*THE JOURNAL OF ASIA TEFL*
Vol. 19, No. 1, Spring, 2022, 216-226
http://dx.doi.org/10.18823/asiatefl.2022.19.1.13.216

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511 

# Listening Final Exam Construction: An Exercise in Technical Expertise and Cooperation from Stakeholders

**Michael Fields**
*English Language Institute, University of Delaware*

## Introduction

This article describes the development and production of a complete suite of listening final exams at the English Language Institute (ELI) at the University of Delaware, a large public university in the USA. The ELI serves incoming international students at both undergraduate and graduate level. The ELI also serves students seeking to improve their English before beginning programs at other universities. In addition, it hosts special groups, such as visiting English teachers and scholars, businesspeople, and foreign university students enrolled in short courses. The English Language Institute runs eight-week courses at six levels, from beginner to advanced level, roughly corresponding to CEFR levels A1 through C1, though courses are not specifically benchmarked against the CEFR. At the end of each session, students take a series of final exams in reading, writing, listening, speaking, and grammar, which count as 20% of a final course grade that determines whether students will proceed to the next level.

The listening final exam had formerly been a 1960s era placement test, an audio-lingual style discrete item test where language was decontextualized and depended heavily on aural recognition of grammar forms, rather than true listening proficiency. The same test was given to all levels, the rationale being that the raw scores generated could be used to track students and show their progress throughout levels. However, a placement test is a different instrument from one which can accurately assess achievement and/or proficiency within a level to assign grades between A and F. Further, the test was not aligned with learning outcomes. The need for designing a new test suite was obvious.

Many EFL departments and institutes continue to rely on teacher-made tests and/or commercially available products. However, it is feasible for such departments and institutes to produce their own, high-quality system-wide tests, which can be used across course sections and levels, and can be used repeatedly, with a high degree of validity and reliability. The advantages of in-house final exam production include the ability to better match tests to course content and learning outcomes. Replacing teacher-made tests with system-wide tests ensures a higher quality of assessment, while replacing commercially available tests offers lower costs. The limiting factors in the development of system-wide tests seem to be a lack of the technical expertise among ESL/EFL teachers, an unwillingness for departments to devote resources to such projects, and a lack of effort to bring stakeholders on board with such a project. The value of gaining the trust and cooperation of all stakeholders (teachers, program coordinators, administration, media center and statistics lab staff) cannot be understated. Without negotiation with these stakeholders and their subsequent cooperation, such large projects are likely not achievable.

 

This article will focus on both the steps in the process of the creation of a test suite (which requires technical expertise), and the role of negotiation and cooperation with various stakeholders. The experience in designing and developing a listening test suite at the University of Delaware's ELI is relevant to ESL/EFL departments and institutes throughout the world for several reasons. Most importantly, such a project can improve the quality of assessment, and thus increase fairness. Secondly, the introduction of thematic situations relevant to test-takers, as described in this article, increases validity in listening assessment. Third, the plan to include an audio-visual component, also described below, is a new step for listening assessment.

## Literature Review

Contemporary listening tests contrast drastically with discrete item models from the audio-lingual era, where tests were based on decontextualized items, each one testing a single grammar or vocabulary point, to gradually build up an image of the test-taker's proficiency (Lado, 1964). Today, listening tests are contextualized, with longer monologues or dialogues and a series of items that attempt to gain insight into listening comprehension itself, not underlying grammar and vocabulary (Buck, 2001; Field, 2009). In order to construct such tests, the listening construct must first be accurately and precisely defined, in terms of what the situations are, what themes and vocabulary can reasonably be expected to be found in these situations, at what rate the speech is delivered, in what accents, and whether we expect test-takers to comprehend main ideas, details, gist, or inferences (Buck, 2001).

Construct validity is increased when the construct is first properly defined, and steps are then taken to ensure that the test measures the entire construct, without including irrelevant measures outside the construct (Fulcher, 2010). Content validity refers to ensuring that the test measures concepts that are consistent with teaching objectives, and that test content matches what students have learned in terms of grammar, situations and themes, and corresponding vocabulary (Fulcher, 2010). Reliability refers to the consistency of measurement, thus different forms of a reliable test should give approximately equal scores to test-takers (Fulcher, 2010). Authenticity refers to the degree to which texts and tasks on a test correspond to language events in a non-testing situation (Brown & Abeywickrama, 2019). Finally, the concept of Overall Test Usefulness (Bachman & Palmer, 1996) considers the overall effect of validity, reliability, and authenticity, as well as practicality, security, washback and impact on the quality of a test.

