Developing a Test of Listening to Support the English Communication Curriculum at Sojo University

by
Elton LaClare *, Alun Roger * and Jon Rowberry *

Abstract

During the first semester of 2012/13, a listening test was developed to support the English Communication 3 course at the Sojo International Learning Center. The development team carefully considered how to design the test so as to align it as closely as possible with the construct being measured while remaining within the practical constraints imposed by the testing context. The test was administered to around 700 students in July 2012 and the results analyzed to determine the test's overall reliability as well as the extent to which specific items appropriately targeted both the population and the construct. Following this analysis, a number of changes were proposed in order to improve the quality of the test for future administrations.

Key Words:　assessment, evaluation, listening, reliability, validity, washback

1. Introduction

The Sojo International Learning Center (SILC), established at Sojo University in Kyushu, Japan in 2010, is a facility responsible for delivering the English program to students from across the university. All students are required to take English Communication courses for three hours per week during each of their first four 15-week semesters. The medium of instruction on SILC courses is English, in sharp contrast to the students' previous experience of English language classes, which are generally conducted in Japanese and rely heavily on grammar translation methods.

During the first two years of the SILC's operation it was noted that, while the vast majority of students reported that their speaking skills had improved, rather fewer had noticed improvements in their listening abilities. Initially, this seemed counter-intuitive given the amount of spoken English the students were being exposed to. Some teachers theorized that perhaps their students were not in fact getting better at listening, but had merely developed strategies to disguise or compensate for their lack of understanding. Such concerns have recently been raised in second language acquisition literature. Field (2008), for example, warns against the assumption that the more learners listen to L2, the more competent listeners they become. In fact, while some may improve, others may not. Those that are weak listeners at the outset may become demoralized and withdraw their cooperation. After all, "it is very difficult to persuade somebody to lend attention to a piece of speech if they believe they are incapable of making sense of what is being said" (Field, 2008, p. 29).

As a result, it was decided that the SILC

*Lecturer, Sojo International Learning Center

curriculum should aim to focus more explicitly on the development of listening skills and SILC lecturers began designing activities, creating resources and developing assessment procedures accordingly. In support of this, a group of teachers was assigned responsibility for developing formal listening assessments to be taken by all students at key stages of the curriculum. The first such test to be developed was for the English Communication 3 (EC3) course, which students take in the first semester of their second year at the university. Thus, the EC3 test was conceived not only as an achievement test to evaluate how effectively students had engaged with the course material, but also as a means of generating a positive washback effect on the curriculum as a whole (Messick, 1996). In addition, it was hoped that developing and validating such a test may be an important step in the longer term objective of demonstrating learning gains in listening over students' two years of study on SILC courses.

2. The EC3 Listening Test

2.1 The Construct

It could be argued that it is unrealistic for a single assessment to serve as both an achievement test tied to a specific course of study and as a wider measure of learner gain in listening skills. Such concerns may turn out to be well grounded and could certainly act as a constraint on test validity (Chapelle, 1999). However, by restricting the test to the content areas and discourse types encountered during the course, it was felt that there was a reasonable chance of capturing learner gain within the domains targeted by the curriculum. Thus the test should not be seen as a tool for measuring improvements in listening *per se* but rather as an attempt to identify gains in listening within the contexts specified in the EC3 syllabus.

Of course, it should be noted that even within these specific language domains, listening proficiency is likely to be the overriding variable. In other words, students who are poor at listening are likely to fare worse on the test than more proficient learners regardless of how well they have engaged with the EC3 curriculum. Moreover, given the wide disparities in listening skills exhibited by SILC students it would be highly unfair to use the EC3 test as the sole or main means of assessing how well students had engaged with the course as a whole. For these reasons, and because the first run of the test was effectively a trial, teachers were asked to substitute adjusted scores (on a bell curve between 60 and 100%) for raw scores, and, in any case, the test accounted for only a very small proportion (10%) of the students' final grades. Other forms of assessment utilized by EC3 teachers include: performance in presentations and assignments, participation in an extensive reading program, and continuous assessment of a variety of in-class and homework tasks.

