Journal of Applied Linguistics and Language Research
Volume 5, Issue 2, 2018, pp. 275-285
Available online at www.jallr.com
ISSN: 2376-760X

Evaluating the Construct Validity of Final Test of Grade Three of Junior High School in Iran

Mojtaba Eghlidi *
Ph.D. candidate, Islamic Azad University, Shahreza Branch, Shahreza, Iran

Omid Tabatabaei
Assistant Professor, Islamic Azad University, Najaf Abad Branch, Najaf Abad, Iran

Abstract
Construct validity "is a fundamental requirement for the effective operation of any assessment system that everybody involved in interpreting assessment results shares at least a basic understanding of the construct or constructs involved" (Green, 2014, p. 81). The aim of the present study was to evaluate the construct validity of final test of grade three of junior high school used in Fars Province, Iran. To achieve this objective 35 EFL teachers teaching English at public schools participated. Also, 50 students' final test sheets in 2016-2017 school year were selected. One instrument was a 17-items researcher-made checklist to evaluate the construct validity by the teachers. Data obtained from this instrument were analyzed using descriptive statistics i.e. Mean and Standard Deviation. The results revealed that participants evaluated 53% of the construct of the test as valid. It means that they evaluated the construct validity of the test as fair, but not strong. The second instrument was the students' final test sheets. The scores were analyzed using item facility (If) and item discrimination (ID) to find the acceptable items. 36.66% of items were acceptable in the sample of the study. Another analysis refers to correlation coefficient among total score and test parts and, also, among test parts themselves. The correlation among test score and test parts was very strong (r>0.8), and it was strong among test parts (r>0.6). Because the test framework is recommended by Curriculum and Textbooks Development Office of Iranian Ministry of Education, the results of this study can be used to enhance construct validity of national-wide tests.

Keywords: construct validity, test, validation, assessment

INTRODUCTION

In the process of each test or assessment material preparation the test-makers in all levels of education have to had a clear construct of what they want to test or assess. Nowadays, the notion of validity and validation is taken into consideration by educationalists (Lynch, 2003). Some educationalists believe that the reliability doesn't work to make judgments about tests and tests scores. They believe that test users such as curriculum developers and policy makers need to more deep interpretations about

* **Correspondence:** Mojtaba Eghlidi, Email: mojtaba.eghlidi@ymail.com
© 2018 Journal of Applied Linguistics and Language Research

the tests and their scores. They claim that reliability is not able to prepare much information to make an appropriate decision making (Lynch, 2003). However, validation is a suitable approach of provision of the suitable evidence to make judgments and decisions. Also, this approach is a level higher than validity. Messick (1989, p. 13) describes validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". The interpretation and use we make of test performance may not be equally valid for all abilities and in all contexts, so it is misleading to speak simply of the validity of test scores. Thus, in test validation, we are not examining the validity of the test content or of even the test scores themselves, but rather the validity of the way we interpret or use the information gathered through the testing procedure. As Richards and Schmidt (2010) define, validation is "the process of accumulating evidence to support the inferences drawn from the scores of a test, using a combination of methods" (p. 622). The keyword here is "accumulating evidence". It means that evidence is needed to interpret the test scores, not the pure test scores to which reliability relies on. Also, this definition proposes a clear-cut path for researchers who want to validate tests. It says that we need evidence to support inferences meaning that, foremost, it is essential to interpret and inference, then support them applying evidence to have a deep interpretation. Therefore, Messick asserts that "it is fundamental that score validation is an empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use" (1995, p. 742).

In the process of validation, we have to consider that the scores are a function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment. In particular, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971).

Messick (1995) considers two types of threat to invalidity that they are operative: construct underrepresentation, and construct-irrelevant variance. The former relates to narrowness of the assessment material which is not able to include important dimensions or facets of the construct. The latter is referred to broadness of the assessment material in which responses are influenced in a manner irrelevant to the interpreted construct. So we need to have a test valid in its construction to remove the effects of invalidity. Therefore, construct validity is a crucial aspect of each test that should be validate. In construct validation the test score is not equated with the construct it attempts to tap, nor is it considered to define the construct, as in strict operationism (Cronbach & Meehl, 1955). So construct validity concerns the extent to which performance on tests is consistent with predictions that we make on the basis of a theory of abilities, or constructs. In a nutshell, a measure estimates how much of something an individual displays or possesses, then the basic question of construct validation is: What is the nature of that something?

