion made by another speaker?	Can be *marked*: "Yes, I agree." Can be *unmarked*: "That's true."	.88	.33
12	**Disagreeing**	Disagree with what another speaker says?	Can be *marked*: "I don't think that's right." Can be *unmarked*: "Well, that depends on your point of view, but I rather think…"	.88	.32
13	**Modifying/ commenting**	Modify arguments/comments made by other speaker?	A: "I think intelligence is important for a teacher." B: "And additionally, if the teacher has a sense of humor, it's nice, definitely."	.56	.50
14	**Asking for opinions**	Ask for opinions?	"What do you think?" "And you?" / "Well?"	.73	.45
15	**Persuading**	Attempt to persuade another person?	Can be *cued*: "Don't you think so?" Can be *uncued*: "Yes, but you can't spend it all!"	.46	.50
16	**Asking for information**	Ask for information?	"What about you? What are your favorite films?" "What are your hobbies?" "Do you know…?"	.80	.40
17	**Negotiating meaning**	Check understanding?	"OK?" / "Is that clear?" "So, do I have to…?"	.54	.50
		Indicate understanding of point made by partner?	Can be *verbal*: "Yes, I know what you mean." / "OK, yes." Can be *non-verbal*: head nod		
		Establish common ground/purpose or strategy?	"Shall we talk about all of them first before deciding?" "So, we both like this one…"		
		Ask for clarification when an utterance is mis-heard/misinterpreted?	"Can you repeat that please?" "What exactly do you mean by wealthy?"		
		Correct an utterance made by other speaker which is perceived to be incorrect/inaccurate?	"No, we've already decided not to take that one." "You mean…" (a lexical or grammatical correction)		
		Respond to requests for clarification?	Can be *cued*: "What I mean is…" Can be *uncued*: "The blue one."		

Table 26 (Cont'd): ***Questionnaire Results from 172 High School Teachers: Language Functions That Japanese High School Teachers Want Their Students to Acquire (N: Valid = 164, Missing = 8)***

Managing Interaction					
18	**Initiating**	Start any interactions?	"Right, so we have to choose the best; what do you think of the blue one?"	.58	.50
19	**Changing topics**	Take the opportunity to change the topic?	"Yes, that would be the best. So what about the worst?" "I don't like going to a gym, but I like to go for a walk. Last weekend…"	.46	.50
20	**Reciprocating**	Share the responsibility for developing the interaction?	"What do you think we should do?" "What do you think?" "Have you ever tried to do it?" May simply be "yes," head nod, "uh huh," "mm hmm" to encourage other speaker to continue.	.65	.48
21	**Deciding**	Come to a decision?	"So, we have decided…" "So, let's choose / we've chosen…"	.54	.50
Others				.04	.19

Figure 17: **Language functions that Japanese high school teachers want their students to acquire**



Appendix 2: Transcription and Segmentation Protocols for Transcription and Segmentation of the TEAP Speaking Trial Speech Samples[5]

Transcription

Unfilled pauses or gaps	Periods of silence. Micro-pauses (less than 0.3 seconds) are shown as (.); longer pauses appear as a time within parentheses.
Colon :	A lengthened sound or syllable; more colons prolong the stretch
Dash -	A cut off
.hhh	Inhalation
hhh	Exhalation
((laughter))	Laughter is signaled by using the double parentheses marker for non-linguistic action and describing the action as "laughter."
(h)	Breathiness within a word
Punctuation	Intonation rather than clausal structure; a full stop (.) is falling intonation, a question mark (?) is rising intonation, a comma (,) is continuing intonation.
Equal sign =	A latched utterance A latched utterance is usually used between utterances by two speakers in adjacent turns.
Open bracket [	Beginning of overlapping utterances
Percent signs % %	Quiet talk
Empty parentheses ( )	Words within parentheses are doubtful or uncertain.
Double parentheses (( ))	Non-vocal action, details of scene.
Arrows ><	The talk speeds up.
Arrows <>	The talk slows down.
Underlining	A word or sound is emphasized.
{neru}	Use curly brackets for Japanese words inserted into an utterance. Write an approximation of the sound inside the brackets. Italize the word.
Italics	Words spoken with noticeable *katakana*-like pronunciation (inserting extra vowels, all syllables evenly stressed; e.g., *dogu* for dog) or other L1 influence (e.g., 'l' for 'r,' 's' for 'th')
Arrow (→)	A feature of interest to the analyst *(This is not for transcribing, but for analyzing the data.)*

[5] The source of the protocols for transcription and segmentation, and the quality control procedures employed during the process of transcription and segmentation are described below. Transcription was carried out by a professional editor and proofreaders using the software Express Scribe.

