Validating a High School Test of English Conversation

Sam Berry
Aoyama Gakuin Senior High School

Reference Data:

Berry, S. (2015) Validating a high school test of English conversation. In P. Clements, A. Krause, & H. Brown (Eds.), *JALT2014 Conference Proceedings*. Tokyo: JALT.

This paper documents the development and validation of the High School Test of English Conversation (HTEC), a paired test of conversational proficiency designed in-house at a high school in Japan. The test is used as part of the new English Expression I course—a Ministry of Education initiative designed to promote communicative language teaching in high schools. As the test is in its infancy, a full validation is considered necessary in order to sanction its continued use and to advise future modifications. A range of validity evidence presents a rather mixed overall test validity profile and indicates the difficulties for classroom assessors in balancing the validity and practicality concerns of speaking tests. The continued use of the HTEC is cautiously backed, with the provision that a number of modifications and suggestions for improvement are met.

本研究では、ある高校で開発された「会話実力ペア・スピーキングテスト」（HTEC）の効果を検証した。HTECは高等学校におけるCLTを推進する文部科学省の「英語表現I」の一環として実施されている。HTECはまだ揺籃期にあり、将来における継続的な実施や、さらなる改善のためにも、十分な検証が必要である。本研究でおこなわれた数多くの検証は、HTECの効果の測定結果にはばらつきのあることを示し、講師にとって、スピーキング・テストの実用性と妥当性のバランスをはかることが難しいことを明らかにしている。より効果的な活用のためにはさらなる改善が必要だということを示しながら、本研究はHTECの継続的な実施を推進している。

THE SHIFT to more communicative approaches in language teaching has brought with it an acute need for tests that accurately assess learners' communicative abilities. Increasingly, test developers are obligated to show that their tests have *test usefulness*, said to comprise the qualities of reliability, construct validity, authenticity, interactiveness, impact, and practicality (Bachman & Palmer, 1996). In the testing of speaking, arguably the most subjective skill to test, two of these qualities, reliability and validity, are considered particularly important.

In Japan, the implementation of valid speaking tests in high schools has been proposed as a way of bridging the widely documented gap between the government's communicative language policies and actual classroom practice (Akiyama, 2003). The problem is that the reliability of school-based assessments tends to be low (Brindley, 1989), and teachers favor the practicality and comparative reliability of pencil-and-paper tests (Akiyama, 2003). An absence of communicative-minded assessments in high schools has an obvious impact on learning and teaching, yet speaking tests that lack reliability and validity can have equally harmful effects (Koyama & Yukawa, 2009).

These developments have motivated the development of the High School Test of English Conversation (HTEC), an in-house paired conversation test at a private senior high school in central Tokyo. The test is in its 1st year and is the first in a series of eight planned oral assessments throughout the 1st year of the new English Expression course. Therefore, there is a need to investigate the validity and overall usefulness of the test in order to determine the feasibility of its continued use and to assist in the development of future tests in the series. The purpose of this paper, therefore, is twofold: to report on an investigation of the HTEC's validity and, in assessing its overall usefulness, to propose further refinements.

Issues in Evaluating Speaking Tests

Bachman and Palmer (1996) identified the components of test usefulness as reliability, construct validity, authenticity, interactiveness, impact, and practicality. More recently, researchers such as Shaw and Weir (2007) have streamlined Bachman and Palmer's understanding of test usefulness to four qualities, collectively identified by the acronym VRIP (validity, reliability, impact, and practicality). Among these, reliability and, particularly, validity have come to dominate discussions of test evaluation.

Traditionally, validity has centered on the question of whether a "test measures what it is supposed to measure" (Akbari, 2012, p. 30). Though this is a seemingly simple question, it is a somewhat general and vague conceptualization and has traditionally been broken into four distinct kinds of validity: face validity, content validity, criterion-related validity, and construct validity (Harrison, 1983; Hughes, 2003). Not all of these have been considered equal. For example, some researchers, such as Bachman (1990), have sought to eliminate face validity, because it was not considered sufficiently empirical or evidentiary. Messick (1989) revised and refined much of the thinking on validity and argued that the classical sources of validity evidence were roughly equal and should be used to present an overall validity *argument*. Messick also pushed forward considerations of consequential validity, entailing issues of washback and test impact. Despite Messick's breakthroughs, adaptations of his work to language testing have yet to produce a "significant and practical" framework for language test validation (O'Sullivan, 2011, p. 20).

