Assessing the Effects of Increased Task Complexity on High-Level Students' Academic Discussion Performance

Jamie Lesley and Travis West

ABSTRACT

This paper reports on closed, convergent, decision-making tasks trialed with students in the highest of four course levels in Rikkyo University's English Discussion Class (EDC) program. This trial involved a deliberate shift away from the standard format of open, divergent, opinion-gap tasks in a bid to add complexity to discussions for more proficient learners. To examine the effects this had on performance, students were assessed under test conditions using both the standard and more complex task format. Results were compared across overall test scores as well as individual scores for each of five test criteria. Data analysis indicated a significant drop in target skills use in the more complex task format, which has implications for how such tasks should be revised if used as part of a future curriculum. Suggestions for how identified weaknesses might be mitigated in terms of task design, application, and assessment are presented and discussed.

INTRODUCTION

Tasks are now widely used to foster language production in the second language classroom (Plonsky & Kim, 2016). The concept of *task* has been interpreted by language teachers and syllabus designers in various ways, and there does not seem to be a clear consensus as to a single accepted definition of the term (Kumaravadivelu, 1993). Long (2015) distinguishes between *target tasks*, which represent activities that occur in the real world, and *pedagogic tasks*, or classroom activities which develop the language skills needed to successfully conduct target tasks. Attempting to draw together many of the definitions of pedagogic tasks available, Ellis (2009) identified four essential criteria:

1. The primary focus is on meaning;
2. There should be some kind of 'gap' (i.e. a need to convey information, or express an opinion);
3. Learners should largely rely on their own resources;
4. There is a clearly defined outcome other than the use of language. (p. 223)

These four features clearly define a task as an activity in which learners must use language as a tool to accomplish a goal, as opposed to studying language in a static way. Variance in task types is related to the nature of the gap involved, and the goal which must be achieved. Prabhu (1987) identified three task types involving a gap of some kind: *information-gap tasks*, *reasoning-gap tasks*, and *opinion-gap tasks*. In the first two, information must be either shared or synthesized by learners in order to complete the task. In the third type, opinion-gap tasks, participants are required to exchange opinions on issues that are typically controversial, and thus likely to elicit different viewpoints. Ellis (2003) classifies each of these tasks according to their outcome. Tasks in which there is no predetermined solution are referred to as *open tasks*. Examples include opinion-gap tasks, debates, ranking activities, and general discussion. *Closed tasks*, on the other hand, require participants to reach a correct solution, as with most information-gap tasks.

Open tasks can be further divided according to the nature of negotiation conducted by the participants – whether this negotiation is *divergent* or *convergent*. Ellis (2003) describes divergent tasks as those in which students are assigned different viewpoints on an issue and asked to defend their position, for example discussing the pros and cons of capital punishment. Convergent tasks, according to Long (2015), require students to agree on a solution to a problem, such as which of four students will receive a scholarship. Another common example is deciding what items to bring

to a desert island. Duff (1985) favors convergent tasks over divergent tasks due to the fact that they are more effective in promoting meaningful negotiation. In another study by Skehan and Foster (2001), convergent tasks produced more interactional modifications and negotiation of meaning than divergent tasks.

A final distinction made by Ellis (2003) regarding task types in TBLT is between *unfocused tasks* and *focused tasks*. In the former, learners interact communicatively, but there is no expectation that specific language forms will be used. Focused tasks, on the other hand, are purposely designed to allow learners the opportunity to practice specific language forms. This can be achieved by designing a task so that it can only be achieved through the use of a target language form, or by making language itself the content of a task. A third option, as discussed in Li, Ellis, and Zhu (2016), is for focused tasks to be implemented after the explicit teaching of language forms, which are then expected to be used by learners during the task. In any case, the primary focus must continue to be on meaning in order to qualify as a task.

In describing a framework for TBLT, Skehan (1996) identifies a *strong form*, in which tasks are considered the primary unit of language teaching, and a *weak form*, in which tasks play an important role in language instruction, but are ingrained in a more complex pedagogic context. In the weak form of TBLT, tasks may be preceded by focused instruction – an approach which resembles more traditional communicative language teaching (CLT), such as the PPP (Presentation – Practice – Production) paradigm. Ellis (2003) refers to this type of instruction as *task-supported language teaching* (TSLT), defined as "teaching that utilizes tasks to provide free practice in the use of a specific linguistic feature that has been previously presented and practiced in exercises" (p. 351). According to Long (2015), TSLT may serve as a bridge between traditional CLT approaches and genuine task-based teaching.

One model for task implementation provided by Willis (1996) specifies three distinct sections: the *pre-task stage*, the *task cycle*, and the *post-task stage*. In the pre-task stage, also referred to as the priming stage, task outcomes are established, planning time is provided, and a preparatory activity similar to the task itself may be conducted. The pre-task planning stage can also be used to shift learners' attention to specific language features, allowing them to attend to form while performing the main task (Ellis, 2016). During the task cycle, learners attempt to achieve the main task outcome through collaboration and meaningful interaction. Task complexity can be modified by adjusting time pressure, providing or denying scaffolding of linguistic features, and introducing elements not anticipated during the pre-task preparation activity (Ellis, 2003). Skehan (1998) stipulates that main tasks have a relationship to real-world activities. However, it is important that cognitive and communicative demands remain manageable for all participants (Candlin, 2009). Finally, in the post-task stage, learners are encouraged to evaluate their performance of the task cycle, both linguistically and in terms of task achievement, and to analyze the language which they have used while attempting to achieve the goals of the main task. As part of this process, the instructor may provide formative feedback, ask students to repeat the task, or conduct similar tasks which provide an opportunity for students to practice issues identified in the feedback (Ellis, 2003).

In TBLT, tasks are not only used as a means of facilitating language acquisition, but also as a means of assessing language proficiency (Nunan, 2004). In language-based testing, there is a basic distinction between *direct* and *indirect assessment*. Indirect assessment, such as a multiple-choice test, represents more traditional criteria, and is centered on providing a sample that can be analyzed for the grammar, lexis, and phonological features of a language (McNamara, 1996). TBLT utilizes direct assessment, which measures the ability of learners to use language in authentic situations using test procedures which are either identical or similar to the criterion procedure or *target task* (Robinson & Ross, 1996). A similar distinction is made between *system-*

referenced and *performance-referenced tests*, with the former assessing general knowledge of language without evaluating the ability to use it in a specific situation (Baker, 1989). However, in a performance-referenced test, learners are required to demonstrate an ability to use language in specific contexts via use of communicative simulations of target-tasks (Robinson & Ross, 1996).