Technical limitations on ESL/EFL instructors' ability to develop in-house assessments with high degrees of reliability and validity are a factor in reliance on teacher-made tests or commercially available tests. Assessment literacy includes such concepts as the understanding of principles of assessment, a familiarity with contemporary assessment theory, and the ability to construct high-performing tests (MacMillan, 2003).

The process of developing in-house listening tests with high degrees of reliability and validity is facilitated when the standard testing cycle is followed (Alderman et al., 1995; Fulcher, 2010). After defining the construct, templates are created, from which individual tests and items are written and edited. A good degree of technical expertise is also required for recording and editing sound files (Buck, 2001), after which pilot forms can be pretested and subjected to statistical analysis, including establishing a reliability coefficient and running an item analysis (Bachman, 2004; Fulcher, 2010). Results will guide in editing and removing items, as well as in making further adjustments if parallel forms are desired. Once final forms are created, the tests can go live. Continuous monitoring of the tests, periodic statistical analysis and further editing can serve to improve them.

The relevant statistical tools to analyze the results of pilots include a reliability co-efficient, a facility index, a discrimination index, and mean item difficulty if multiple forms are desired (Alderman et al., 1995; Fulcher 2010). The Kuder-Richardson, or KR-20 is a mathematical measure of reliability based on internal consistency. The score can range from 0 to 1, where scores above .5 are considered to indicate reasonable reliability, scores of .6 to .7 indicate desirable reliability, and scores over .8 indicate excellent

 

reliability. The facility index is a measure of how easy or difficult any single item is for the baseline test-takers, and can range from 0 to 1.00. The discrimination index measures the degree to which a single item corresponds to overall test performance by any test-taker. In other words, we would expect that more high achieving test-takers answer any single item correctly than lower-achieving test takers, and vice versa. Discrimination is measured from -1 to 1, with 1 being perfect discrimination, and any score at or above +.25 an acceptable score. This score is a powerful and robust measure for editing items, especially used in combination with the facility index in order to make decisions about which items to eliminate. Mean item difficulty shows the average facility index for the entire test and can be used to ensure parallel forms. Once the mean is established for one form of the test, other forms of the test should fall within a close range.

Penny Ur (1984) makes the case that other than telephone conversations, almost all interactive listening events include visual paralinguistic cues. This leads to the question of whether an innovative listening test should be in an audio-only mode or an audio-visual mode. The benefits of an audio-visual mode include a greater level of task authenticity and validity. However, on a practical level, video recording adds another level of complexity to test construction, and delivery of tests with a visual component is also more complex, as it is necessary to have proper equipment and a good venue for audio-visual format, to ensure that all test-takers can see, as well as hear, clearly, and that nothing goes wrong in the delivery. There is also the chance to inadvertently give away answers to test questions through non-verbal cues on the recording. The cognitive load on test-takers, who must balance tasks of listening, watching, reading test questions and marking answers may become distracting, leading to reductions in reliability (Suvorov, 2018).

## Test Format and Design: Cooperating with Administration

The project began through a discussion with the institute's administrative director, who had a long term goal of replacing the outdated placement test with something more modern and appropriate. Several searches for commercially produced tests had not turned up anything that met our needs, thus the decision to create a suite of in-house final exams. Agreement on the need between administration and the assessment committee gave this project a firm foundation.

However, the design of the test was something that required negotiation. Our administrative director still thought of a single test that would be given to all students at every level, as our previous test had done. The assessment committee originally suggested that a separate test be designed for each level. We finally agreed on three tests, Beginner (for levels one and two, roughly corresponding to CEFR A1 - A2), Intermediate (levels three and four, roughly corresponding to CEFR levels B1 - B1+) and advanced (levels five and six, roughly corresponding to CEFR levels B2 - B2+/C1). This meant that there would need to be two sets of pass marks and scoring systems for each test, which created another level of statistical challenge. Further, we would produce three parallel forms of each test. While the assessment committee had not originally conceived of tests that would span levels, this compromise to cooperate with administration was the only way in which the entire project could go forward.

