Article / 論文

九州産業大学語学教育研究センター紀要
Language Education & Research Center Journal

LERC
lerc.kyusan-u.ac.jp

An Analysis of a New *Listening & Speaking* Final Test

Ian DAGNALL

Abstract

This study presents a statistical analysis of the first administration of a new *Listening & Speaking I* final test developed by the LERC curriculum development team. Course-level achievement tests are an important part of language assessment and curriculum development but are rarely validated after administration. In this study, the test was analysed using descriptive statistics, item statistics, and measures of consistency. Criterion-referenced test analysis was undertaken as the new test assessed learners' understanding of course content. Norm-referenced analysis was also undertaken to reflect the statistics automatically generated in Moodle Quiz and teachers' familiarity with this type of analysis. A secondary aim of the study was to compare the actionable insights provided by norm-referenced and criterion-referenced analyses of the test results. The results indicated that the curriculum development team succeeded in creating a consistent test, and the psychometric properties of the test meant that it would be suitable for use as a pre/post-test with future cohorts. Further, the questions would be suitable to add to a bank of questions for future placement and achievement tests. Criterion-referenced test analysis provided more nuanced insights into item revision, but the study showed that on a single post-course administration, norm-referenced analysis can also provide insight into item suitability. While the score distribution suggested that most students had a strong grasp of the course content, the analysis highlighted the limitations of a single post-course test in providing insights into learning, and it is suggested that a pre/post-test regime is used for future administrations of the test.

Cite as: Dagnall, I. (2024). An analysis of a new *Listening & Speaking* final test. *Language Education and Research Center Journal, 19*, 3-29.

本研究は、LERC カリキュラム開発チームによって開発された新しい「Listening & Speaking I」最終テストの初回実施に関する統計分析である。コースレベルの到達度テストは、言語評価とカリキュラム開発の重要な一部であるが、実施後に検証されることはほとんどない。本研究では、記述統計、項目統計、一貫性の尺度を用いてテストを分析した。新しいテストは学習者のコース内容の理解を評価するものであるため、目標基準準拠テスト分析が行われた。また、Moodle Quiz で自動的に生成される統計と、教師がこのタイプの分析に慣れていることを反映させるため、集団基準準拠テスト分析も行われました。研究の第二の目的は、集団基準準拠テスト分析と目標基準準拠テスト分析から得られる実用的な洞察を比較することである。その結果、カリキュラム開発チームは、一貫性のあるテストを作成することに成功し、テストの心理学的特性は、将来の集団の事前／事後テストとして使用するのに適していることが示された。さらに、その問題は、将来のクラス分けテストや到達度テストの問題集に加えるのに適している。目標基準準拠テスト分析により、項目の改訂についてのより微妙な洞察が得られたが、本研究では、コース終了後の 1 回の実施において、集団基準準拠テスト分析も項目の適合性について洞察できることが示された。得点分布から、ほとんどの学生がコースの内容をしっかりと理解していることが示唆されたが、分析により、コース終了後の 1 回のテストでは学習に関する洞察に限界があることが浮き彫りになった。

Background

The development of reliable tests is an important part of the development of language learning courses. As well as evaluating student performance in a course, testing is an integral part of curriculum development. Test analysis can feedback into needs analysis, course objectives, course materials, and teaching (Brown, 1991; Brown & Hudson, 2002). However, several shortcomings to teacher-developed tests have been highlighted. First, many teachers lack proper training in the principles of assessment and are unfamiliar with the appropriate interpretation of test performance (Douglas, 2014; Green, 2021). Second, item quality in teacher-developed tests is often lower than that of standardised tests and reliability is often not assessed. Further, projects to develop new tests at classroom and programme level are often undertaken as linear projects: the specifications are decided, items are written, the test is administered to a group of learners, and grades are awarded (Green 2021).

Rather than this linear approach, Green (2021) has suggested a seven-phase assessment production cycle as a more effective approach to test design. This cycle includes reflection on the extent to which a test is fulfilling its purpose, and systematic analysis of the

test to inform effective improvement of the test. The specifications phase includes deciding on the purpose of the test, the constructs to be tested, the number of items, the time allotted for the test, the response format, marking, and administration details. The item writing phase involves the creation of the test questions, and in the item review phase, these items are checked by other members of the test development team. In the piloting phase, the test is given to a similar group to the target population, and the results are analysed in the pilot review and modifications are made to the next version of the test. The operational assessment phase is the use of the test in the field, and in the assessment review phase, the quality of the test is checked through statistical analysis. The purpose of the statistical analysis is to describe the distribution of scores, assess the difficulty of the test, decide which items to keep, which items to modify, and which items to discard in revised versions of the test, and to establish the reliability of the test (Brown & Hudson, 2002; Brunfaut & Harding, 2014; Green, 2021).

There are a number of different approaches which can be taken to test analysis, including classical test theory (CTT), criterion-referenced testing (CRT), item response theory (IRT), and generalisability theory (G-theory). However, due to the complexity of IRT and G-theory and a lack of training in these methods, they are rarely used by language teachers in course-level assessment. While less complex than IRT or G-theory, many post-graduate EFL courses do not provide any instruction on CRT analysis (Green, 2013; Brown, 2021).

Which approach to take to test analysis depends largely on the type of test being analysed. Language tests can be generally divided into norm-referenced tests (NRTs) and criterion-referenced tests (CRTs). In language testing, NRTs are used to assess general language ability. This kind of test is often used to assess students' proficiency or for placement decisions. Examples of NRTs in language testing include tests such as TOEIC and TOEFL. NRTs are designed to measure a student's performance relative to the other students who took the test and are analysed and revised with the intention of creating a normal distribution of scores.

