# Critical Evaluation of the 'Matura' Test: CEFR Alignment Project for the Austrian National Examination in English (B2 Level)

Minzi Li[1*] Yongqiang Zeng[2] Ligerui Chen[3]
1.Guangdong University of Foreign Studies, Baiyun Dadao Bei #2, Guangzhou, Guangdong, 510420 China
2.Guangdong Teachers College of Foreign Language and Arts, Tianhe Shougouling #463, Guangzhou, Guangdong, 510640 China
3.Queensland University of Technology, Brisbane, Queensland, QLD 4000, Australia
* E-mail of the corresponding author: whlmz@163.com

**Abstract**
This study attempted to evaluate the usefulness a CEFR (Common European Framework of Reference) alignment project 'Matura' listening test with respects of the item analysis, validity (i.e., content validity) and reliability (i.e., internal consistency reliability, scoring reliably). 93 students randomly selected from six secondary-school of different school-levels across three different regions in Austria completed the listening test. SPSS was employed for the statistical analysis. Facility value, discrimination index together with other descriptive statistics were reported for the item analysis. Content validity by a panel of expert judgments as well as the reliability of the listening test was further examined. Findings showed that: 1) the test paper was of average to high difficulty with its peak score locating around the mid-point, which allow higher education institutions to set an appropriate cut-score for decision making. Relatively widespread test scores further indicated its potential to efficiently discriminate learners of varied listening proficiency; 2) the majority of items in the listening test performed well and were qualitatively reasonably good. However, we did identify several problematic items with possible causes related to construct-irrelevance, too many possible answers, mismatch sign-posted words, fast audio speed, testing simply the background knowledge, and the heavy cognitive load required for "Not Given" option; 3) regarding validity and validity issues, construct-under representation and the less authentic listening materials were then spotted by careful content analysis. Though the statistical output reported relatively high reliability, results should be interpreted with caution. Based on findings, general evaluation of the test as well as implications for further improvement were revealed.

**Keywords:** test evaluation, usefulness, content validity, reliability, Common European Framework of Reference, listening comprehension, item analysis
**DOI:** 10.7176/JEP/11-11-01
**Publication date:** April 30th 2020

## 1. Introduction

The Austrian 'Matura' is a high-stake CEFR (Common European Framework of Reference) alignment end-of-secondary-school-leaving examination in English and has been widely employed in the country. The examination, consisting of four subtests (writing, reading, listening, and use of English), is developed as a proficiency test as its primary goal is to ascertain whether learners' listening proficiency have reached the B2 level to enrol the university. Though the curriculum goal for the final two years of secondary school in Austria is CEFR B2, the Matura test will not test the specific content extracting from individual English courses. Results, indicating learners' language proficiency, will serve the selection purposes by higher education institutes.

The CEFR is a descriptive language standard measure drawing an overall picture of language-level classification for certification across European nations, as well as providing easier access for co-operation among educational institutions. This language framework is constructed based on an action-orientated approach to language learning (Council of Europe, 2001) and involves both general competences of an individual and communicative language competence in the light of Bachman's categorization (Bachman,1990; Bachman & Palmer, 1996). Accordingly, the CEFR describes an ascending continuum of levels for language proficiency and categorizes into three broad levels: basic users (levels A1, A2), independent users (levels B1, B2), and proficient users (levels C1, C2) with each comprises two sub-groups (Council of Europe, 2001).

The Matura listening test is designed on the basis of CEFR descriptors (can-do statements) as sub-skills listed in the test specification directly correspond to the CEFR descriptors at the B2 level. Considering that language learners' proficiency of the Matura test is highly associated with the decision made by higher education institutions whether language learners are qualified to enrol the university system (the fixed attainment of B2 level), it is, therefore, recommend test-developers/item writers to ensure the quality of the Matura test (high-stake test).

According to Bachman (1990), developing a high-quality proficiency test can be a complex endeavour. Bachman and Palmer (1996) claimed that "the most important quality of a test is its usefulness" (p. 17). The pre-test evaluation is deemed an essential step for the test development as it examines whether the quality criterion has been met or can be improved (Alderson, et al., 1995). Buck (1991) asserted that even experienced item-writers/test developers can hardly identify some problems within the test. It is claimed that only when scores are found both reliable and valid the test itself fulfils the purpose being useful (Douglas, 2010). *Test reliability* refers to "the extent to which test scores are consistent" (Alderson et al. (1995), p. 294). When comparisons were made between assessments, scores (e.g., scores on different forms of a test, marked by different raters, or given on different conditions from the same test) bring us insights into *reliability*. Considering that the test reliability is likely to be affected by factors such as the quality of the test, or scoring methods (Brown & Abeywickrama, 2010), discussion on reliability in the study will be elaborated from those two aspects (i.e., internal consistency reliability, scoring reliability). Another important indicator, *Validity*, specifies a "test measures exactly what it proposes to measure" (Brown & Abeywickrama, 2010, p.30). Concerning diversified types of validity with each of no absolute definitions according to the synthesis literature, this study only concerns the content validity of the test due to the research purpose. According to Hughes (2003), *content validity* refers to whether test items and scores of test-takers are representative of those speculated in its language domains. Those two important indicators will be employed for the evaluation of the test.

Furthermore, *item analysis* is a statistical technique used to ascertain the usefulness of test items in the test. Statistical analysis of items together with the classical test theory provides important information to evaluate a test and diagnose where individual items may be contributing to the inconsistency of results. Based on the interpretation of statistical data, problematic items can be identified, revised, replaced, or rejected to improve the quality of an assessment. It has been claimed that the Facility Value (F.V.) and Discrimination Index (D.I.) are two main statistical indicators to evaluate the quality of test items (e.g., Alderson, et al., 1995).