Indeed, the lack of a practical framework for test developers, and classroom assessors in particular, is a criticism that could be leveled at much of validity literature, although it has been suggested that Weir's (2005) framework (see Figure 1) may offer the beginnings of a workable and practical framework. Without one, there is a risk of test developers focusing only on the types of validity evidence that paint a positive picture of their tests (O'Sullivan, 2011).

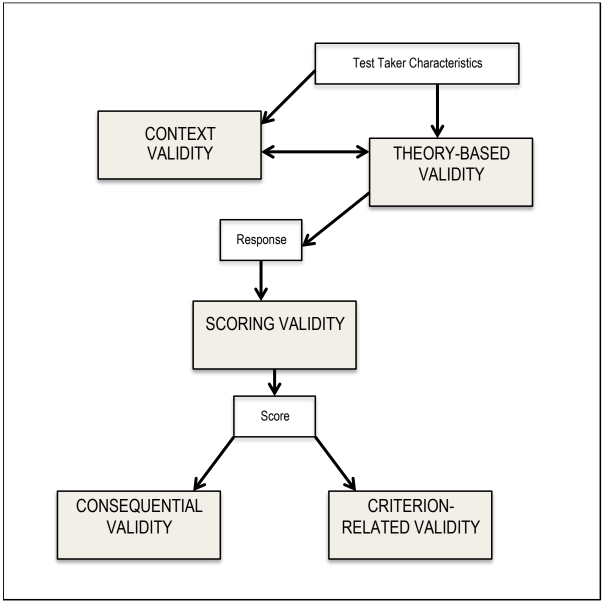


Figure 1. Weir's (2005) framework for validating speaking tests. Used with permission.

For clarity, it is perhaps worth mentioning here that Weir employed several different terms for the traditionally accepted validities. Context validity might be thought of as content validity, theory-based validity as construct validity, and scoring validity as reliability. Consequential validity refers to issues of impact and washback. Within this field, Bailey (1996) highlighted three conditions for promoting positive washback: congruence between tests and educational goals, increased self-assessment and learner autonomy, and detailed score reporting.

For classroom assessors, a lack of resources and time may make it impractical to gather enough evidence to adhere fully to frameworks such as Weir's but as Underhill (1987) advised, we should "gather as much information as possible about different types of validity" (p. 105).

Research Question

To what extent can the HTEC claim to be valid in the following test properties?

- content-related validity
- construct-related validity
- scoring validity
- consequential-related validity
- criterion-related validity
- face validity
- practicality

Method

The HTEC consists of a single open task. In pairs, students are given one of five starter topics, for example, "Who is your best friend?" or "Do you have any brothers and sisters?" From the starter topic, the students should maintain a conversation on the topic for 2 minutes by responding to and following up on each other's responses. Both the starter topic and the partner are chosen randomly to encourage practice in class and to discourage the memorization of set dialogues. Students are marked on four categories: fluency, pronunciation, grammar and vocabulary, and interactive competence. See Appendix A and B for rating scales and mark sheets. One hundred and fifty-six 1st-year students (ages 15-16) took the test. There was a range of levels, and both female and male students took the test.

The test properties raised in the research question were investigated. Content validity of the HTEC was investigated in two ways. First, the discourse of one of the tests was analyzed for occurrences of conversational features, such as follow-up questions and back channels. Second, a short questionnaire was administered to course teachers concerning the nature of the relationship between the test and the objectives of the test.

Following procedures outlined by Alderson, Clapham, and Wall (1995), the HTEC's construct validity was evaluated through two internal correlations. First, the four scoring categories were correlated with each other. Next, the correlations between each scoring category and the total score were calculated.