On the other hand, CRTs are designed to assess the extent to which a student has mastered the objectives of a particular course. CRTs are usually used for diagnostic purposes at the beginning of a course and for achievement decisions at the end of a course (Brown, 1991; Brown, 2021; Brown & Hudson, 2002; Douglas, 2014). In a CRT, students should be

familiar with the exact content on which they will be tested, and their performance is measured on how much of this content they know, with no reference to the performance of other students. Unlike NRTs, CRTs are not expected to produce a normal distribution of scores. Students who know all the content should score 100%, and if a course has been successful, most students should perform well on the test (Brown & Hudson, 2002).

Norm-referenced test analysis

The analysis of NRTs is usually undertaken using CTT (Brown, 2021). CTT involves the calculation of descriptive statistics (mean, standard deviation, range, score distribution skewness, score distribution kurtosis), item statistics (item facility and item discrimination), and a coefficient of internal consistency. Analysis of the descriptive statistics allows test developers to ensure that the test scores are normally distributed, while item statistics allow the evaluation of the effectiveness of individual test items (Brown, 1991; Brown, 2021; Brown & Hudson, 2002; Douglas, 2014). Reliability - the extent to which students' scores on a test reflect their real ability - can be measured using coefficients of internal consistency (Brown & Hudson, 2002; Yan & Fan, 2021).

NRT item analysis usually involves the calculation of item facility (IF) statistics and item discrimination (ID) statistics. The IF statistic is the percentage of test takers who answer an item correctly and indicates the difficulty of a test item. It is calculated by dividing the number of students who answered an item correctly by the total number of students (Brown & Hudson, 2002; Green, 2013):

$$IF = N_{correct} / N_{total}$$

Where:

$N_{correct}$ is the number of students who answered the item correctly.

N_{total} is the total number of students.

The result of the calculation is a value ranging from 0.00, for items where all the students answered incorrectly, to 1.00, for items where all the students answered correctly. The most useful information about students' proficiency levels comes from items with an item facility value of between 0.2 and 0.75. An IF of greater than 0.75 indicates that an item is too easy and an IF of less than 0.2 indicates that an item is too difficult (Brown, 2021; Khalifa & Weir, 2009).

The ID statistic compares students' performance on a particular item to their performance on the test as a whole. Students' performance on the test as a whole is expected to be a more reliable indicator of their ability than their performance on an individual item, and ID analysis allows test developers to select items which best separate high performing and low performing students. (Brown & Hudson, 2002; Green, 2013). It is calculated using the following formula:

$$ID = IF_{upper} - IF_{lower}$$

Where:

IF_{upper} is the item facility for the top 25% to 33% of students on the test.

IF_{lower} is the item facility for the bottom 25% to 33% of students on the test.

ID values can range from 1 to -1. An ID value of 1 shows that all the higher-performing students answered correctly, and all the lower performing students answered incorrectly. An ID of 0 indicates that an equal number of higher and lower-performing students answered the item correctly. A negative ID shows that more of the lower-performing students answered the item correctly and suggests that there is a serious problem with the item. Items with an ID value of 0.4 or greater are considered to be discriminating well, and items with an ID of 0.3 to 0.39 can generally be accepted as discriminating between the high and low-performing students in the same way as the overall test but might be subject to improvement. Where the ID is between 0.2 and 0.29 the item is not discriminating well and should be reviewed, particularly if it has a low IF. Items with an ID value of below 0.19 are considered to be poor items which should be revised or discarded (Brown, 1990; Green, 2013; Khalifa & Weir, 2009). By selecting items with IF values of 0.2 to 0.75 and ID values of greater than 0.4, test developers can maximise the variance in scores on the test, and hence make the test more reliable (Brown, 2021).

Test reliability measures show how well each part of the test relates to the other parts of the test. There are a number of different methods for calculating test reliability. Measures of the reliability of single administration NRTs include Kuder-Richardson (K-R) 20, K-R 21, and Cronbach's alpha. For tests consisting of dichotomously scored items, K-R 21 is the easiest to calculate as it only requires the number of items (k), the mean (M) and the variance (Var). It can be calculated using the following formula:

$$\text{K-R 21} = [k/(k\text{-}1) * [1\text{-}(M*(k\text{-}M)/(k*Var))]$$

K-R 21 assumes that all items are of equal difficulty, which is often not the case in tests, and generally underestimates reliability (Brown & Hudson, 2002; Riazi, 2016).

K-R 20 is also used for tests consisting of dichotomously scored items but is slightly more complex to calculate as it requires calculating the sum of the proportion of students passing each item multiplied by the proportion students failing each item (Brown & Hudson, 2002; Riazi, 2016). It can be calculated using the following formula:

$$\text{K-R 20} = [k/k\text{-}1] * [1\text{-}(\Sigma p*q)/Var]$$

Where:

k is the number of items.

M is the mean.

Var is the variance.

p is the item facility.

q is $1 - p$.

Cronbach's alpha can be used with tests in which items are not scored dichotomously. It is more complex to calculate than K-R 20 or K-R 21 and is usually calculated using statistical software such as SPSS. All three methods result in a value of between 0 and 1, and a higher value shows higher reliability. A reliability estimate of 0.8 means that 80 percent of the total variance in scores is due to score variance and 20 percent is due to error variance. Reliability estimates of 0.70 or greater are considered acceptable for NRT tests (Brown, 2021; Brown & Hudson, 2002; Riazi, 2016).

Reliability can also be assessed using the standard error of measurement (SEM). The SEM represents a confidence interval around a student's score and describes the range of scores in which a student could be expected to score with repeated administrations of the test. It is calculated as follows:

$$SEM = S_x\sqrt{1 - R}$$

Where:

S_x is the standard deviation of the test.

R is the reliability of the test (calculated with a reliability measure such as K-R 20).

