Evaluating a High School Discussion Test

Sam Berry
Aoyama Gakuin Senior High School

Reference Data:

Berry, S. (2017). Evaluating a high school discussion test. In P. Clements, A. Krause, & H. Brown (Eds.), *Transformation in language education*. Tokyo: JALT.

This paper documents a validation of a group oral discussion test used as a classroom assessment in a high school in Japan. The test has been developed in response to Ministry of Education initiatives to promote the development of communicative abilities, including discussion skills, in compulsory high school English courses. A range of validity evidence suggests the test has promise, but has significant problems with both perceived and actual fairness. A case is made for the continued use of the discussion test, providing that further steps are taken to investigate and address the test's poor test–retest reliability.

本研究では、日本のある高等学校で使用された「グループ・オーラル・ディスカッション・テスト」(HTED)の効果を検証した。テストは高等学校の英語必修科目を通して、ディスカッションを含めたコミュニケーション能力を積極的に養う方針を打ちだした文部科学省の新学習指導要領を受けて開発された。本研究で行われた多くの検証は、テストが有効的に利用できる可能性を示す一方で、実際のもしくは認識可能な公平性において著しい問題があることを示した。テスト-再テストの信頼性を検証し、正すための処置がとられることを条件に、テストを引き続き使用することを推薦する。

Including discussion activities in English classes has long been seen as a way to encourage active learning (Pauk & Owens, 2001), develop students' English communicative skills (Kobayashi & Kitsuno, 2016), and foster critical thinking skills (Munezane, 2008). In Japan, too, as part of its revised course guidelines, the Ministry of Education, Sports, Science and Technology recommended the inclusion of discussion activities in English Expression and English Communication courses taught in high schools, and further called for an increased focus on active learning, including discussion skills, at the tertiary level (MEXT, 2009).

Despite this, discussion activities are not widely used in Japanese high schools or universities (Kobayashi & Kitsuno, 2016), and several studies have commented on the reluctance of Japanese learners to engage in group discussion activities (Miller, 1995; Sakamoto & Naotsuka, 1982). This reluctance has been attributed to cultural factors (Sato, 1990), task type (Stroud, 2014), and textbooks (Kaneko & Kimizuka, 2007). Another likely reason may be that in high schools, where tests are often the largest motivating factor (LoCastro, 1996), a failure to match classroom activities with assessment methods may mean students will not take such activities seriously. That is, without assessment tasks requiring group oral discussion skills, discussion-based activities are unlikely to be successful in the high school classroom.

Teachers may be wary of using group oral assessments due to the addition of many potential and uncontrollable influences on test-taker performance, such as test-taker assertiveness (Ockey, 2006), shyness (Bonk & Van Moere, 2004), willingness to communicate (Berry, 2004), learner acquaintanceship (O'Sullivan, 2002), gender (O'Sullivan, 2000), and language level (Iwashita, 1996).

Despite this wariness, there are several benefits in promoting the inclusion of discussion tests in high school curriculums. First, as a large number of students can be tested quickly, and because testers do not need specialized interlocutor training, group oral tests can be very practical (Ockey, 2001). Furthermore, Hilsdon (1995) argued that group oral tests can provide positive washback for classrooms. This is especially true in teaching contexts that are aimed at promoting more communicative teaching and learning (Shohamy, Reves, & Bejarano, 1986)

There is also a case to be made that group oral tests have the potential to be fairer and more valid measures of oral skills than traditional interview-style tests. Fulcher (1996) argued that group discussion tasks may be more authentic than interview-style tests, which result in inauthentic discourse (Van Lier, 1989). This is important, as more authentic tasks arguably lead to more valid score interpretations (Ockey, 2001). Similarly, group oral tests remove the influence of the interviewer on test performance and score as demonstrated in

Brown (2003). Finally, Van Moere (2006) has suggested group tasks may be less intimidating for students, a fact that may be particularly true in high school contexts.

