Reliability and content validity of an English as a Foreign Language (EFL) grade-level test for Turkish primary grade students

Ipek Ozer, Shawn M. Fitzgerald, Ebed Sulbaran, Diana Garvey

School of Foundations, Leadership, and Administration, Kent State University, Kent, 44240, United States

Abstract

The present study examined the reliability and content validity of an English as a Foreign Language (EFL) grade-level test for Turkish 3rd grade primary students. While the Content Validity Index (CVI) was found to be low (.52), the reliability coefficients varied between .77 and .91 for the sub-subsections, indicating high reliability of the items. The item difficulty and discrimination indices were also consistent with the CVI as 50% of the items were identified as being in need of revision.

© 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of Dr Zafer Bekirogullari.

Keywords: Test Evaluation; Validity; Reliability; Item Analysis.

1. Introduction

English as a Second Language (ESL) is explained by Gunderson (2009) as the process of learning a language in which the official language in the community is English. English as a Foreign Language (EFL) differs from ESL due to the non-native use of a language which gives students a limited experience to learn the language. As an assessment method, standardized tests play an important role with EFL curriculum evaluation as well as student evaluation. According to Brown and Abeywickrama (2010), a good standardized test is a product of empirical research and development that goes beyond simply agreeing on particular standards or benchmarks. This type of test also includes systemic procedures for administration and scoring. Most schools around the world use standardized tests to assess students at each level of their educational experience. In some cases, standardized
tests are developed and administered by particular entities, such as a Board or Ministry of Education, while other tests are administered by departments within schools (Akiyama, 2004).

Numerous schools have utilized standardized EFL tests in order to assess student ability and progress in this domain and several empirical research articles have provided evidence of these practices. Mbali and Douglas (2012) described the process and procedures used to measure proficiency in the oral English skills of 6th graders in one South African school. In this context, a structured assessment was created and framed using the Common European Framework of Reference (CEFR) for assessing language skills. The main objective of the study was to evaluate and consider how to improve speaking skills in their schools (Mbali & Douglas, 2012). Likewise, the Bi-National center in Uruguay developed a norm-referenced test of American English for students in the last year of elementary school (Freurquin, 2003). While test construction of this test did not necessarily follow a prescribed standard, such as the one followed by Mbali and Douglas (2012), as item development, analysis, as well as content validity, was conducted based on local school requirements and curricula, the goals were similar---to evaluate speaking skills in their school.

In order to accurately assess the EFL development of students within one school system in Turkey, school administrators and a research team developed a pilot test of one standardized assessment designed for this purpose. The main purpose of this study was to evaluate the quality of an EFL test for 3rd graders and offer empirical evidence related to its use as a standardized test.

2. Method

2.1. Participants

The participants of this study were 3rd grade-level students from a private primary school in Turkey (N = 71). Thirty-eight of the students were female (53.5%) while 33 were males (46.5%). During the 2011-2012 academic year, test-takers were taught the same level of English with the curriculum provided by the Turkish government. The test was administered at the end of the academic year, in April 2012.

2.2. Measures

The aim of the EFL test was to assess the proficiency levels of students in both written and oral English. The written section of the EFL test included four major sections: (I) Listening, comprehension, and vocabulary (i.e., 4 sub-sections, 25 items); (II) Reading, comprehension and vocabulary (i.e., 3 sub-sections, 15 items); (III) Reading comprehension and use of English (3 sub-sections, 20 items); and (IV) Writing (i.e., a paragraph with at least 10 sentences). Table 1 contains a description of the item formats and total number of items included on the written section of the test. The speaking or oral part of the EFL test had 6 sub-sections: personal questions, general questions, pair-work about the picture, pronunciation, general frequency, and general accuracy. The total score for written part is equal to seventy (i.e., each test item worth 1 point; writing section worth 10 points), whereas the oral part is 30 points. In addition, only open-ended items were continuously graded (i.e., grading scores include .00, .25, .50, .75, and 1.00), and other three item formats were dichotomously graded (i.e., 0 or 1).