Video recordings of 32 students were used to investigate inter-rater, intra-rater, and test-retest reliability. The tests were watched and scored by four raters. Consequential validity was investigated through a series of follow-up interviews with teachers involved with the English Expression course and with Japanese teachers of English.

Two measures were used to investigate concurrent validity: the listening section of the TOEIC bridge test and ratings assigned to students during longer 15-minute discussions. Face validity was investigated with a 10-item questionnaire administered to all 156 students who took the test. Observations on the practicality of the HTEC will be discussed in the next section.

Results and Discussion

Content Validity

The content validity questionnaire administered to four teachers provides various insights into the HTEC's content validity. The test was felt to strongly match the goals and content of both the course and the particular course unit, and it was felt that the test required students to display a command of a wide range of features of conversational proficiency. However, it was felt that 2 minutes was not sufficient time for students to display a range of grammatical and lexical knowledge.

Additionally, pronunciation perhaps should not have been weighted as highly as the other categories. It was felt that in terms of defining conversational ability, pronunciation was not as important as interactional competence, which is effectively the only category distinguishing the rating scale as a scale for conversational proficiency as opposed to general proficiency. Related to this, the equal weighting of pronunciation was thought perhaps not to cohere with our course's focus on world English. Finally, it was suggested that conversational proficiency depended on things other than the ability to use language. Intelligence, general knowledge, personality, and quality of content were all features not reflected in the rating scale.

The discourse analysis showed the that test required students to display a range of target conversational behaviors covered in class, particularly follow-up questions and clarification back-channels. However, the nature of the task meant the students' ability to open and close conversations was not tested, which perhaps should be addressed in future administrations.

Construct Validity

High correlation between scoring categories might indicate that the categories are essentially testing the same thing. Table 1 shows the correlations among individual scoring categories. Correlations given are Pearson product-moment correlation coefficients.

Table 1. Correlations Between Scoring Categories

Category	Pronunciation	Grammar + vocabulary	Interactive competence
Fluency	.80	.67	.54
Pronunciation	-	.66	.68
Grammar + vocabulary	-	-	.42

Alderson et al. (1995, p. 184) recommended that test developers consider dropping components that correlate around .9. To this end, the correlation between fluency and pronunciation (.80) seems particularly high. To some extent, pronunciation and fluency might be expected to correlate more highly than other categories, as some features of pronunciation, such as linking and ellipsis, are intrinsically linked to fluency. However, the rating scales made no reference to these features and concentrate on pronunciation fundamentals, such as clarity and L1 interference. This may suggest that the raters were marking pronunciation intuitively rather than using the rating scales and consequently may need to adhere more strictly to the scales. Alternatively, it might be necessary to reconsider and rewrite the pronunciation-scoring category. Another possibility might be to merge the categories into a *pronunciation and fluency* category, which could potentially make marking easier and ensure that students proficient in one of the two categories are not rewarded twice. Although this seems a practical step, there does not seem enough theoretical grounding to justify merging the traits into a single category, and a reduction in the number of scoring categories would lead to a reduction in the precision of feedback on performance.

Correlations between scoring categories and the total score (see Table 2) can indicate the relative importance of each trait toward the construct.

Table 2. Correlations Between Scoring Category and Total Score

Category	Correlation
Fluency	.88
Pronunciation	.92
Grammar + Vocabulary	.80
Interaction	.79

As the total score includes the score for each category, and as there are only four categories contributing to the total score, these correlations should be considered artificially inflated. The correlations were thus recalculated to correlate the scoring categories with test total *minus* the score for that particular scoring category. These recalculated correlations are shown in Table 3.

Table 3. Correlations Between Scoring Category and Total Score Minus Scoring Category

Category	Correlation
Fluency	.78
Pronunciation	.86
Grammar + Vocabulary	.65
Interaction	.61

Two worrying trends emerge from the data. First, the high correlation (.86) of pronunciation to the total score suggests that in our understanding of conversational proficiency, it is the trait of pronunciation that is most important. This is a concern as the dependency of the pronunciation-scoring category has already been questioned, and it is hard to justify why pronunciation should be the most important component of conversational proficiency.