## Defining the Construct: Cooperating with Level Coordinators

Once the general format of the test was determined, the next step was to define the construct. Cooperation and inclusion are important at this stage, and meetings were held with each level coordinator to prioritize learning outcomes for listening, as well to make lists of themes and vocabulary that would be encountered at each level. Most coordinators were very willing to take the time to cooperate in this project, as there was a general agreement on the need to replace the existing final test. However, several level coordinators expressed frustration at the time commitment required at this stage of planning, as an

 

imposition on an already full workload, and expected the assessment committee to undertake this without their input. A further challenge at this point was that in our institute's upper levels, courses divide into separate tracks: academic, culture, business or general English. Yet one test would be required to meet the needs of all of these different tracks.

The resulting construct definitions ran to about six pages for each test, and cannot be reproduced here. Following are some of the most important specifications that went into the construct definitions. It was determined that for the beginner level tests, themes and vocabulary should focus on everyday situations that are most familiar and relevant to students' personal lives, such as home, family, school and work, telling time, shopping, travel and holidays, hobbies, colors, numbers, and describing people. Verb tenses would be limited to present simple, present progressive and past, with some use of past progressive in the more difficult sections of the test. All sections of the test would consist of short (1 – 2 minute) dialogues, each with a male and female voice that are clearly distinguishable. Dialogues would be spoken in clear and slightly slow standard American English. Idiomatic usage would be avoided, and natural repetition would highlight keys.

For the intermediate level tests, themes and vocabulary should reflect situations in which learners will find themselves in everyday interactions, as well as general topics from radio and television. While themes and vocabulary should go beyond the everyday situations in the beginning level tests, to common social encounters, they should not include academic, work or business situations, and no specialized vocabulary should be included. Listening texts should be spoken at a normal rate of speech, and include some features of native speech (reductions, elisions, contractions and etc.) but be clear at all times. There should be some use of common idiomatic language. Test sections should include both dialogues and monologues.

To address the differing needs of the various tracks at the upper level, it was decided that each form of the advanced test should contain one academic section, one business-oriented section, and one science and/or technology-related section. Specialized vocabulary and any special knowledge would be avoided, so that each topic would be readily accessible to all students. Texts should be spoken in standard American English, at normal speech rates, with many features of native speech. Common idiomatic usage should be included. Texts would consist of both dialogues and monologues.

All students at the University of Delaware's English Language Institute are required to take the final listening exam on a single day, and results are needed quickly for reporting purposes. Because of the need for quick turn-around of large numbers of tests, it was decided that all items would be multiple choice format, with three options for each item.

## Building a Validity Argument: Continued Cooperation with Stakeholders

### Construct Validity

A construct validity argument seeks to affirm that, in the development and production of the test, the constructs measured match those of the construct definition, they are measured completely, that measurement of elements outside the construct is minimalized, and that linguistic skill (in this case, listening) is not measured through the means of another skill (for example, there is no undue burden of a high reading level, or students do not have to answer questions orally to demonstrate listening ability). In this test suite of listening comprehension for learners of American English in personal, general and academic/professional/business settings, the following features ensured high levels of construct validity.

A checklist was created to compare points on the initial construct definition to features of the final test suite, and these points were cross-referenced one by one. It can be clearly demonstrated that all points on the construct definition have been taken into consideration on the tests. Texts were written in standard American English conversational style, on topics relevant to students and young people. Collaboration with level coordinators provided feedback to ensure realistic-sounding speech styles. Situations were

 

contextualized, such as visiting a friend, shopping for groceries, planning a future meeting, discussing photos, listening to radio programs, discussing academic assignments, and listening to university lectures. In addition to providing clues as to what students can expect to hear, contextualization enables the measurement of true listening comprehension, not lexico-grammatical knowledge, as many discrete point listening tests do. Texts include a variety of styles: short exchanges, longer exchanges, narratives, explanations requiring comprehension of main ideas and details, interrupted monologues and longer, uninterrupted monologues. Contextualization and different styles of speech also add authenticity as the situations are similar to a wide variety of native speech acts encountered in non-testing situations.