- The transcription protocol was based on Atkinson and Heritage (1984) and Hutchby and Wooffitt (1998). The protocol was revised through iterative consultation among the transcriber, Eiken researcher, and Dr. Nakatsuhara.
- The segmentation protocol employs the AS-unit (Foster, Tonkyn, & Wigglesworth, 2000) as the basic unit of segmentation. The protocol is based on that described in Foster et al. (2000), but some modifications were made through iterative consultation with the research assistant in charge of segmentation.
- Both transcription and segmentation followed an iterative consultation process between the research assistant involved, the principal Eiken researcher, and the principal consultant, Dr. Nakatsuhara. Rather than employing multiple ratings with inter-rater reliability checks, the transcriber made one complete transcription, which was checked by both the Eiken researcher and Dr. Nakatsuhara to ensure consistency of interpretation before continuing with the remaining speaking samples. Several complete transcriptions were checked by the Eiken researcher at points during the transcription process, and any differences in interpretation were discussed and resolved. As an extra quality-control procedure, the entire set of transcriptions was checked by the research assistant in charge of segmentation, who watched the entire set of recorded samples before proceeding to segmentation. Any discrepancies between the transcriber and the research assistant were checked by the Eiken researcher and resolved through discussion.
- The research assistant in charge of segmentation completed segmentation of several complete transcripts, highlighting points of uncertainty, and these were checked by the Eiken researcher and differences resolved through discussion. The research assistant then proceeded with segmentation of the remaining samples. She continued to highlight any points of uncertainty, which were periodically resolved through discussion with the Eiken researcher.

Appendix 2 (Cont'd): Transcription and Segmentation Protocols for Transcription and Segmentation of the TEAP Speaking Trial Speech Samples

Segmentation

<table>
<tr><td>| |
| Uh, in my elementary school, I learned English, | but I didn't learn (0.7) *grammar*. |</td><td>Vertical bars are used to separate the utterances into AS-units. An AS-unit is defined as: *A single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clauses associated with either* (Foster, Tonkyn, & Wigglesworth, 2000). Applications of the AS-unit to segmentation follow the examples in Foster et al., 2000.</td></tr>
<tr><td>$ $
|so $I was$ (0.7) I belong to cooking club|</td><td>Dollar signs enclose all instances of repair, meaningless repetition, or false starts.</td></tr>
<tr><td>Equal sign =</td><td>Latched clauses within the same speaker's utterance that are separated by a well-formed separate comment, etc., which has been interpolated between a main clause and its subordinate clause. It is used to show that the main clause and its subordinate clause belong to the same AS-unit. (This symbol is usually used in the transcription protocol between utterances by two different speakers in adjacent turns.)
|But when I heard (.) her song, $[the$ (1.1) =|%I don't know \ how to say% | $the (0.5) lysic?$ No, | (1.7) =the words in her (0.8) songs were really good |
This would be counted as two AS-units, not three. The clauses highlighted in yellow belong to one AS-unit. In the above example, we want to count the main clause and subordinate clause as a single AS-unit with syntactic subordination, as below:
|But when I heard (.) her song, the words in her (0.8) songs were really good. |
At the same time, the case of repair is itself clearly signaled with a well-formed, syntactically complex clausal structure, which we would want to credit the speaker with, rather than excluding this entire AS-unit as a case of repair. It is important that the speaker's intonation pattern clearly demonstrates that the main clause and subordinate clause are clearly intended to be integrated into one message.</td></tr>
<tr><td>￥
|by speaking English a lot, in their earlier age, ￥maybe (.) they will not hesitate ￥ to speak English ￥ when they become high school student, or university students, |</td><td>Yen signs signify subordinate clauses within AS-units (the definition of subordination follows Foster et al., 2000)</td></tr>
</table>