Tasks are used as a means of facilitating language production, but do not directly measure a learner's language ability. For this, a tool or method is required. Such methods can be placed within three categories: (a) *direct assessment of task outcomes*, (b) *external rating*, and (c) *self-assessment* (Ellis, 2003). Closed tasks can be assessed using (a), since they culminate in a solution which is either right or wrong, or completed or not completed. Open tasks are assessed through the two remaining options: external rating and self-assessment. External ratings are widely used in TBLT, and involve a rater who observes and assesses task performance by a learner or group of learners (Nunan, 2004). A tool used by raters is the *performance scale*, which describes learner behavior based on a list of specific descriptions (also referred to as *bands*). These can be framed in linguistic terms, behavioral terms, or both. An alternative to this is the *checklist*, which includes boxes corresponding to specific language use or task-completion criteria which may be marked while observing the performance of a task. These two methods can also be combined, i.e. a scaled checklist in which the rater subjectively marks the completion of a task criterion based on a scale.

Having established some defining features and categories of TBLT, as well as different ways in which tasks can be evaluated, the importance of evaluation to the present study should briefly be emphasized, particularly in relation to use of tasks in an EFL curriculum and an academic discussion skills course taught at the university level. Before details of this teaching context are introduced, a short overview of curriculum evaluation, its purpose, and value will be provided, since it is to this end that the primary motivations of this paper are oriented.

In his consideration of language curriculum, Brown (1995) highlights five components of *needs analysis*, *objectives*, *testing*, *materials*, and *teaching* that designers should address when planning. These same five aspects should be systematically revisited post-implementation to evaluate a curriculum's effectiveness. Such evaluations may be conducted for *summative* or *formative* purposes and, in either case, are typically oriented towards a curriculum's *products* (i.e. its goals) and/or its *processes* (i.e. the means by which the goals are pursued and achieved). Information gathered on these aspects can be collected in various ways and analyzed using *quantitative* and *qualitative* methods. Combining these findings can allow for a range of specific and holistic assessments to be made, which in turn can lead to improvements in the curriculum.

Brown (1995) argues that the process of curriculum development is never truly finished. This is because modifications to existing designs and applications should routinely be made in response to weaknesses identified through data analysis of what is being achieved by the curriculum's use. In this way, a curriculum may continue to evolve to best meet the needs of its participants and to maximize learning outcomes in a principled manner (Nation & Macalister, 2010). And so it is with this belief that the following paper is written to document some of the steps taken by Rikkyo University's Center for English Discussion Class (EDC) towards strengthening the course for its higher-level students through a deliberate shift in the type and focus of tasks used to practice target skills.

Teaching Context

The EDC Center oversees a compulsory, first-year undergraduate course situated within Rikkyo University's general curriculum of study. Approximately 4,500 freshmen take this course for 90 minutes each week in both semesters of the year in micro-classes of 7-9 students. A communicative approach is taken to help meet the main aims of building speaking fluency and developing the ability to express ideas and organize speaking turns within the context of a group discussion

(Hurling, 2012). Focus is also placed on improving strategic competence by encouraging students to find and fix communication breakdowns as they occur (Schaefer, 2018).

EDC employs a strongly unified curriculum that standardizes all aims, methodology, materials, and assessments across all levels. These levels are determined by TOEIC placement test scores: level I (680-999), level II (480-679), level III (280-479), and level IV (less than 280). Skills taught across the four levels are generally the same with some exceptions in level I (see Table 1). Specifically, certain skills are combined and taught earlier, as with Supporting Opinions (Reasons plus Examples) and Changing Topics (Choosing Topics plus Closing Topics). Other skills are also unique to level I as with Definitions and Reconsidering Opinions. However, despite some variations in skills focus, class content in all levels always maximizes student-to-student interaction and places a high priority on meaningful repetition of target forms and formative feedback to help improve subsequent use of the skills.

Table 1. Discussion skills taught in spring and fall semesters by level

Spring Semester			Fall Semester		
Lesson	Levels II-IV	Level I	Lesson	Levels II-IV	Level I
2	Opinions		2	Connecting Ideas	
3	Reasons	Supporting Opinions	3	Closing Topics	Reconsidering Opinions
6	Examples	Definitions	6	Different Viewpoints	
7	Joining a Discussion		7	Information	
10	Choosing Topics	Changing Topics	10	Balancing Opinions	
11	Possibilities		11	Comparisons	

Students' in-class performance is assessed following grading rubrics for regular lessons and discussion test lessons. In regular lessons (excluding the first and last lessons of the course), the rubric covers five criteria with 0-4 points possible for each and a maximum available score of 20:

- *Attendance* (with a sliding scale for tardiness)
- *Quiz* scores (based on a homework reading activity)
- *Discussion skills* use (with listener and speaker behaviors, e.g. asking for and giving reasons)
- *Communication skills* use (with listener and speaker components to help convey/confirm attention and comprehension, and to resolve miscommunications as they arise)
- *Participation* (based on active involvement and the ability to balance listener and speaker contributions without dominating discussions)

Each lesson, students complete two extended group discussions (referred to as D1 and D2) that typically last 10 and 16 minutes respectively. It is during these group discussions that the last three grading criteria for skills use and participation are applied. As the course progresses, more skills are introduced, and thus the challenge to use them all appropriately and consistently increases

incrementally over the semester.

Separate descriptors are used in three designated test lessons with five grading criteria of *communication skills* (for reactions and negotiation of meaning), *content* (i.e. speaking turns), *questions*, *discussion skills – listener*, and *discussion skills – speaker*. Each of the five criteria may receive a maximum of five points for a total test score out of 25. All instructors receive regular training on how to use the rubrics to standardize application, maintain continuity, and raise inter-rater reliability (Doe, 2012). Tests occur three times in each 14-week semester in lessons 5, 9, and 13. During these tests, skills that have been taught to that point in the course are evaluated under more controlled conditions. In the first test, two target discussion skills are prioritized and formally assessed. In the second test, the number of assessed skills rises to four, while in the third and final test, all six discussion skills are examined. Test procedures and questions are standardized each time with criterion-referenced score sheets and test-specific grading criteria (see Appendix A-B) that are used by all instructors.

The discussion test score sheet divides all gradable test performance into one of two halves as listener- or speaker-side behaviors. Communication skills span both of these halves, although they are totalized as a single combined category. The questions criterion is only on the listener side. Content is only on the speaker side. To help weight the main target forms of the course more heavily in discussion tests, Discussion skills occupy two test criteria with separately scored listener and speaker categories. Within each criterion, various ways to score exist (see Table 2) and, to do so, such items must be clearly marked. A target behavior is considered marked when its pragmatic function (e.g. asking for an opinion) or the pragmatic function of the content to which it is attached (e.g. paraphrasing yourself) is clear and accurate to the point of being understandable to other group members. If so, then, the target behavior is marked on the score sheet.