The *Facility Value* gives hint about the difficulty degree learners perceived on items, which is represented by a number ranging from 0 to 1. To be specific, an item facility of 1 suggests that all test-takers respond correctly to the item, while 0.0 represents none of the students answers the item correctly. According to Khalifa and Weir (2009), an item can be reviewed as too difficult if it demonstrates a facility value is .25 or below, and too easy if it displays a facility value of .75 or above. Alderson et al. (1995) suggested that the facility value above .80 may be deemed as an easy item. However, there is no clear-cut rule as the 'rule of thumb' varies in accordance with test purposes or what the test is intended to measure (Hughes, 2003). Items in the proficiency test are likely to generate higher facility values as items too difficult or easy in a proficiency test may inform no useful information to distinguish test-takers of different language competences, and thus fail to fulfil the purpose of the test. In this regard, Alderson et al. (1995) argued that an item with a facility value of .95 or above provides no information and is therefore not very useful.

Despite the difficulty level of an item, how well an item discriminates between high performing students and low performing students should be taken into account, which can be informed by another indicator *Discrimination Index*. An item with low discrimination indicates that less able test-takers outperform capable test-takers, which sends the message to item writers/developers that the item may attract the wrong person and requires a further revision (Alderson, et al., 1995). Though no absolute value of what a perfect discrimination index of an item might be as it may sensitive to test purposes and task types (Hughes, 2003), it is generally acknowledged that a value above .4 may be considered acceptable (Alderson, et al., 1995). Since the discrimination index can be calculated in different ways, the D.I. of items in the study were examined on the basis of "Corrected Item-Total Correlation" (CITC) presenting by the *point biserial discrimination*.

## 2. Purpose

The purpose of this study is to evaluate the usefulness a high-stake CEFR-alignment Austrian 'Matura' listening test with respect of the item analysis, validity (i.e., content validity) and reliability (i.e., internal consistency reliability, scoring reliably) and make brief suggestions concern how the test might be further improved.

## 3. Research Method

### *3.1 Participants*

93 students randomly selected from six secondary-school of different school-levels (high-level school and average-level school) across three different regions in Austria completed the listening test. Among which, 7/8 of the total population (83 students) are of average level, while only around 1/8 (10 students) are high-level students.

### *3.2 Instrument*

(1) The listening pretest of 'Matura'

The test evaluated in this study is a listening pretest of 'Matura'. The test has four tasks with each of a separate audio. All listening materials involves linguistic activities that take place in daily-life context (e.g., climate change, a popular social platform, iPhone and the bicycle tour) that may frequently be encountered by learners in order to ensure learners' familiarity with the topics to a large degree. Each task contained 6 to 10 items with a total of 30 items. Items in the test are of three formats: multiple matching, True/ false/ not given items, and short-answer questions with the words limit of four. Each item is of equal weight and score 1 point per item. According to the marking instruction, spelling and punctuation mistakes will not be penalized if it does not affect the "communication".

(2) The content analysis grid

The content analysis grid presented the CEFR descriptors at B2 level as well as some general listening sub-skills which the test intend to measure in the test specification (Table 6), which is a crucial part of the test construct and directly corresponding to task and individual items of 'Matura' listening test.

### *3.3 Procedure and Data Analysis*

Prior to the study, all the participants were informed about the purpose and content of the study. Test-takers' responses were scored by two experienced markers. To ensure the marking consistency, the absolute agreement percentage was calculated (96.4%). Then SPSS was performed for the statistical analysis.

Furthermore, a content analysis grid was completed to examine whether items in the test show adequate coverage of the CEFR descriptors and listening sub-skills illustrated in the test specification. In addition, the content analysis evaluation was performed by a panel of expert judgements for assessing whether the actual test would meet the requirements outlined in the test specification.

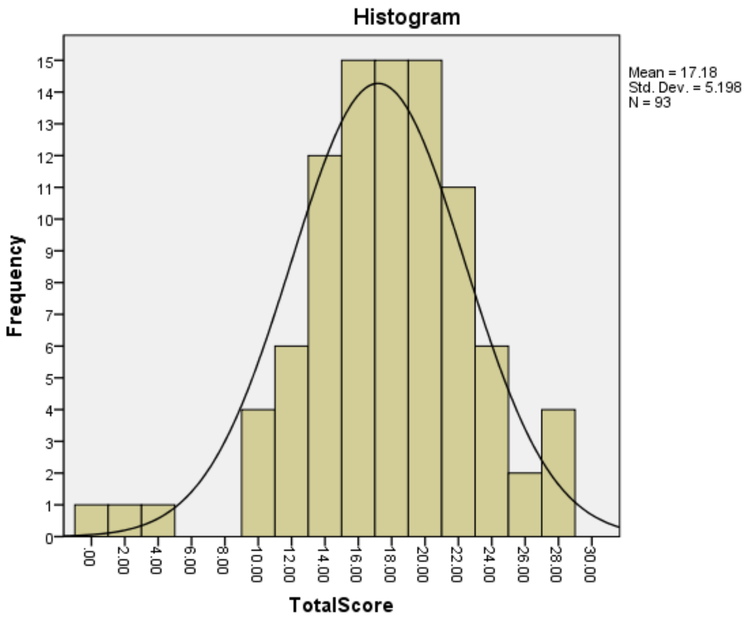
## 4. Findings and Discussion

### *4.1 Overall test results*

All test performance (scores) were entered into SPPS 25.0 for statistical analysis. The overall test results were given below (Table 1, Figure 1).