Appendix 3: Rater Discussion Summary

(By Jamie Dunlea)

TEAP Speaking Trial Rating Feedback Session

Aug. 26, 2011, 14:00–16:00
Raters: R1, R2, R3
Eiken: Jamie Dunlea (J)

Numerical scores referred to in the discussion correspond to the CEFR as follows:
0 = A1
1 = A2
2 = B1
3 = B2

Recording (1)

(Part 1, 0:00–2:20)

J: What confirmed your decision first?
R3: Pronunciation, as a starting point—one of those things you don't need to reevaluate.
R2: Same.
R1: Accuracy and grammar—she makes some mistakes but she corrects them herself. My initial impression was "this is really high," but then I thought that some of the lexical items she used were memorized—chunks that she might have used before—might have access to that level of language.
R2 & R1: Clearly above 1.
R3: I tried to keep it open at this stage.
R1: I would circle the scores as I went through, so unlike what Peter said I was giving scores as I went along.
R3: I tend to be pretty vague.
J: So everyone felt at the initial stage that this is not a 1?
R3: I think of the global first and try to find out what it's made of later.
R1: I tried to go through the analytic scores first, going against my intuition, which would be to give a global score.
R2: I tried to stick to the empirical categories—so hard to separate score into five parts.

(Part 2, 2:20–4:32)

J: Anything in there that might have modified your initial impression?
R2: Probably. Just confirmed it for me—raised interaction score?
R1: Not enough language production in this part (too limited) to move someone's score up—most of the lexical items are given to them and there's not much they can do.
R3: She didn't do anything technically wrong.
R1: She actually got hung up a little bit.
R3: Didn't lead to misunderstanding [as specified in] the descriptors.
J: Were you able to find something in the descriptors to associate with the features?
R3: I was looking at the CEFR.
R1: I found that two descriptors would describe the same speaker at the same time.

J: Having both features of the two descriptors?
R1: Yes. Sometimes found it a little difficult.
R2: I had the same trouble.
J: Bits of both descriptors.
R3: Can get a little difficult if you try to analyze everything she says.
J: Did you find this part useful in confirming the impression from part 1?
R3: Grammar.
R2: And interaction.

(Part 3, 4:33–6:34)

J: What's happening in this section?
R3: Her weakest section—hesitation, searching for vocab.
R1: If I had to put this on a continuum, this would be high 2's—I think she's good, but she's a good communicator, not the grammatical range of a 3—the level of complexity in her speech hasn't pushed her up to a 3.

(Part 4, 6:35–End)

J: How did this part affect your decision making?
R2: Tough call, but she stayed at 2.
R1: She had such good strategies/discourse fluency/markers—I think I got a hundred percent of her meaning, though she makes grammatical mistakes.
J:) Strong agreement with everyone on interaction.
R2: No strain at all.
R3: Doesn't seem to restrict her meaning—in the B2 band.
R1: Probably something that might need to be cleared up in the descriptors/hard to resist these good communicators into discrete categories.
J: If there were points that were relevant to you that you didn't find in the descriptors, please tell us, and stuff that wasn't helpful.
R3: Might need better descriptors than "synonyms…"

Recording (2)

(Parts 1 and 2, 0:00–5:06)

J: What was guiding your decisions with this?
R2: For pronunciation, intelligible but occasional (vs. constanR1) strains.
R3: Monotonous.
R1: At the end of part 1, I had the idea that she'd moved out of 1, gone into safe 2 territory; felt that it'd be unlikely to move beyond a 2.
J: How about the other categories beyond pronunciation?
R1: Across the board—confirming my hypothesis.
J: We've got strong agreement here.
R3: I wouldn't put in marks down, though I'm thinking about it as I listen.
R2: Same for me—for grammar and lexicon you can't really give a 3.
J: What pushed up your interactional effectiveness to 3?
R2: We'd have to watch till the end.

(Part 3, 5:06–7:23)

R1: That was a beautiful sentence there, she talked about not being good at English—not complex but no way she would go down back to 1.