The smaller the SEM, the smaller the band of scores in which a student's true score lies, which indicates a higher level of reliability (Brown, 2021; Brown & Hudson, 2002, Douglas, 2014; Riazi, 2016).

Criterion-referenced test analysis

CRT analysis aims to achieve a test which distributes students into categories (masters or non-masters, pass or fail) according to their knowledge of specific instructional objectives. Similar to NRT analysis, this is undertaken using item facility values, calculated in the same way as for NRT analysis, and discrimination indices. CRT discrimination indices are designed to discriminate between masters (students who have learned a sufficient amount of the language material or skills outlined in the course objectives) and non-masters (Brown & Hudson, 2002; Riazi, 2016). For courses where students take a pre-test and a post-test, the most basic index item discrimination index is the difference index (DI). This is calculated by subtracting the item facility on the pre-test from the item facility for the same item on the post-test. DIs can range from +1.00 (none of the students knew the material at the beginning of the course but all the students knew that material at the end of the course) to -1.00 (all the students knew the material at the start of the course but unlearned it by the end of the course) (Brown, 1991; Brown; 2003; Brown & Hudson, 2002).

CRT discrimination indices which require only a test at the end of the course include the *B*-index, item phi (Φ), and the item agreement statistic (A). The *B*-index is similar to the NRT item discrimination statistic but includes all the students who took the test in the calculation. It shows how well an item distinguishes between students who passed the test and students who failed the test. It is calculated using the following formula:

$$B\text{-index} = IF_{pass} - IF_{fail}$$

Where:

IF_{pass} is the item facility of students who passed the test.

IF_{fail} is the item facility of students who failed the test.

B-index values can range from +1.00 (all of the students who passed the test answered the question correctly while none of the students who failed the test answered the question correctly) to -1.00 (none of the students who passed the test answered the question correctly while all of the students who failed the test answered the question correctly). High positive

values show that an item is discriminating well. Negative values indicate that there is a problem with item and that it should be reviewed (Brown, 2003; Brown & Hudson, 2002).

Item phi shows the correlation between students' performance on an item and their performance on the test as a whole and is calculated as follows:

$$\Phi = (P_{iT} - P_i P_T / \sqrt{P_i Q_i P_T Q_T})$$

Where:

P_i is the proportion of examinees who answered the item correctly.

Q_i is the proportion of examinees who answered the item correctly (1-Pi).

P_T is the proportion of examinees who passed the test.

Q_T is the proportion of examinees who failed the test (1-*PT*).

P_{iT} is the proportion of examinees who answered the item correctly and passed the test.

Item phi values will generally be similar to *B*-index values and can be interpreted in the same way (Brown & Hudson, 2002).

The item agreement statistic (A) shows the probability of agreement between a student answering an item correctly and whether they passed or failed the test. It can be expressed using the following formula:

$$A = 2P_{iT} + Q_i - P_T$$

Where:

P_{iT} is the proportion of total students who answered the item correctly.

Q_i is the proportion of students who answered the item incorrectly.

PT is the proportion of students who passed the test.

The A statistic can also be used on items which are not scored dichotomously, in which case, Q_i is the proportion of students who achieved a passing score on the item. The range of values for the A statistic is 0.00 to 1.00, with higher values indicating items are discriminating better (Brown & Hudson, 2002).

Results for the *B*-index and item phi are often similar, while the item agreement statistic is likely to be different because it does not reference students who failed the test.

(Brown, 1991; Brown & Hudson, 2002; McCowan & McCowan, 1999). For all the indices, a higher value indicates that an item is contributing more towards master / non-master decisions, but low values do not necessarily indicate bad items. The low value might be a result of the learning materials being confusing with regards to the target of the particular item, or an indication that students are not yet ready to learn that particular objective (Brown, 2003). When selecting items for a revised form of a test, it is important that test developers consider a range of item statistics and how items fit the objectives or content being measured (Brown & Hudson, 2002).

While test consistency is usually referred to as reliability with regards to NRT, in the case of CRT, consistency is described by the term dependability and refers to the consistency of classification of students into masters or non-masters (Brown & Hudson, 2002; Yan & Fan, 2021). Brown (1990) suggests a shortcut to the phi (lambda) dependability index as a method of establishing the dependability of CRTs with only one administration. The formula is as follows:

$$\Phi(\lambda) = 1 - (1/k - 1)\ [(X_p(1 - X_p) - S_p^2) / ((X_p - \lambda)^2 + S_p^2)]$$

Where:

λ is the cut point.

k is the number of items.

X_p is the mean of proportion scores.

S_p^2 is the standard deviation of proportion scores.

Kane (1986) suggests that the reliability for a CRT should be above 0.5, and that tests which show a reliability of lower than 0.5 should be lengthened or the criteria specifications should be improved.

Test development and analysis in practice

There are relatively few published studies into course-level EFL test development and analysis. Brown (1991) reported on the development of a set of CRTs for the EFL programme at the English Language Institute at the University of Hawai'i. The analysis compared the usefulness of NRT, CRT and Item Response Theory approaches. Due to the large number of tests, individual item analyses were not included in the results, but the study suggested that using the NRT analysis was useful because the development team were

familiar with how to interpret the results from NRT analysis, and modifying the tests in line with the NRT analysis would allow the creation of effective placement tests. The study found that CRT analyses were more useful in achieving a suitable level of difficulty for the of end of course tests. The tests were consistent as both NRTs and CRTs, but CRT dependability for the same test was sensitive to cut point, that is, the same test could have high or low dependability depending on the level of the pass / fail score.

Yoshida (2007) reported on the development of a course-level English language vocabulary test at a Japanese university. The analysis used a mix of CRT statistics (B-index, phi (lambda)) and NRT statistics (item discrimination, K-R 21). Item analysis was based on the item facility, B-index and item discrimination values. After the first analysis, 19 items were removed, of which 14 were items with an item facility of 1.00 (all the students answered correctly) and five had negative B-index values (more students who failed the test answered the item correctly than students who passed the test). A further 14 easy items were removed after a second round of analysis.