The dialogues and monologues were recorded by native speakers, at a normal speech rate for the intermediate and advanced tests (slightly slower for the beginning test), and features of authentic speech, such as reductions, contractions, elisions, and interruptions, were incorporated. A normal amount of repetition of key ideas is scripted into the dialogues, mimicking authentic speech. These facets of native-like speech increase both construct validity and authenticity.

Items test for comprehension of main ideas and details at all levels. At the intermediate and advanced levels, there are also items designed to test for pragmatic comprehension and ideas implied but not directly stated. Inclusion of various aspects of listening comprehension, rather than simply focusing on listening for details, raises construct validity.

The linguistic input is purely auditory. All items are multiple choice, with no requirement for any kind of language production on the part of test-takers. Though it is necessary for students to read the items, they are written at a level which precludes any challenges to comprehension. These factors together aim to avoid the testing one skill through the use of another, thereby isolating the listening skill. This avoidance of *muddied measurement* also bolsters construct validity.

## Content Validity

Content validity was assured by meeting with level coordinators and reviewing textbooks. At each level, lists were made of what themes and topics could be encountered, and from these lists, vocabulary relating to these themes was identified. At the beginning level, the ALTE Social and Tourist Activities list (ALTE, 2002) was also consulted to generate possible themes and vocabulary. At the beginning level, grammar was limited to what could reasonably be expected for students to comprehend.

Content validity was enhanced by giving level coordinators the opportunity to comment on finished texts. In many cases, single vocabulary items, sentence structures, or even complete tasks were changed or removed entirely because the coordinator felt that it did not match the learning outcomes, or that students could not reasonably be expected to demonstrate comprehension.

## Writing Texts and Items: Cooperating with Level Coordinators

After the creation of construct definitions, test specifications and templates were created for each test, so that parallel forms could be constructed. Tables 1, 2, and 3 list the topics of monologues and dialogues at each level, beginning (A1-A2), intermediate (B1-B2) and advanced (B2+-C1). At the beginning level, topics reflect everyday encounters, and are informed by the ALTE Social and Tourist Activities list (ALTE, 2002). Intermediate topics include general and some easier academic English, while the advanced level specializes in English for business, science and technology, and an academic lecture. Times of monologues and dialogues, also listed, increase with level, and the number of items per section are also listed. The topics listed here reflect one form of the tests only. In other forms of the tests, topics were changed.

 

TABLE 1
*Beginning Level Template of Topics*

| **Time** | **Topic** | **Item number** |
|---|---|---|
| 1 minute | Describing people | **3** |
| 1 minute | Time (whole hours) / daily routines | **3** |
| 1 minute | Rooms in a house/locations and prepositions | **3** |
| 1 minute | Weather/parts of the day | **3** |
| 2 minute | Buying presents (clothes) for family members | **4** |
| 2 minute | Time (partial hours)/hobbies and free time | **4** |
| 2 minute | Cultural event/ giving directions on a city street | **6** |
| 3 minute | Planning a trip to Florida/sightseeing | **7** |
| 3 – 4 minute | false monologue | **7** |

*Note.* A *false monologue* is a monologue interrupted by questions. The function of these questions is twofold. First, they break the monologue into more manageable parts, and allow students to refocus if they have lost the thread of the monologue. Secondly, the questions allow students to focus on the important information, which will appear in the items. The three monologues all take place in a classroom setting: a teacher giving an introduction to a class; a teacher giving instructions to the class for what to expect with a substitute teacher; a teacher describing a class field trip.

TABLE 2
*Intermediate Level Template of Topics*

| **Time** | **Topic** | **Item number** |
|---|---|---|
| 4 minute | Dialogue: General English: Two students discuss their plans for the future/two former students talk about what they have done since graduating/two young people compare living in a city and a small town | **8** |
| 4 minute | Monologue: General English: An authority figure (neighborhood policeman, librarian, park ranger) gives useful information | **8** |
| 4 minute | Dialogue: Two students discuss an upcoming project or review for a test (technology, engineering, psychology) | **8** |
| 4 minute | Monologue: A narrative (German grandmother's life story, Mexican grandfather's life story, cross-country travelogue) | **5** |

*Note.* As this is a test of general English, in both dialogues, pragmatics are tested as well as transactional language. For example

*24. Which is true about Phil and Lisa's discussion of driverless cars?*
*A. Phil and Lisa agree with each other.*
*B. Phil and Lisa have different opinions.*
*C. Phil changes Lisa's mind about driverless cars.*