Table 2. Discussion test criteria

Criteria	Items
Communication Skills	Comprehension (checking / confirming understanding, reactions) Paraphrasing (yourself / others) Clarification (asking for / providing explanation or repetition)
Questions	Any other strategies used to negotiate meaning Content Questions (e.g. asking for reasons) Organizational Questions (e.g. asking others to join the discussion)
Content	Speaking Turns Agreeing / Disagreeing
Discussion Skills – L	Listener-side Discussion Skills
Discussion Skills – S	Speaker-side Discussion Skills

Unlike regular lessons, test groups are assessed in isolation from each other to strengthen rating accuracy, i.e. by allowing the instructor to focus on just a single discussion as opposed to concurrent ones across multiple groups. Test durations still mirror those of regular lessons, and students are again provided with predetermined discussion questions. However, two pairs of test questions are used to broaden the range of talking points, although it at the group's discretion how many of the four questions they choose to discuss. The first pair of test questions are always linked thematically to the previous lesson's topic, while the second pair match the current lesson's topic.

When writing discussion questions, EDC program managers follow principles intended to guide the creation process (Lesley & Brereton, 2018). These principles dictate that the main function of discussion questions is to facilitate an extended group discussion between students on a given topic for a given length of time within which use of given target skills can feature repeatedly. To assist this, questions should generate a variety of responses to ensure a genuine need to communicate (otherwise, the impetus for students to share their ideas might be diminished or removed). Furthermore, questions should be graded to the course level and framed within students' pre-existing knowledge and vocabulary so that attention can be directed more towards use of target skills than comprehension of question content. If the subject matter or lexis are too complex, this might divert focus away from target forms and meaningful interaction, which is counter-productive to the stated aims of the course.

In EDC lessons, a key feature of the discussion format and the accompanying questions is how they steer students towards a free exchange of ideas, which, as was described earlier, equates to an open, potentially divergent, opinion-gap task. For level I students, whose proficiency generally falls between TOEIC 680-990, such tasks might be quite easily achieved, especially at the upper end of the placement test score range where near-native speaker levels are approached or exceeded. This concern was heightened at the end of the 2016-2017 academic year when the ability to opt out of taking EDC was removed for students with placement scores of TOEIC 900 and above.

In response to this anticipated intake of higher-level English users, steps were taken at the beginning of the 2017-2018 academic year to add complexity to the standard class discussion, by trialing a modified task format with selected level I classes. This format required students to not only demonstrate use of target skills, but to do so to collaborate and reach group decisions. Accordingly, this came with a shift from an open, feasibly divergent, opinion-gap type of task to one with a distinctly closed, convergent, decision-making focus. Such modified tasks were used in regular, non-test lessons of selected classes in the 2017-2018 and 2018-2019 academic years with some success, and positive feedback from students who participated in their use, albeit with some hesitation from instructors, who believed that focus on the more difficult task appeared to come at the expense of skills use (Lesley, 2018). It was therefore decided that a next step regarding this aspect of the program would be to pilot the new decision-making tasks under test conditions to more accurately investigate the effect of a more complex task type on students' discussion performance. What follows is a description of that process and a presentation of the results and implications.

METHODOLOGY

Research Questions

To help consider the effects of a more complex task on level I students' group discussions, the following research questions were created:

1. What effects, if any, does a closed, convergent decision-making task have on discussion performance in terms of students' overall test scores?
2. What effects, if any, does a closed, convergent decision-making task have on discussion performance in terms of students' individual scores for each of the five test criteria? (*Communication skills*, *Questions*, *Content*, *Discussion Skills - Listener*, and *Discussion Skills - Speaker*)

Participants

Eight level I classes from the spring semester and eight from the fall semester were selected for the study. However, due to class cancelations resulting from a typhoon, two classes were

unavoidably removed from the fall total, preventing an even distribution of classes across both semesters. The study was scheduled for the last two lessons of the course: lesson 13 (L13) – the final discussion test – and lesson 14 (L14) – the full course review. These were chosen based on sequential proximity, as well as L14 having the greatest flexibility in lesson content and therefore able to accommodate a repeat staging of a discussion test lesson, unlike in lessons preceding this. The students came from six of the university's ten colleges: *Arts* and *Intercultural Communication* were represented in both semesters, *Law and Politics*, and *Sociology* only in the spring, with *Business* and *Economics* only in the fall. All students were first-year undergraduates. Some had experience of living or studying abroad while others did not. Six of the fourteen classes were taught by EDC program managers and the remaining eight were taught by instructors. The aims and format of the study were explained to all classes in lessons 12 and 13 in each semester. Consent was obtained from everyone with permission to retain students' test scores from lessons 13 and 14 for the purposes of program development and for publication.

Materials

L14 materials were created with a dual purpose. First, they needed to fulfil the given aims prescribed for all levels by the course syllabus to provide specific feedback on the performance on the discussion test from the previous lesson and to review all the discussion and communication skills in the course. For the purposes of this study, the materials additionally needed to help prepare students to complete a decision-making discussion task conducted and scored under test conditions by introducing the test topics in the first half of the lesson and providing opportunities for them to be considered and discussed before the test itself in the second half.

To meet the combined needs of the lesson, a two-page handout was created based on the activities in the textbook for L14 (see Appendices A and B). The first page included a designated fluency-building warmer and a practice section intended to be used during the initial 40-45 minutes of the lesson. The second page contained the discussion test content of a pre-task section with input items (as in a regular lesson discussion) and a scenario with directions for what the group should discuss and decide based on the given content. Beneath the test questions was a box for students to take notes in if required. This same format was replicated in both semesters.

In the spring semester, the discussion test was on the topic of how to have a good discussion, while in the fall the discussion focused on developing skills and personal qualities at university. In both cases, students were given a preparation section containing nine items, with space for a tenth of their own choosing, from which to select the three most suitable responses to the given scenario and decision. The spring discussion question was "What is most important for having a good discussion?" In the fall, it was "What skills and personal qualities are most important for university students to develop?" The aim of the discussion was to decide in their groups the best choices by consensus.

Given the restraints of the fixed syllabus, it was not possible to trial these materials with students prior to their use in L14, and thus it was not known if the materials could provide sufficient scope for an extended discussion that would occupy the required test length. Because of this, a second open-ended test question was included in both sets of materials to be used in the event of a group reaching an early decision. In the spring, this question was "What things are important when making a group decision?" and in the fall, it was "What are some good ways to develop these skills and personal qualities?" In both cases, the instructions on the materials made it clear to students that these second questions were optional talking points to be considered only after the first question had been discussed and the main decisions had been made.

To score the discussions in L14, instructors and program managers used the same discussion test score sheet, grading code, and grading criteria as in L13 (see Appendices C, D, and

E). This was to ensure continuity between how the two tests were scored in both lessons.