Development of a new Level 2** Listening & Speaking I **final test

A new Level 2 *Listening & Speaking I* course is being developed by a working group of teachers at Kyushu Sangyo University's Language Education and Research Center (LERC) as part of a redesign of the Level 2 *Listening & Speaking* curriculum. A full pilot of the new course was undertaken in the first semester of the 2023 academic year. The pilot involved all Year 1 Level 2 *Listening & Speaking I* classes, and six Year 2 Level 2 *Listening & Speaking III* classes.

The primary aim of the new course was to help students speak in detail about familiar topics, and during the first semester students studied eight speaking topics. The course utilized a flipped approach to learning. Before each class, students completed a number of homework tasks to help scaffold speaking tasks in the class. The homework consisted of studying model answers to the week's topic questions, followed by e-learning activities in which students studied useful vocabulary (Language Practice 1) and grammar (Language Practice 2) related to the topic questions. Students were then presented with detailed answer guides to the topic questions which provided the scaffolding for students to write their own answers to the topic questions. At the beginning of each class, students took a short test to assess their understanding of the Language Practice 1 and Language Practice 2 content. The questions used in the weekly test were randomly drawn from the week's Language Practice 1

and Language Practice 2 homework activities. Examples of the types of questions used in the homework e-learning activities can be found in Appendix A.

Assessment for the new course consisted of six parts. The primary focus of the course was speaking, so two speaking tests, each worth 15% of the students' final grade, were included in the course. The flipped approach used by the course meant that homework completion was integral to students being able to participate fully in classes. As such the Language Practice 1 and 2 homework activities were worth a combined 20% of the final grade as were the weekly homework tests. The students' answers to the week's topic questions were worth 10% of the final grade. In order to encourage the students to revisit previous homework activities, students took a final test drawn from all questions from the Language Practice 1 and 2 homework activities in the final class of the semester. This test was worth 10% of the students' final grade. The final 10% of the students' grade was determined by their performance on the KSU achievement test.

The purpose of the final test was threefold. The first purpose was to encourage the students to review vocabulary and grammar items presented during the course. The second purpose was to assess students' mastery of the grammar and vocabulary items presented during the course. The final purpose was to provide feedback to the curriculum development team regarding areas in which students might need more support in learning the vocabulary and grammar items presented during the course.

In terms of the test specifications, the curriculum development team specified that the final test should be designed to assess students' knowledge of the key grammar and vocabulary items from the course. A cut score of 60 percent was decided. It was decided that the test should take a maximum of 25 minutes and consist of 40 questions – five from each of the eight topics studied during the semester – and that all questions should be equally weighted. The test would be taken on students' smartphones in the final class of the semester and would be graded automatically in Moodle Quiz.

As the test was designed to measure students' mastery of the course material, the items were selected from the weekly in-class tests of the homework e-learning. Questions were selected based on item analysis statistics automatically generated in Moodle. The nature of the homework activities meant that most of the homework items had high facility index values, but where possible, items with facility index values of 0.40 to 0.75 and discrimination index values of greater than 0.30 were selected. These items discriminated well between

students who had learned the content and students who had yet to learn the content in the weekly class tests. One or two vocabulary items and three or four grammar items were selected for each topic. Two main question types were used: fill-in-the-blanks and drop-down multiple choice. The fill-in-the blank questions were designed to test productive knowledge of language points, and the multiple-choice questions were designed to test receptive knowledge. Three questions for each topic were fill-in-the-blank questions with the students required to read a Japanese sentence and fill in the blanks in the English translation. It was decided that using all fill-in-the-blank questions would be too time-consuming and repetitive for the students, so the first two questions in each section were drop-down multiple-choice questions. The decision to use drop-down multiple-choice was to prevent directly displaying incorrect answer choices to the students. These questions were either word order questions or discrete word choice, and no Japanese translation of the English was provided. The items were reviewed by the members of the curriculum development team and minor modifications were made to some items.

No direct piloting was undertaken due to a lack of time; however, the items were selected from items which had good item statistics when used in the e-learning homework. To ensure that students were familiar with the test content and the format of the test, an important feature of criterion-referenced tests, a review activity containing 40 questions in the same style as the final test but also including grammar review information was assigned as a homework activity in the week prior to the final test. The operational assessment was the use of the test at the end of the 2023 *Listening & Speaking I* course, and the assessment review phase was the statistical analysis discussed in the next part of this paper.

Aims

The previous discussion has highlighted how test analysis is an integral part of the test development cycle and how it can feedback into the materials development cycle, but that many classroom and course-level tests are developed in a linear way and with no analysis of the test or score distributions. As such, the main aim of this research project was to undertake a statistical analysis of the 2023 Level 2 *Listening & Speaking I* final test. The results of the analysis will be important in the further development of the Level 2 *Listening & Speaking I* course.

It is also clear from the previous discussion that many teachers are more familiar with NRT statistical analysis and that NRT statistics have been applied successfully to CRT

analysis. Therefore, a secondary aim was to assess the usefulness of NRT statistics, such as those automatically generated by Moodle Quiz, in analysing the outcomes of the Level 2 *Listening & Speaking I* final test compared to an analysis using CRT statistics.

In order to facilitate the analysis, the following research questions were formulated.

1. What were the descriptive statistics and score distribution of the 2023 Level 2 *Listening & Speaking I* final test?
2. What were the item statistics of the 2023 Level 2 *Listening & Speaking I* final test?
3. How consistent was the 2023 Level 2 *Listening & Speaking I* final test in measuring students' knowledge of the e-learning homework content?
4. Which approach, between the norm-referenced test analysis and criterion-referenced test analysis, provided more actionable insights for revising the 2023 Level 2 *Listening & Speaking I* final test?