There is arguably then a need for group discussion assessments in high school curriculums. In response to this, a group discussion test, the High School Test of English Discussion (HTED), has been developed at a private senior high school in Tokyo. The hope is that a reliable, fair, and trustworthy discussion test will help to bridge a gap between classroom activities and classroom assessment and, moreover, will in turn foster a more positive attitude towards participation in group discussions. The test is in its early stages, and consequently a full evaluation was considered necessary in order to determine the usefulness of the test and to advise on decisions over its continued use.

Evaluating Speaking Tests

The evaluating of language tests in general often refers to a test's "usefulness" (Bachman & Palmer, 1996). Shaw and Weir (2007) updated and streamlined Bachman and Palmer's work on test usefulness and coined the acronym VRIP, referring to test validity, reliability, impact, and practicality. Reliability generally refers to the degree to which we can trust the test score, and validity is generally used to refer to how well the test measures what it is supposed to measure (Akbari, 2012).

Validity is often broken down into four distinct areas: construct validity, content validity, criterion-related validity, and face validity. Messick (1989) argued that test makers and test evaluators should use these "validities" to present a validity argument. However, particularly in the case of oral assessments, there remains a need for a practical test validation framework, for without one, test makers may concentrate disproportionately on validity evidence that supports their test (O'Sullivan, 2011).

Weir (2005) attempted to provide workable frameworks for reading, writing, listening, and speaking tests, yet his comprehensive models may be beyond the scope of most classroom assessors. More realistically, following Underhill's (1987) advice, classroom assessors perhaps should seek to collect as much validity evidence as possible and use this evidence to present a "validity argument" as proposed by Messick (1989). Among the "validities" identified by Weir (2005), scoring validity, consequential validity, criterion-related validity, and face validity may be of particular relevance and may be most accessible to classroom assessors. Scoring validity traditionally refers to reliability and for oral assessments typically encompasses interrater reliability (how closely different raters agree on the test taker's performance), intrarater reliability (the extent raters agree with their own scores for the same test-taker performance over time), and test–retest reliability (how similar are the test taker's scores through multiple administrations of the test). Consequential validity concerns issues of washback and impact. In criterion-referenced validity evidence, the test takers' scores are compared or contrasted with other measures of ability. Face validity generally refers to how the test is perceived by the test takers.

About the HTED

The HTED is one of eight oral assessments employed throughout the 1st year of the compulsory English Expression 1 course. The other assessments include paired conversations, interview-style tests, and presentations. The HTED itself is a 5-minute discussion in groups of four. The students are given one of four discussion topic picture cards (see Appendix A for an example): how to improve the school, the best destination for the school trip, the best present for a classmate, or the most appropriate stores for a new shopping center near the school. All of the topics are practiced in class, and both the topics and group members are chosen at random. Students are scored in three categories: language, interaction (including active listening, participation, etc.) and discussion language and skills (see Appendix B for mark sheet).

Research Aim

The aim of the study is to investigate the following test properties of the HTED: scoring validity, consequential validity, criterion-related validity, face validity, and practicality, and in doing so, to make a judgment on the overall usefulness of the test and its continued use.

Method

Participants were 142 first-year high school students who took the test as part of the compulsory English Expression I course. The test properties of the research question were investigated in the following ways.

Video recordings of 28 students who had taken the test were used to investigate scoring validity. The decision to use this relatively small number was made due to time considerations and time demands on the raters. The tests were watched and rated separately by two markers, both of whom are teachers of the course. These ratings were used to investigate interrater and test–retest reliability. The tests were rated again after 1 month, and these scores were used to investigate intrarater reliability. These figures are reported as Pearson product-moment correlation coefficients. Consequential validity was investigated through informal follow-up interviews with teachers on the English Expression I course and through a student questionnaire (see Appendix C).

Criterion-related validity was investigated through two measures: the student scores on the Test of English for International Communication (TOEIC) Bridge test and the student scores in a paired-discussion test taken in the previous term.

The student questionnaire (see Appendix C) was administered to the 142 students who took the test to investigate face validity. The survey was anonymous and questions were given in Japanese and English. The survey was based on an example provided by Fulcher and Davidson (2007) and consisted of six multiple-choice questions and two open-ended questions. Finally, observations were made to determine the practicality of the test, and these will be discussed in the following section.