Procedure

The first 40-45 minutes of the lesson were spent reviewing the course contents and providing feedback on discussion test 3 performance using page one of the handout. In some cases, the designated fluency activity was completed, although this was not essential to the lesson, and not used by all instructors. However, the practice activity questions were used by everyone. The manner in which the questions were used was not prescribed, since this was somewhat dependent on the skills focus required of the L13 test review. Despite this, the practice activity served its purpose to provide test-related feedback and to help introduce the topic-specific components of the decision-making task. In this way, the L13 review and L14 test preparation stages of the lesson were combined.

With approximately half of the lesson time remaining, students were assigned to test groups and the test procedure was explained. While one group took their test, the other group(s) waited outside the classroom. Before each test began, students were given a few minutes to read the pre-task directions and make their selections based on the input material. Thereafter, the group completed the task, discussed their respective selections, and tried to reach a consensus decision all while using the full range of target skills in the course. All discussions were conducted under usual test conditions and the test duration was fixed according to the number of group members (i.e. 12 minutes for three people, 16 minutes for four, and 20 minutes for five). The group was entirely responsible for managing and maintaining their own interaction for the given time allowance without instructor intervention or support. The test questions and the timer were also visible to students throughout. During each test, the instructor graded the students' interactions using the same score sheet as used in L13. At the end of the discussion, brief feedback was given to the group to comment on their performance and exemplify some of its strengths and weaknesses.

Data and Analysis

The data from the discussion tests taken in L13 and L14 in each semester were first cleaned to ensure that only students who attended both lessons were included in the sample. Those that attended one lesson but not the other were removed so that a comparison could be made across both weeks' scores. All students were coded and their test scores from each lesson compiled using Excel to show overall scores out of 25 points and individual scores out of five points for the five discussion test criteria comprising the overall scores (i.e. Communication Skills, Questions, Content, Discussion Skills – Listener, and Discussion Skills – Speaker). Averages of these scores were then calculated and the means compared. In order to examine whether there was a significant difference in overall test scores, paired samples t-tests were conducted using SPSS.

RESULTS

The results show that the implementation of closed, convergent, decision-making tasks had a significant effect on students' test scores. First, the means of overall test scores are compared, and the results of the paired samples t-tests are presented. This is followed by a descriptive analysis of individual category scores.

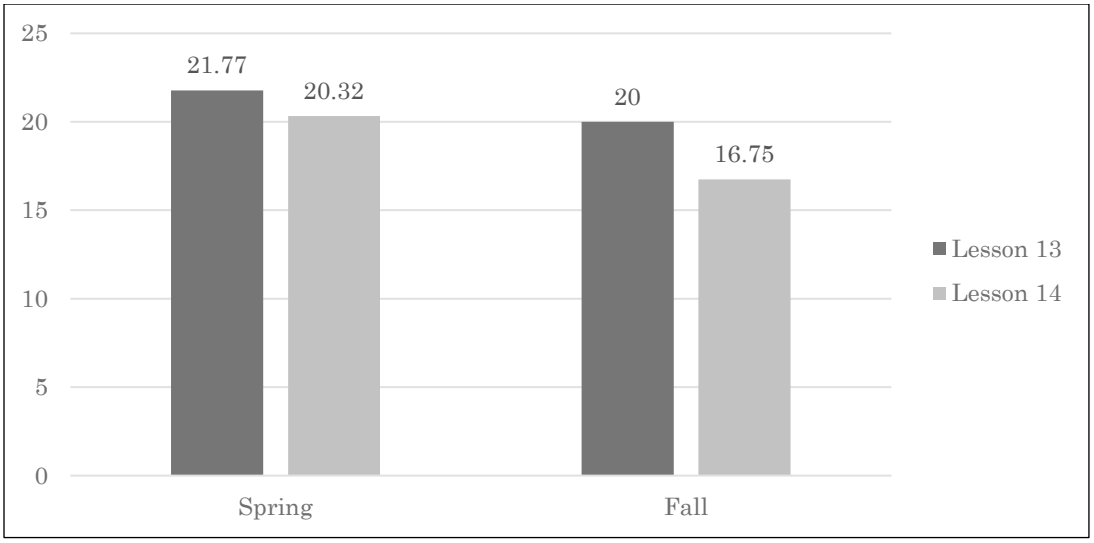

Figure 1. Comparison of total test scores in the spring and fall semesters for L13 (standard test format) and L14 (experimental test format)

The mean scores presented in Figure 1 show that the overall test scores dropped in the L14 test in both the spring and fall semesters. This drop was more pronounced in the fall semester, where the L13 mean fell from 20 points to 16.75 in L14. The difference in mean score was not as noticeable in the spring semester – the mean score in L13 was 21.77, and 20.32 in L14. Descriptive statistics regarding the differences between test scores in the spring semester (see Table 3) and fall semester (see Table 4) are presented below.

Table 3. Descriptive statistics of differences in spring semester test scores

	Mean	N	SD	SEM
L13	21.77	47	2.286	.333
L14	20.32	47	2.563	.374

Table 4. Descriptive statistics of differences in fall semester test scores

	Mean	N	SD	SEM
L13	20.00	36	2.839	.473
L14	16.75	36	4.436	.739

Table 5. Paired samples t-test of differences in test scores

		Paired Differences							
					95% CI of the Difference				
		Mean	SD	SEM	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Spring Semester	1.447	2.348	.343	.757	2.136	4.224	46	.000
Pair 2	Fall Semester	3.250	3.894	.649	1.932	4.568	5.008	35	.000

The t-tests (see Table 5) comparing the overall test scores revealed that the difference in means between the spring semester L13 and L14 scores, $M = 1.45$, $SD = 2.35$, was significantly different, $t(46) = 4.22$, $p < .001$. There was also a significant difference in means between the fall semester L13 and L14 scores, $M = 3.25$, $SD = 3.89$, $t(35) = 5.01$, $p < .001$. In order to better understand what may have caused these differences in mean scores, it is important to analyze the individual category scores from each of the tests. From this analysis, it can be determined if the increase in task complexity affected some skills more than others. Since the discussion skills in the spring semester were different than the skills taught in the fall semester, these scores will be discussed separately.