2.2 Test Content

The EC3 curriculum focuses on topics such as 'health' and 'describing people', as well as a 'story-telling' unit in which the focus is on building communication skills through a process of exchanging personal information and relaying or responding to personal anecdotes. There is also an extensive reading element to the course in which students are required to read books of their own choosing and introduce or discuss them in class. As far as possible, the items in the test were modelled on these content areas and discourse types. For example, test takers were required to: recognize items of vocabulary relating to parts of the body, follow a consultation between a doctor and patient, identify people correctly from physical descriptions, understand conversations about books and daily university life, and follow monologues in which the speaker describes a personal anecdote. There were a total of 46 items, and the students were given an hour to complete the test.

2.3 Test Design

Before building the EC3 test, the development team met frequently to discuss the most appropriate format and structure for meeting the test's objectives. Buck (2001), drawing on the work of Buchman and Palmer (1996), proposes a framework for evaluating listening assessments based on the properties of reliability, construct validity, authenticity, interactiveness, impact, practicality and efficiency. This framework emphasizes the importance of reliability and validity while acknowledging that the usefulness of a test is determined by the extent to which it is able to serve the practical purpose for which it has been designed.

While keen to ensure that the quality of the test was as high as possible, the practical constraints imposed by the testing situation meant that many of the decisions taken were, to a certain extent, compromise positions.

2.3.1 Test Rubric

The majority of commercial listening tests rely on the target language not only for the listening texts but also for instructions and question options. The problem with this is that it may increase the influence of construct-irrelevant variance by assessing reading ability, test taking technique or other constructs, as well as adding to the overall cognitive load for the test-taker. Since all of the test takers had native or near-native Japanese reading skills, it was decided that all instructions and questions would be in the native language rather than in English.

2.3.2 Test Structure

The items were placed in order of presumed difficulty starting with short, single item recordings mostly focusing on isolated phrases or items of vocabulary, through longer dialogues with several items attached to each recording and finishing with three extended monologues, each with five items attached. The test was delivered online through networked PCs with headphones via the Center's Moodle site. Since the level of the majority of test-takers was low, questions were supported by visual prompts where possible in order to provide additional context.

2.3.3 The Listening Texts

The listening texts were scripted and items created on paper before making the recordings. This afforded us a high degree of control over distractors and enabled the creation of texts which were accessible even to the lower level learners. However, this approach has been criticized for tending to generate recordings which lack the typical features of spoken language. Buck (2001), for example, argues that such texts are in fact written texts delivered orally rather than true oral texts and recommends the use of semi-scripted or unscripted recordings in order to enhance authenticity. In fact, as will be seen in the evaluation section below, there was a tendency to overestimate the level of difficulty of many of the items. As such, the use of semi-scripted items for future iterations of the test may be one way of increasing the level of challenge and better targeting our population.

We used both native-English speakers and Japanese speakers of English for recording test passages. As the test developers were themselves native-English speakers, the most practical option was to self-record. However, Japanese speakers were used for recordings in which the speakers were in the role of students.

All of the items were based on audio rather than video recordings. The use of video is an attractive option as it can provide the test-taker with a good deal of contextual information to help make sense of the recordings. Video is also more authentic insofar as real world interactions are based on visual as well as aural input. Moreover, a study by Progosh (1996) found that test-takers expressed a preference for tests delivered by video over those delivered only aurally. However, there are numerous practical constraints on the use of video, such as the need to authenticate by using appropriate settings and props and the importance of having the speakers rehearse before recording. In the absence of properly trained actors, good quality recording equipment, and suitable locations, there is a danger that recordings will be stilted and, therefore, a source of confusion rather than support for test-takers. The use of video in listening assessments has been quite widely researched but with rather ambiguous results (Buck, 2001; Gruber, 1997). Indeed, Ockey (2007) suggests that while visual cues may be helpful for some students, they could prove distracting for others. Therefore, it is far from proven that video is a superior medium for delivering listening assessments, especially when the quality of the recordings is questionable. Moreover, in purely practical terms, video files are extremely large and place significant strain on computer systems, thereby compromising download speeds. For these reasons, we quickly decided to limit ourselves to the use of audio recordings.