Table 1. Descriptive statistics：overall results.

| | *Mean* | *Mode* | *Median* | *Range* | *Standard Deviation* | *Skewness* |
|---|---|---|---|---|---|---|
| *Overall Results* | 17.18 | 16 | 17 | 28 | 5.20 | -.64 |



Histogram of overall scores of students

Figure 1. Histogram of overall test scores.

The histogram (Figure 1) showed overall scores of 93 test-takers on the listening test, suggesting the test paper was of average to high difficulty (mean value of 17.18). The minimum score was 0 and the maximum score was 28. None of the test-takers obtained the full mark (score 30). Several observations were further identified: First, the histogram displayed an "asymmetrical distribution" (Larson-hall, 2010, p. 65) and most of the test scores were clustered in the higher end of the distribution. Correspondingly, the distribution of test scores in Figure 1 negatively skewed which can be seen by the negative value -.64 of skewness, together with a slightly smaller mode value (16) than mean (17.18) value. However, a negatively skewed distribution is not unexpected for a criterion-referenced test with a set attainment level (Bachman, 2004; Brown & Hudson, 2002). Considering that a small group test-takers were from the same school-level, the negative skew may possibly suggest that a larger proportion of students are at the needed proficiency level. Second, we noticed that a peak (between 16 to 18) in frequency for test scores corresponding to the mode, mean and median (Figure 1). The peak score locating around the mid-point may allow higher education institutions or other stakeholders to set an appropriate cut-score for decision making. Taking the example of this study, the pass-level cut-score could be set at 17 as it close to the mode (16) and the mean value (17.18) (Bachman, 2004). Third, the value of range (28), standard deviation (5.20), together with the variance (27.02) indicated that the distribution of the overall test results is rather wide. The histogram displayed a heterogeneous distribution of scores and a relatively good spread of results within the range. With a careful inspection of data, we found that this may possibly due to the fact that no students scored between 4 and 8. Though the test set fixed attainment of B2 proficiency level, a relatively widespread of test scores can be considered acceptable as it is necessary for this test to differentiate learners whose language proficiency is above or below the B2 level in order to serve the selection purpose of deciding whether learners are qualified to enrol the university programmes.

### *4.2 Item analysis*

#### 4.2.1 Facility value

Facility value (FV) indicates whether test-takers perceive items easy or difficult. The higher the facility value demonstrates, the easier an item will be perceived by test-takers (Alderson, et al., 1995). Given that the listening test attempts to fulfil the selection purpose (the fixed attainment of B2 level), items should have a similar difficulty of the overall mean score (.57). However, it would be demanding work to ensure that all items have an FV of 0. 57 among test populations. Several 'rule of thumb' for the acceptable F.V. (e.g., Bachman, 2004; Green, 2013) was put forward in the literature. Since there is no clear-cut rule for a satisfactory FV, the widely accepted FV range from .40 to .75 was adopted in the study, which allows for one standard deviation below and above the mean score. Among all items, 16 out of 30 were situated in the accepted FV range. However, 6 items reported an FV below .40 and were considered too difficult. 8 items reported the FV above .75 and were thus deemed too easy. Table 2 showed facility values for each task. Test-takers viewed Task 2 as the most difficult section as its average FV (.40) was much smaller than the mean score (.57) and contained more items with low FV (i.e., Question 8, 10, 11 in Task 2). Besides, easy items are likely to cluster in Tasks 2 and 3 (i.e., Question 13, 16, 20, 21 in Task 2; Question 22, 24, 25, 30 in Task 3). It is interesting to found that though Task 3 contained items either too difficult (i.e., Question 18, 19) or too easy (i.e., Question 22, 24, 25, 30), the average FV for Task 3 (.58) is close to the satisfactory FV (.57).

Table 2. Facility value for each Task.

| *Test Task* | *Average Facility Value* | *Number of items* | *Items with FV <.40* | *Item number* | *Items with FV >.75* | *Item number* |
|---|---|---|---|---|---|---|
| *Task 1* | .63 | 6 | 0 | | 0 | |
| *Task 2* | .40 | 7 | 3 | 8, 10,11 | 0 | 13,16,20,21 |
| *Task 3* | .58 | 8 | 2 | 18,19 | 4 | 22,24,25,30 |
| *Task 4* | .65 | 9 | 1 | 28 | 4 | |
| *Total* | .57 | 39 | 6 | | 8 | |

#### 4.2.2 Discrimination Index

Item discrimination (D.I.) is important for norm-referenced assessments (Brown & Hudson, 2002) and allows us to discriminate efficiently between high-performing language learners and low-performing language learners. Alderson. et al (1995) clarified that an item with the D.I. above 0.4 is considered acceptable. Popham (2000) further suggested that 0.3 to 0.39 for D.I. may still be viewed as reasonably good. However, a negative D.I. is not acceptable as it hints that test-takers of lower language proficiency are more likely to answer the item correct than test-takers of higher language proficiency (Alderson, et al., 1995). Findings revealed that most items in the study have D.I. values within the range of 0.3 to 0.4 (see Table 4 for the results). To delete items with D.I below 0.4 would be counterproductive. Instead, we may focus items with a D.I below 0.3, especially those with a negative D.I (see Table 3 for the results). Tasks 3 and 4 both contained four items with low D.I values, while Tasks 1 and 2 only had few unsatisfactory discriminating items. It should note that close attention should be paid to Question 19 regarding its negative D.I value. (see Table 4 for the results).