TABLE 3
*Advanced Level Template of Topics*

| **Time** | **Topic** | **Item number** |
|---|---|---|
| 5 minute | Dialogue: Two students prepare a project or review for an exam (history of the internet, social psychology, business) | **8** |
| 5 minute | False monologue (radio interview with the author of a new book) Colorado River, Marketing, Self-driving cars | **8** |
| 10-12 minute *(heard once only)* | Monologue (university lecture) Modern zoo, Smart Phones, Franchising | **12** |

*Notes.*
1. In section 1, pragmatics are tested in addition to transactional language.
2. The false monologue in section 2, in the form of brief questions being asked to a writer on a talk-radio program, allow students to refocus if they have lost the thread of the longer and more detailed responses. The questions also serve to break the monologue into more digestible parts, and to give test-takers a clue about what they are listening for.
3. In section 3, the university lecture, attitudes implied but not directly stated by the speaker are also tested. For example

*27. The opinions provided by the lecturer suggest that*
*A. older zoos should be modernized.*
*B. zoos should all be closed.*
*C. zoos should do more to attract the public.*

 

After texts and items were developed, level coordinators were asked for their feedback. One of the essential issues to deal with in the development of texts and items was vocabulary. We needed to ensure that vocabulary was accessible to test-takers at all levels. In order to do this, test writers familiarized themselves with the textbooks used at each level, and lists of commonly occurring lexis were made. (Of course the tests were not intended to be specific to any single textbook, but only reflect a proficiency level, so vocabulary was not limited to these lists; rather, they were used as a starting point.) It was also critical to ensure that items were worded simply enough so that they would not be challenging for test-takers to read and understand. On both of these points, feedback from level coordinators was essential, and so as test forms were completed, they were sent back to coordinators to review and give feedback, edits and comments.

The biggest challenge was again the amount of cooperation that could be expected from each coordinator. Some gave detailed feedback and corrections. In other cases, feedback and edits were minimal. Ultimately, in such a large and important undertaking, administration should ensure that all relevant staff see it as a part of their normal workload, and an additional imposition.

## Recording the Sound Files: Working with Staff in the Media Center

Four teachers were selected and given class release time to record the dialogues. Recording was done in the university library's media center, which includes professional-level recording studios, and staff who can assist with the technology. The technology made it possible to record sections separately, edit out mistakes, and finally splice entire tests together. After piloting, it was also possible to remove unwanted sections of the test and re-arrange others.

Establishing a single contact in the media center who was familiar with the project was a key point in cooperation. This contact was also able to give a short training session in use of the technology. Clear communication with the teachers doing the recordings was also a key form of cooperation. Familiarizing teachers with the texts in advance increased the feeling of natural speech, and teachers were also briefed on features of natural speech, such as speed, pauses, interruptions, reductions and linking. Listening to the teachers' concerns about phrasing, and making last-minute edits while in the studio is another example of cooperation among stakeholders.

## Piloting: Cooperating with SALC Staff, Invigilators, and All Teachers

Once question papers and sound files were created, tests were piloted. This was done in the Self-Access Learning Center (SALC), which can accommodate up to 75 students at a time, and has an adequate sound system installed. Instructions needed to be given to SALC staff and invigilators, as this was a new procedure for them, and all teachers also had to be informed. The entire student body, approximately 600 students, would serve as our baseline group, and scores would be normed to their performance. The piloting took place over three sessions, with one form of the test being piloted in each session. Thus each form of the test was benchmarked to a baseline group of about 200 students, which was an adequate sample size for statistical analysis.

## Statistical Analysis and Editing: Cooperating with the Statistics Lab

In order to carry out a proper statistical analysis, cooperation with a staff member of the statistics lab was critical in collecting the necessary data. As the staff were generally not well informed on the required statistics for language testing, explaining the needs of the project from the outset was also critical. For each test, the following data were collected: a KR-20 score, the facility index of each item, the

 

discrimination index of each item, the mean item difficulty for the entire test (expressed both as a percentage and as a raw score) and the standard deviation from the mean. Following is a discussion of how the above statistical measures were used in validation, assessment of reliability, deleting or repairing poorly functioning items, and creating parallel forms of the tests.

