Cumhuriyet Üniversitesi Fen Fakültesi
Fen Bilimleri Dergisi (CFD), Cilt:36, No: 3 Ozel Sayı (2015)
ISSN: 1300-1949

Cumhuriyet University Faculty of Science
Science Journal (CSJ), Vol. 36, No: 3 Special Issue (2015)
ISSN: 1300-1949

A quantitative test analysis: Implementing an interpretive approach to validity argument

Forough RAHIMI

Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Received: 01.02.2015; Accepted: 06.06.2015

Abstract. Validity and test fairness have concerned many researchers and scholars in the field of language testing due to its importance in decision making process (Kunnan, 2008; Shohamy, 2001, 2007; Bachman, 2005) as well as the consequences that an unfair practice may bring to test takers (Kane, 2004, 2006; Xi, 2005a, 2005b). This study was a quantitative investigation of test validity and fairness using an argument-based approach developed by Xi (2010). The main purpose of this study was to quantitatively assess and determine the relationship between a series of inferences in an argument-based framework that can potentially lead to degrees of unfairness and lack of validity in language testing. To this end samples of IELTS test were analyzed and the obtained results determined degrees of fairness which existed among the inferences in the interpretive model.

Keywords: test validity, test fairness, interpretive approach, quantitative analysis.

1. INTRODUCTION

Testing is not a neutral process and always brings consequences for test takers. This fact highlights the importance of test fairness as one main challenge in language testing research. This study was an attempt to detect and explain the potential sources of unfairness by describing the relationship between some variables which exist in an argument-based framework. The fairness argument consisted of a series of inferences that are connected to each other in chain manner. These inferences and the relationship among them determine degree of fairness of a test. The main purpose for the application of this framework was to extract potential sources and degrees of unfair testing observed in samples of IELTS test.

The study was organized around 5 research questions which were formed based on the components of Xi's (2010) argument-based framework. These interwoven components were a series of factors building a fairness argument in the heart of a validity argument, ranging from domain description to evaluation, generalization, explanation, extrapolation and utilization. It is worth mentioning that the inference of utilization could not be assessed quantitatively so the researcher tried to examine the relationship between variables in inferences of domain description to extrapolation only.

The data for this study was collected mainly by tests' content analyses and questionnaire survey.

1.1. Conceptual framework of the study

The framework which is going to be used for this study is adopted from Xi (2010), which is an approach linking fairness directly to validity. Xi's (2010) framework to test fairness is an approach which links fairness directly to validity, and develops a fairness argument through a validity argument.

* Corresponding author. *Email address: rahimi.forough@yahoo.com*

Special Issue: The Second National Conference on Applied Research in Science and Technology

This argument-based approach is illustrated by six inferential steps and the mechanisms under which they can be organized conceptually to link an observation in a test to score-based interpretations and uses. These steps include:

1. ***Domain description:*** The first link is from the target domain to observations on the test. The warrant supporting this inference is that the target domain of language use in the English-medium institutions of higher education provides a basis for the observations of performance on the test to reveal relevant knowledge, skills, and abilities.

2. ***Evaluation*****:** The second link from observations on the test to observed test scores hinges on the warrant that observations of performance on the test are obtained and evaluated appropriately to provide observed scores reflective of intended academic language abilities, not other irrelevant factors.

3. ***Generalization***: The third link is from the observed score to the expected (universe) score. The pertinent warrant is that the observed scores on the test are generalizable over similar language tasks in the universe, test forms and occasions.

4. ***Explanation*****:** The fourth link between the expected scores and the theoretical score interpretation bears on the warrant that expected scores can be accounted for by underlying language abilities in an academic environment.

5. ***Extrapolation*****:** The fifth link connects the theoretical score interpretation and target score interpretation. The warrant is that the theoretical construct of academic language abilities accounts for the quality of language performance in English-medium institutions of higher education. At these two links (Explanation and Extrapolation), meaning can be attached to the expected scores in two potential ways to support valid interpretations of the assessment results. The expected scores can be interpreted by drawing on a theoretical construct (e.g. a communicative competence model) that underlies consistencies in test takers' performances. For assessments for which specific domains of generalization can be defined, this representation of the meaning of assessment results is further contextualized in the target domain to which the test scores are intended to be generalized. In some instances, in the absence of a strong construct theory, the generalization of test performance to the intended domain may sustain the link from the expected scores to the target score interpretation.

6. ***Utilization*****:** The last link connects score-based interpretations and test use. The warrants are that test scores and other information provided to users are relevant, useful and sufficient for evaluating the adequacy of international students' English proficiency for studying at English medium institutions, for determining the appropriate ESL coursework needed, and for selecting international teaching assistants, and have beneficial consequences for the teaching and learning of English.