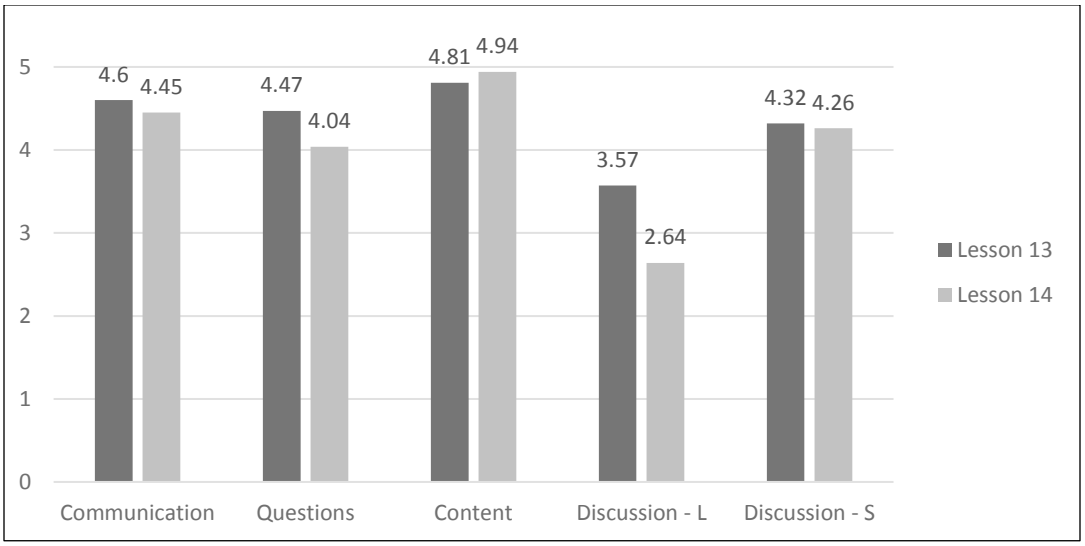

Figure 2. Comparison of individual category performance in the spring semester.

As with the overall test scores, the majority of individual category scores in the spring semester dropped slightly in L14. The biggest difference appears in the Discussion Skills – Listener category, where the L13 mean was 3.57, while the L14 mean was 2.64. The difference in mean score of the Questions category was also affected, dropping from 4.47 in L13 to 4.04 in L14. There was one exception to this trend, however, which can be seen in the Content mean score. This mean actually rose in L14, from 4.81 to 4.94. Differences in the mean scores of Communication Skills and Discussion Skills - Speakers were not significant.

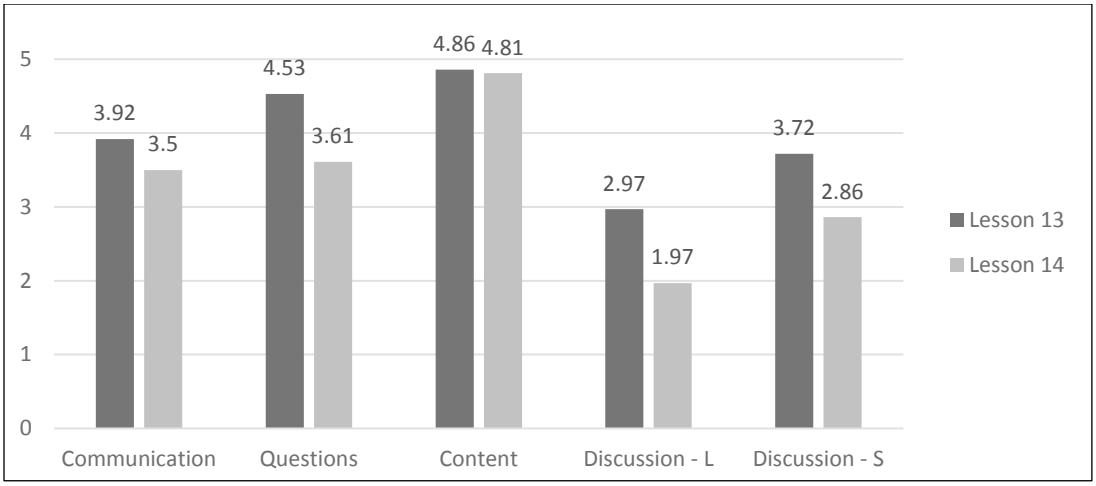

Figure 3. Comparison of individual category performance in the fall semester.

Individual category scores in the fall semester show a higher rate of discrepancy when compared to the spring semester scores. Every category showed a significant decrease in mean score between L13 and L14, with the exception of Content, which was nearly identical. The largest drops can be seen with the Questions category, and both of the Discussion Skills (Listener and Speaker). The largest decrease in mean score occurred with the Discussion Skills – Listener, where the mean in L14 (1.97) was a full point lower than L13 (2.97). The decrease in mean score for Communication Skills was not as pronounced, but still significant, dropping from 3.92 in L13 to 3.5 in L14.

DISCUSSION

These results show that there was a general downward trend in both the overall test scores and the individual category test scores between the standard L13 task and the L14 task with heightened complexity. Therefore, the answer to both of the research questions in this study is that the increase in task complexity led to a decrease in the frequency of target language use. This decrease was apparent in the lower average test scores, which were observed in both overall test scores, and individual category scores, with only one exception: the Content score from the spring semester. One surprising result was that the average Communication Skills score decreased as a result of implementing the L14 discussion task. This seems to contradict the suggestion made by Long (2015) that convergent tasks lead to higher instances of negotiation of meaning.

The analysis in this study was able to compare which of the five test categories were affected. However, it could not determine how each individual category was affected, especially in terms of target language variety. This would require a more in-depth analysis of individual discussion skills which was not possible in the space provided here. A detailed study including a conversation analysis of learner discourse recorded during discussion tasks would help examine the frequency and variety of target language phrases for each individual discussion skill.

An implication of this study is that in order to maintain average test scores equivalent to the L13 discussion tasks, there must be a change in the L14 task as it currently stands. This change might occur in terms of target language, task assessment, or task complexity. In the case of target language, Discussion Skill phrases could be considered which better facilitate successful convergent-type task completion. Studies involving convergent tasks have shown that these types of tasks generate greater interactional demands on students (Duff, 1985; Skehan & Foster, 2001).

If convergent tasks were continued to be used in level I classes, target language could be adjusted to take these interactional demands into account. A few examples that might be useful for this include practicing how to facilitate a meeting, how to manage time within it so that the discussion remains focused, how to introduce and move between topics in an agenda, how to confirm decisions, or how to take notes and document meeting minutes. Imposing greater organizational structure on discussions would change the roles of the participants within it, which hopefully would lead to improved motivation for using the skills for a collaboratively reached end goal.

In terms of task assessment, a new rubric could be devised specifically to correlate with closed, convergent, discussion tasks. Such a rubric could include criteria such as task completion, or student speaking time. It could be argued that task complexity was the determining factor in the lower test scores which resulted from the implementation of the L14 tasks because the language and the assessment rubric were consistent in both the L13 and L14 tasks. However, since increasing complexity was one of the goals of this study, it would be preferable to adjust the target language or assessment rubric first. Any new rubric would need to be designed, piloted, and revised as necessary to help create usable, valid, and reliable grading instruments. This would, of course, take time to undertake and investigate effectively and should begin with a thorough needs analysis. It would also require new means of support for instructors through training so that newly developed materials, classroom approaches, and assessment practices are not only understood but, perhaps most importantly, are embraced by teaching staff as a positive and necessary change for the good of the curriculum and the learning outcomes they are able to achieve.