Sampling and Methods

Participants

The new 2023 Level 2 *Listening & Speaking I* course contained 348 Year 1 students across 21 classes. 90 Year 2 students across six classes taking Level 2 *Listening & Speaking III* also completed the same material. The students' English proficiency, based on their performance in the KSU placement test, was equivalent to A2 on the Council of Europe's Common European Framework of Reference for Languages (CEFR). 400 students completed the final test. In order to comply with university research ethics guidelines, at the beginning of the 2023 academic year, all students completed a data consent form which included the use of data collected from the homework e-learning activities and the final test, from which data for this research were gathered. The data from the eight students who did not consent to their data being used for research purposes were deleted before statistical analyses were conducted. Further, all data were anonymized and identifying data were deleted.

Material

The materials used in this study were data from 392 student scores on the 2023 Level 2 *Listening & Speaking I* final test. The test consisted of 40 questions. Examples of the questions used in the test can be found in Appendix B.

Methods

Data for the analysis were downloaded from the final test Moodle Quiz results page. The NRT statistics generated automatically by Moodle Quiz were not used as they contained data from students who had not given their consent. Instead, the raw data were downloaded from Moodle Quiz and data from students who had not consented to their data being used were deleted. Much of the analysis required dichotomous data, so the raw scale scores generated by Moodle Quiz were first converted into binary data.

Descriptive statistics and item statistics were calculated. Item facility values were calculated, and item difficulty was assessed. In line with item analysis criteria suggested by Brown (2021) and Khalifa and Weir (2009), items with an IF value of 0.76 or higher were classified as easy, between 0.75 to 0.20 were ideal, and 0.19 or lower were difficult. Item discrimination statistics were calculated for NRT and CRT analyses. For the NRT analysis, items with an ID value of 0.40 or higher were classified as good, 0.39 to 0.30 as acceptable but should be checked, 0.29 to 0.20 as need revision, and lower than 0.19 as poor items. For CRT item analysis, *B*-index, A statistic, and item phi were calculated for each item. The three values were considered along with the item facility. Items with an IF value of more than 0.76 combined with item phi or B-index values of lower than 0.10 were easy items for this cohort and were categorised as making little contribution to cut decisions. Items with IF values of less than 0.75 and B-index or item phi values of lower than 0.10 were categorised as need checking, as were items with an A-index value of less than 0.60 and a B-index or item phi value of less than 0.30. NRT reliability was estimated using K-R 20, K-R 21, and Cronbach's alpha. CRT dependability was estimated using Phi (Lambda) at cut scores of 60, 70, 80, and 90. All analysis was conducted in Microsoft Excel.

Analysis

In this section, the results are presented from the statistical analysis of the 2023 final test for the Level 2 *Listening & Speaking I* course. Descriptive statistics for the scores can be found in Table 1, and Figure 1 shows a histogram of the test scores.

Analysis of the descriptive statistics and histogram shows that the distribution of the scores is skewed. The median score (29) is greater than the mean (28.78,) indicating that there is a higher concentration of scores towards the higher end (right side) and a longer tail extending to the left side (lower scores) than would be expected in a normal distribution. This

suggests a left-skewed distribution. This is confirmed by the skew value. The negative skew value indicates a left-skewed distribution. The skew value of -0.39 exceeds two standard errors of skewness, which is approximately 0.225 (2*[√6/392]). Since - 0.39 is outside the range of -0.225 and +0.225, the distribution can be considered significantly skewed. The kurtosis value of 0.08 is within two standard errors of kurtosis indicating the relative peakedness of the distribution shows no significant deviation from a normal distribution (Brown, 1997).

Table 1

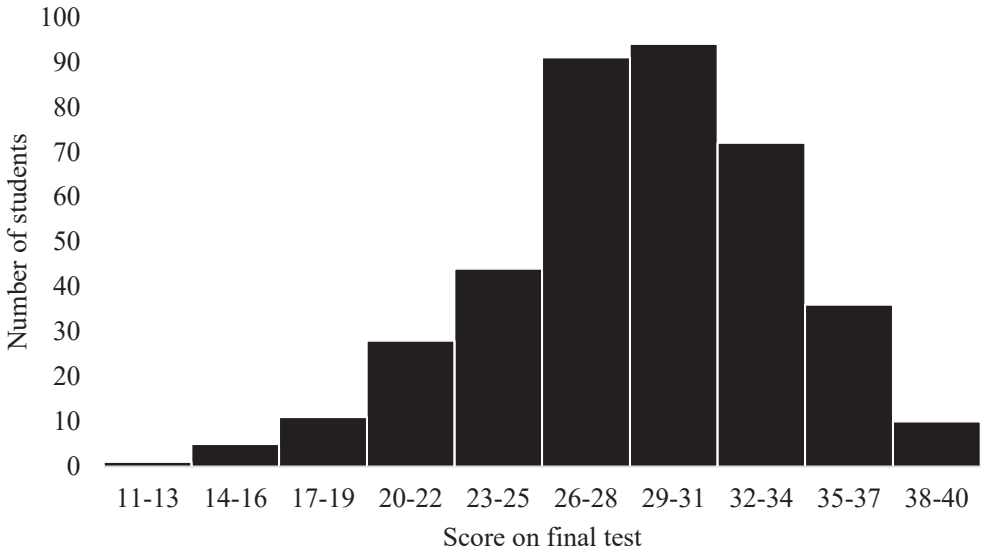
Descriptive Statistics for the Listening & Speaking I Final Test

Statistic	value
N	392
k	40
M	28.78
Median	29
Mode	31
Low	13
High	40
Range	28
S	4.89
Skew	-0.39
Kurtosis	0.08

Note. N is the number of students; *k* is the number of items in the test; *M* is the mean; *S* is the standard deviation.

