Esfandiari et al. Language Testing in Asia (2018) 8:7
https://doi.org/10.1186/s40468-018-0062-7

Language Testing in Asia

RESEARCH Open Access

A quantitative analysis of TOEFL iBT using an interpretive model of test validity

Mohammad Reza Esfandiari[1], Mohammad Javad Riasati[1], Helia Vaezian[2] and Forough Rahimi[3*]

* Correspondence: rahimi.forough@yahoo.com
[3]School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
Full list of author information is available at the end of the article

Abstract

Background: Validity is a notable concept in language testing which has concerned many researchers and scholars in the field of language testing due to its importance in decision making process. Tests' results always introduce consequences to test takers' lives which emphasizes the need to ensure their validity. Detecting and delineating the potential sources that may threaten the validity of standardized tests of English proficiency is therefore of great importance.

Methods: This study was a quantitative investigation of test validity using an interpretive model. The main purpose of this study was to quantitatively assess and determine the relationship between a series of inferences in an interpretive argument-based framework that can potentially threaten the validity of the test.

Results: To this end samples of TOEFL iBT were analyzed and the obtained results determined its validity expressed in terms of the chain of inferences in the interpretive model.

Conclusion: The findings indicated that there is a direct relationship between the performance on the TOEFL iBT and the target domain of English language use, observed test scores are reflective of intended academic language abilities, observed test scores can be generalized to similar language tasks in different occasions and test forms, and expected test scores can be accounted for by underlying language abilities in an academic environment.

Keywords: Test validity, Interpretive model, TOEFL iBT, Language testing

Background

Since language testing is not a neutral practice, it may trigger some consequences for test takers. This issue highlights that test validity should be taken into serious account in language testing research. There is a set of relationships, intended and unintended, positive and negative, between testing, teaching, and learning in any educational setting. The possibility that inappropriate score-based interpretations and uses of tests may introduce unfair consequences to different groups of test takers is a great problem for language teachers and researchers. It is of high significance for educational programs and institutes to ensure maximum validity in standardized tests. This is the reason for the existence of extensive studies on this topic. Test score and score-based interpretations are important because they are a basis for making decisions about individuals. Therefore, decisions should be fair and sensitive to the values of the decision makers (Schmidgall 2017). Standardized tests are widely used in education for

decision-making purposes, a fact that requires ensuring their validity in stages of test development, administration, scoring, interpretation, and decision-making. Kunnan (2008) believes that fairness must be ensured in every stage of test development (from test design stage through the research stage).

Cronbach (1988) and Messick (1988) were pioneers in validity studies. This concept which was initially regarded as a characteristic of measuring tools (Chapelle 1999) is now considered as a feature of score uses and interpretations or inferences derived from test scores (Messick 1988). The theoretical framework of the present study is based on Messick's (1989) unified validity framework and Chapelle's (1994) validity table. Messick extended insight into validity research. He argued that validity is a unitary concept, which is based on construct-referenced evidence. Messick (1980) maintained that construct-referenced evidence is a "unifying concept that integrates criterion and content considerations into a common framework" (p. 1015) and that content- and criterion-referenced evidence are insufficient in validation. Messick argues that content-referenced evidence is subjective because it is based on expert judgments and ignores the psychological processes of test takers, internal structures of the test, and differences in performance across test takers (Messick 1988, p. 8). He further stated that the relationship between criterion-referenced evidence with future performance on a criterion is not exact and precise because the criterion should be validated like the test itself (Messick 1988, p. 9).

Messick (1989) mentioned some important points such as misuse of the test, social consequences of tests, and test fairness. Messick's validity argument serves as underlying theoretical framework of this study because it considers consequences of test use in his framework. He believed that undesirable social consequences are results of invalid test interpretations. If the adversarial social consequences are due to a test's invalidity, then the validity of the test use becomes questionable. We need to include the effect of tests on students, institutions, and society as one type of validity evidence.

Chapelle's (1994) validity table interprets Messick's framework in detail. Chapelle's validity table has three columns (evidence, argues in favor, and argues against) and four rows which are related to four types of justification (construct validity, relevance/utility, value implications, and social consequences). Table 1 provides the main layout of her validity argument.

The interpretive model of validity

Test users can decide if the use of a test is valid in a specific situation via paying attention to both negative and positive attributes. This validity table does not explicitly

Table 1 Analysis of justifications for the use of tests

Justifications/evidence	Argues in favor	Argues against
Construct validity		
Relevance/utility		
Value implications		
Social consequences		

Adopted from Chapelle (1994), p. 177

report and refute negative evidence. Each test that we want to validate will have its own table, with different types of evidence in rows.