This study confirms an issue identified in the literature on creating focused tasks that designing a task which obliges students to use target forms can prove difficult (Ellis, 2003; Li, Ellis, & Zhu, 2016). While it is easy to create tasks which render the production of specific language forms natural or useful, Loschky and Bley-Vroman (1993) point out that it is much more difficult to make it *essential*. This may explain, to some extent, the discrepancy between the L13 (open, divergent, opinion-gap) task performance, and the L14 (closed, divergent, decision-making) task performance. In the L13 task, students were given only one goal – to use the target language. However, L14 provided the students with a second goal of making a group decision – a task which could be completed with limited use of target language.

A final point regarding the results of this study relates to attentional resources. According to VanPatten (1990), such resources are not unlimited, and learners tend to focus more on meaning than form when processing ability becomes restricted. In both versions of the discussion task, students were prompted by instructors to use target language (i.e., to focus on form). However, only in the L14 task were students asked to take on the additional job of making a group decision within a time limit. This may have had the effect of drawing students' attention away from using target language and shifting matters towards the more meaning-focused task of making a decision. This diversion of attention away from form could have accounted for the lower frequency in target language use in the L14 tasks. Moving forward, the L14 task may be improved by taking learners' attentional resources into account.

Limitations

Despite attempts to make the test conditions comparable between the two weeks' of discussions, there were several limitations and confounding variables to account for. Firstly, the topics of the discussions in lessons 13 and L14 were inconsistent. In the spring semester, the L13 test topics were learning values and happiness but the L14 topic was how to have a good discussion. The latter coincided with the textbook content for the final lesson of that semester, but not the previous week's test lesson. Similarly, the fall semester L13 test topics were crime and punishment and public behavior but the L14 focus was developing skills and personal qualities. In addition, L13

contained a passage read for homework to cover the main topics of that lesson, while L14 had no homework reading passage in the textbook. These mismatches could have been avoided by designing tasks with more obvious topic crossover from one lesson to the next.

The L14 discussion topics were quite difficult to create plausible decision scenarios for. It was also difficult to design tasks that obligated the production of predetermined linguistic features. The tasks could essentially be completed without needing to use the target skills. In both L14 tasks, the same type of decision was presented of having to select the three most suitable choices from a given list of pre-task items. There is no evidence to suggest that this is the most appropriate format for either making a decision and/or using the target skills.

The relative low stakes of L14 as the last of the course was also not insignificant. The slightly contrived nature of the topics and the tasks coupled with some students' general lack of engagement and lower sense of urgency in discussions was noted from several instructors. The L14 tasks may have recreated the conditions of a discussion test, as in the previous week's lesson, but they were unable to replicate the same pressure on students to perform as a genuine discussion test might be expected to generate. Since group members were aware that the task was not a real test, the atmosphere was a little less focused than hoped for. Students knew that their overall grades were not going to be significantly affected by their performance, and thus both the test conditions and participants were noticeably laidback.

Another complication related to the focus of L14 was the need for teachers to fulfil the required lesson aims of the prescribed syllabus, i.e. to give feedback on the discussion test 3 performance and to review all the skills in the course. This was a potential distraction from preparing students more fully to discuss the new L14 task topic. The attempt to combine the test feedback and skills review with the introduction of the pre-test items appeared to be the best option at the time, but ideally these requirements would have been better separated.

Other factors may also have adversely influenced the findings. For example, students were not kept with the same group members in both lessons. In addition, students were not kept in the same size of group for both lessons, meaning that some may have taken part in a group discussion with three members for 12 minutes in L13 but been in a group of four people for a duration of 16 minutes in L14. It is not currently known what effect the number of people in a group and the duration assigned to their discussion has on performance. Moreover, only students from six of the university's ten colleges were represented in the data sample. A better distribution of selected classes may have ensured that all colleges were included in the study.

Possible Curriculum Revisions

Despite clear limitations, the results of this study appear to add strength to the argument that level I students at EDC are capable of doing more complex and more challenging types of discussion tasks. Analysis of test scores suggests some decrease in use of target skills, but if this decrease can be mitigated, it may add credence to suggestions for possible changes to the level I curriculum. Indeed, support from students for a revised discussion format with decision-making components trialed in selected level I classes has already been reported (Lesley, 2018). Occasional reservations from level I students have also been noted in end-of-semester surveys regarding how EDC skills are currently practiced and assessed (Brereton, Schaefer, Bordilovskaya, &, Reid, 2019). These concerns suggest that level I students can feel constrained by the need to routinely mark aspects of their utterances to other classmates that they believe fully comprehend everything without such marking being necessary. To help remove unwanted claims that students have to comply to get a good grade, despite not feeling appropriately challenged, a raise in task difficulty could be a positive avenue to explore. Reframing matters more as a business skills / meeting skills course where groups complete discussions in order to decide something might just be this avenue.

Revising how target skills are practiced and assessed in EDC's level I course would, of course, need to be carefully handled if pursued. Nation and Macalister (2010) identified five guiding steps for introducing change to curriculum design:

1. Make sure the change is really needed;
2. Plan the type of change so that it is not too great and not too small;
3. Make sure that enough people see that the kind of change is possible;
4. Use a wide range of change strategies;
5. Be prepared for the change to take a long time. (p. 173)

With these steps in mind, an appraisal of the existing EDC's level I course raises some important considerations. The broad range of level I TOEIC placement test boundaries of 680-990 should also not be overlooked. For students that are closer to the lower boundary limit, adding complexity to group discussions is arguably an unnecessary burden, especially on those for whom the current discussion format and skills focus is challenging enough. However, for those approaching near-native English speaking skills, such curriculum revisions may be both warranted and welcomed.

CONCLUSION

This paper began by outlining key features of TBLT, task design, application, and evaluation. It also noted the importance of routine evaluations of curricula to ensure that learning outcomes are being met. EDC's teaching context was then introduced with reference to the specific changes in student intake and concerns about level I being too easy for higher-level learners who were no longer able to opt out of taking the course. In response, the program piloted more challenging discussion tasks by shifting from the standard open, divergent, opinion-gap format in favor of a closed, convergent, decision-making type from the 2017-2018 academic year. It was hoped that this modification would result in more difficult but more engaging applications of target skills for these more proficient learners. However, the effect on target skills was not accurately known, despite a general impression that a drop in skills use was likely. Conducting discussions of each type under controlled test conditions was one way to help determine this. Results suggest that this impression was correct – the frequency of target skills use in the closed, convergent, decision-making tasks was lower than in the open, divergent, opinion-gap tasks. This drop was inferred from the resulting test scores, which were lower in terms of overall scores and individual test scores. Further investigation into how best to use more complex tasks in level I is therefore desirable.