In attempting to answer this question, we must identify and define what the "something' is that we want to measure, and when we define what this is, we are, in effect, defining a construct. Carroll (1987) asserts that a construct of 'mental ability' is defined in terms of a particular set of mental tasks that an individual is required to perform on a given test.

A few years ago, textbooks of Iranian public system have been changed based upon Communicative Language Teaching Approach (CLT) principles. Ministry of Education of Iran claims that the textbooks can lead the students to get ready for international communication. The books prepared for Junior High Schools are in a series called "Prospect" with the sub-title of English for Schools. They are used for grades seven to nine (1st to 3rd grade of the Junior High School). For each grade there are just two hours a week (90 minutes). The Ministry of Education's educationalists claim that based upon the rules and principles of CLT on which the books were designed the students should be trained in all language skills in an integrated manner. According to this point of view, the assessment method is communicative, and it integrates all language skills and subskills in a communicative manner.

Based upon the notion of construct validation the following question generated by the researchers:

- How well does the final test of grade three of Junior High School validate in its construction?

LITERATURE REVIEW

Cronbach and Meehl (1955) assumed that the construct would be implicitly defined by the theory, and therefore, that measures of the construct could be validated by validating the theory, with the postulated relationship between the assessment scores and the construct considered part of the theory. They presented construct validity as an alternative to the criterion and content models to be used, whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined. Constructs can be viewed as definitions of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behavior.

Messick (1995) introduced construct validity as a comprehensive approach to validation. He asserts that:

> "In essence, construct validity comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables. In its simplest terms, construct validity is the evidential basis for score interpretation. (1995, p. 743).

Historically, primary consideration in construct validation has been taken into internal and external test structures. It means that the focus of studies in this scope has been on the evaluation of theoretically expected patterns of relationships among item scores or between test scores and other measures (Messick, 1995).

In examining the relationship among different observations of language performance, not all of which will be tests, the test developer involved in the process of construct validation is likely to collect several types of empirical evidence. These may include any or all of the following:

1. the examination of patterns of correlations among item scores and test scores, and between characteristics of items and tests and scores on items and tests;
2. analyzing and modeling of the processes underlying test performance; e.g. studying test taking strategies through verbal protocols
3. studies of group differences;
4. studies of change over time, or
5. investigation of the effects of experimental treatment (Messick, 1989).

In the following paragraphs some empirical studies carried out in the scope of construct validation of different tests well be depicted.

Moore and Morton (2012) worked on construct validity of IELTS Academic Reading Test. Investigation was made of the suitability of items on the test in relation to the reading and general literacy requirements of university study. This was researched in two ways – through a survey of reading tasks in the two domains, and through interviews with academic staff from a range of disciplines. Tasks in the two domains were analyzed using a taxonomic framework, adapted from Weir and Urquhart (1998), with a focus on two dimensions of difference: level of engagement, referring to the level of text with which a reader needs to engage to respond to a task (local vs global); type of engagement referring to the way (or ways) a reader needs to engage with texts on the task (literal vs interpretative).

Heydari et.al's (2014) investigation of the construct validity of a nationwide large-scale English proficiency test called TOLIMO showed that the test demonstrated construct validity in examinee's ability level in structure and writing.

Fallahian Sichani and Tabatabaei (2015) evaluated Construct Validity of MSRT Reading Comprehension Module in Iranian Context. The findings showed that explanatory factor analysis did not reveal similar findings as those in the judgmental phase of the study. The items in the MSRT reading comprehension tests didn't confirm that MSRT reading parts assess the reading skills in the Iranian context. This study highlighted the importance of designing and using more reliable and valid tests, for researchers and designers, and those who are interested in the use of such tests.

Zoghi, Rostami and Gholami (2016) evaluated the construct validity of Iranian National Test of English at High Schools. The compared the construct validity of two test types administered in 2000 and 2014. Results revealed that there was a significant difference between 2000 and 2014 versions of final national tests of English for grade three high schools' students in Iran in terms of test items, item facility, item difficulty, and item discrimination. The second finding showed that the total score of the 2014 test correlated with every subtest. Similarly, different subtests of the 2000 version correlated with each other. In addition, EFL teachers believed that the final national test of English for grade three high school do not have construct validity.

METHOD

Participants

In this study, 35 English teachers who are teaching at public schools in Eghlid, Fars, Iran participated.

Instruments and Data Collection Procedure

Actually the first and main evidence for the present study was the final exam of reading and writing skills that administered for test-takers of grade three of the Secondary Program of Iran's Ministry of Education. This exam administered in Fars Province in Khordad of 1396 (June 2017) (1395-96 school year (2016-2017)). The researchers used the final test scores of the final exam of grade nine of Junior High School in Eghlid. In order to work on the tests 50 test sheet were selected through selecting five schools. It means that five schools were selected out of 23 Junior High Schools in Eghlid through random sampling. From each school 10 test sheet were selected randomly.