(Adopted from Xi 2010, pp. 156-157).

These six inferences, if supported, increasingly add meaning and value to the elicited test performance, thus supporting score-based decisions.

1.2. Research questions

1. To what extent do the observations of the performance on the IELTS tests match the target domain of English language use in English medium institutions?

2. To what extent do the observations of performance on the IELTS tests match observed test scores?

3. To what extent can the IELTS observed test scores be generalized to similar language tasks in the universe, test forms and occasions?
4. To what extent can the IELTS expected test scores be accounted for by underlying language abilities in an academic environment?
5. To what extent does the theoretical construct(s) of academic and/or social language abilities account for the quality of language performance in English-medium institutions or English speaking countries?

2. METHOD

2.1. Participants

The participants of this study were 140 members from three main groups including IELTS candidates, teachers and raters. To this end 100 candidates of IELTS were chosen randomly to take part in the questionnaire survey. Besides, 20 IELTS teachers and 20 IELTS raters were chosen based on convenient random sampling technique to participate in teachers' and raters' questionnaires, respectively.

2.2. Instruments

The instruments used in this study included the candidates' questionnaire for those who took IELTS, the teachers' questionnaire for those preparing candidates for IELTS, and raters' questionnaire.

In addition to data collected via questionnaire, a careful test content analysis was conducted based on Xi's (2010) six inferences model. The model, builds an argument to test interpretations and uses which are made based on test scores. This chained model examined and analyzed links among interwoven factors that illustrated relationships between the target domain to observations on the test, observations on the test to observed test scores, the observed score to the expected score, the expected scores and the theoretical score interpretation, the theoretical score interpretation and target score interpretation, and score-based interpretations and test use.

2.3. Data analysis

The quantitative analyses provided descriptive statistics for variables addressed in research questions 1 to 5, and utilized Pearson correlations, ANOVA and post hoc analysis. The results were significant enough to reject null hypotheses.

2.3.1. Quantitative results for the argument: Part 1

This section presents quantitative analysis and statistical results for the inference of domain description. This inference detects the relationship between test takers' performances on the test and their target domain of language use.

Table 1 represents descriptive statistics for two variables in the inference of domain description in the IELTS tests.

Table 1. Descriptive statistics for performance on the test and target domain of language use in the IELTS test

	Mean	Max	Std.	N
performance	20.00	40	6.30	125
target domain	33.10	55	10.23	125

The Means and Standard deviations for performance on the test and the ability in the target language domain are not statistically different in this test. In order to estimate the relationship between the observations of the performance on the tests and the target domain of English language, Pearson correlation was utilized and estimated. The results of correlation analyses are presented in Table 2 for the IELTS test.

Table 2. Correlation between the performance on the tests and the target domain in the IELTS test

		performance	target domain
performance	Pearson Correlation	1	.762**
	Sig. (2-tailed)		.000
	N	125	125
target domain	Pearson Correlation	.762**	1
	Sig. (2-tailed)	.000	
	N	125	125

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

Sig=.0<.05

As shown in Table 2 the correlation between the observations of the performances on the tests and the target domain of language use is 0.762 for the IELTS test which indicate a high relationship between the variables. Considering the significance level (Sig=.0<.05) it can be concluded that the relationship between these variables is statistically meaningful. Therefore the first null research hypothesis is rejected and it can be concluded that there is a high relationship between test takers' performances on these high-stakes tests and the way they will perform in target domains of language use in future.

2.3.2. Quantitative results for the argument: Part 2

This section presents the descriptive statistics and also the correlation result which are related to variables in the inference of evaluation. From this inference the second research question and null hypothesis of this study were formed. The inference of evaluation determines the relationship between observed performance on the test and observed test scores. Table 3 shows the descriptive statistics for the variables under investigation in the IELTS test.

Table 3. Descriptive statistics for observed performance on the test and observed test scores in the IELTS test.

	Mean	Max	Std.	N
Operf	34.00	55	10.57	125
ALA	24.61	40	9.16	125

As shown above, the Means and Standard deviations of the two variables of observed performances and observed test scores do not mark statistically important difference in these tests. In order to estimate the relationship between the observations of performance on the tests and observed test scores and if these scores are reflective of intended academic language abilities, Pearson correlation was utilized and estimated. The result of correlation analysis is presented in Table 4.

Table 4. Correlation between observed performance on the test and observed test scores in the IELTS test

		Operf	ALA
Operf	Pearson Correlation	1	.711**
	Sig. (2-tailed)		.000
	N	125	125
ALA	Pearson Correlation	.711**	1
	Sig. (2-tailed)	.000	
	N	125	125

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

Sig=.0<.05

As shown in Table 4, the correlation between the observations of the performances on the tests and observed test scores which are reflective of intended academic language abilities is 0.711 in the IELTS test. The results indicate a high relationship between the variables. Considering the significance level (Sig=.0<.05) it can be concluded that the relationship between these variables is statistically meaningful. Therefore the second null hypothesis is rejected and it can be concluded that there is a high relationship between test takers' performances on these high-stakes tests and their scores which are indicators of intended language abilities.