2.3.4 The delivery platform

There are a number of problems inherent in attempting to deliver a listening test through a central sound system (e.g., tape recorder, CD player, PC with speakers). Firstly, ambient noise from inside or outside the testing area may affect the performance of the test-takers. Secondly, the audibility of the recordings often varies from one part of the testing area to another, creating unequal conditions and unwanted variance in test scores. Finally, any technical mishaps that occur will affect the entire group of test-takers.

Aside from the problems mentioned above, there is the difficult issue of who ought to control the pacing of the test and the number of repetitions allowed for each test passage. Although there are sound reasons for placing this authority with the test administrator, there are considerable benefits to outsourcing a measure of control to the test-taker. Test-takers require varying lengths of time to process instructions and read and respond to questions. Similar diversity exists in the number of exposures to a recording that are required to successfully decode the meaning. Therefore, it makes sense to allow each test-taker to navigate the test according to his or her own needs. Forcing the entire group to proceed in lockstep can lead to frustration in some test-takers and boredom in others. However, it is important that external controls are in place to ensure that the test is completed in a timely fashion. An overall time limit is advisable as well as recommendations for the test takers to help them maintain a pace that will allow them the opportunity to respond to all of the questions.

Due to the factors mentioned above, it was determined that the test would be delivered online. Each student would navigate the test on a personal computer and listen to the recordings via earphones. While the test administrator would control the start and finish time of the test, all other aspects were left in the hands of the test-takers. As a precaution against technical difficulties, a short ‘test recording’ was placed on the instruction page of the test. This allowed each test-taker to check the functionality of their earphones prior to commencing. It should be pointed out that, though computers were used, the test was still conducted in an area supervised by the test administrator.

A variety of different platforms were considered for administering the test (Moodle, Survey Monkey, iSpring). However, in the end Moodle was selected for various reasons. Although not the most user-friendly option, Moodle offered the best capabilities for collecting data on the performance of each test-taker. Programs such as iSpring were visually appealing and efficient in terms of downloading speed, but could not provide the response stream required to assess the validity of test items. The Gradebook feature included in Moodle indicates whether a test-taker’s response was correct or incorrect, and, with customization, distinguishes among incorrect responses ― a capability essential in enabling the subsequent analysis of test items.

Although Moodle does provide detailed response information for each test-taker, this information is not easily accessible in bulk fashion. Because of this shortcoming, it was necessary to manipulate the scoring structure by assigning incorrect responses with pre-determined partial scores corresponding to the A, B, C or D multiple-choice options. This facilitated analysis of the results using Rasch Winsteps 3.72.0 (Linacre, 2009) in order to establish the validity of the assessment instrument (see section 3 below). These partial scores were subsequently stripped out of the data to generate the actual scores reported to the test-takers.

Although Moodle provided a platform to deliver the test and enabled collection of the necessary data, there were considerable problems surrounding the actual administration. The burden of 100 or more test-takers attempting to access the test simultaneously proved too great for the university's server. As a result, the time required to download question pages increased dramatically, to the point that it was nearly impossible to proceed. Because of this problem, it was necessary to remove the time limit imposed upon the test both within Moodle and externally. It was also necessary to stagger the tests so that no more than two groups were accessing it at the same time. This caused significant problems for some test-takers as well as the test administrators. However, since the time the test was delivered, computing infrastructure has improved and a testing rota has been devised that should help minimize such problems in future iterations.

2.3.5 Question types

Multiple-choice offers numerous advantages in a context such as that of the SILC. Most importantly, it is fair. Ideally, the test items will be well-designed, but inevitably on a first run of a test some will work better than others. Indeed, it is possible that some items may not function at all. However, given a relatively homogenous population, it is likely that students will be equally advantaged and disadvantaged in the event of items which do not work well. Also, while it is very time-consuming to make good multiple choice items, once they have been made they are extremely easy to score, especially if done via computer or bubble cards. Moreover, the resulting data lends itself to analysis in order to determine which test items work and which do not.