The second instrument as the evidential basis for the research was a 17-items researcher-made questionnaire to see how the teachers evaluate the construct validity of the test using a Five-Likert Scale. The Cronbach Alpha equals 0.73. It means that the questionnaire has an acceptable reliability. In order to hand out the questionnaire the researchers invited the teachers to a workshop and after explaining the objectives of the study and the necessity of evaluating the final exam of grade nine the questionnaires were distributed in person. After 15 minutes the questionnaires were collected.

Data Analysis Method

This study is a mixed-method research. The collected data was entered in SPSS software and prescriptive analyses such as mean (M), percentage and standard deviation (SD).

Provincial Exam of Grade Three

For a better understanding of the claims of the research it is essential to introduce the rubric of Provincial Exam of Grade Nine. First of all, it should be said that the main framework of the testing is one that Curriculum and Textbooks Development Office recommends to the Education offices around the country. The reason is that in Iran education is centralized and it is limited to top-down method. It means that the textbook, materials, and assessment methods and format of exams are dictated to the Education offices by the Ministry of Education. It can be said that if the present study will be done in a broader range, around the country, the result will be the same, because the textbook, class hours, assessment framework, and scoring method is the same in all cities and villages in the country.

The exam is administered for measuring reading and writing skills with the sub-skills of grammar, vocabulary, includes 7 parts i.e. A to G. It contains 30 questions and is administered in 80 minutes. Part A includes 4 questions with the method of matching. The test-takers should match the sentences with the pictures. This part is just for vocabulary, and actually is not a reading comprehension. Part B includes a dialogue in which some parts were missed. But the missed parts are in parentheses with the

method of multiple-choice items. This part measures grammar sub-skill. In the third part, part C, we see 4 multiple-choice item for vocabulary and grammar. In part D 4 essay type questions for grammar exist, two for yes/no and two foe Wh-questions, respectively, with picture prompts. Part E is a four-item simple completion task for vocabulary items. Actually it is a matching task with the given words that the test-takers have to fill in the blanks with them. Part F is an error recognition task for grammar. There are four errors in a short paragraph that the test-takers are asked to find and write the correct form. And the last part, part G, is a reading comprehension item. This item compromises 3 tasks. Two T/F tasks, two fill in the blanks, and two essay type questions are 6 questions we see in reading comprehension. The last point to consider is that the instructions are in Farsi. It seems that the writing is limited to write the answers to the questions, not writing as process or product.

Table 1. Test rubrics

Part	Skill/Sub-skill	Testing Method	N. of Questions
A	vocabulary	matching	4
B	grammar	MCH	4
C	voc. & grammar	MCH	4
D	grammar	essay	4
E	vocabulary	matching	4
F	grammar	error recognition	4
G	reading	T/F, completion, essay	6

Table 2. Number and percentage of test skill/sub-skill

Skill/Sub-skill	Total N. of Questions	Percentage
vocabulary	8	26.66%
grammar	12	40%
voc. & grammar	4	13.33%
reading	6	20%

RESULTS AND DISCUSSION

Data gathered were analyzed by SPSS software and for the first part of the study i.e. analysis of the students' scores the following results in the case of item facility (IF) and item discrimination (ID) were obtained. Scientifically, items should be rejected if the IF is <.33 or >.67. It means that if <.33 of the subjects answered the item the item is too difficult. Also, if >.67 of the test-takers answered the item the item is too easy. It means these items are not able to measure the appropriate knowledge of the subjects and are not acceptable. To calculated the ID, first a high group and low group must be established. Brown (1996) recommends that ID should be between %25-35% of the total group. For this study, 30 % (n=25) was used. The acceptable items of each part are depicted in the following table.

Table 3. Acceptable items of the test

Part	Skill/Sub-skill	N. of Questions	Acceptable Items
A	vocabulary	4	1
B	grammar	4	2
C	voc. & grammar	4	2
D	grammar	4	1
E	vocabulary	4	1
F	grammar	4	1
G	reading	6	3
	Total	30	11

As Table 3. depicted 11 items out of 30 items were acceptable. It means 36.66 of items were acceptable in the sample of the present study. Lord (1952) suggests that if >30% test items is acceptable the test will be fair, buy not perfectly. So the test acceptability is fair. However, this index is near the low-level of fairness boundary. It seems that the test should be revised to get a higher level of acceptability. Alderson, Clapham and Wall (2000) suggest that a good way of evaluating the construct validity of a test is to correlate its various test components with each other. The following table will demonstrate the correlation between the scores and he different parts of the test and the parts of the test with each other based upon the analysis of the students' scores.