Table 3. Point biserial discrimination. (D.I.).

| *Test Task* | *Number of items* | *Items with negative discrimination* | *Item number* | *Items with Discrimination 0-0.3* | *Item number* |
|---|---|---|---|---|---|
| *Task 1* | 6 | 0 | | 2 | 1,3 |
| *Task 2* | 7 | 0 | | 1 | 8 |
| *Task 3* | 8 | 1 | 19 | 3 | 14,16,18 |
| *Task 4* | 9 | 0 | | 3 | 26,28,29,30 |
| *Total* | 39 | 6 | | 9 | |

4.2.3 Items for review

Table 4 showed a detailed item statistic calculated with SPSS 25.0. Findings revealed that the test had several items in which the sample population found too difficult or easy. Only seven items did discriminate test-takers well between low and high listening proficiency. Five items (i.e., Question 8, 11, 16, 19, 28) require further revisions were listed in Table 5.

Table 4: Item analysis: facility value (mean), standard deviation, corrected item-total correction, Cronbach's alpha if item deleted.

| ***Item Number*** | **Facility Value** | **Standard Deviation** | **D.I/Corrected item-total correction** | **Cronbach's alpha if item deleted** |
|---|---|---|---|---|
| ***Q1*** | 0.66 | 0.48 | 0.30 | 0.81 |
| ***Q2*** | 0.65 | 0.48 | 0.47 | 0.81 |
| ***Q3*** | 0.70 | 0.46 | 0.24 | 0.82 |
| ***Q4*** | 0.72 | 0.45 | 0.44 | 0.81 |
| ***Q5*** | 0.63 | 0.48 | 0.37 | 0.81 |
| ***Q6*** | 0.47 | 0.50 | 0.30 | 0.81 |
| ***Q7*** | 0.43 | 0.50 | 0.47 | 0.81 |
| ***Q8*** | 0.16 | 0.37 | 0.27 | 0.82 |
| ***Q9*** | 0.48 | 0.50 | 0.32 | 0.81 |
| ***Q10*** | 0.31 | 0.47 | 0.42 | 0.81 |
| ***Q11*** | 0.24 | 0.43 | 0.33 | 0.81 |
| ***Q12*** | 0.68 | 0.47 | 0.44 | 0.81 |
| ***Q13*** | 0.52 | 0.50 | 0.45 | 0.81 |
| ***Q14*** | 0.86 | 0.35 | 0.28 | 0.81 |
| ***Q15*** | 0.48 | 0.50 | 0.43 | 0.81 |
| ***Q16*** | 0.92 | 0.27 | 0.21 | 0.82 |
| ***Q17*** | 0.59 | 0.49 | 0.39 | 0.81 |
| ***Q18*** | 0.10 | 0.30 | 0.22 | 0.82 |
| ***Q19*** | 0.11 | 0.31 | -0.03 | 0.82 |
| ***Q20*** | 0.76 | 0.43 | 0.30 | 0.81 |
| ***Q21*** | 0.85 | 0.36 | 0.36 | 0.81 |
| ***Q22*** | 0.81 | 0.40 | 0.38 | 0.81 |
| ***Q23*** | 0.52 | 0.50 | 0.36 | 0.81 |
| ***Q24*** | 0.91 | 0.28 | 0.44 | 0.81 |
| ***Q25*** | 0.80 | 0.41 | 0.32 | 0.81 |
| ***Q26*** | 0.68 | 0.47 | 0.24 | 0.82 |
| ***Q27*** | 0.71 | 0.46 | 0.37 | 0.81 |
| ***Q28*** | 0.11 | 0.31 | 0.02 | 0.82 |
| ***Q29*** | 0.54 | 0.50 | 0.25 | 0.82 |
| ***Q30*** | 0.80 | 0.41 | 0.28 | 0.81 |

Table 5. Items require further revision.

| ***Task of the Test*** | **Item Number** | **FV** | **D.I** |
|---|---|---|---|
| ***Task 2*** | 8 | .16 | .27 |
| ***Task 2*** | 11 | .24 | .33 |
| ***Task 3*** | 16 | .92 | .21 |
| ***Task 3*** | 19 | .11 | -.25 |
| ***Task 4*** | 28 | .11 | .02 |

4.2.4 Item review

**Task 2**

Task 2 required test-takers to listen to a monologue of social networking sites and complete the short- answer questions with a word limit of four. As previously mentioned, items in Task 2 were considered the most difficult.

With careful inspection, several problems were identified in this section. Though test specifications clarified that spelling may not cause the deduction of scores if it is believed 'can be understood', no detailed instructions or exemplars were given. In addition, markers may interpret and apply the rule with a higher degree of subjectivity. Therefore, scores may sensitive to spelling mistakes and possibly bring a construct-irrelevant problem of testing spelling in the listening test. Likewise, Question 12 also brought about the construct-irrelevant issue by simply testing students' memory rather than utilizing other listening sub-skills. Test-takers could get the answer directly from the tape as the sentence in listening materials is not paraphrased. Besides, discrepancies between the test instruction and given examples were identified (e.g., Task 2-Question 7, 13). Test developers/ item writers should offer typical examples of performance standards for each task and give detailed descriptions of instructions to inform test-takers of what was intended to measure in the test. Unfortunately, we found that the instructions of task 2 require learners to complete the task items in the word limit of four, but the word count of illustrated samples is far beyond the standard. This may cause confusion to both test-takers and markers and would potentially lead to scoring reliability issues. Question 8, by far the most difficult item perceived by test-takers, is presented below.