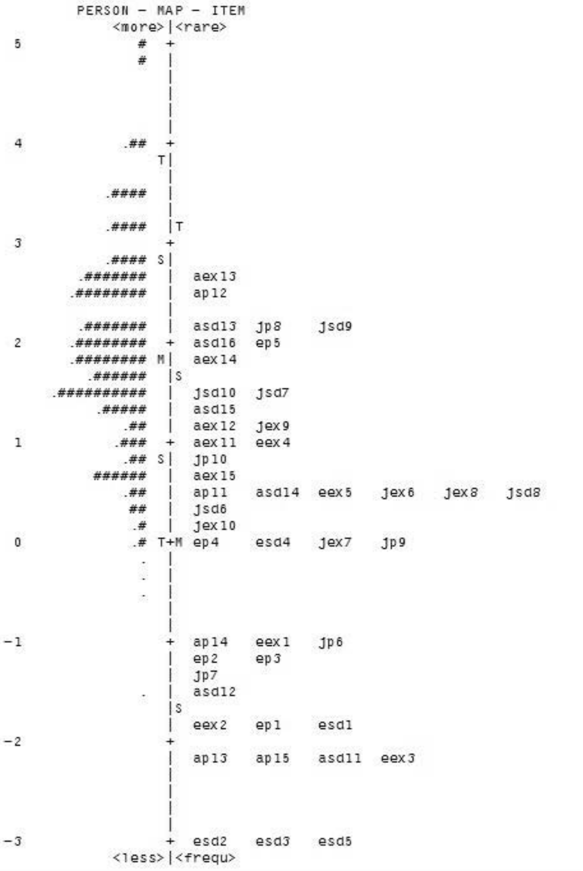
Of course, multiple-choice is not the only option, and, as a question type, it suffers from serious limitations. The most obvious of these is the role of guessing. Research conducted by Wu (1998) suggests that the format promotes "uninformed guessing" (pg. 38), often resulting in candidates "giving the right answers for the wrong reason" (pg. 38). Wu also found that advanced listeners were unfairly advantaged by the format relative to the less able. Another limitation is that multiple choice tests are poorly suited to generating formative feedback. In the event of a wrong answer, no information is provided to indicate where or why the breakdown in understanding occurred. As such, the feedback provided is of limited value to either the test-takers themselves or their teachers (Field, 2008).

There are a wide variety of alternatives for assessing listening which may overcome these limitations to varying degrees. Historically, one of the most widely used forms of assessment was dictation. Dictation tests are simple to create and administer but are difficult to score. Also they clearly test writing ability and knowledge of grammar as much as listening so may be inappropriate, particularly for lower level learners. Another possible option is aural cloze, various formats of which have been implemented with some success (see for example, Templeton, 1977 and Lewkowicz, 1991) but such tests are difficult to administer. The use of partial dictation (Cai, 2012) or listening recall tests (Henning et al., 1983) allays some of these issues and may be particularly useful with lower levels, but these test formats are as yet unproven. Other task types were also considered such as matching or gap-fill type tasks but these proved difficult to realize because of technical limitations. Therefore, given the size of the population and the need for fair and efficient scoring, multiple-choice was seen as the only practical option for this first iteration of the test.