2.3.3. Quantitative results for the argument: Part 3

This section presents the descriptive statistics and also the correlation result which are related to variables in the inference of generalization. This inference addresses the third research question and null hypothesis of this study. Inference of generalization investigates whether observed test scores can be generalized to similar language tasks in the universe, test forms and occasions or not.

Table 5 shows the descriptive statistics for the variables of observed test scores and universe or expected scores in different occasions of target language use in the IELTS test.

Table 5. Descriptive statistics for the inference of generalizability in the IELTS test.

	N	Mean	Std.	Minimum	Maximum
strongly disagree	12	5.08	2.74	3.00	10.00
disagree	30	8.13	3.20	5.00	15.00
decidedun	19	8.57	1.53	6.00	12.00
agree	39	11.06	2.37	5.00	15.00
strongly agree	20	12.92	2.64	7.50	15.00
Total	120	9.65	3.43	3.00	15.00

In order to compare differences among groups, ANOVA was run, the result of which is presented in Table 6 for the IELTS test.

Table 6: ANOVA for the IELTS test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	633.558	4	158.389	23.679	.000
Within Groups	769.242	115	6.689		
Total	1402.800	119			

Sig=0<α=0.05

The results of ANOVA show that differences among groups are statistically significant at Sig=0<α=0.05, indicating that observed test scores can be generalized to other language tasks in the universe, occasions and test forms. Therefore, the null hypothesis which incorporates that observed test scores cannot be generalized to similar language tasks in the universe, test forms and occasions is rejected.

2.3.4. Quantitative results for the argument: Part 4

This part shows quantitative results for the inference of explanation which demonstrates the link between expected scores and the theoretical score interpretation. In other words, it determines whether expected scores can be accounted for by underlying language abilities in an academic environment or not.

First of all descriptive statistics for two variables (expected test scores and language abilities in academic environment) are presented in Table 7 for the IELTS test.

Table 7. Descriptive statistics for expected scores and underlying language abilities in academic environment in the IELTS test.

	Mean	Max	Std.	N
ex.test.score	12.33	20	3.84	125
acad.envir	31.48	40	8.53	125

As shown in this Table, the Means and Standard deviations for the two variables of expected test scores and language abilities in academic environment do not indicate statistically significant difference in these tests. In order to see if there is any significant relationship between expected test scores and language abilities in academic domain, Pearson correlation was conducted. The results are presented in Table 8 for the IELTS test.

Table 8. Correlation between expected scores and underlying language abilities in academic environment for the IELTS test.

		ex.test.score	acad.envir
ex.test.score	Pearson Correlation	1	.807**
	Sig. (2-tailed)		.000
	N	125	125
acad.envir	Pearson Correlation	.807**	1
	Sig. (2-tailed)	.000	
	N	125	125
**. Correlation is significant at the 0.01 level (2-tailed). Sig=.0<.05			

As can be seen in the table the correlation between the two variables of expected test scores and academic language abilities is reported 0.807 for the IELTS test. The results are statistically significant at 0.05 level. This incorporates that there is a strong relationship between expected test scores and underlying language abilities in academic environments. The results reject the fourth null hypothesis of the study.

2.3.5. Quantitative results for the argument: Part 5

This section presents the quantitative analysis and statistical results related to the inference of extrapolation. This inference addresses the relationship between the theoretical construct of academic language abilities and the quality of language performance in target domains of language use. Descriptive statistics for the variables of this inference are presented in Table 9 for the IELTS test.

Table 9. Descriptive statistics for academic language abilities and language performance in target domain for the IELTS test.

	Mean	Max	Std.	N
ac.lan.ability	27.02	30	8.19	125
qual.lan.perform	28.72	45	9.44	125

As this Table illustrates, the Means and standard deviations for the two variables of theoretical construct of academic language abilities and the quality of language performance in target domains of language use do not indicate statistically significant difference in both tests. In order to find the relationship between these two variables, Pearson correlation was run. The results are reported in Table 10 for the IELTS test.

Table 10. Correlation between academic language abilities and language performance in target domain for the IELTS test.

		ac.lan.ability	qual.lan.perform
ac.lan.ability	Pearson Correlation	1	.912**
	Sig. (2-tailed)		.000
	N	125	125
qual.lan.perform	Pearson Correlation	.912**	1
	Sig. (2-tailed)	.000	
	N	125	125
**. Correlation is significant at the 0.01 level (2-tailed). Sig=.0<.05			

As shown in this Table, the correlation between variables of academic language ability and the quality of language performance in target domains is .912 for the IELTS test. The results are statistically significant to reject the fifth null hypothesis of this study. The result indicates that there is a strong relationship between these variables.