Task 2: Question 8 (short-answer questions)

Q8： As far as their own lives are concerned people can…

________________________________________

Key: post a personal profile/ details [about themselves]

Tape transcript:
You can also *post a personal profile* showing what you are doing now and read peoples' details.

| **Facility Value** | **Std. Deviation** | **Discrimination Index/CITC** | **Cronbach's Alpha if item Deleted** |
|---|---|---|---|
| **.16** | .37 | .27 | .82 |

**i. Unclear instructions.** The test instruction did not specify explicitly whether test takers should use the original words in the listening materials. This may cause confusion since provided example of keys answers listed both words extract from the listening texts and a partial paraphrased version. In this regard, student would have been allowed to answered question partially but did not know. According to Buck (2003), test-takers should know what constitutes a sufficient response in open-ended questions as they not only need to be told what is expected but also how much is expected.

**ii. Multiple answers.** To avoid ambiguity and ensure consistent marking reliability, constructed-response questions should constrain possible answers into a limited number and fully present in the answer key (Buck, 2001). The alternative answers for Item 8 can be: "find a list of members", or "read other people's details" etc. In fact, too many possible answers can be taken into account for this item. We, therefore, can possibly assume that some test-takers failed to write the 'correct' answer simply due to the confusion caused by too many possible answers.

Task 2: Question 11 (short-answer questions)

Q11： Besides school reunions, people consult Friends Reunited because they want…

________________________________________

Key: look up/ find their childhood sweethearts/ girl-/boyfriends

Tape transcript:
Friends United has also led many successful school reunions and people meeting up with each other after many years in particular look up their *childhood sweethearts*.

| **Facility Value** | **Std. Deviation** | **Discrimination Index/CITC** | **Cronbach's Alpha if item Deleted** |
|---|---|---|---|
| **.24** | .43 | .33 | .81 |

**i. Mismatch sign-posted words.** According to the data, item 11 had a relatively low facility value (.24) and was less capable to discriminate language learners of varied listening proficiency. With careful inspection, this can be explained by the mismatch sign-posted words between the item and listening materials. Though the sign-posted words have no necessary to be identical in the item and the audiotape, test-takers should be given fair warning words to process the listening comprehension (Hughes, 2003). For the warning word given in Questions 11, "besides" in the item certainly cannot sending the same messages to test-takers with the "in particular" showed in the listening text, which may cause confusion to test-takers as there is no 'cue' for them to predict the answer is coming.

**Task 3**

Task 3 required test-takers to listen to a monologue about the iPhone and complete the item with no more than four words. Likewise, spelling is not penalized if meaning can be understood. Though Task 3 share the same question types with Task 2, the average facility value of this task is much higher (.58 ＞.40) and close to the ideal facility value (.57), which can be explained by the clear-informed instructions (i.e., test specifications explicitly informed test-takers that answers should be extracted directly from the audio) and ways in which questions posed and organized (i.e., all items were adopted Wh- questions with an exception of Question 16). According to Alderson et al. (1995), items posed with Wh-questions in open-ended questions are generally considered easy since this format questions may inform test-takers directly what is expected. However, item writers/test developers should take careful steps when employing the Wh-questions as test-takers may possibly believe they know what they are supposed to perform in the test but in fact, they do not, which further discussed in Question 19 below.

Task 3: Question 16 (short-answer questions)

Q16：Name three special features that the iPhone will have.

a) ________b)_________c)__________

Key: iPod capability/iTunes/video playback/3.5-inch screen/ camera/ internet access/ touch screen/ email

Tape transcript:

It's packed with feature, including *iPod capability, and iTunes, a camera, email…*

| Facility Value | Std. Deviation | Discrimination Index/CITC | Cronbach's Alpha if item Deleted |
|---|---|---|---|
| **.92** | .27 | .21 | .82 |

Question 16 was an easy item as 92% of test-takers earned the score. The D.I. value of this item was relatively low of .21. The possible explanations were given below:

**i. Construct-irrelevance.** With careful inspection, we found that Question 16 can be completed simply by extracting the same word from the audiotape without utilizing any sub-skills involved in listening comprehension. That is, even if test-takers do not understand what "features" in item means, they only need to listen for the word "features" in the audio and pick out the answer. It is more like a dictation or memory test rather than a test assessing listening proficiency.

**ii. Involvement of background knowledge & Multiple answers.** According to Buck (2003), test-takers with more background knowledge about the topic of texts are likely to achieve better performance in the test. For Question 16, test-takers may still get the right answer even if they do not understand listening texts as they could guess the answer according to the background knowledge relevant to the Apple Company or mobile phones. Besides, a wide variety of possible answers can be generated on the basis of the background knowledge, thus leading to the situation that any guesses of mobile phone features are likely to be correct.

Task 3: Question 19 (short-answer questions)

Q19：Why is the man interviewed excited about the iPhone?

______________________________________________

Key: operates on OS X/ load other applications on/ nice device

Tape transcript:

Interviewee (the man): it seems very *nice devices*. I am pretty excited about it. It looks like basically I am waiting for since…disappear…I hope with the *OS. 10*. Based system you can actually *load other applications* on there…becomes more general…devices…

| Facility Value | Std. Deviation | Discrimination Index/CITC | Cronbach's Alpha if item Deleted |
|---|---|---|---|
| **.11** | .31 | -.03 | .82 |

The low facility value of .11 and the negative discrimination index of -.25 of Question 19 clearly suggested that the item attracted the wrong person. Negative D.I. indicated that low performing learners may outperform the high performing learners on an item (Alderson, et al., 1995). Possible reasons were given below:

**i. Artificial Restriction** (Bachman, 1990). Due to the word limit, the answer for Questions 3 should be presented in the note version. However, items presented in the Wh-questions required test-takers to answer the questions in a complete sentence, such as starting a response with "Because…". This may bring difficulty to proficient language learners. On the contrary, less capable language learners certainly will not take this into account, and thus explain why lower proficiency learners outperformed the high proficiency learners.