Messick's expansion of validity to include values and social consequences of testing practices (Messick 1989) is related to the social and political dimensions of tests, which paved the way for developing argumentative frameworks for validity studies.

In an interpretive framework, validity is typically established by evidence that supports the soundness of score-based interpretations and uses for the whole test-taking population. The interpretive model 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. The first link is from the target domain to observations on the test which is named "domain description." The warrant supporting this inference is that the target domain of language use is the basis of the observations of performance on the test (indication relevant knowledge and abilities).The second link from observations on the test to observed test scores, named "evaluation," is based on the warrant that observations of performance on the test are evaluated appropriately to provide observed scores reflective of intended academic language abilities and knowledge. The third link, named "generalization," connects the observed score to the expected score based on the warrant that the observed scores on the test are generalizable to similar occasions. The fourth link, named "explanation," connects the expected scores and the theoretical score interpretation based on the warrant that expected scores can be accounted for by underlying language abilities in an academic environment. The fifth link, named "extrapolation," 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 academic settings. The sixth link, named "utilization," connects score-based interpretations and test use. The warrants are that test scores and other information provided to users are relevant and useful for evaluating the students' English proficiency for studying at English-medium institutions and have beneficial consequences for the teaching and learning of English (Xi 2010). If these six inferences are supported, they can add meaning and value to the elicited test.

Background of the study

Test validation methods are an important part of language testing research. Validation methods have generally been predisposed in three domains (Rahimi et al. 2016): developments in psychometric and statistical methods in education (Bachman 2004; Bachman and Eignor 1997), qualitative methods in language testing influenced by second language acquisition (Banerjee and Luoma 1997; Bachman & Cohen 1998), conversation analysis and discourse analysis (Lazaraton 2002), and cognitive psychology research studies amalgamated into language testing research (Green 1997). The 1950s and 1960s witnessed studies on validation of discrete-point language tests while the research shifted attention to communicative language tests during the 1970s (Clark 1975). This illustrates the fact that validity studies in earlier times was confined to limited concepts such as content and construct validity rather than score interpretations as the basis for predictions (Sackett et al. 2017). Additionally, test-taking processes and strategies and test consequences were ignored.

The 1980s witnessed a change of attention from concurrent validity studies to test-taking processes and factors affecting test performance which promoted focus on score interpretation based on empirically grounded explanations of scores. The improvements terminated in Messick's (1989) explication of validity. Different types of validity became changed into a unitary concept of construct validity, underlining the importance of uniting different types of evidence to support a particular test use. Messick also formally expanded validity to incorporate social values and consequences, arguing that evaluation of social consequences of test use as well as the value implications of test interpretation both "presume" and "contribute to" the construct validity of score meaning (p. 21). Messick's unitary concept of validity influenced language testing through Bachman's 1990 work (Cumming and Berwick 1996; Kunnan 1998).

However, Messick's model is highly abstract and provides practitioners limited guidance on the process of validation (Rahimi et al. 2016). Therefore, Bachman and Palmer (1996) proposed the notion of test usefulness to make Messick's work more manageable to language testers. After this principal research trend change, there were some elaborate studies on topics such as theories of validity, impact, ethics, principles of critical language testing (Shohamy 2001), policy and social considerations (McNamara 2006), and fairness (Kunnan 2004). These studies significantly improved and covered areas such as factors affecting test performance, generalizability of scores, and consequences of test use (Bachman 2000; Cumming and Berwick 1996; Kunnan 1998). Furthermore, there was a rise of quantitative and qualitative and triangulated methodologies (Xi 2005b) and validity argument through a coherent analysis of all the evidence (Cronbach 1988).

Toulmin (2003) saw validation as constructing an interpretive argument stage and developing and assessing a validity argument. The interpretive argument is illustrated through the chain of inferences linking test performance to a decision and their underlying assumptions. If the assumptions are true, they support the relevant inference. If the network of inferences is supported, it gives sense to test performance and the corresponding score and, therefore, a score-based decision is justified. This approach expanded the validation studies on TOEFL (Chapelle et al. 2008). It is essential that more elaborate validation research studies be conducted to address the concerns of teachers and students about the purpose, quality, and quantity of testing in education (Lederman and Burnstein 2006).

These points clearly show that validity research direction shifted toward establishing validity evidence (Rahimi et al. 2016). Kane (2006) maintained that evaluating the credibility of interpretations and uses describes the validation process. This definition allowed the researchers to establish evidence of the appropriateness of inferences in a test. As time progressed in the field of validity studies, the construct validity approach was widely accepted as a general model for validation (Anastasi 1986; Embretson 1983; Guion 1977; Messick 1980, 1988, 1989).