Figure 1

Histogram of Test Scores for the Listening & Speaking I Final Test



Item statistics for the test are displayed in Table 2. Analysis of the item facility values shows that none of the items were difficult for this group of test takers, 21 items were of moderate difficulty, and 19 items were easy. The NRT item discrimination values indicate that nine items were effective, seven items require checking, eight items require revision, and 16 items should be discarded. The CRT item analysis shows that nine items made little contribution to cut decisions but were easy items, and four items made little contribution to cut decisions and were of moderate difficulty so require checking.

Table 2

Item Statistics for the Listening & Speaking I Final Test

				NRT		CRT			
Item	Question Type	IF	Remarks	ID	Remarks	B-index	A	Item Φ	Remarks
1	Multiple-choice (word choice)	.79	easy	.29	needs revision	.25	.77	.21	
2	Multiple-choice (word choice)	.97	easy	.05	poor	.03	.84	.06	little contribution to cut decisions
3	Fill-in-the-blanks	.66	moderate	.41	good	.34	.69	.25	
4	Fill-in-the-blanks	.40	moderate	.48	good	.29	.50	.21	
5	Fill-in-the-blanks	.38	moderate	.22	needs revision	.14	.45	.10	requires checking
6	Multiple-choice (word order)	.74	moderate	.39	moderate	.39	.77	.32	
7	Multiple-choice (word order)	.99	easy	.04	poor	.07	.86	.21	little contribution to cut decisions
8	Fill-in-the-blanks	.44	moderate	.37	moderate	.28	.53	.20	
9	Fill-in-the-blanks	.92	easy	.11	poor	.12	.83	.16	little contribution to cut decisions
10	Fill-in-the-blanks	.79	easy	.32	moderate	.35	.80	.31	
11	Multiple-choice (word choice)	.56	moderate	.27	needs revision	.19	.59	.13	requires checking
12	Multiple-choice (word choice)	.86	easy	.17	poor	.08	.77	.08	little contribution to cut decisions

13	Fill-in-the-blanks	.41	moderate	.55	good	.33	.52	.24	
14	Fill-in-the-blanks	.49	moderate	.38	moderate	.47	.61	.33	
15	Fill-in-the-blanks	.73	moderate	.28	needs revision	.14	.70	.11	
16	Multiple-choice (word order)	.77	easy	.34	moderate	.39	.79	.33	
17	Multiple-choice (word order)	.97	easy	.04	poor	.11	.86	.23	
18	Fill-in-the-blanks	.91	easy	.13	poor	.18	.84	.22	
19	Fill-in-the-blanks	.85	easy	.23	needs revision	.28	.82	.28	
20	Fill-in-the-blanks	.49	moderate	.32	moderate	.34	.57	.24	
21	Multiple-choice (word order)	.57	moderate	.29	needs revision	.38	.64	.27	
22	Multiple-choice (word order)	.93	easy	.17	poor	.33	.89	.46	
23	Fill-in-the-blanks	.82	easy	.28	needs revision	.30	.80	.28	
24	Fill-in-the-blanks	.58	moderate	.54	good	.45	.67	.32	
25	Fill-in-the-blanks	.36	moderate	.48	good	.33	.48	.24	
26	Multiple-choice (word choice)	.97	easy	.06	poor	.02	.84	.04	little contribution to cut decisions
27	Multiple-choice (word choice)	.84	easy	.19	poor	.20	.79	.20	
28	Fill-in-the-blanks	.60	moderate	.51	good	.50	.69	.36	

29	Fill-in-the-blanks	.49	moderate	.17	poor	.02	.50	.02	requires checking
30	Fill-in-the-blanks	.70	moderate	.19	poor	.12	.67	.01	requires checking
31	Multiple-choice (word order)	.99	easy	.00	poor	.01	.85	.06	little contribution to cut decisions
32	Multiple-choice (word order)	.97	easy	.05	poor	.09	.86	.21	little contribution to cut decisions
33	Fill-in-the-blanks	.65	moderate	.51	good	.46	.72	.34	
34	Fill-in-the-blanks	.61	moderate	.28	needs revision	.20	.63	.14	
35	Fill-in-the-blanks	.80	easy	.16	poor	.15	.75	.14	little contribution to cut decisions
36	Multiple-choice (word order)	.89	easy	.19	poor	.21	.83	.24	
37	Multiple-choice (word order)	.99	easy	.02	poor	.05	.86	.17	little contribution to cut decisions
38	Fill-in-the-blanks	.48	moderate	.40	good	.29	.55	.21	
39	Fill-in-the-blanks	.70	moderate	.37	moderate	.36	.73	.28	
40	Fill-in-the-blanks	.74	moderate	.40	good	.49	.79	.39	

Consistency estimates – Cronbach's alpha, Kuder-Richardson 20 (K-R 20), Kuder-Richardson 21 (K-R 21), and standard error of measurement for NRT reliability analysis, and phi (lambda) dependability index for CRT dependability – are presented in Table 3.

K-R 20 indices and Cronbach's alpha values were greater than 0.70, which is generally considered to be acceptable for tests of around 50 items (Riazi, 2016). K-R 21 indices are typically lower than K-R 20 (Brown & Hudson, 2002), so the lower K-R 21value is not unexpected. The SEM indicates that the true score of a student who scored 29 on the test lies between 26.53 (29 – 2.47) and 31.47 (29 + 2.47) with 68% certainty. CRT dependability values were greater than 0.50 at all cut scores, indicating that the test was dependable at these cut scores.