**ii. Inconsistency between the instruction and marking criteria.** According to the instruction, test-takers should fill-in the answer with sufficient information mentioned in the audiotape with no more than four words to earn the score. However, the illustrated sample showed test-takers could earn the point with partial information mentioned in the audiotape. This discrepancy between the instruction and marking criteria may lead to the fact that test-takers would have been allowed to only answer one of the reasons but did not know. This may explain why Question 19 attracted the wrong person as high-performing students may struggle with a response to the item and try to condense all information into the limited words whereas low-performing students may have been at ease writing "nice device" even it seemed not syntactically fit.

**iii. Audio speed.** We noticed that the audio speed for Question 19 is relatively fast. It is claimed that audio speed may have an impact on learners' performance on the test as learners under this occasion are likely to decode words rather than comprehend the meaning (Buck, 2001).

**Task 4**

Task 4 required test-takers to listen to a passage about the bicycle tour and judge the given statements accordingly by ticking the True, False and Not Given options. The majority of items within the task reported a relatively high F.V.(with an exception of Question 28) and a relatively low D.I, which may possibly be accounted by the fact that the true/false/not given-type questions are frequently considered as a three-option multiple-choice question with a relatively high chance for test-takers to guess the correct answer (Hughes, 2003). Burger and Doherty (1992) further asserted that true/false/not given questions should be avoided in a listening test since it only concerns what is said rather than what is not mentioned, and test-takers have no way to refer back. What's worse, one will never know what part of any scores is from random guessing (Hughes, 2003). Regarding Question 28, low F.V. (.11) and D.I. (.02) suggested the item was extremely difficult for test-taters and it failed to discriminate learners of varied listening proficiency. Alderson et al. (1995) claimed that the "Not Given" option would always be a demanding choice for test-takers in the listening test. Through a careful survey, we did notice that 70% of the sample population chose the wrong answer "True" rather than the correct answer "Not Given". Insights into the item and audio transcript, possible reasons were given below:

**i. Heavy cognitive load of "Not Given".** The "Not Given" option is challenging for test-takers on a listening test as learners have to invest more cognitive resources in performing the task (Alderson, et al, 1995). Unlike the reading test, where test-takers have the chance to refer back to the original text, listening happens in real-time and test-takers cannot hold it verbatim in memory or rewind the audio transcript (Buck, 2001). It is, therefore, very difficult for test-takers to make the decision (true/false/not given) as they are likely to experience a higher degree of cognitive load when processing the listening materials.

**ii Construct-irrelevance.** The answer for this item can be simply "infer" by test-takers in the audio transcript. To be specific, the tape mentioned that participants camped on the opposite riverbank to the slave houses in a quarter mile away, clearly sending a message to test-takers that houses can be seen from participants. In this regard, the true meaning of listening materials can be simply got by the inference from the text rather than sub-skills more centered upon the listening comprehension, thus causing the problem of construct-irrelevant issues.

Task 4: Question 28 (True/false/not given)

Q28： From their campsite on the banks of the Ohio the participants can see the houses where the slaves worked. T/ F/ NG Key: NG

Tape transcript:

The riders set up 10th outside the national underground railroad freedom centre, not less than a quarter mile away on the banks of the Ohio, where slaves once worked there.

| Facility Value | Std. Deviation | Discrimination Index/CITC | Cronbach's Alpha if item Deleted |
|---|---|---|---|
| **.11** | .31 | .02 | .82 |

*4.3 Content validity*

*Validity* specifies a "test measures exactly what it proposes to measure" (Brown & Abeywickrama, 2010, p.30), which has been widely deemed as an important indicator in evaluating a test. Considering diversified types of validity with each of no absolute definitions on the basis of synthesis literature, this study only concerned the content validity of the test due to the research purpose.

Bachman (1990) defined *content validity* as "… the test is representative of those specified in its domains" (p. 289). Hughes (2003) clarified that content validation can be achieved if items on a test and resulting scores are representative of whatever content or language abilities the test was intended to measure. To be specific, content validity requires decisions on whether there has been adequate coverage of the test content specified, which frequently applied by the comparison between the test specification and the test content (Alderson, et al., 1995).

Content validity, therefore, is highly interrelated to the construct validity as higher content validity can be obtained by defining more accurate construct validity (Hughes, 2003). Bachman (1990) also mentioned that content validation may consider as a key process to define the construct (Bachman, 1990). Though two types of validity are highly interrelated, the evaluation of this study drawn a distinction in between and directed the attention only to the issues of content. That is to say, the discussion with regard to the relationship between the construct and the communicative language ability, or the appropriateness of the construct concerning the test purpose and target language users were no taken into account.

Furthermore, considering that no consensus definition has been made about what exactly listening ability is and various kinds of taxonomies of sub-skills of listening comprehension has been proposed (Brindley & Slatyer, 2002; Field, 2008), decisions (i.e., comparison between the test content and specifications) are frequently made by a panel of experts under a thorough discussion of what sub-skills should operationalize by which test item

(Alderson, 2000). In this study, the evaluation was made by four experts. The disagreements were then compared and resolved through discussion.