ACKNOWLEDGMENTS

The authors wish to thank Matthew Hartley, Marc Hasstedt, Hugh Kirkwood, Matthew M. Y. Schaefer, and Davey Young for their help administering discussion tests in lesson 14, and Christopher Nicklin, for his support with the statistical treatment of the data.

REFERENCES

Baker, D. (1989). *Language testing: A critical survey and practical guide*. London, UK: Edward Arnold.

Brereton, P., Schaefer, M. Y., Bordilovskaya, A., & Reid, S. (2019). An analysis of student survey results on classroom experiences and learning outcomes within a large-scale English discussion program. *New Directions in Teaching and Learning English Discussion*, *7*, 272-297.

Brown, J. D. (1995). *The elements of language curriculum - A systematic approach to program development.* Boston, MA: Heinle, Cengage Learning.

Candlin, C. N. (2009). Towards task-based language learning. In K. Van den Branden, M. Bygate, & J. M. Norris (Eds.), *Task-based language teaching: A reader* (pp. 21-40). Amsterdam, The Netherlands: John Benjamins.

Doe, T. (2012). Assessment: Improving rater reliability on the EDC discussion test. *New Directions in Teaching and Learning English Discussion*, *1*(1), 1.11-1.18.

Duff, P. A. (1985). Another look at interlanguage talk: Taking task to task. *University of Hawai'i Working Papers in English as a Second Language 4*(2), 1.-61.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford, UK: Oxford University Press.

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*(3), 221-246.

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research, 20*(3), 405-428.

Hurling, S. (2012). Introduction to EDC. *New Directions in Teaching and Learning English Discussion*, *1*(1), 1.2-1.10.

Kumaravadivelu, B. (1993). The name of the task and the task of naming: Methodological aspects of task-based pedagogy. In G. Crookes & S.M. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 69-96). Clevedon, UK: Multilingual Matters.

Li, S., Ellis, R., & Zhu, Y. (2016). Task-based versus task-supported language instruction: an experimental study. *Annual Review of Applied Linguistics, 36,* 205-229.

Lesley, J. (2018). *Developing tasks in high-level discussion classes.* Case study presentation at the JALT Cue-SIG Conference, September 15, 2018, Tokyo, Japan.

Lesley, J., & Brereton, P. (2018). *Designing principled academic discussion Tasks*. Workshop at the JALT International Conference, November 26, 2018, Shizuoka, Japan.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. West Sussex, UK: Wiley Blackwell.

Loschky, L., & Bley-Vroman, R. (1993). Grammar and task based methodology. In G. Crookes & S. M. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 123-167). Clevedon, UK: Multilingual Matters.

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

Nation, I. S. P. & Macalister, J. (2010). *Language curriculum design*. New York, NY: Routledge.

Nunan, D. (2004). *Task-based language teaching*. Cambridge, UK: Cambridge University Press.

Plonsky, L. & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics, 36*, 73-97.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford, UK: Oxford University Press.

Robinson, P., & Ross, S. (1996). The development of task-based assessment in English for academic purposes programs. *Applied Linguistics*, *17*(4), 455-476.

Schaefer, M. Y. (2018). Communicative skills for strategic competence, *New Directions in Teaching and Learning English Discussion*, *6*, 273-279.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183-205). Cambridge, UK: Cambridge University Press.

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, *12*(3), 287-301.

Willis, J. (1996). *A framework for task-based learning*. London, UK: Longman.

Appendix A – Spring Semester Lesson 14 Test Materials

Lesson Goals:

- ✧ Review the semester 1 discussion skills
- ✧ Review the semester 1 communication skills
- ✧ Discuss how to have a good discussion

3-2-1 Fluency

Talk to a partner. Say as much as you can. Don't worry about grammar or vocabulary!

1. What was good about your first semester at university?
2. What was challenging about your first semester at university?
3. What is challenging about doing group discussions in English?

Practice

Discuss the following topics. Use all of the discussion and communication skills.

● *Participating*	How can discussion group members achieve equal participation?
● *Listening*	What does a good listener do in a discussion?
● *Checking Understanding*	What is the most common cause of communication breakdowns in group discussions?
● *Disagreeing*	What makes it difficult for some people to disagree with others?
● *Language Accuracy*	Which is more important – speaking accurately or speaking fluently?
● *Giving Opinions*	Is it important to always give your honest opinion?
● *Time Management*	How can you manage time efficiently during discussions?
● *Group Decisions*	What is difficult about making decisions in a group?

How to Have a Good Discussion

Pre-Task: A group of first-year, Level 1 English Discussion Class (EDC) students are meeting to discuss what advice to pass on to next year's EDC students. What is most important for having a good discussion? (Choose three.)

___ A: Listening carefully to other people's ideas

___ B: Giving your honest opinion

___ C: Changing your opinion about a topic

___ D: Participating equally

___ E: Asking questions about other people's ideas

___ F: Not being afraid to disagree

___ G: Making sure everyone understands

___ H: Using language accurately

___ I: Managing time carefully

___ J: (your own idea)

Meeting Task: Discuss and decide in your group…

1. What is most important for having a good discussion? Choose three within the time limit.

   i. ____ ii. ____ iii. ____

2. What things are important when making a group decision?

Appendix B – Fall Semester Lesson 14 Test Materials

Lesson Goals:

- ✧ Review the semester 2 discussion skills
- ✧ Review the semester 2 communication skills
- ✧ Discuss developing skills and personal qualities at university

3-2-1 Fluency

Talk to a partner. Say as much as you can. Don't worry about grammar or vocabulary!

1. What things were most memorable about your first year at university? (e.g. people, places, activities / events)
2. What skills and personal qualities have you developed at university?
3. What do you want to do next year at university?

Practice

Discuss the following topics. Use all of the discussion and communication skills.

● *Being good at communicating with others*	Is it important to be good at communicating with others?
● *Being independent*	What are some good ways for university students to become more independent?
● *Having time management skills*	Why can time management be difficult for some university students?
● *Working well with different types of people*	Is it important to be able to work well with different types of people?
● *Being responsible*	How can university students become more responsible?
● *Being a leader*	Is it important for all university students to develop leadership skills?
● *Having foreign language skills*	How can students use foreign language skills after graduating from university?
● *Having specialized knowledge*	How can students use specialized knowledge after graduating from university?
● Job hunting	What is the best way to get a full-time job?

Developing Skills and Personal Qualities at University

Pre-Task: A group of first-year, Level 1 English Discussion Class (EDC) students are meeting to discuss what advice to pass on to next year's EDC students. What skills and personal qualities are most important for university students to develop? (Choose three.)