Table 3

Reliability and Dependability Measures for the Listening & Speaking I Final Test

NRT				CRT			
				Phi (lambda)*			
alpha	K-R 20	K-R 21	SEM	60	70	80	90
.75	.75	.68	2.47	.83	.69	.77	.90

Note. Phi (lambda) values were calculated for cut scores of 60, 70, 80 and 90.

Discussion

This section discusses the results from the analysis section in relation to the four research questions outlined earlier.

What were the descriptive statistics and the score distribution of the 2023 Level 2 Listening & Speaking I final test?

Analysis of the descriptive statistics and histogram showed that more students achieved higher scores than would be expected with a normal distribution. Given the nature of the test, this is to be expected, and suggests that the test was working as intended. However, there are implications for how item statistics and reliability should be calculated and interpreted. NRT

test analysis statistics assume a normal distribution, so data from NRT analysis should be treated with caution.

What were the item statistics of the 2023 Level 2 Listening & Speaking I final test?

The item facility values showed that there were no difficult items in the test and there was a relatively even split between moderately difficult and easy items. This suggests that the students had a good level of knowledge of the vocabulary and grammar items taught in the course, and the content of the questions was an appropriate representation of content from the course. For a criterion-referenced achievement test, these results are to be expected and show that the test was functioning as planned (Brunfaut & Harding, 2014). Due to the single post-test, no inferences regarding learning can be made from this administration of the test.

The item discrimination analysis showed that all items had positive discrimination values, which indicates that the questions were all testing the same construct and that no items contained major errors or incorrect answers. However, in the norm-referenced analysis, only nine items had item discrimination values of greater than 0.40 indicating that they would be suitable for inclusion in the next iteration of the test without revision. 16 items had ID values of less than 0.19 and would need to be discarded and replaced with new items if the aim of the test was to produce a normal distribution. These results are similar to those found by Yoshida (2007), however, removing the items from the test is likely to lower the test consistency, and there is a limited pool of possible replacement items from the homework activities.

In the CRT analysis, 13 items were making little contribution to pass / fail decisions. Nine of these items had high item facility values. In a CRT, it is not unusual for items to have high facility values and therefore offer little discrimination between high scoring and low scoring students. It suggests that most students had a good understanding of the language points being tested rather than a problem with the questions. These items should be checked by the curriculum development team, but it is possible that the items do not need revising. It should be noted that eight of these items were multiple-choice items, and the curriculum development team should consider whether this style of question is adequately testing students' knowledge. The remaining four items, items 5, 11, 29 and 30, were not easy items in terms of their IF values. Items with lower IF values would be expected to discriminate well between high performing and low performing students, however, the low discrimination values indicate that these items are not functioning as expected. These items should be

checked by the curriculum development team to ensure that the domain being tested is clear, and that the CEFR level of the target language is appropriate for students on this course.

How consistent was the 2023 Level 2 Listening & Speaking I final test in measuring students' knowledge of the e-learning homework content?

The results (Table 3) showed that the test was reasonably reliable as both a NRT and a CRT. This test was relatively easy for this group of students and a large number of the items showed little variance in scores. It is difficult to achieve high reliability in a test without a large spread of scores, and lower reliability values are not unusual when the range of proficiencies of the students has been restricted by a placement test (Brown, 1991; Brunfaut & Harding, 2014), so the reliability indices for this test are reasonably good.

Similarly, the CRT dependability values appear to be acceptable. At a cut score of 60, which is the score used for pass / fail decisions in this course, the dependability value of 0.83 is good. When tests are used for pre/post-tests, it is common for the cut score for the pre-test to be set at a higher level (Brown, 1991). The test was most dependable at a cut score of 90, so this might be the most appropriate cut score if the test is used as a pre-test with future cohorts.

Which approach, between the norm-referenced test analysis and criterion-referenced test analysis, provided more actionable insights for revising the 2023 Level 2 Listening & Speaking I final test?

The NRT statistics provided a good general overview of the test. From the NRT statistics it was possible assess the difficulty of the test and the reliability of the test. While the reliability of the test was acceptable, revising the test to achieve acceptable NRT item statistics would require a more difficult test. This would be difficult to achieve with questions drawn from the current course material if the test was to keep the same specifications of 40 items with five items for each topic. This is a potential drawback of using only NRT statistics to revise a CRT. However, as suggested by Brown (1991), NRT analysis could be used to create a bank of validated questions for the creation of a placement test for the course.

CRT analysis provided more nuanced insights into item discrimination. Given the nature of the test, this is not surprising. A CRT achievement test is likely to have a number of relatively easy items, and the availability of different types of item statistics provided more information with which to make decisions on which items to keep and which items to revise or discard.

In this analysis, there was not a big difference in the insights to be gained from the two approaches. Both analyses suggested that the test was easy and that the items were generally working as expected. However, this analysis only considered the first administration of the test and there was no data from a pre-test to analyse how much learning had occurred during the course. With data from a pre-test, it is likely that CRT analysis would provide greater insights for both revising the test and informing curriculum development. A pre-test would allow calculation of difference index statistics, which can be used to infer how much students have learned during the course and the appropriateness of the difficulty of the course material. However, this analysis validated the test and means that it should be suitable for use with future cohorts, both as a pre-test and a post-test.

Conclusion

Statistical analysis of the 2023 Level 2 *Listening & Speaking I* final test shows that the curriculum development team created a consistent test. The distribution analysis suggests that this group of students had a good level of knowledge of the language content presented in the course, and the item analysis shows the test functioned as planned. The psychometric properties of the 2023 *Listening & Speaking I* final test suggest that with minor modifications, it would be suitable for use as a pre/post-test for future cohorts provided that the course content remains unchanged, and that the questions could be added to a bank of questions to use as part of a placement test for the course.