**Content Analysis**

A content analysis grid (Table 6) was completed to examine whether items within the test displayed adequate coverage of the CEFR descriptors and listening sub-skills illustrated in the test specification. Furthermore, the content analysis of the test was then performed to evaluate whether the actual test met the requirements outlined in the test specification (see results in Table 7).

Table 6. Content analysis grid

| | *Topic* | *Input* | *Testing Sub-skills* | *Corresponding CEFR descriptors* | *Task type* | *Audio speed* | *Accents* |
|---|---|---|---|---|---|---|---|
| *1* | Climate Change | Interview<br>Authentic<br>Expository | Specific info/detail<br>Gist | 1,2,4,7 | multiple matching | Fast | BE accent |
| *2* | Social networking websites | Monologue<br>Expository | Specific info/detail<br>Writing notes | 1,4 | short answer questions | Medium | BE accent |
| *3* | Launch of iPhone | News report<br>(Monologue)<br>Expository<br>Authentic<br>Background noise | Specific info/detail<br>Writing notes | 1,4 | short answer questions | Fast | BE + American accents |
| *4* | Cycle Trip | Monologues<br>Expository<br>Authentic<br>Background noise | Specific info/detail<br>Gist<br>Infer meaning | 1,2,4,6,7 | True/false/not given | Medium-fast | American accents |

Table 7. Content analysis of the test: strengths, weaknesses.

| | *Strengths* | *Weaknesses* |
|---|---|---|
| *Test methods* | Diversified task types (i.e., constructed-response questions, selected-response questions) were employed for the test. | Short answer questions were not listed in the test specification. |
| *Listening Texts* | Authentic listening materials with different topics from a variety of sources (e.g., radio, documentary) were used in the test. Some with background noises.<br>Topics of texts relevant to the purpose of the test. | Listening passages were mainly expository.<br>No abstract content included in the texts.<br>Task 2 appeared to be less authentic, contradicting to what is specified in the test specification. |
| *Descriptor Coverage* | Items in Tasks 1 and 4 covered an adequate range of CEFR descriptors. | "Understanding conversation between native speakers" (descriptor 3) and "Listening to announcements and instructions" (descriptor 5) are not applied to the test items. |
| *Sub-skills Coverage* | Items of Tasks 2 and 3 included sub-skills of both listening and writing, which is important for university study. | Attention was paid to only testing sub-skills of "listening for specific information or detail information". |
| *Audio* | Both standard British and American accents were used.<br>Long monologues offer preparation for listening comprehension. | Only standard and native accents were used in the audio. In Task 4, the time length for audio is much longer than specified, which increased the difficulty level for test-takers. |
| *Instructions* | Test-takers were given both verbal and written instructions. | Instruction in Task 2 specified the answer should "no more than 4 words", however, illustrated example is far beyond the standard. |

Despite several strengths addressed in the content analysis table above, several problems were identified (see results in Table 7). First, a major problem should concern is the "construct under-representation" (Buck, 2001, p.249) since attentions were largely paid to testing the sub-skill of "listening for specific information or detail information", left other sub-skills of listening comprehension being neglected. Though it may not necessary to cover all descriptors and sub-skills within one test, a greater variety of aspects being tested would certainly increase the content validity of a test (Bachman & Palmer, 1996). Besides, we would recommend replacing Task 2 concerning its less authentic listening materials. Additionally, item type of Task 3 can be changed from the short-answer question to the sentence completion according to the test specification. Furthermore, it is of great necessity to rewrite the test specification and cover a wider variety of of sub-skills and CEFR descriptors corresponding to each Task since it may offer test developers/item writers useful information when conducting replica tests (Buck, 2001).

### *4.4 Reliability*

When comparisons were made between assessments–scores on different forms of a test; scores from the same assessment given on different conditions; scores awarded by different markers – bring us insights into reliability. According to Alderson et al. (1995), test reliability refers to "the extent to which test scores are consistent" (p. 294). Information regarding reliability of the test was also examined as presented in columns titled with "Cronbach's alpha" and "Cronbach's Alpha if Item Deleted" (see Table 4 for the results). The former column gives the overall picture of reliability value, while the latter offers insights into how well each item contributes to a reliable test. Reliability value is represented by a number ranging from 0 (worst) to 1 (ideal). According to Hughes (2003), genuine test reliability can only be found in between. Considering that test reliability is likely to be affected by factors such as the quality of the test, scoring methods (Brown & Abeywickrama, 2010), discussion on reliability in this study was elaborated from the perspectives of the test itself and scoring.

#### 4.4.1 Internal Consistency Reliability

Unclear instructions, ambiguous items, or construct-irrelevance etc. may lead to lower reliability of a test (Alderson, et al.,1995). Such problems were identified in the test and have been fully discussed in the sections of Item Analysis, Item Review and Content Validity above. Despite problems with items, findings suggested the internal consistency in the test results is relatively high. Internal consistency is often reported using a statistic called *Cronbach's alpha (α)* which is based on the mutual correlations between all items of an assessment (Alderson, et al., 1995). The results reported a Cronbach's Alpha of .82 (see Table 8 for the results), which is above the accepted level of reliability (e.g., Lado (1961) suggested .80 to .89 for tests of listening; Bachman (2004) suggested the acceptable reliability of listening should be above .80). We detected several problematic items (Question 19, 28) and if they were to be deleted, the adjusted correlation would even slightly decrease from .82 to .81. Since the deletion of these items didn't increase the internal consistency, we would not recommend deleting these items in the test. It should be pointed out that the high internal consistency in the results should be interpreted with caution. High reliability may not always come from the high-quality of a test as items with a high degree of specificity measured in the same sub-skills may also prompt the acceptable reliability value (Alderson, et al., 1995; Hughes, 2003).