Results and Discussion

Scoring Validity

Interrater Reliability

The interrater reliability was 0.82. For an oral assessment, this is somewhat higher than is usually reported in the literature. Luoma (2004) suggested oral assessments with interrater agreement in the 0.8-0.9 range might be considered very strong. Shohamy et al. (1986) compared group oral ratings with other oral assessments (interview task, picture description, and reporting task) and reported an interrater reliability of 0.71, which was the lowest of the four tasks. In a study of a four-student group oral test in a university in Japan, Van Moere (2006) reported an interrater agreement of 0.74.

One reason for the high rater agreement in the HTED may be that only three scoring categories were used, as opposed to the five scoring categories in Van Moere's (2006) study. This is further supported by Luoma (2004), who suggested the cognitive load of raters, and presumably the ability to assign accurate judgments, begins to be stretched at four or five scoring categories. Furthermore, the two raters of this study are both trained and experienced examiners for both the International English Language Testing System (IELTS) and Cambridge speaking exams, and being the sole practitioners of the HTED at this time, perhaps they have a more intuitive understanding of test performance. Although the scoring validity is encouraging in the sense that the two teachers currently responsible for the test are achieving a high level of agreement, the fact that the investigations of scoring validity rely on just two raters' judgments calls for caution when attempting to draw conclusions on a wider scale.

Interestingly, as Table 1 indicates, the interrater agreement by individual scoring category is not as high as the overall agreement. This suggests that although the raters agree generally on what a good performance on the discussion test is, they do not necessarily agree on what that entails. This may indicate a need for refinement in the scoring criteria.

Table 1. Interrater Reliability

Criterion	Score
overall	0.82
language	0.75
interaction	0.78
discussion language and skills	0.74

Intrarater Reliability

The intrarater reliability is 0.76, which again is within an acceptable range (Brown, 2004; Alderson, Clapham, & Wall, 1995).

Test–Retest Reliability

The test–retest reliability is 0.60. This is in line with Van Moere's (2006) study, which reported a test–retest reliability of 0.61. Luoma (2004) pointed out that correlations in the 0.5-0.6 range might be considered worryingly weak.

Test–retest is sometimes considered a tricky form of test reliability (Hughes, 2003) because it is often difficult to recreate the circumstances and motivation of the first test administration. In the case of the HTED, in order to encourage students to perform at their best on both occasions, some students (*n* = 20) were asked to take the test twice on the same day and, following Van Moere (2006), were told either of the two scores may be used. Although these steps were taken to maintain similar circumstances between the two tests, each student who retook the test encountered different partners for the second administration. This seems to suggest that a large part of the score variance may be caused by the influence of other group members. This is supported by the test–retest correlations by individual scoring category as shown in Table 2. Although the test taker's language performance is consistent across the two tests, the extremely low coefficient in the interaction category (0.14) implies virtually no correlation at all.

Table 2. Test–Retest Reliability

Criterion	Score
overall	0.60
language	0.91
interaction	0.14
discussion language and skills	0.32

This clearly casts doubt on the fairness of the test and consequently on the validity of the test overall. It is outside the scope of this paper to investigate the specific reasons for this, but student comments and the results of the student questionnaire (discussed further in the face validity section) suggest students view learner acquaintanceship (i.e., how well they know their fellow group members) as the biggest threat to the fairness of the test. Follow-up research could try to determine exactly what factors are influencing the score variance. For example, is it actually the degree of familiarity, or could it be the perceived level of fellow group members or the personalities of fellow group members?

Although it is worth investigating exactly what interpersonal factors are causing the score variance, the bigger problem perhaps for classroom assessors is how to deal with them. For example, it does not seem practical, or even possible, to ensure students are placed in groups of roughly equal "acquaintanceship." It may be more practical to look at the rating criteria and modify or reduce the weighting of the interaction category, particularly elements that are likely to reward students who are in groups with friends. The concern then, however, is how to ensure that the test remains a test of discussion abilities rather than merely one of proficiency.