3. Evaluation

3.1 Item Targeting

A very common issue with non-specialist test item writers is that of targeting. In order for a test to be a valid measurement tool it must be accurate in its ability to differentiate test-taker ability. A test with too many easy or difficult items will fail to discriminate ability effectively. As such, it was expected that out of 46 items, around 15 would be either too easy or too difficult and would have to be omitted from future versions of the test (or perhaps used in a special false beginner test for a certain section of the population). One key analysis from this pilot study was to identify which of the items and item types performed within an acceptable range of difficulty for the Sojo EC3 population. Winsteps 3.72.0 (Linacre, 2009) was used to process the data from 520 individual response strings, and a variable map of person-item targeting was produced (see Figure 1).



```
              PERSON - MAP - ITEM
                 <more>|<rare>
 5                  #  +
                    #  |
                       |
                       |
                       |
                       |
 4                .##  +
                      T|
                       |
                .####  |
                       |
                .####  |T
 3                     +
                .#### S|
             .#######  |  aex13
            .########  |  ap12
                       |
             .#######  |  asd13  jp8    jsd9
 2          .########  +  asd16  ep5
            .######## M|  aex14
              .######  |S
          .##########  |  jsd10  jsd7
               .#####  |  asd15
                  .##  |  aex12  jex9
 1               .###  +  aex11  eex4
                  .## S|  jp10
               ######  |  aex15
                  .##  |  ap11   asd14  eex5   jex6   jex8   jsd8
                   ##  |  jsd6
                   .#  |  jex10
 0                 .# T+M ep4    esd4   jex7   jp9
                    .  |
                    .  |
                    .  |
                       |
                       |
-1                     +  ap14   eex1   jp6
                       |  ep2    ep3
                       |  jp7
                    .  |  asd12
                       |S
                       |  eex2   ep1    esd1
-2                     +
                       |  ap13   ap15   asd11  eex3
                       |
                       |
                       |
                       |
-3                     +  esd2   esd3   esd5
                 <less>|<frequ>
```

Figure 1. Variable Map of EC3 Listening Items and Persons

It is immediately clear from the variable map that there are problems with the item targeting. However, item wastage proves to be on par with estimations. The bottom section of the map indicates 17 items that are clearly too easy and display logit scores of -1 to -3. As the lowest ability student has an ability value of -0.4 logits, it is clear that these items are much too easy. Interestingly, of these 17 items, eight are related to picture tasks, six to short dialogues, and three to extended dialogues. It may be that the picture format is an inherently easier cognitive task or that its associated topic (describing people) is especially familiar to test-takers. This issue will be investigated in a subsequent analysis. Interestingly, there is also a lack of items that target the more able students located toward the top of the map. There are around 65 individuals who found all items on the test at least +0.5 logits easier than their estimated true ability. Clearly, this must be addressed in future iterations. Even the extended monologues, which were assumed to be more difficult, failed to appropriately target this section of the population. The apparent ease of this test may be explained in the subsequent technical analysis of misfit and reliability. If the test contains too many poorly performing items or malfunctioning distractors, then this may in part explain the item targeting issue.

3.2 Misfit and reliability

An overall reliability (Cronbach Alpha) can be calculated using the Kuder-Richardson (KR-20) formula, a common classical test method. For this test the KR-20 reliability was 0.82. However, Rasch analysis can provide specific feedback on individual items, which allows much more fine-tuning, especially in instances like this where a small team of test makers requires as little item wastage as possible. A study of outfit mean-squares, Z-standardized scores, and item characteristic curves was conducted and a table of actionable results and information created based on these analyses (see Table 1 below).

Table 1. List of Good Items with Associated Actions Required for Improvement

ITEM	MEASURE	OUT. MSQ	OUT. ZSTD	ACTION
4 ep4	0.05	1.55	3.69	EL - ICC&PTMEACORR suggest low convergent valid, no differentiation, only 1 possible answer? Pics not good?
5 ep5	1.97	1.12	2.41	OK
8 jp8	2.09	1.29	5.23	JR - ICC&PTMEACORR suggest low convergent valid, no differentiation, only 1 possible answer? Pics not good?
9 jp9	-0.07	0.67	-2.64	JR - underperforming item. Keep and trial again
10 jp10	0.78	1.03	0.37	OK
11 ap11	0.42	1.15	1.36	OK
12 ap12	2.45	1.28	4.46	AR - ICC&PTMEACORR suggest low convergent valid, no differentiation, only 1 possible answer? Pics not good?
19 esd4	0.05	0.71	-2.43	EL - sub dimension in ICC, modify or replace
21 jsd6	0.38	0.83	-1.68	OK
22 jsd7	1.55	1.03	0.57	OK
23 jsd8	0.42	0.98	-0.18	OK
24 jsd9	2.17	1.07	1.34	OK
25 jsd10	1.58	1.01	0.29	OK
28 asd13	2.2	1.1	1.82	OK
29 asd14	0.57	1.21	2.11	AR - sub dimension in ICC, modify or replace
30 asd15	1.38	0.89	-1.9	OK
31 asd16	1.98	0.96	-0.72	OK
35 eex4	1.07	0.76	-3.76	EL - underperforming item. Keep and trial again
36 eex5	0.42	0.83	-1.7	EL - sub dimension in ICC, modify or replace
37 jex6	0.52	0.94	-0.61	OK
38 jex7	-0.06	1.02	0.2	OK
39 jex8	0.46	0.83	-1.68	OK
40 jex9	1.18	0.89	-1.72	JR - sub dimension in ICC, modify or replace
41 jex10	0.15	0.77	-1.97	JR - underperforming item. Keep and trial again
42 aex11	1.07	1.2	2.65	OK
43 aex12	1.22	0.92	-1.21	OK
44 aex13	2.66	1.24	3.37	AR - sub dimension in ICC, modify or replace
45 aex14	1.81	0.92	-1.54	OK
46 aex15	0.75	1.01	0.1	OK