(Part 4, 7:24–End)

R1: Bye-bye 3…
R1: As soon as she entered this section, she seemed to hit a wall.
R2: I think the test is successful at doing that.
R3: Part 3 and 4—you could actually see it.
R1: That's why I found the role play very unsatisfying—I don't feel that it's part of this upward continuum.
R3: Do you think you can get this into 10 minutes? Seems like no one can.
J: They're all like 12–13 minutes—let's discuss that in the end.
R3: If you're going to this constantly, this would cause problems over time.
R2: No two students are taking the same test. All the teachers are giving different tests.
R3: How far would you let the student talk beyond what they're supposed to?
J: We need to decide that later on.
R2: She has parts of both 2 and 3 for interactional effectiveness.
R3: But she doesn't go beyond the descriptors.
J: Was there anything in part 4 that confirms or changes your decision?
R2: Watching it the second time, I don't know… especially in part 2, turn taking, clarifications, etc. She didn't have any problems there, but I can understand someone giving 2 as well.
R3: It's a solid 2.
R1: If someone has decent pronunciation, he/she can have high fluency and interactional effectiveness but low grammar/lexicon.
R1: I think there's a lot riding on the word "effective"…
J: Maybe we need to discriminate between the two levels—we need to make them more explicit.
R2: To me, whether it's video or audio makes a big difference.
J: Definitely not B2, but stronger than A2.

Recording (3)

(Parts 1 and 2, 0:00–5:39)

J: So?
R1: I initially thought that she was a 1, but when I imagined her with all the gaps cut out, her grammar is fine, so for grammar and lexical range I put her down as low 2.
J: Would that be fluency that's affecting that?
R1: That is fluency, and that's teasing us at judging grammar and lexical range.
R2: Looking at the descriptors for interactional effectiveness; she seemed to be fine doing the interview.
R1: I was struck by her failure to acknowledge answers; she didn't look up, etc.
R2: More things like that in the descriptors would be helpful—I used a combination of the descriptors and the points you [J] said in the training session.
J: I noticed that both Tim and Peter agree on the fluency and interactional effectiveness.
R3: Soon as I heard this, I knew I had to strain to hear her—she doesn't speak up.
R2: Impedes conversation.

R3: If I was talking to her, I would have to keep asking back.
J: That may be something we need to [clarify] in the descriptors.
R3: High 1, maybe a 2 level.
R2: Maybe we're used to listening to these kinds of speech.
R1: That's something we can't overcome [as teachers]… WE can understand them, so you might be getting a false positive in some ways.
J: We may need to make clearer who the intended listener is.
R3: Very *katakana*-like monotonous speech.

(Parts 3 and 4, 5:40–End)

J: Anything you'd like to add?
R3: I think my score for grammar came from the final section.
J: How did you feel about the fluency scale? Did it work? First and third descriptors are about pausing.
R1: I didn't think about it until you mentioned it again, so that wasn't part of my evaluation—the relationship between fluency and interactional effectiveness is very murky when we think about it in regular terms—I think these two are highly integrated.
R3: It's about the flow—whether it's continuous or start-stop, start-stop, so I was trying to figure out whether the speaker is looking for words, etc.; so if it was choppy, thinking about why it was choppy.
R1: I think we all have built-in descriptors for fluency—it's really hard to take them apart.
R3: [Fluency and interactional effectiveness] are both describing "is the person a good communicator?"
R2: The quantitative descriptors were useful—if you could put those in each of the descriptors, that would be good.
R3: And putting certain words in bold, to pick them up.
R1: Get rid of the CEFR standards.
R2: Have you thought about breaking down interactional effectiveness like the CEFR?
J: We could.
J: Non-verbal strategies were sometimes effective—did you find those important or relevant?
R3: For Room 3, the faces were cut off and distracting to see, so I was listening to the audio.
R2: I didn't [look at the non-verbal strategies], because you told me not to.
J: Would we be better focusing on the interviewee?
R3: I think we should focus on both, but should be standardized.
R1: Mark crosses on the floor [for the placement of the camera and desks, etc.]. I think separating the two people would be unnatural.
R3: It's the question of standardizing everything.
J: Definitely we need to give each test taker a standardized amount of time for each section—ultimately our goal is to standardize the timing for all sections.
R2: I think we need high-school students to take the test—they're Sophia students, so they're already standardized.