Consequential Validity

Combining observations on washback in the literature, we believe the following claims about the HTED are tentatively supported:

- There is a link between the test and the goals of the course (Bailey, 1996).
- The abilities of skills we wish to encourage are being tested (Hughes, 2003).
- The skills are tested directly (Wall, 1996).
- The test is criterion-referenced (Hughes, 2003).
- The test is perceived to be important (Weir, 2005).
- The test is understood by both students and teachers (Bailey, 1996; Hughes, 2003).
- Students spend a lot of time in class practicing the skills necessary for the test (Weir, 2005).

Bailey (1996) further suggested positive washback could be fostered through the avoidance of single score reporting in favor of detailed score reporting. In the case of the HTED, although students receive a score for each scoring category, the categories are perhaps too vague to realistically provide useful feedback to the students. For instance, it is difficult to imagine what inferences a student would make about receiving a "6" in "language." It may be worthwhile to include a more specific checklist or tick boxes on the mark sheet for teachers to provide more specific feedback.

Bailey (1996) also pointed out that an important facet of retaining the consequential validity of a test is to ensure the results are believable and credible to test takers. Considering the test–retest validity results and some of the results discussed in the face validity, this is an obvious weakness of the test.

Criterion Validity

As shown in Table 3, there is clearly an extremely weak correlation (0.31) between the discussion test and the TOEIC Bridge test. The TOEIC Bridge test is a simplified version of the TOEIC Reading and Listening test designed for beginner to intermediate learners and is taken by all students upon entry to the high school. Although some studies have suggested TOEIC test scores may be an appropriate indicator of oral proficiency (Lee, 2006; Woodford, 1982), many studies have found moderate to weaker correlations between TOEIC scores and oral proficiency (Cunningham, 2002; Hirai, 2002; Liao, Qu, & Morgan, 2010). It is perhaps unsurprising that TOEIC scores do not correlate with group discussion test scores, but it does cast doubt on the appropriateness of the TOEIC Bridge test and TOEIC tests, in general, as placement and achievement tests on courses with discussion skill-based and communicative goals.

Table 3. Correlation Between HTED and Two Measures of Student Ability

Measure	HTED
TOEIC Bridge	0.31
Pair discussion rating	0.64

The relationship between the HTED scores and the scores of a Pair Discussion test is perhaps more enlightening. The Pair Discussion test is a test based on part two of the Cambridge Preliminary English Test. The correlation between the tests is moderate (0.64), and may tentatively support findings that group oral assessments test some different skill sets to individual and pair tests (Shohamy et al., 1986). This would further seem to lend support to Fulcher's (1996) recommendation that group oral tests be used as part of a series of oral assessments.

Face Validity

The results of the face validity survey are expressed in percentages and are presented in Table 4.

Table 4. Face Validity Student Survey Results, Percentages (*N* = 142)

Statement	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
1. The test gives an accurate idea of my ability to have a discussion in English.	5.6	59.9	16.9	15.4	2.1
2. I believe I would get a different score if I had different partners.	15	42.1	20.7	19.3	2.9
3. I believe I would get a different score if I had a different topic.	21.6	45.5	17.9	13.4	1.5
4. The test gave me a good idea of what I need to improve.	30.4	62.2	5.9	1.5	0
5. I believe the discussion test is related to activities we do in class.	43.3	48.9	4.3	3.5	0
6. I believe the discussion test is good for me.	32.5	59.0	7.9	0.7	0

The results seem to speak for themselves and suggest that the students see the test as useful and worthwhile. The positives are tempered somewhat by the results of Statements 4 and 5, which relate to the perceived fairness of the test. The majority of students believe that their score would be different had they been grouped differently, a claim that is strongly supported by the test–retest results.

Indeed, this is a good example of how the different validities interact and affect each other. The poor scoring reliability clearly affects the face validity of the test, which is seen by many students to lack fairness. In turn, this is likely to further impact the consequential validity of the test, because students and teachers may not be receiving reliable information about test performance. It may be possible to remedy the poor test–retest reliability by having students take the test on multiple occasions or by attempting to ensure groups of equal "acquaintanceship." Such steps would likely have a positive effect on the HTED's face validity but conversely would negatively affect the test practicality.