Additionally, scores for interactional competence correlated comparatively weakly (.61) with the total score. Considering that the other three scoring categories (fluency, pronunciation, grammar and vocabulary) could feasibly be categories of any oral test, and that interactional competence is the only category that details features specific to conversation, this is an obviously disappointing result. Remedies may involve refining the scoring criteria or experimenting with the weighting of individual categories, but it is certain that further research is required.

Scoring Validity

Table 4 shows the inter-rater reliability between the four raters. Taking an average of the correlations gives an overall inter-rater reliability correlation of .62, which, according to Luoma (2004), would be considered fairly weak. However, the extremely low correlations of Rater 2 with the other raters suggest that this rater was far out of sync. Indeed, removing Rater 2's results produces an overall inter-rater reliability of 0.74, which might be considered quite strong (Lado, 1961). This perhaps suggests a problem with Rater 2's interpretation or understanding of the rating scales rather than major problems with the rating scales themselves.

Table 4. Inter-Rater Reliability for Overall Score

Rater	1	2	3	4
1	-	.36	.71	.73
2	.36	-	.53	.62
3	.71	.53	-	.79
4	.73	.62	.79	-

Next, the ratings given by the raters for each scoring category were investigated (see Tables 5-8). This should highlight areas of rater disagreement, and point to areas of the rating scale in need of refinement or clarification.

Table 5. Inter-Rater Reliability for Fluency

Rater	1	2	3	4
1	-	.56	.74	.77
2	.56	-	.44	.47
3	.74	.44	-	.75
4	.77	.47	.75	-

Table 6. Inter-Rater Reliability for Pronunciation

Rater	1	2	3	4
1	-	.31	.47	.74
2	.31	-	.00*	.34
3	.47	.00	-	.72
4	.74	.34	.72	-

Table 7. Inter-Rater Reliability for Grammar and Vocabulary

Rater	1	2	3	4
1	-	.18	.60	.66
2	.18	-	.39	.55
3	.60	.39	-	.35
4	.66	.55	.35	-

Table 8. Inter-Rater Reliability for Interactive Competence

Rater	1	2	3	4
1	-	.04	.39	.17
2	.04	-	.59	.67
3	.39	.59	-	.72
4	.17	.67	.72	-

The weak correlations by scoring category, particularly in grammar and vocabulary and interactional competence, conflict with the stronger correlations for the total overall scores. This suggests that perhaps raters were marking holistically, that is, that raters had an overall impression of each test-taker's conversational competence and then used the scoring categories to derive a total they saw as fitting, rather than the other way around. Moreover, the lack of consistency across the individual categories had a negative effect on the accuracy of the results and consequently on the entire reliability of the test. This again points to a need for rating scale refinement, in collaboration with all raters, and a more thorough rater training process.

Rater comments written on the score sheets provide further insights. Rater 1's correlations with other raters for interactional competence (.04, .39, .17) are extremely weak and statistically insignificant. On the scoring sheets, Rater 1 recorded several observations on matters related to body language, such as "good eye contact," "highly distracting gestures," and "poor eye contact." It would seem that in Rater 1's interpretation of the interactional competence-scoring category, body language was the most important component. Again, further rater training, rater scale refinement, and greater collaboration in the rating scale development process may go some way to remedy this.

The overall intra-rater reliability was .69. This might be described as moderate (Brown, 2004, p. 158) or quite strong (Alderson et al., 1995, p. 79). The test-retest correlation is .68, which again might be described as moderate.

Consequential Validity

To investigate consequential validity, we shall refer to Bailey's (1996) three conditions for promoting positive washback. First, there was a high degree of congruence between tests and educational goals, particularly when compared to the previous reliance on presentations to measure oral ability. All course teachers have reported an increase both in the use of communicative pair-work activities and particularly in the enthusiasm of students to participate in such activities. This suggests that the test may have acted as a bridge not just between the course goals and classroom practice, but also between policy goals of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2011) and classroom practice.