Table 1. Test item types for the written section of the test.

|       | True-False | Multiple Choice | Matching | Open Ended |
|-------|------------|-----------------|----------|------------|
| Part I (n = 25) | 10         | -               | 15       | -          |
| Part II (n = 15) | -          | 5               | 10       | -          |
| Part III (n = 20) | -          | -               | -        | 20         |
| Part IV    | -          | -               | -        | 10         |
2.3. Procedure

The purpose of this study was to evaluate the quality of an EFL test for 3rd graders and to provide empirical evidence related to its use as a standardized test. The quality of this EFL test was examined using commonly applied techniques to establish validity (i.e., the degree to which evidence and theory support the interpretation entailed by proposed use of tests) and reliability (i.e., the desired consistency of test scores). Therefore, the reliability, content analysis, and item analyses of the existing EFL test will be discussed.

The aim of the analysis was to estimate how consistently the examinees performed across the test items, specifically how they performed across the sub-tests. The EFL test was administered only once to a group of examinees, therefore the procedure designed to estimate reliability is called internal consistency (Crocker & Algina, 2006). Reliability, the internal consistency of sub-scale items, was assessed by calculating the coefficient alpha or Kuder-Richardson (KR) 20 statistics depending on the measurement properties of the items. The coefficient alpha was reported for continuously scored, and the KR 20 was used for dichotomously scored items.

According to Crocker and Algina (2006), two major questions have to be answered when discussing the validity of an instrument: (1) Is the scale measuring the construct intending to be measured, and (2) Is there sufficient evidence to support the intended uses or interpretations of the test. The initial stage of establishing the validity of any instrument begins with an assessment of content validity (Bachman, 1990; Bachman, Davidson & Milanovic, 1996; Bachman & Palmer, 1996; Malmgreen, Graham, Shortridge-Baggett, Courtney, & Walsh, 2009). Bachman and colleagues (1996) define content analysis as “application of a model of test design to a particular measurement instrument, using judgements of trained analysts” (p. 125). In the existing study, content validity was assessed by a group of experts from a public university in the United States (i.e., two experts specializing in EFL and two experts specializing in evaluation and measurement). Waltz and colleagues (2010) defined the Content Validity Index (CVI) as the extent of agreement between the experts and to compute the CVI, the specialists independently rated the relevance of each item to the objectives using a 4-point rating scale (i.e., (1) not relevant; (2) somewhat relevant; (3) quite relevant; and (4) very relevant; Waltz et al., 2010, p.165).

In addition to reliability and validity analysis, when evaluating a test, besides analyzing the agreement level of experts in content assessments of EFL test items, and it is also essential to evaluate the difficulty and discrimination of the test items. Item difficulty ($p_i$) is defined as the proportion of examinees who answer the item correctly, and the index of item discrimination, the degree to which the items discriminate among examinees, is obtained by using the entire upper and lower 50% of the examinee group, based on the test scores.

3. Results

Data presented in Table 2 suggest that part I and part II of the test were fairly easy for students compared to overall averages for parts III and IV of the test. It is also worth noting that there appeared to be slightly less variability in test scores for the earlier sections of the test relative to the later sections. All of the parts were negatively skewed (i.e., the scores tailed off at the lower end of the scale). The overall means for the total written and oral scores were low. A variety of assessments were applied starting with reliability analysis, the rating and justification process (i.e., content validity), elicitation of expert views and finally statistical item analysis ($N = 71$).
Table 2. Test statistics (N = 71)

|               | M   | SD  | Max. | Min. |
|---------------|-----|-----|------|------|
| Part I        | 18.80 | 4.89 | 25.00 | 6.00 |
| Part II       | 11.54 | 2.76 | 15.00 | 3.00 |
| Part III      | 12.41 | 4.37 | 20.00 | 0.00 |
| Part IV       | 5.87  | 2.93 | 10.00 | 0.00 |
| Written Test Score | 48.62 | 13.40 | 68.50 | 15.25 |
| Oral Test Score | 23.51 | 6.27 | 30.00 | 1.80 |
| Total Score   | 72.13 | 18.78 | 98.50 | 17.65 |

3.1. Reliability

The EFL test results seemed to suggest a high degree of internal consistency for subsection items. The KR-20 reliability indices (i.e., dichotomously scored items) for parts I and II were .872 and .773 respectively. The coefficient alpha for part III was .906.