Test Practicality

Although Hughes (2003) and Weir (2005) suggested that practicality concerns should typically come after validity concerns, Davies (1990) pointed out that a test cannot exist if it is not practical. This is arguably even truer for classroom assessors. Indeed, the HTED has been developed within the context of several practicality constraints. For example, the test length of 5 minutes was chosen not because it was deemed to be enough time to elicit a reliable sample of student speech, but because it was estimated to be the maximum length of time possible in which to test classes of up to 28 students. Obviously, decisions like these have consequences on other areas of the test's validity. Some of the observations regarding the practicality of the HTED are listed below:

- It was possible to test a class of 28 students in one 50-minute period.
- Test scores were given back to students the following week.
- During classes, the tests were administered and graded by a single teacher.
- The rater needed only a set of topic cards, mark sheets, grading criteria, and a timer.
- There were no additional costs.

Because of the space required to seat four students together, it was felt that it would be easier to hold tests in a larger classroom than the classrooms where lessons are normally held. This may be problematic during certain times of the academic year and may not be possible in other teaching contexts.

Conclusion

This study has presented a range of evidence into the validity and overall usefulness of a high school group discussion test. On the positive side, it is felt that the HTED is a highly practical and useable test within the high school context. The consequential and criterion-related validity evidence paint a positive picture of a test that is providing good opportunities for washback and that appears to be testing skills not assessed in other tests. These positives are tempered somewhat by the mixed face-validity evidence, particularly in regards to the test's perceived fairness. Furthermore, the scoring validity could be considered low to moderate and considering its likely impact on other validities, might be considered a serious threat to the HTED's overall validity.

Interpretations of this study's results should bear in mind the following limitations. First, as noted, the test's scoring validity was investigated using a very small sample (n = 28), and moreover, much of the empirical data relies on correlation coefficients, which can be strongly affected by smaller sample sizes. In addition, the scoring validity data is reliant on just two raters. Finally, at this stage there has been no attempt to investigate rater severity.

Even with these limitations acknowledged, it remains clear that the HTED requires further efforts to address the concerns raised in this study. Yet, the test shows initial promise and fits well with both course aims and government policy goals, and it is hoped that future versions of the HTED will result in a worthwhile addition to the oral assessment options on the English Expression course.

Bio Data

Sam Berry teaches at Aoyama Gakuin Senior High School. His professional interests include testing, task-based learning, CLIL, discussion skills, vocabulary instruction, global issues, and extensive listening.

References

Akbari, R. (2012). Validity in language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 30-36)*.* Cambridge, England: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13,* 257-279.

Berry, V. E. (2004). *A study of the interaction between individual personality differences and oral performance test facets* (Unpublished doctoral dissertation). King's College, University of London, England.

Bonk, W. J., & Van Moere, A. (2004). *L2 group oral testing: the influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores.* Paper presented at the Language Testing Research Colloquium.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20,* 1-25.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York, NY: Longman.

Cunningham, C. (2002). *The TOEIC test and communicative competence: Do test score gains correlate with increased competence* (Unpublished master's thesis). University of Birmingham, England.

Davies, A. (1990). *Principles of language testing*. Oxford, England: Blackwell.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13,* 23-51.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, England: Routledge.

Hilsdon, J. (1995). The group oral exam: Advantages and limitations. In J. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 189-197). Hertfordshire, England: Prentice Hall International.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *6*(3), 2-8.

Hughes, A. (2003). *Testing for language teachers*. Cambridge, England: Cambridge University Press.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, *5*(2), 51-66.

Kaneko, T., & Kimizuka, J. (2007). Teaching how to manage discussions in English at Japanese college [1]: How discussion techniques could be taught effectively. *Studies in Teaching Strategies, Ibaraki University, 26,* 75-87.