Furthermore, there was increased self-assessment and learner autonomy. Part of the test-preparation procedure involved students self-rating their performance in practice tests. The exact influence of this has not been investigated, although it might be posited that students giving consideration to their strengths and weaknesses is a beneficial learning strategy to be encouraged. Additionally, as discussed in the next section, students were able to identify specific areas of conversational proficiency to improve.

Finally, the test allowed for detailed score reporting as opposed to single scores. Although students received an overall score total, they also received their score for each of the four scoring categories. This provided a range of feedback for both teacher and students, and, it is hoped, validated in the eyes of students those classroom activities that prioritize skills such as fluency or pronunciation.

Criterion Validity

The extremely low level of correlation (.18) between the HTEC and the students' TOEIC bridge listening scores indicates no relation of any significance between the two measures and is perhaps at first glance quite worrying. The TOEIC test is designed as a test of communication and claims high correlation with direct measures of speaking ability such as the Language Proficiency Interview (see Woodford, 1982). In defense of the HTEC, the topics and language of the business-oriented TOEIC test should and do differ greatly from a test designed for 1st-year high school students. Additionally, more recent studies (see Buck, 2001) have seen a growing movement that questions the suitability of TOEIC test scores as an accurate predictor of communicative ability.

For the second measure of concurrent validity, ratings from group discussions were used, and these produced a far more encouraging correlation coefficient (.69) with the HTEC. This positive result should be treated with some caution, however. First, although the topics of the group discussions were standardized (family, hometown, club, school, experiences abroad), the raters, level of group members, discussion length, and the degree of teacher intervention during the discussions were all nonstandardized, and as such cannot be treated as (nor were they intended as) a reliable measure. These limitations acknowledged, we might tentatively conclude that the HTEC, to a certain extent, may appear to predict student performance over a longer time period, in a similar task on similar topics.

Face Validity

The face validity questionnaires were anonymous, and the questions were written in English and Japanese. Eight of the questions were multiple-choice, and the results are presented in Table 9. All responses are expressed in percentages.

Table 9. Face Validity Questionnaire Results (*N* = 156)

Statement	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
1. The test gives an accurate idea of my English conversational ability.	5	40	37	18	0
2. I think the time of the test was too short.	5	10	43	40	2
3. The test is related to activities we do in class.	40	46	9	4	1
4. If I had had a different partner, I would have got a better score.	14	21	27	17	21
5. If I had had a different question, I would have got a different score.	10	34	25	25	6
6. The test gave me a good idea of what I need to improve.	20	60	9	10	1
7. Overall, I believe the test was good for me.	40	49	2	8	1
8. Overall, I think the test is a fair test.	8	51	28	10	3

The results of the face validity questionnaire are fairly self-explanatory and do not warrant an in-depth explanation of each point. In general, three main trends of interest emerge. First, there is some work needed to convince students that the test gives an accurate impression of their conversational ability. Second, although there are issues with perceptions of fairness relating to the assigned topic and partner, overall, a majority of students felt the HTEC was a fair test. Finally, a large majority of students felt the test had relevance to class activities and felt the experience was beneficial.

Test Practicality

The HTEC was designed primarily from the standpoint of practicality. Although this conflicts with Weir's (2005) view that practicality concerns should come *after* validity concerns, the HTEC, as perhaps with most high school and classroom-based tests, is constrained by practical concerns. In short, without being able to be effectively administered using the available resources, the test could not exist. For instance, the test's length of 2 minutes was chosen as it was estimated to be the maximum test length to enable classes of up to 32 students to be tested in one 50-minute period and leave time for the teacher to set up the test space and give out and collect work for students to do while waiting. Additionally, due to institutional constraints, tests must be administered by classroom teachers and must be marked live.

Overall, the HTEC was felt to have a very high degree of practicality. All class tests (excluding absent students) were completed during the 50-minute testing period, and test scores were returned to students the following week. An additional room for testing was not required, and raters needed only marking sheets, a rating scale, a timer, and one set of question prompt cards. All tests were administered and rated by the class teacher (no second rater), and there were no additional costs associated with test administration.