3. CONCLUSION

The importance of high-stakes testing has been continuously emphasized in the literature (Chapelle, 1999; Davies, 2003; Cheng, 2005). Researchers (Bachman, 1990; McNamara and Roever, 2006; Messick, 1980, 1988, 1989, 1996) highlighted the point that the production, execution, and successive use of language test score have significant effects on individuals, institutions and society. Bachman and Palmer (1996) maintain that the act of testing has consequences for the participants. There are consequences for the individuals, organizations and society involved when decisions such as entrance, selection or admission are made based on the examinations scores rather than other criteria. This concern highlights the importance of fairness as a test quality which appears to be universally desirable in assessment domains. The results of the present study which applied an argumentative approach to fairness marked a number of conclusions.

This study indicated that there is a direct relationship between the performance on the IELTS tests and the target domain of English language use in English medium institutions. This finding rejects the first null hypothesis of the study. Besides, the results showed that there is a direct relationship between the performances on the IELTS tests observed test scores and these scores are not reflective of intended academic language abilities. The results reject the second null hypothesis of the study for two tests. The study also proved that the IELTS observed test scores can be generalized to similar language tasks in the universe, test forms and occasions. The findings reject the third null hypothesis of the study for both tests. The findings pointed out that the IELTS expected test scores can be accounted for by underlying language abilities in an academic environment. The results reject the fourth null hypothesis of the study for both tests. In addition, the results indicated that the theoretical construct of academic language abilities accounts for the quality of language performance in English-medium institutions. The findings reject the fifth null hypothesis of the study for both tests. Moreover, it was shown that the IELTS score-based interpretations are relevant, useful and sufficient for evaluating the adequacy of test takers' English language proficiency for studying at English medium institutions, for determining the appropriate ESL coursework needed, and for selecting international teaching assistants, and have beneficial consequences for the teaching and learning of English. The results reject the sixth null hypothesis of the study for both tests.

One limitation of this study was that the inference of utilization which connected score-based interpretations and test use could not be assessed quantitatively in this study. Further research studies can address this inference using qualitative techniques.

REFERENCES

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

[2] Bachman, L. F. (2005). Building and supporting a case for test use. Language Assessment Quarterly,2, 1–34.

[3] Bachman, L.F. & Palmer, A. (1996).Language testing in practice: Designing and developing useful language tests, Oxford University Press, Oxford.

[4] Chapelle, C. (1999). Validity in language testing. Annual Review of Applied Linguistics, 19, 254-274.

[5] Cheng, L. (2005). Changing language teaching through language testing: A Washback study. Studies in Language Testing: Volume 21, Cambridge University Press, Cambridge.

[6] Davies, A. (2003). Three heresies of language testing research. Language Testing, 20(4), 355 –368.

[7] Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. Measurement: Interdisciplinary Research and Perspectives, 2, 135–170.

[8] Kane, M. T. (2006). Validation.In Brennan, R. L. (Ed.), Educational measurement, 4th edn. (pp.18–64). Washington, DC: American Council on Education/ Praeger.

[9] Kunnan, A. J. (2008). Large-scale language assessment. In E. Shohamy & N. Hornberger (Ed). Encyclopedia of language and education, 2nd Edition, Volume 7: Language Testing and assessment (pp. 135-155). Amsterdam: Springer Science.

[10] McNamara, T. F. & Roever, C. (2006).Language testing: The social dimension. Oxford: Blackwell.

[11] Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35(11), 1012-1027.

[12] Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 33 - 45). Hillsdale, N.J. :: Lawrence Erlbaum Associates.

[13] Messick, S. (1989). Validity. In Linn, R. L. (Ed.), Educational measurement, 3rd edn. (pp. 13–103). New York: American Council on Education and Macmillan.

[14] Messick, S. (1996). Validity and washback in language testing. Language Testing, 13(3), 241–256.

[15] Shohamy, E. (2001). The power of tests: A critical perspective of the uses of language tests. London: Longman.

[16] Shohamy, E. (2007). Language tests as language policy tools. Assessment in Education: Principles, Policy & Practice, 14, 117-130.

[17] Xi, X. (2005a).An argument-based approach to investigating fairness for the new TOEFL test. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

[18] Xi, X. (2005b). Do visual chunks and planning impact performance on the graph description task in the SPEAK Exam? Language Testing, 22(4), 463–508.

[19] Xi, X. (2010).How do we go about investigating test fairness? Language Testing, 27(2), 147–170.