Kobayashi, Y., & Kitsuno, J. (2016). Discussions without argument in English classrooms. In P. Clements, A. Krause, & H. Brown (Eds.), *Focus on the learner* (pp. 348-354). Tokyo: JALT.

Lee, I. (2006). The effectiveness of TOEIC scores on English oral proficiency. *Modern English Education, 7*(1), 33-52.

Liao, C. W., Qu, Y., & Morgan, R. (2010). The relationships of test scores measured by the TOEIC listening and reading test and TOEIC speaking and writing tests. *TOEIC Compendium*, *10*(13), 1-15.

LoCastro, V. (1996). English language education in Japan. In H. Coleman (Ed.), *Society and the language classroom* (pp. 40-58). Cambridge, England: Cambridge University Press.

Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Miller, T. (1995). Japanese learners' reactions to communicative English lessons. *JALT Journal, 17.* 31-52.

Ministry of Education, Sports, Science and Technology (MEXT). (2009). *Koutougakkou gakushu shidou yoryo gaikokugo eigoban kariyaku* [Study of course guideline for foreign languages in senior high schools; provisional version]. Retrieved from www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/1298353.htm

Munezane, Y. (2008). Courtroom drama and jury discussion in the classroom. *The Language Teacher*, *32*(9), 3-8.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28,* 373-386.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19,* 277-295.

O'Sullivan, B. (Ed.). (2011). *Language testing: Theories and practices*. Houndmills, Basingstoke, Hampshire, England: Palgrave Macmillan.

Ockey, G. (2001). Is the oral interview superior to the group oral? *Working Papers on Language Acquisition and Education, 11,* 22-41.

Ockey, G. J. (2006). *Making a case for the group oral discussion test: The effects of personality on the group oral's score-based inferences* (Unpublished doctoral thesis). University of California.

Pauk, W., & Owens, R. J. (2013). *How to study in college*. Boston, MA: Cengage Learning.

Sakamoto, N., & Naotsuka, R. (1982). *Polite fictions: Wby Japanese and Americans seem rude to each other.* Tokyo: Kinseido.

Sato, C. J. (1990). Ethnic styles in classroom discourse. In R. C. Scarcella, E. S. Anderson, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 107-120). New York, NY: Newbury House.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge, England: Cambridge University Press.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal, 40,* 212-220.

Stroud, R. (2014). Investigating student group discussion participation. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings* (pp. 404-412). Tokyo: JALT.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge, England: Cambridge University Press.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23,* 489-508.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23,* 411-440.

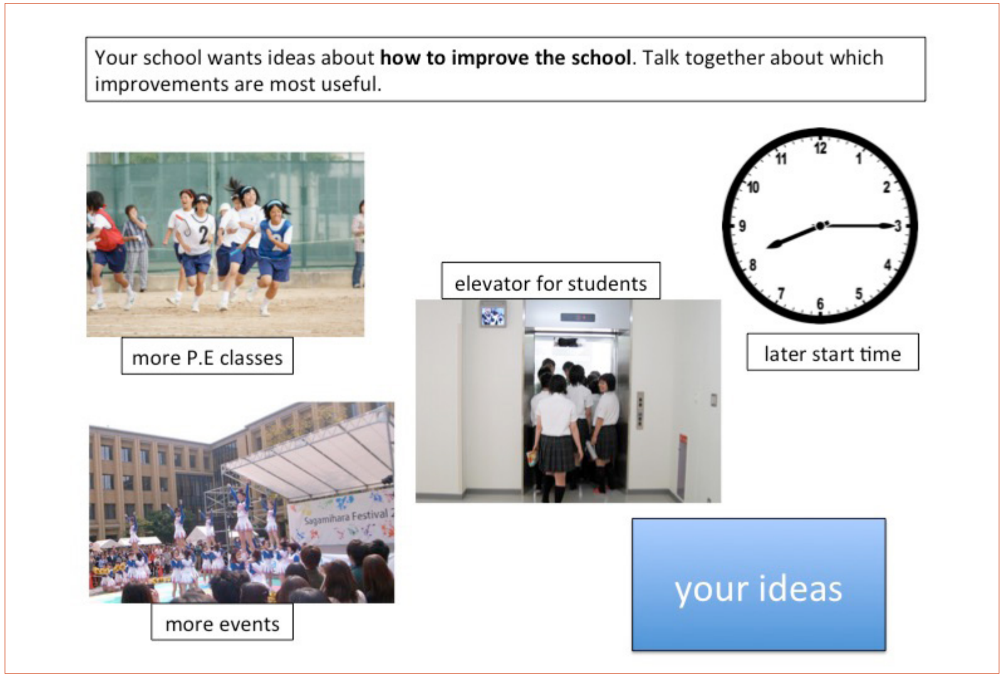
Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13,* 334-354.