___ A: Being good at communicating with others

___ B: Being independent

___ C: Having time management skills

___ D: Working well with different types of people

___ E: Being responsible

___ F: Being a leader

___ G: Having specialized knowledge

___ H: Having foreign language skills

___ I: Being good at writing job applications

___ J: (your own idea)

Meeting Task: Discuss and decide in your group…

1. What skills and personal qualities are most important for university students to develop? Choose three within the time limit.

   i. ____ ii. ____ iii. ____

2. What are some good ways to develop these skills and personal qualities?

Appendix C – Discussion Test Score Sheet

LEVEL 1

Date: mm / dd / 2019 Instructor: first family

Fall 2018: EDC Discussion Test 3

Period: 1 2 3 Level: 1 Group: 1 2 3 # of Students: 3 4 5

Student 1: first family *Total /25*

Listener		Speaker	
<Communication Skills> *Score /5*			
React □□□□□□			
CU-L □□□□□□		CU-S □□□□□□	
Para-L □□□□□□		Para-S □□□□□□	
Clar-L □□□□□□		Clar-S □□□□□□	
Jpn (−1) □ □ □ □ □			
<Questions> *Score /5*	<Content> *Score /5*		
CQs □□□□□□ □□□□□□	ST □□□□□□ □□□□□□		
OQs □□□□□□ □□□□□□	Ag □□□□□□ Dis □□□□□□		
<D. Skills-L> *Score /5*	<D. Skills-S> *Score /5*		
CI □□□ RO □□□	CI □□□ RO □□□		
DV □□□ Inf □□□	DV □□□ Inf □□□		
BO □□□ Com □□□	BO □□□ Com □□□		
<Notes>			

Student 2: first family *Total /25*

Listener		Speaker	
<Communication Skills> *Score /5*			
React □□□□□□			
CU-L □□□□□□		CU-S □□□□□□	
Para-L □□□□□□		Para-S □□□□□□	
Clar-L □□□□□□		Clar-S □□□□□□	
Jpn (−1) □ □ □ □ □			
<Questions> *Score /5*	<Content> *Score /5*		
CQs □□□□□□ □□□□□□	ST □□□□□□ □□□□□□		
OQs □□□□□□ □□□□□□	Ag □□□□□□ Dis □□□□□□		
<D. Skills-L> *Score /5*	<D. Skills-S> *Score /5*		
CI □□□ RO □□□	CI □□□ RO □□□		
DV □□□ Inf □□□	DV □□□ Inf □□□		
BO □□□ Com □□□	BO □□□ Com □□□		
<Notes>			

Student 3: first family *Total /25*

Listener		Speaker	
<Communication Skills> *Score /5*			
React □□□□□□			
CU-L □□□□□□		CU-S □□□□□□	
Para-L □□□□□□		Para-S □□□□□□	
Clar-L □□□□□□		Clar-S □□□□□□	
Jpn (−1) □ □ □ □ □			
<Questions> *Score /5*	<Content> *Score /5*		
CQs □□□□□□ □□□□□□	ST □□□□□□ □□□□□□		
OQs □□□□□□ □□□□□□	Ag □□□□□□ Dis □□□□□□		
<D. Skills-L> *Score /5*	<D. Skills-S> *Score /5*		
CI □□□ RO □□□	CI □□□ RO □□□		
DV □□□ Inf □□□	DV □□□ Inf □□□		
BO □□□ Com □□□	BO □□□ Com □□□		
<Notes>			

Student 4: first family *Total /25*

Listener		Speaker	
<Communication Skills> *Score /5*			
React □□□□□□			
CU-L □□□□□□		CU-S □□□□□□	
Para-L □□□□□□		Para-S □□□□□□	
Clar-L □□□□□□		Clar-S □□□□□□	
Jpn (−1) □ □ □ □ □			
<Questions> *Score /5*	<Content> *Score /5*		
CQs □□□□□□ □□□□□□	ST □□□□□□ □□□□□□		
OQs □□□□□□ □□□□□□	Ag □□□□□□ Dis □□□□□□		
<D. Skills-L> *Score /5*	<D. Skills-S> *Score /5*		
CI □□□ RO □□□	CI □□□ RO □□□		
DV □□□ Inf □□□	DV □□□ Inf □□□		
BO □□□ Com □□□	BO □□□ Com □□□		
<Notes>			

Appendix D – Discussion Test Grading Code

Communication Skills	
Active Listening: Reactions	*React*
Acting Listening: Checking Understanding (Listener)	*CU-L*
Paraphrasing Others	*Para-L*
Asking for Explanation / Asking for Repetition	*Clar-L*
Checking Understanding (Speaker)	*CU-S*
Paraphrasing Yourself	*Para-S*
Explaining / Repeating	*Clar-S*
Use of Japanese [minus points]	*Jpn*
Questions	
Content Questions	*CQs*
Organizational Questions	*OQs*
Content	
Speaking Turn	*ST*
Agreeing	*Ag*
Disagreeing	*Dis*
Discussion Skills	
Connecting Ideas	*CI*
Reconsidering Opinions	*RO*
Different Viewpoints	*DV*
Information	*Inf*
Balancing Opinions	*BO*
Comparisons	*Com*

Note: The above abbreviations for Discussion Skills are used in combination with –L (Listener) and –S (Speaker) to refer to the respective sides of the skill.

Appendix E – Discussion Test Grading Criteria

Communication Skills

Listener-side Phrases:	1/2 point each (maximum of 6 phrases counted)
Speaker-side Phrases	1/2 point each (maximum of 6 phrases counted)
Japanese Use	-1 point (maximum of 5 counted)

Note: If the overall Communication Skills total results in a half point, please round-up the score to the nearest whole number (e.g. 3.5 becomes 4).

Questions

Content Questions:	1 point each
Organizational Questions:	1 point each (maximum of 2 counted)

Content

Speaking Turns:	1 point each
Agree/Disagree:	1 point each (maximum of 1 counted)

Discussion Skills (Listener)*

5 points:	5 Discussion Skills, at least one *Balancing Opinions* and one *Comparisons*
4 points:	4 Discussion Skills, at least one *Balancing Opinions* or one *Comparisons*
3 points:	3 Discussion Skills
2 points:	2 Discussion Skills
1 point:	1 Discussion Skill

Discussion Skills (Speaker)*

5 points:	5 Discussion Skills, at least one *Balancing Opinions* and one *Comparisons*
4 points:	4 Discussion Skills, at least one *Balancing Opinions* or one *Comparisons*
3 points:	3 Discussion Skills
2 points:	2 Discussion Skills
1 point:	1 Discussion Skill

Note: "Discussion Skill" in the above criteria refers to Discussion Skills such as *Connecting Ideas* and *Information*. It does not mean individual Discussion Skill phrases.