## KR-20

The Kuder-Richardson, or KR-20 measures reliability based on internal consistency. Note that these scores represent reliability of the pilot versions of the tests, and are presumed to be higher on the final versions, after poorly functioning items were eliminated or repaired. The KR 20 results for the listening test suite are shown in table 4. It was the intent of the test development team that all KR-20 coefficients would be over .70. Out of eighteen total calculations for each form of the test at each level, all but one achieved a score of at least .60, indicating desirable reliability, and the one test that did not achieve this score (Intermediate form 1) came in at .555. A total of thirteen tests achieved at least .70, indicating desirable reliability, with five achieving over .80, or excellent reliability. It is expected that after editing, these scores will continue to increase.

TABLE 4
*Reliability Coefficients (KR-20) of Pilot Test Forms*

| | **Beginner 1** | **Beginner 2** | **Intermed. 3** | **Intermed. 4** | **Advanced 5** | **Advanced 6** |
|---|---|---|---|---|---|---|
| **Form 1** | .731 | .833 | .555 | .653 | .717 | .627 |
| **Form 2** | .788 | .894 | .734 | .752 | .763 | .792 |
| **Form 3** | .870 | .878 | .785 | .856 | .607 | .624 |

All items required a facility index in the range of .60 to .99 in order to be meaningfully translated into the letter grades of A through D (100% through 60%) that students receive. Any items scoring outside this range were marked for editing or removal.

Discrimination Index The discrimination index measures the degree to which a single item corresponds to overall test performance by any test-taker. In other words, we would expect that more high achieving test takers answer any single item correctly than lower-achieving test takers, and vice versa. Discrimination is measured from -1 to 1, with 1 being perfect discrimination, and any score at or above +.25 an acceptable score. This score is a powerful and robust measure for editing items, especially used in combination with the facility index in order to make decisions about which items to eliminate.

To establish truly parallel forms of the test, the mean item difficulty was taken into consideration. In our context, mean item difficulty should be between 70 and 80 at levels 2, 4 and 6, with 75 being the perfect mean item difficulty. For levels 1, 3, and 5, it was determined to accept ten points lower, so a mean difficulty of between 60 and 70, with 65 as the perfect score. As can be seen in table 5 (these represent statistics for the pilot forms of the test), mean item difficulty ranged considerably around these numbers for different forms and levels of the test. Therefore in the editing process items that may be otherwise functioning well needed to be removed to increase or decrease the mean item difficulty to within the acceptable range (between .73 and .78 for levels 2, 4 and 6). Mean item difficulty was a good starting point in the editing and elimination of items, with other factors following.

TABLE 5
*Mean Item Difficulty for Pilot Test Forms*

| **Level** | **Mean Form 1** | **Mean Form 2** | **Mean Form 3** |
|---|---|---|---|
| **1** | 70 | 49 | 67 |
| **2** | 83 | 72 | 76 |
| **3** | 55 | 59 | 56 |
| **4** | 58 | 71 | 72 |
| **5** | 78 | 73 | 69 |
| **6** | 75 | 75 | 73 |

 

Finally, the standard deviation from the mean was used to design cut scores between letter grades. Thus, if the mean is defined as a C, then one standard deviation in either direction defines a B and D, and two defines an A and F.

Pilot tests had been written with one or two extra items in each section, so that poorly functioning items could be eliminated in developing the final versions of the tests. Three considerations needed to be taken into account: items may be poorly functioning because their facility was outside the range required, items may fail to discriminate adequately, or there may be a need to eliminate items (even those performing perfectly well) to increase or decrease the mean item difficulty, in order to ensure true parallel forms of the test. Additionally, since the number of items per section was predetermined, only one or two items could be deleted, no matter how well or poorly any of the items in the set functioned.

Managing all of these factors to create strong items sets requires balancing all these factors. First, tests were marked for their mean item difficulty, in relation to the standard (between .73 and .78 for levels 2, 4 and 6). Tests with a low mean difficulty would need to have some of the more difficult items removed, while tests with a high mean difficulty would need to have some of the easier items removed. Next each item was checked for its facility and discrimination, and items that fared poorly on both of these indexes were flagged for removal. If only one index was problematic, it was sometimes possible to edit the item to improve its score. Items were then deleted from the final version of the test based on poor performance on two scales, as well as the need to remove items to increase or decrease the mean item difficulty and bring all tests into alignment. While many poorly functioning items were repaired through editing, in the end it was necessary to leave some items in place which may have either fallen outside the facility index or did not discriminate adequately. The result was a suite of tests in which mean item difficulty levels were prioritized, in order to create truly parallel forms of the tests. It is believed that, while not every individual item may support the facility and discrimination levels that we intended, each item supports at least one of these, and overall, the tests do perform within range in terms of facility and discrimination.