Summary of Findings

Placing the results onto some kind of validity scale proves difficult. What may be acceptable for one test may not be acceptable for another. For instance, high stakes test developers should obviously prioritize validity concerns over issues of practicality, whereas for classroom assessments, practicality issues may be the most important factor in shaping many of the decisions taken during the test development process. The process of test validation may never be able to answer the question "Is this test valid?" but rather attempts to answer in terms of degree and relies on how we interpret these degrees. With that in mind, in response to the research question, the validation of the HTEC cautiously offers the following observations.

First, the teacher questionnaires and discourse analysis suggested the test appeared to have quite good content-related validity. On the other hand, issues with the performance of scoring categories and raters mean the construct validity and scoring validity could be described only as moderately satisfactory. Teacher feedback suggested the test might claim good consequential-related validity. The data on criterion-related validity is promising, yet inconclusive at this stage. The positive observations on face validity and practicality were highly satisfactory within the HTEC's context.

Limitations

Consideration of the results should bear in mind the following limitations of the study. First, the low number of participants (*N* = 32) for the scoring validity research discussed in this paper calls for extreme caution when interpreting or generalizing the data. In addition, the study did not seek to investigate rater severity. Particularly in a high school context, where rumors of *strict raters* and *kind raters* are not difficult to imagine, ensuring and demonstrating rater fairness is of high importance to a test, particularly its face validity. Finally, the study did not investigate parallel form reliability.

Although these limitations mean caution should be exercised when interpreting the data, it is felt that the validity evidence collected represents a principled and unbiased approach to classroom assessment validation based on Weir's (2005) framework and a practical base from which to evaluate further administrations and other future oral assessments.

Conclusion

This paper presented a range of validity-related evidence used to determine the usefulness of the HTEC and ultimately make a decision on its continued use, using Weir's (2005) framework for speaking test validation as a guide. The results presented a mixed validity profile of the test with some promising results (face validity and practicality) and some areas in need of refinement (construct and scoring validity). If, however, as MEXT course guidelines suggest, a more communicative approach to language teaching will lead to more communicative students, then oral assessments in high schools such as the HTEC may contribute in part to this goal.

However, such claims should be made extremely tentatively. For high schools, teachers, and students, two components of testing—practicality and fairness—are perhaps of most importance. Although in this paper I have argued that the HTEC can claim a high degree of practicality, a similar case cannot be made in support of the test's fairness. This is expressed most clearly in in the fairly unsatisfactory results of the test's scoring validity.

Despite this, the HTEC is in its infancy and consequently it must be expected that there is room for improvement, and certainly, in our context at least, the test shows promise. Therefore, the continued use of the HTEC is cautiously recommended, with the proviso that a range of targets for improvement is met.

References

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

Akiyama, T. (2003). Assessing speaking: Issues in school-based assessment and the introduction of speaking tests into the Japanese senior high school entrance examinations. *JALT Journal*, *25*, 117-141.

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

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

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

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

Brindley, G. (1989). *Assessing achievement in the learner-centered curriculum*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

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

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

Harrison, A., 1983. Communicative testing: Jam tomorrow? In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 77-85). London: Academic Press.

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

Koyama, T., & Yukawa, E. (2012). Validity of the YTK speaking test: Construct validity of a performance-based English speaking test for elementary school students in Japan. *Bulletin of Kyoto Notre Dame University, 42*, 25-42.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book.* New York: McGraw-Hill.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2011). *Senior high school government course guidelines (foreign language activities)*. Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/__icsFiles/afieldfile/2011/04/11/1298353_9.pdf

O'Sullivan, D. B. (Ed.). (2011). *Language testing: Theories and practices*. Oxford: Palgrave.