After removing the very easy items, there were 29 appropriately targeted items remaining. Of the 29 surviving items, 19 were functioning effectively (well enough for high stakes use). Items 9, 35 and 41 were all under-fitting the Rasch model (low <0.8 Out.MSQ), but this simply means that they are not as efficient as possible due to the fact that they do not provide any new information in the measurement of the construct. This stage of test development requires as many functioning items as possible, and so issues such as these can be dealt with at a later stage, once the item bank is sufficiently large. The remaining eight items (highlighted in blue in Table 1) all displayed contradictory ICC curves and will need to be examined for possible distractor issues or off-dimension elements before being used again.

3.3 Dimensionality

A further key issue in assessing the instrument's usefulness in future gains studies and accurate measurement of student performance is construct validity. How much variance exists within the test that cannot be explained by the construct of listening? Is this variance a threat to the statistical validity of the measure? Items, answers and distractors were written in the L1 to ensure that there would be no requirement for English reading or grammatical knowledge. As such, it is to be expected that only one major dimension would exhibit in the data. A clear uni-dimensional reading would be an important first step in the validation process of SILC listening gains assessments. A Rasch-residual based Principal Component Analysis (R-PCA) was performed on 520 response strings and the results displayed in Table 2 below.

Table 2. Rasch Standardized Residual Variances for EC3 Listening Test

Variance	Eigen-value	% Total	% unexp	% Model
Raw (measure)	20.6	31.0		31.5
Raw (person)	8.7	13.0		13.2
Raw (item)	12.0	18.0		18.3
Unexplained	46.0	69.0	100	68.5
1st Contrast	**2.2**	3.3	4.7	
2nd Contrast	1.8	2.8	4.0	
3rd Contrast	1.7	2.6	3.7	

Listening accounts for 31% of the variance in the data. The modelled value for this variance was 31.5%, a negligible difference of 0.5%, suggesting that the instrument does tap into the construct as efficiently as modelled with a test of this type and length. The real indicator of a potential sub-dimension lies within the *Contrasts* rows at the lower end of the table. Linacre (2008) states that any contrast with an eigenvalue of >2 signifies the possible presence of a sub-dimension. There exists only one small sub-dimension in this data set. Contrast 1 has an eigenvalue of 2.2, which accounts for 3.3% of all variance within the data. Interestingly, this possible sub-dimension exists in items 15, 16 and 17. There is something about these items that contrasts with items 8, 43 and 45 in such a way as to suggest they measure something other than listening. Fortunately, these three items each had very low logit measures (15 = -2.20, 16 = -1.85, 17 = -3.00) and were removed from future use. Thus we can conclude that the remaining items are uni-dimensional and the further two contrasts do not pose a significant statistical threat to the validity of the listening test.

3. 4 Student Perceptions

Following the test, students were asked to provide feedback on the test via the end-of-semester course survey. A summary of the 524 responses is shown in the chart below.