## Future Directions

Covid-19 and the subsequent phenomenon of teaching online has made the use of these new listening tests impractical just as they were beginning to be used live. Due to concerns about test security once they are loaded onto an online learning management system, it was decided to delay any further use of the test suite until a return to campus. Currently the testing committee and administration is looking into a secure means to deliver these tests online. At the time of this writing, a platform is being developed in-house which we hope will allow the secure online delivery of these tests, which should go live in the very near future.

It is the goal of the administrative director to reformat the entire test suite as an audio-video project, allowing test-takers to contextualize auditory input. A listening final exam suite that incorporates a visual component would in many ways be a more valid instrument, and also an innovation in testing listening. However, adding another layer of complexity brings with it a new set of challenges. Creating audio-visual recordings of a high quality requires a different skill set, in terms of both acting and filming. Costumes, props and backgrounds or sets would be required. Ensuring that answers to test questions are not inadvertently given away through visual signals requires specific attention to the gestures of actors at all times, who may need to be directed to do things in a slightly unnatural way. Finally, playing the videos would require greater technology, and ensuring that all test-takers can see and hear well makes administering the test more difficult.

 

## Conclusions

There are many advantages of producing in-house system-wide tests for use as final exams. The collaboration of many stakeholders in their creation, and the rigors of statistical analysis give these a far greater degree of validity and reliability than most teacher-made tests. Likewise, commercially produced tests do not match with any specific curricula and are expensive for ESL programs as they must be purchased regularly.

In order to successfully create in-house system-wide tests, there must exist within the department or program the technical expertise in test creation, which is lacking in many ESL/EFL programs. Then, there must be the will on the part of the program head to allot resources to test creation. The greatest resource needed is time, as the creation of these tests is time-consuming and requires course release for an extended period by one or several faculty members.

Finally, what is most important is to ensure cooperation from all stakeholders, to avoid the entire process from being derailed. The project should have the firm support of the departmental or program administration. Level coordinators should be willing to spend time to meet with test creators and read and comment on drafts of the tests. There should be adequate preparation within the program to pilot these tests, and teachers should be informed and cooperative throughout the piloting process. Open lines of communication should be available between the ESL/EFL department and the statistics lab.

When the resources can be obtained, and the cooperation of all stakeholders ensured, and with the proper technical expertise to carry out such a project, the development of in-house institutional final exams will enhance the quality of any ESL/EFL program. It can introduce more parallelism and equality in measured outcomes between sections of a course. It can help guarantee that all relevant learning outcomes are measured adequately. It can increase the reliability and validity of assessment.

## The Author

*Michael Fields* is Assistant Professor at the University of Delaware, USA. He teaches in both the English Language Institute and the MA TESL program, and serves as head of the assessment committee. He has participated in a study to benchmark the TOEFL to the CEFR, and is a test writer for Cambridge exams. He publishes and presents regularly on assessment issues.

English Language Institute
University of Delaware
Email: mrfields@udel.edu

## References

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

ALTE. (2002). *The ALTE Can Do Project*. https://www.cambridgeenglish.org/images/28906-alte-can-do-document.pdf

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices*. Pearson.

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

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment companion volume.* https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

 

Field, J. (2009). *Listening in the language classroom*. Cambridge University Press.

Fulcher, G. (2010). *Practical language testing*. Routledge.

Lado, R. (1964). *Language testing: The construction and use of foreign language tests*. McGraw-Hill.

MacMillan, J. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for Theory and Practice. *Educational Measurement: Issues and Practice, 22*(4), 34-43

Suvorov, R. (2018). Test takers' use of visual information in an L2 video-mediated listening test. In G. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp.146-160). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.50.10suv

Ur, P. (1984). *Teaching listening comprehension*. Cambridge University Press.

*(Received October 29, 2021; Revised February 24, 2022; Accepted March 18, 2022)*

 