While the test analysis showed no major faults with the test, it did highlight questions that should be checked and revised, confirming the importance of a circular approach to test development (Brown & Hudson, 2002; Brunfaut & Harding, 2014; Green, 2021). Students invest a lot of time in learning English, and it is important that tests which form part of students' grades are analysed and improved after each administration.

The analysis in this study showed that relying on NRT statistical analysis to analyse criterion-referenced tests might lead to undue modifications to the test. While teachers may be less familiar with CRT approaches to test analysis, they can provide more actionable insights into how the kind of achievement tests that teachers routinely design can be modified and made more effective.

The single post-course administration of the test meant that the results provide no insight into learning gains. For future cohorts, a pre-test post-test regime would provide the course designers with information on the appropriateness of language targets and any learning gains made during the course, and better inform the future development of the course.

References

Brown, J. D. (1990). Short-cut estimators of criterion-referenced consistency. *Language Testing, 7*(1), 77-97. https://doi.org/10.1177/026553229000700106

Brown, J. D. (1991). A comprehensive criterion-referenced language testing project. *University of Hawai'i Working Papers in English as a Second Language 10*, 95-125. Retrieved from https://scholarspace.manoa.hawaii.edu/items/05fd1f48-d157-4e79-a48c-96e7e334aadf/full

Brown, J. D. (1997). Skewness and kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(1), 20-23. Retrieved from https://hosted.jalt.org/test/PDF/Brown1.pdf

Brown, J. D. (2003). Criterion-referenced item analysis (The difference index and the *B*-index). *Shiken: JALT Testing & Evaluation SIG Newsletter, 7*(3), 18-24. Retrieved from https://hosted.jalt.org/test/bro_18.htm

Brown, J. D. (2021). Classical test theory. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 447-461). Oxon, UK: Routledge.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing* (Cambridge Applied Linguistics). Cambridge: Cambridge University Press.

Brunfaut, T., & Harding, L. (2014). *Developing English language tests for Luxembourg secondary schools: The Test Design and Evaluation (TDE) project, 2011-2014.* Lancaster University. http://portal.education.lu/Portals/22/English/Documents/BrunfautandHarding2014.pdf

Douglas, D. (2014). *Understanding language testing*. Oxon, UK: Routledge.

Green, A. (2021). *Exploring language assessment and testing*. Oxon, UK: Routledge.

Green, R. (2013). *Statistical analyses for language testers.* London: Palgrave Macmillan. https://doi.org/10.1057/9781137018298

Kane, M. T. (1986). The role of reliability in criterion-referenced tests. *Journal of Educational Measurement*, *23*(3), 221-224. Retrieved from https://www.jstor.org/stable/1434609

Khalifa, H., & Weir, C.J. (2009). *Examining reading: studies in language testing 29*. Cambridge: Cambridge University Press.

McCowan, R. J., & McCowan, S. C. (1999). Item analysis for criterion-referenced tests. *Center for Development of Human Services.* Retrieved from https://eric.ed.gov/?id=ED501716

Riazi, A. M. (2016). *The Routledge encyclopedia of research methods in applied linguistics*. Oxon, UK: Routledge.

Yan, X., & Fan, J. (2021). Reliability and dependability. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 477-494). Oxon, UK: Routledge.

Yoshida, H. (2007). Analyzing an achievement test. *関西大学外国語教育フォーラム巻, 6*, 37–51. Retrieved from https://www.kansai-u.ac.jp/fl/publication/pdf_forum/6/04_yoshida_37.pdf

Appendix A

E-Learning Homework Sample Questions

Language Practice 1

Complete the sentences with an appropriate English word. One blank represents one word.
空欄に英単語を入力しましょう。各空欄は 1 つの英単語を表しています。

1. 私の故郷は長崎県の町である島原です。

My _____ is Shimabara, which is a town in Nagasaki Prefecture.

(correct answer: My **hometown** is Shimabara, which is a town in Nagasaki Prefecture.)

Language Practice 2

a. Read the Japanese sentence and complete the English sentence.
日本語の文章を読んで、英文を完成させましょう。

1. ６年前に英語の勉強をはじめました。

I started (playing the piano / studying English / learning Korean / learning to cook / living in Fukuoka) (last year / when I was 15 / when I was 8 / six years ago / two months ago)

(correct answer: I started **studying English** **six years ago**.)

2. Complete the sentences using the ~*ing* form of a verb from the list. You do not have to use all the verbs.
リストにある動詞の中から正しい動詞を選び、空欄を埋めましょう。動詞の「~*ing*」形を入力しましょう。リストにある動詞をすべて使う必要はありません。

play	**study**	**work**
go	**cook**	**read**
do	**watch**	**live**

1. I started _____ in a convenience store last week.
(correct answer: I started **working** in a convenience store last week.)
2. I started _____ karate about five years ago.
(correct answer: I started **doing** karate about five years ago.)

Appendix B

2023 Level 2 Listening & Speaking I Final Test Sample Questions

Topic 1 Self-introduction

Choose the correct word or complete the sentence by filling in the blank with an appropriate English word (one space represents one word).

正しい単語を選び、または表示されていない単語（1 つのスペースは 1 つの単語を表す）を入力して文章を完成させましょう。

1. I (play / do / go) yoga.

(correct answer: I **do** yoga.)

2. I have (gone / known / lived) in Fukuoka for 6 months.

(correct answer: I have **lived** in Fukuoka for 6 months.)

3. 所有物

_____

(correct answer: possession.)

4. バレエを始めたのは 8 歳の時です。

I _____ _____ ballet when I was eight.

(correct answer: I **started** **doing / learning** ballet when I was eight..)

5. 鉄道に興味を持ったのは、幼い頃です。

I _____ _____ _____ trains when I was very young.

(correct answer: I **got / became** **interested** **in** trains when I was very young.)