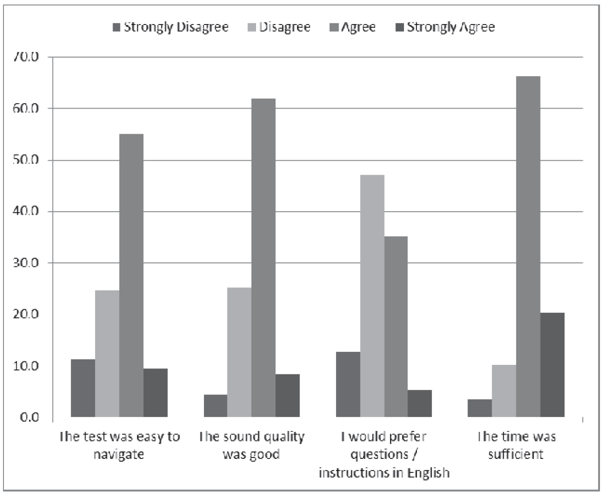


Chart 1: Student feedback on the EC3 test

66% of the test takers agreed that the test was easy to navigate while 70% agreed that the sound quality was good. These results were lower than hoped but probably reflect the technical difficulties relating to the challenges posed by the test platform outlined previously. Interestingly, over 40% of the students said that they would prefer the questions and rubric to be in English rather than in their own language. This surprisingly high figure may reflect the English-medium ethos of the center as a whole; indeed a number of teachers also expressed dissatisfaction that the test was in Japanese. However, as noted above, the test overall was rather too easy for the population so it may be simply that the more able test takers would have liked a higher level of challenge. It will be interesting to see whether the proportion of students expressing a preference for English remains high even after the inclusion of more difficult items. However, we would also like to trial a version of the test with instructions and rubric in English and see how this impacts test scores and the perceptions of the test-takers.

4. Conclusions and Next Steps

Despite the numerous challenges inherent in the development of a test of this type for such a large and diverse population, it seems that the EC3 test was reasonably successful in meeting its objectives. As a first administration of the test it was not surprising that a number of items did not function as hoped. However, of the initial bank of items, 19 seem to be functioning well, while it may be possible to retain a further 10 items after modification. It will, of course, be necessary to create more items before running the test again but the experience gained during the first administration should mean that the test designers are better able to target items appropriately. These items will mainly target higher level students. One way of doing this will be to base the items on semi-scripted or unscripted recordings which will also enhance the authenticity of the test by ensuring that recordings retain more of the typical features of spoken language. It may also be possible to experiment with a variety of different task-types such as gap-filling or matching tasks in order to mitigate the limitations of the multiple-choice format.

In the longer term, and once there is an established a bank of high quality items, it will be necessary to conduct further analysis on the items in order to establish precisely which skills and competencies they target and to map these against the EC3 curriculum itself. This will enable us to identify any gaps as we continue to add to the bank of items, as well as helping us to understand in which skills and competencies test-takers are stronger or weaker. In this way, it is hoped that the EC3 test can become an important tool not only for measuring the achievement of individual students but also for providing feedback on the course in order to inform future curriculum development.

References

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

Cai, H. (2012) Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing (Online First)* published 27 November 2012: http://ltj.sagepub.com/content/early/2012/08/16/0265532212456833

Chapelle, C. A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254-272.

Field, J. (2008) *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Gruber, P. (1997). Video media in listening assessment. *System 25* (3), 335-345.

Henning, G., Gary, N., and Gary, J. (1983) Listening recall: a listening comprehension test for low proficiency learners. *System, 11,* 287-293.

Lewkowicz, J. (1981). Testing listening comprehension: a new approach? *Hong Kong Papers in Linguistics and Language Teaching 14,* 25-31.

Linacre, J. M. (2009). Winsteps® (Version 3.72.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13* (3), 241-256.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24* (4), 517-537.

Progash, D. (1996) Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal 14 (1)*, 34-44.

Templeton, H. (1977). A new technique for measuring listening comprehension. *ELT Journal 31 (4)*, 292-299.

Wu, Y. (1998) What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15(1)*, 21-44.