Table 4. Correlation among test parts and total score

	Parts	A	B	C	D	E	F	G	Total Score
A	Pearson Correlation	1	.682**	.862**	.832**	.824**	.804**	.689**	.897**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
B	Pearson Correlation	.682**	1	.660**	.739**	.683**	.780**	.640**	.828**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
C	Pearson Correlation	.862**	.660**	1	.819**	.799**	.770**	.716**	.881**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
D	Pearson Correlation	.832**	.739**	.819**	1	.727**	.823**	.720**	.924**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
E	Pearson Correlation	.824**	.683**	.799**	.727**	1	.799**	.697**	.861**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
F	Pearson Correlation	.804**	.780**	.770**	.823**	.799**	1	.811**	.946**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
G	Pearson Correlation	.689**	.640**	.716**	.720**	.697**	.811**	1	.863**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
Total Score	Pearson Correlation	.897**	.828**	.881**	.924**	.861**	.946**	.863**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	

** Correlation is significant at the 0.01 level (2-tailed).

Table 4. demonstrated the correlation coefficient between the test items and total score and items among each other. Miller and Miller (2012) classify the correlation coefficients to interpret them. They suggest that if a coefficient is between 0.6 and 0.8 the correlation is strong, and above 0.8 it is very strong. As revealed from the results of Table 4. All the correlation coefficients of the parts of the test are >.6. Because the Sig two-tailed equals 0.000 we can say that the correlation is significant. It means that there was a strong relationship among the test parts. The strongest correlation is between parts A and C (r=0.862, Sig 0.000). Both of them are vocabulary items. The weakest correlation is between parts B and G (r=0.640, Sig 0.000). Part B is to measure grammar knowledge and part G is reading comprehension. The result is reasonable because the nature of these two items are different. On the other hand, correlation coefficient among the total score and other parts is very strong because all the coefficients are above 0.8. Total score had the strongest correlation with part F (r=.0.946). Table 5. shows the results of the teachers' evaluation of the construct validity of the test.

Table 5. Teachers' evaluation of the construct validity of the test

N	Items	M	SD
1	The content of the test is similar to what the students are familiar with in their student and work books.	3.81	.740
2	The test is suitable for all levels of students' knowledge.	3.36	.850
3	It tests the content of the book properly.	2.67	.954
4	It measures vocabulary knowledge properly.	3.93	.712
5	It measures grammar knowledge properly.	2.81	.671
6	It measures the writing skill properly.	1.60	.701
7	It triggers different writing strategies.	1.29	.457
8	It measures the reading skill properly.	2.98	.715
9	It triggers different reading strategies.	2.45	.504
10	It's a real communicative test.	1.62	.623
11	The test fits the CLT principles.	2.00	.625
12	The aims of the test correspond closely with the aims of the teaching program.	2.45	.504
13	It fits the principles of integrated tests.	3.21	.645
14	The test uses different methods to elicit students' knowledge.	3.88	.670
15	The face of the test helps students to take it better.	3.02	.563
16	The questions related to skills and sub-skills distributed equally. (equal numbers)	1.29	.457
17	The parts have relationships to each other.	2.81	.634
	Total	2.65	0.865

As revealed from the above table items 4 (It measures vocabulary knowledge properly; M=3.93), 14 (The test uses different methods to elicit students' knowledge; M=3.88), 1 (The content of the test is similar to what the students are familiar with in their student and work books; M=3.81), and 2 (The test is suitable for all levels of students' knowledge; M=3.36) are four items that the respondents evaluated them as the most valid items related to test construct. They reported that items 7 (It triggers different writing strategies; 1.29), 16 (The questions related to skills and sub-skills distributed equally; M=1.29), 6 (It measures the writing skill properly; M=1.60), and 10 (It's a real communicative test; M=1.62) were the least items that the test designers considered in

making the test. The overall mean of the checklist is (M=2.65) with the standard deviation of (SD=0.865).

CONCLUSION

Ebel and Frisbie (1991) assert that construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean. There are on-going challenges for designers of tests with a communicative orientation both in terms of mapping-out the precise nature of the constructs of communicative competence which underlie test performance, and in generating validity evidence that these constructs are being operationalized through a particular set of test tasks.