The intercorrelations of the four parts and the correlations between the parts and the total test score are presented in Table 3. The four parts of the test were intended to test different aspects of language, and strong correlations were not expected (Wall, Clapham, & Alderson, 1994). The correlations between Part I, II, III and IV ranged from .63 to .82, suggesting moderate to high correlations between different parts of the test. However, even though the analyses suggests these test subsections share some common variance, as would be expected since they are all testing English language, there is enough unshared variance to suggest that different skills are likely being tapped. The highest correlation was found between the parts I and III—listening and use of English (.817).

Table 3. Pearson product-moment correlation coefficients (N = 71)

|          | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|----------|-----|-----|-----|-----|-----|-----|-----|
| (1) Part I | 1   |     |     |     |     |     |     |
| (2) Part II | .665** | 1   |     |     |     |     |     |
| (3) Part III | .817** | .636** | 1   |     |     |     |     |
| (4) Part IV | .797** | .629** | .751** | 1   |     |     |     |
| (5) Written Score | .942** | .793** | .919** | .884** | 1   |     |     |
| (6) Oral Score | .764** | .700** | .678** | .703** | .798** | 1   |     |
| (7) Total Score | .927** | .800** | .882** | .865** | .979** | .903** | 1   |

Note. ** p < .001

3.2. Content validity

To establish content validity of the EFL test, four experts were asked to evaluate the instrument. Based on the feedback from the expert United States panel, revision of the test items was indicated. Two faculty members and two graduate students reviewed the EFL exam for the curricular validity, and discussed the extent to which items
are relevant to the objectives of the curriculum (Chapelle, 1999). The four experts also used the following
guidelines to review the EFL Exam for content validity criteria: (1) Clarity in wording, (2) Relevance of the
items, (3) Use of Standard English, (4) Absence of biased words and phrases, (5) Formatting of items, and (6)
Clarity of the instructions (Fowler, 2002). Using the process outlined by Waltz and colleagues (2010), the
Content Validity Index (CVI), which quantifies the extent of agreement between experts, was calculated. The
CVI for the test as a whole was reported to be .52, which means almost half of the items were evaluated as (1) not
relevant or (2) somewhat relevant. According to Polit and Beck (2006), a CVI of .80 or better indicates good
content validity. Therefore, the CVI was reported low, and revision of the items (e.g., re-wording of items;
changing pictures; edits) was recommended.

Feedback from the focus group discussions resulted in major changes in items related to the pictures (Items
#1, #3, #4, #5 in Section 1; picture used in Section 3; and picture used in Section 4), complete re-wording of two
items (Items #19, #20), and changing the multiple choice options for true-false items. Examples of changes in
wording or pictures cannot be provided to protect the copyright of the EFL test which remains under further
development.

3.3. Item analysis

The means of the 60 test items (i.e., writing part is not included) were interpreted as a measure of item
difficulty (i.e., the percentage of students who answered an item correctly). Item difficulty can range from 0.0
(i.e., none of the subjects answered the item correctly) to 1.0 (i.e., all of the subjects answered the item correctly;
Crocker & Algina, 2006). Since the EFL test was designed as a placement test, a p-value for difficulty range from
.3 to .9 was considered to be desirable (Kim & Shin, 2006). The results of the EFL test analysis showed that the
item difficulty scores ranges from 0.08 to 1.00. Based on these values, 15 test items were determined to be too
easy (p > .9); whereas nine items were observed to be too difficult (i.e., p-value lower than .3). Thus, 25% of the
items were found to be easy for students.