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

Woodford, P. E. (1982). *An introduction to TOEIC: The initial validity study*. Princeton, NJ: Educational Testing Service.

Appendix A

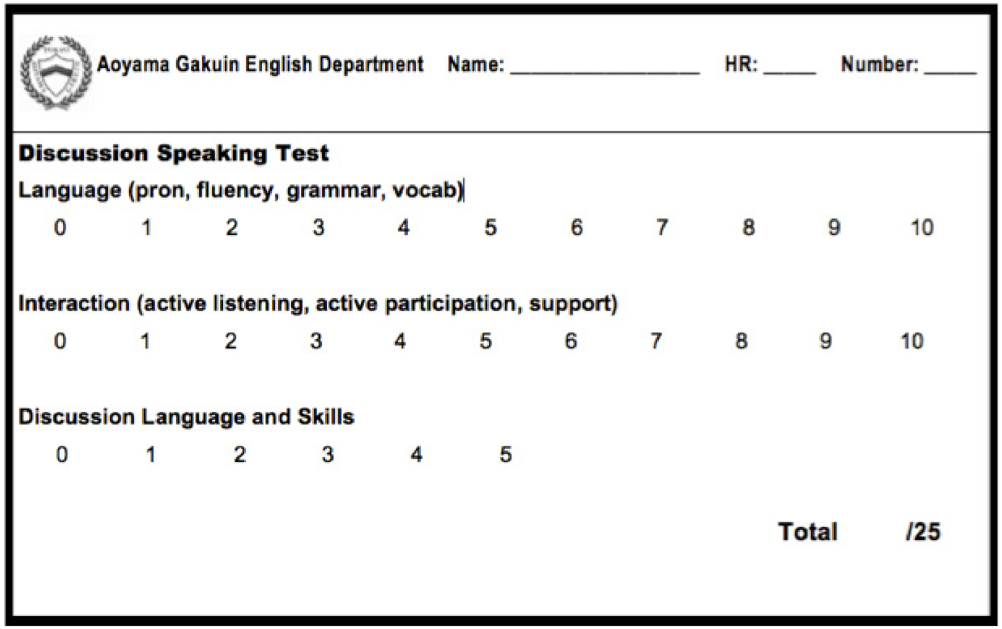
HTED Sample Topic Card



Your school wants ideas about **how to improve the school**. Talk together about which improvements are most useful.

more P.E classes
elevator for students
later start time
more events
your ideas

Appendix B
HTED Mark Sheet



Aoyama Gakuin English Department Name: ________________ HR: _____ Number: _____

Discussion Speaking Test

Language (pron, fluency, grammar, vocab)

0 1 2 3 4 5 6 7 8 9 10

Interaction (active listening, active participation, support)