The aim of the present study was to evaluate the construct validity of the Final test of grade three of junior high school in Fars Province, Iran. As the evidential bases final test scores of students and a researcher-made questionnaire to evaluate the construct validity were applied. The results of analysis of the first instrument revealed that 36.66% of the tests items were acceptable. This is fair, but it must achieve a higher level of acceptability. To achieve this goal, the test designers have to revise or reject the items of the test that they were not acceptable. An interesting result of the study was the strong relationship among the test parts. All the correlation coefficients were >0.6. The greatest correlation was between vocabulary parts. It is reasonable. Another reasonable result was the weakest correlation between grammar and reading. Interestingly, total score had very strong correlation with all the parts (r>0.8). The strongest correlation of total score was with part F i.e. grammar (r=0.946). According to the scores the test has a fair construct validity. However, this is not enough to judge and make decision.

Another instrument results showed that teachers evaluated 53% of the construct of the test as valid. It means that the construct validity of the test is fair, but it is not strong. The teachers claimed that the test is able to measure students' knowledge in different level of proficiency. They asserted that the test is a good measurement for vocabulary knowledge. It means that the test is not, thoroughly, able to measure other skills and sub-skills. Here the construct validity is under question according to teachers' evaluation. On the other hand, the teachers evaluated some items of the checklist as the weakest items considered in test design. They believed that the test has failed to focus on the real writing tasks and triggering their appropriate strategies, although the aim of the book designers is to lead learners to write English as properly and communicatively as possible. Also, they evaluated the reasonable item equivalence. When we look at the test rubric there is not any real writing tasks. The percentage of grammar items is the highest, 40%. Although the Ministry of Education of Iran recommends the teachers to have less focus on grammar, the test framework designed by its educationalists is quite different. The respondents claimed that tis test is not a real communicative test, because it did not consider the communicative tests principles. It means that although Ministry of Education claims that the book and its test is communicative, the test design does not have a communicative framework. So this test should be revised to be a communicative

test, and to be able to measure writing. We recommend that the framework of the test should shift to real communicative one using CLT and communicative test designing principles. Then there will be a valid construction. However, the process of validation should be in progress to make a real valid test; as Hodson (2014) suggests, "while validation will never be finished, since resources are finite and test uses require continual revalidation, a program can address stakeholders' concerns and drive continuous improvement of qualifications" (p. 1). So the results of this study can be used to enhance construct validity of national-wide tests, because the test framework is recommended by Curriculum and Textbooks Development Office of Iranian Ministry of Education.

REFERENCES

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

Brown, J. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Carroll, J. B. (1987). 8 new perspectives in the analysis of abilities. *The Influence of Cognitive Psychology on Testing, 13*, 267-284.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall, Englewood Cliffs.

Fallahian Sichani, E., & Tabatabaei, O. (2015). Construct Validity of MSRT Reading Comprehension Module in Iranian Context. *English Language Teaching*, *8*(9), 173-186.

Green, A. (2014). *Exploring language assessment and testing*. New York: Routledge.

Heydari, P., Bagheri, M., Zamanian, M., Sadighi, F., & Yarmohammadi, L. (2014). Investigating the construct validity of "structure and written expression" section of tolimo through factor analysis. *International Journal of Language Learning and Applied Linguistics World, 5*(1), 430-438.

Hodson, P. (2014). Practical validation: Organisational approaches to large-scale evaluation and continuous improvement. *IAEA Singapore*, 1-10.

Lord, F. M. (1952*).* The relationship of the reliability of multiple-choice test to the distribution of item difficulties. *Psychometrika, 18,* 181-194.

Lynch, K. B. (2003). *Language assessment and program evaluation. Edinburgh, Scotland: Edinburgh University Press.*

Messick*, S*. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY American Council on education and Macmillan.

Messick*, S.* (1995). Validity of psychological assessment. Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist,* 5(4), 741-749.

Miller, I., & Miller, M. (2012). *John E. Freund's Mathematical Statistics with Applications (8th ed.).* Prentice Hall: Pennsylvania State University.

Moore, T., & Morton, J. (2012). Construct validity in the IELTS Academic Reading Test: A comparison of reading requirements in IELTS test items and in university study. *Studies in language testing. 34.*

Richards, J. C., & Schmidt, R. (2010). *Longman dictionary of language teaching & applied linguistics* (4rd ed.). Essex: Pearson Education Limited.

Zoghi, M., Rostami, G, & Gholami, H. (2016). An evaluation of the construct validity of Iranian national test of English at high schools. *Journal of Applied Linguistics and Language Research, 3*(1), 185-196.