Item discrimination is the degree to which the items discriminate among examinees. The higher the
discrimination index, the better the item because such a value indicates that the item discriminates in favor of the
upper group, which would be expected to get more items correct. An item that everyone gets correct or that
everyone gets incorrect (i.e., the item difficulty will be close to 1) will have a discrimination index equal to zero
or close to zero. The decision criteria using item discrimination values is listed as follows: (1) If \( D \geq .40 \), item is
functioning satisfactorily, (2) If \( .30 \leq D \leq .39 \), little or no revision is required, (3) If \( .20 \leq D \leq .29 \), item is
marginal and needs revision, and (4) If \( D \leq .19 \), item should be eliminated or completely revised (Crocker &
Algina, 2006). Overall, according to the discrimination indices, 17 items (\( D \) lower than .19) were found to be
revision was necessary. Most of these items also had high item difficulty scores (i.e., easy items). Only 12 items
in the written part were found to be satisfactory.

4. Discussion

A variety of analysis was applied to evaluate the quality of the EFL exam. The reliability, content validity and
statistical item analysis were discussed. The EFL test results showed high reliability coefficients, however the test
found invalid with regards to the content validity. The content validity was resulted in recommending the revision
of the EFL test items. Not surprisingly, the CVI was also in parallel with the item difficulty and discrimination
analysis results. Thus, almost half of the test items were identified as being in need of revision. In addition to
revision of test items, the instructions for some sections were found to be necessary to edit. Since, the EFL test is
specific to 3rd graders, the directions for the sections and sample items were advised to be written in an easier
format.
Given that many schools develop and use their own standardized tests, ones that closely match the curriculum objectives set locally or internally, it is imperative that administrators and staff develop these test in a technically appropriate manner to ensure both valid and reliable results and the fair use of test results. The initial steps in the systematic, quantitative approach for validating an English as a Foreign Language (EFL) grade-level test for Turkish 3rd grade primary students are presented in this study. Further analysis and study to refine the test and assess other dimensions of validity and reliability for different grade-levels is in progress.

References

Akiyama, T. (2004). Introducing EFL speaking tests into a Japanese senior high school entrance examination. University of Melbourne, Department of Linguistics and Applied Linguistics.

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

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. Language Testing, 13, 125-150.

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

Brown, D., & Abeywickrama, P. (2010) Language Assessment Principles and Classroom Practices, 2nd Ed Pearson Education U.S.

Chapelle, C. A. (1999). Validity in language assessment. Annual review of applied linguistics, 19, 254-272.

Crocker, L., & Algina, J. (2006). Introduction to classical and modern test theory. Thomson Learning.

Fowler, F. J. Jr., (2002). Survey research methods (3rd ed.). Thousand Oaks, CA: Sage Publications.

Freurquin, F. (2003). Development of A standardized test for young EFL learners. Spaan Fellow Working Papers in Second or Foreign Language Assessment, 1, 1-24.

Gunderson, L. (2009). ESL (ELL) Literacy Instruction: A Guidebook to Theory and Practice, 2nd ed. Taylor & Francis US.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. Research in nursing & health, 29, 489-497.

Shin, S. K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. Language Testing, 22, 31-57.

Kim, H. J., & Shin, H. W. (2006). A Reading and Writing Placement Test: Design, Evaluation, and Analysis. Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 6, 1-36.

Kunnan, A. J. (2000). Fairness and justice for all. Fairness and validation in language assessment, 9, 1-14.

Malmgreen, C., Graham, P. L., Shortridge-Baggett, L. M., Courtney, M., & Walsh, A. (2009). Establishing Content Validity of a Survey Research Instrument: The Older Patients in Acute Care Survey—United States. Journal for Nurses in Staff Development, 25, E14-E18.

Mbali, C., & Douglas, J. (2012). Implementing a speaking test at Grade 6 (primary school) in South Africa. Southern African Linguistics & Applied Language Studies, 30, 525-537.

O’Malley, J., & Valdez, L. (1996) Authentic Assessment of English language Learners: Practical approaches for teachers. Reading, MA: Addison Wesley.

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. Language Testing, 11, 321-344.

Waltz, C., Strickland, O. L., & Lenz, E. (Eds.). (2010). Measurement in nursing and health research. Springer Publishing Company.