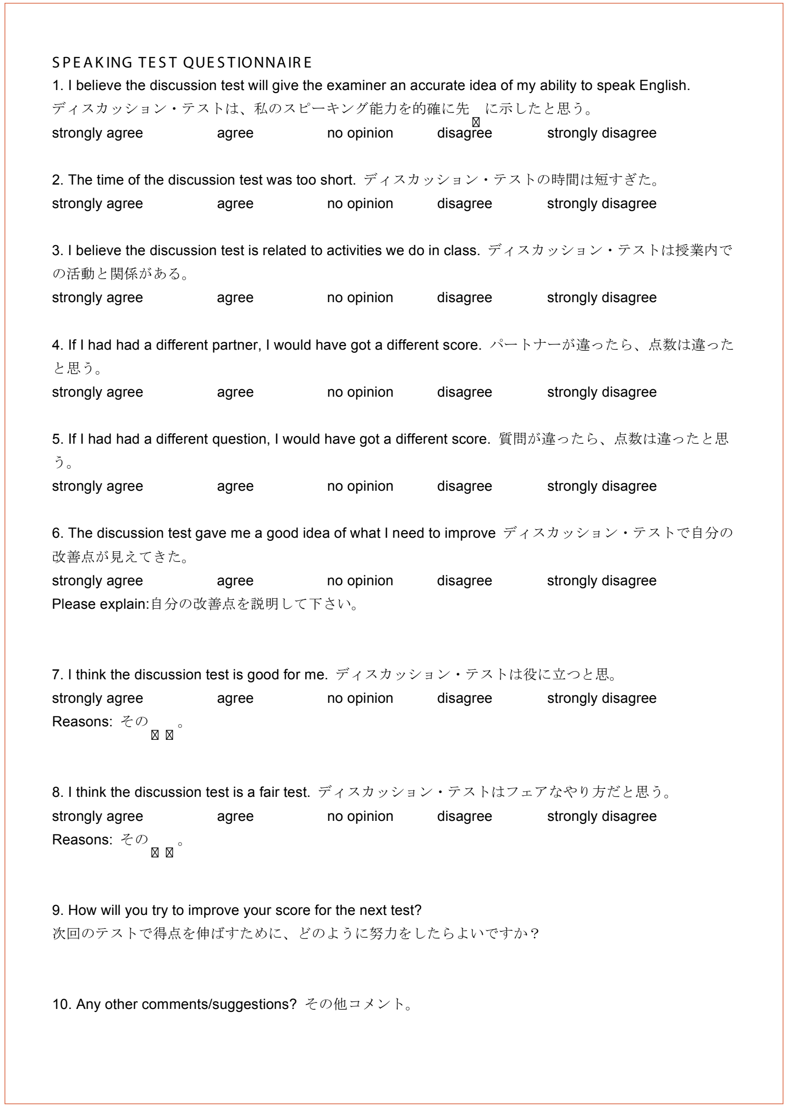
0 1 2 3 4 5 6 7 8 9 10

Discussion Language and Skills

0 1 2 3 4 5

Total /25

Appendix C
Face and Consequential Validity Student Questionnaire



SPEAKING TEST QUESTIONNAIRE

1. I believe the discussion test will give the examiner an accurate idea of my ability to speak English.
ディスカッション・テストは、私のスピーキング能力を的確に先⊠に示したと思う。
strongly agree agree no opinion disagree strongly disagree

2. The time of the discussion test was too short. ディスカッション・テストの時間は短すぎた。
strongly agree agree no opinion disagree strongly disagree

3. I believe the discussion test is related to activities we do in class. ディスカッション・テストは授業内での活動と関係がある。
strongly agree agree no opinion disagree strongly disagree

4. If I had had a different partner, I would have got a different score. パートナーが違ったら、点数は違ったと思う。
strongly agree agree no opinion disagree strongly disagree

5. If I had had a different question, I would have got a different score. 質問が違ったら、点数は違ったと思う。
strongly agree agree no opinion disagree strongly disagree

6. The discussion test gave me a good idea of what I need to improve ディスカッション・テストで自分の改善点が見えてきた。
strongly agree agree no opinion disagree strongly disagree
Please explain:自分の改善点を説明して下さい。

7. I think the discussion test is good for me. ディスカッション・テストは役に立つと思。
strongly agree agree no opinion disagree strongly disagree
Reasons: その⊠⊠。

8. I think the discussion test is a fair test. ディスカッション・テストはフェアなやり方だと思う。
strongly agree agree no opinion disagree strongly disagree
Reasons: その⊠⊠。

9. How will you try to improve your score for the next test?
次回のテストで得点を伸ばすために、どのように努力をしたらよいですか？

10. Any other comments/suggestions? その他コメント。