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

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

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

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

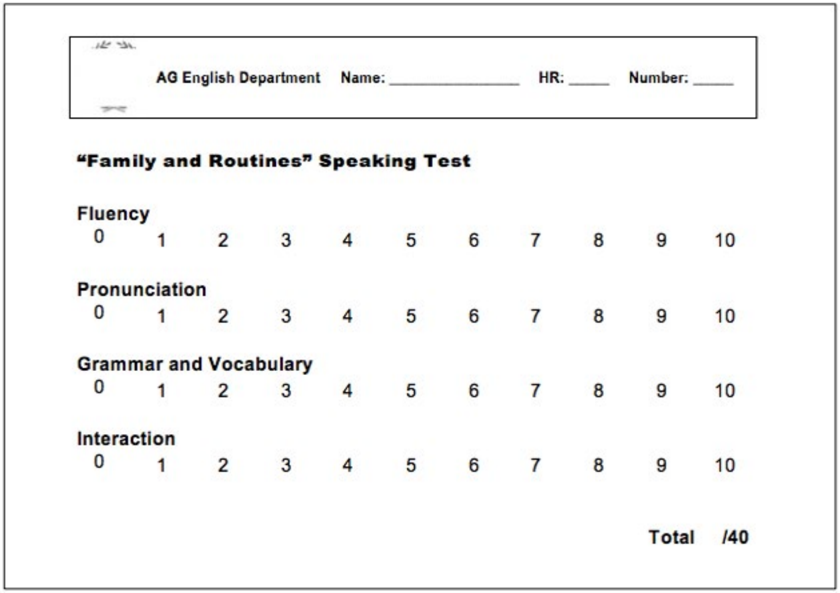
Appendix A

Rating Scales Used in the HTEC

	Fluency	Pronunciation	Grammar and Vocabulary	Interactional Competence
10	Displays all the positive features of 9 at all points throughout the test.			
9	Smooth delivery of answers, comments, reactions and command of most question forms. Pauses are likely for content rather than language.	Very clear. Successful use of stress and intonation.	Generally accurate, and shows a willingness and ability to use some complex language forms.	Shows an awareness of turn-taking norms. Makes appropriate reactions, follow-up questions and comments to partner's turn. Helps to assist understanding when necessary. Helps to repair/maintain conversation when necessary. Appropriate body language, eye contact.
8	Features of 7 and 9			
7	Pauses, hesitations, and restarts to formulate questions and answers, though generally they do not cause listener strain. Longer turns, less common questions may be delivered more slowly.	Some signs of L1 influence, and some inaccuracies, but these do not lead to misunderstanding. Although the student makes attempts at producing intonation and stress, there may be some inappropriacies or a lack of intonation and or sentence stress, which may cause delivery to sound a little flat at times.	Overall command of basic language relevant to the task. Frequent, but minor errors that do not lead to misunderstanding.	Some reactions, and attempts at follow-up questions, but these may be monotonous, or occasionally inappropriate.
6	Features of 5 and 7			
5	Pauses, hesitations, and restarts that cause considerable strain. Most output is delivered slowly.	Strong L1 influence, volume or inaccuracies may cause misunderstanding. Speech may be very flat, lacking intonation, or sentence stress, possibly caused by a slow rate of speech.	Persistent Errors even in basic structures. Errors may lead to misunderstanding, or the need for substantial empathy from the listener. May lack lexical resources to fulfill task. Turns may consist of very short one or two word answers.	A lack of reactions and follow-up questions. A lack of participation which is damaging the interaction and/or causing the partner to work hard to maintain/repair interaction.
4	Features of 3 and 5			
3	Very little comprehensible output / A shortage of output to get a reasonable impression.	Very little comprehensible output / A shortage of output to get a reasonable impression.	Very little comprehensible output / A shortage of output to get a reasonable impression.	Very little comprehensible output / A shortage of output to get a reasonable impression.
2	Features of 1 and 3			
1	Practically no ratable language. This may be due to a refusal or inability to speak.			

Appendix B

Mark Sheets Used in the HTEC



AG English Department Name: ______________ HR: _____ Number: _____

"Family and Routines" Speaking Test

Fluency

0 1 2 3 4 5 6 7 8 9 10

Pronunciation

0 1 2 3 4 5 6 7 8 9 10

Grammar and Vocabulary

0 1 2 3 4 5 6 7 8 9 10

Interaction

0 1 2 3 4 5 6 7 8 9 10

Total /40