Table 8. Reliability Statistics.

| ***Reliability Statistics*** | |
|---|---|
| *Cronbach's Alpha* | N of Items |
| *.818* | 30 |

#### 4.4.2 Scoring Reliability

Considering the practicality and operationalization in the large-scale testing context, objective marking was employed in the study. Task 1 (multiple matching) and Task 4 (True/ false/ not given items) of selected-response questions reported a high scoring reliability. Task 2 and Task 3 (short-answer questions with the words limit), constructed-response questions where spelling is accepted unless it does not interfere with communication, reported a lower-scoring reliability. This may be explained along two lines. The first explanation pertained to the scoring rules of flexible spelling. Flexibility in spelling may possibly lead to a larger variance in scoring as markers may interpret such rating criteria with a higher degree of subjectivity. Second, with careful inspection, we found that alternative answers for Sections 2 and 3 were not fully provided. Therefore, markers are of great uncertainty to ensure their consistent marking as they were not given sufficient information to direct themselves to make decisions and some may simply stick rigidly to the given answers.

## 5. Conclusion and Implication

This study intended to offer valuable insights into the evaluation of the CEFR-alignment project (i.e., Matura listening test) with respects of the item analysis, validity (i.e., content validity) and reliability (i.e., internal consistency reliability, scoring reliability), which in our view, may advance our understanding of how to develop a high-quality proficiency test. First, the descriptive statistics showed the overall test results. The test paper was of average to high difficulty with its peak score locating around the mid-point, which allow higher education institutions to set an appropriate cut-score for decision making. Together with widespread test scores, further indicated that the listening Matura test has the potential to efficiently measure whether learners' language proficiency has been attained the set B2 Level. At the same time, there is room for improvement in terms of the usefulness of the test, on which the current study can offer suggestions. Second, through a careful inspection of item analysis, the majority of items in the listening test performed well and were qualitatively reasonably good. However, we did identify several problematic items with possible causes related to construct-irrelevance, too many possible answers, mismatch sign-posted words, fast audio speed, testing simply background knowledge, and the heavy cognitive load required for "Not Given" option. Finally, regarding the content validity and validity issues, construct-under representation and the less authentic listening materials were then spotted by careful content analysis. Though the statistical output reported relatively high reliability, results should be interpreted with caution. All these results enable us to direct attention to factors that may contribute to the high-quality test. We hope these results will stimulate more interest in the field.

Concerning the implication of this study, our findings offer a number of specific suggestions to improve the quality of the test. The usefulness (high-quality) of a test is likely to be maximized when the following conditions are met:

1. It is recommended that further studies should have a broad selection of key elements in a test since presenting every sub-skills stated in specifications is impossible and no necessary while illustrating only specific sub-skills may bring the problem of construct under representation.
2. High reliability should be interpreted with caution as it may not always come from the quality of a test. Items with a high degree of specificity measured in the same sub-skills may also lead to a satisfactory reliability value.
3. Item writers/test developers should take a careful step in employing the True/ False/ Not given type questions in a listening test considering it is frequently deemed as a three-option multiple-choice question, which may increase the possibility for test-takers to guess the correct answer. In addition, this type question requires learners to invest more cognitive resources in performing the task since listening happens in real-time and test-takers have no chance to refer back to the text or hold it verbatim in memory, especially for the "not given" information.
4. Test-takers should be provided with clear-articulated instructions and informed explicitly what constitutes a sufficient response as they need to be told what is expected and how much is expected.
5. Fair warning words should be given for test-takers to predict the answer is coming.
6. Too many possible answers should be avoided in the constructed-response questions. Indeed, it is necessary to re-emphasize the importance of fully presenting typical examples of performance standards for each item with detailed descriptions of instructions to avoid ambiguity and ensure consistent marking reliability.

## Funding

This study was supported by the MOE Project of Humanities and Social Sciences (Project No. 20YJA740031) and MOE Project 2019 Vocational Education Teaching and Research (Project No. FLE0001).

## References

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

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

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

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Brindley, G. & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. Language Testing, *19* (4), 369-394.

Brown, J. D. & Hudson, T. (2002). *Criterion-referenced Language Testing.* Cambridge: Cambridge University Press.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, *8*(1), 67-91.

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

Buck, G. (2003). *Expert estimates of test item characteristics.* Paper presented at the Language Testing Research Colloquium, Princeton.

Burger, S., & Doherty, J. (1992). Testing receptive skills within a comprehension-based approach. *Comprehension-based second language teaching*. Ottawa: University of Ottawa Press.

Brown, H., & Abeywickrama, Priyanvada. (2010). *Language assessment: Principles and classroom practices.* White Plains, N.Y.: Pearson Education.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder Education.

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

Green, R. (2013). *Statistical Analyses for Language Testers.* Baskingstoke: Palgrave and Macmillan.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Khalifa, H., & Weir, C.J. (2009). *Examining Reading: Research and Practice in assessing second language reading.* Studies in Language Testing 29. Cambridge: Cambridge University Press.

Lado, R. (1961). *Language testing: The Construction and Use of Foreign Language Tests.* London: Longman.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS* (Second language acquisition research). London: Routledge.

Popham, W. J. (2000). The mismeasurement of educational quality. *School Administrator*, *57*(12), 12-15.