Research questions

The present study aimed at implementing an interpretive argument of validity to evaluate TOEFL iBT base on the inferences and their relevant warrants. It sought to provide answers to the following research questions:

1. Do the observations of the performance on the test match the target domain of language use?
2. Do the observations of performance on the test match observed test scores?
3. Can observed test scores be generalized to similar occasions?
4. Do expected test scores account for underlying language abilities in an academic setting?
5. Does the theoretical construct of language abilities account for the quality of language performance in related settings?

Methods

This study was a quantitative investigation of test validity using an interpretive model. The study was conducted on some officially released test samples of TOEFL iBT. This involved a highly systematic test content analysis, plus a questionnaire survey. The content of test samples available to the researcher was examined and analyzed, and then, the participants took part in a questionnaire survey about the content and purpose of TOEFL iBT and indicated their opinions about the validity of this test. For the purpose of this study, three questionnaires (a candidates' questionnaire, a teachers' questionnaire, and a raters' questionnaire) (Additional files 1, 2, and 3) were developed, validated, and utilized.

Participants

The participants of this study were 140 members from three main groups including TOEFL iBT candidates, teachers, and raters. Since accessing the participants was subjected to some limitations, the sample size was determined by availability. To this end, 100 candidates of TOEFL iBT were chosen randomly to take part in the questionnaire survey. Besides, 20 TOEFL iBT teachers and 20 TOEFL iBT raters were chosen participate in teachers' and raters' questionnaires, respectively. The utilized sampling technique was convenient sampling due to limitations in accessing the participants.

Instruments

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

All the questionnaires were developed in English and had four parts. The first part was about the respondents themselves. The second and third parts' items of the questionnaires were constructed based on six inferences of the framework. The last part presented some open-ended questions through which participants could produce some free writing and reflect their opinions. The closed-ended items were designed on a 5-point Likert scale of agreement and frequency. The construct validity and the reliability of the questionnaires of the study were examined through conducting a pilot study with a sample of 30 questionnaires completed by participants at some available language institutes in Shiraz (Rahimi et al. 2014). Cronbach's alpha was employed to ensure the reliability of questionnaires. The results indicated an alpha of 0.86 for candidates' questionnaire, 0.75 for teachers' questionnaire, and 0.88 for raters' questionnaire which indicated that these questionnaires are reliable to be employed in this study.

In addition to data collected via questionnaire, a careful test content analysis was conducted based the interpretive model. 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.

Results and discussion

The quantitative analyses provided descriptive statistics for variables addressed in research questions 1 to 5, and utilized Pearson correlations, ANOVA, and post hoc analysis.

Quantitative results for "domain description" inference

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 2 represents descriptive statistics for two variables in the inference of domain description in TOEFL iBT.

As illustrated in Table 2, the means for test performance and target domain of language use are 20.40 and 32.81, respectively, while the standard deviations are 6.40 and 8.23 for these two variables, respectively. 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 3.

As shown in Table 3, the correlation between the observations of the performances on the tests and the target domain of language use is .623 for TOEFL iBT which indicates 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, addressing the first research question of the study, 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.

Quantitative results for "evaluation" inference

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 of this study was formed. The inference of evaluation determines the relationship between observed performance on the test and observed test scores. Table 4 shows the descriptive statistics for the variables under investigation.

As shown in Table 4, the means and standard deviations of the variable of observed performances are 33.15 and 9.40, respectively, and 24.51 and 7.28 for the variable of

Table 2 Descriptive statistics for performance on the test and target domain of language use in the TOEFL iBT test

	Mean	Max	Std.	*N*
Perf.	20.40	40	6.40	140
TD	32.81	55	8.23	140

Table 3 Correlation between the performance on the tests and the target domain in the TOEFL iBT test

		Performance	Target domain
Perf.	Pearson correlation	1	.623**
	Sig. (two-tailed)		.000
	N	140	140
TD	Pearson correlation	.623**	1
	Sig. (two-tailed)	.000	
	N	140	140

Sig = .0 < .05
**Correlation is significant at the 0.01 level (two-tailed)

observed test scores, respectively. 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 5.

As shown in Table 5, the correlation between the observations of the performances on the tests and observed test scores which are reflective of intended academic language abilities is .697 in the TOEFL iBT 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, addressing the second research question, 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.

Quantitative results for "generalization" inference

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 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 6 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 TOEFL iBT.

In order to compare differences among groups, ANOVA was run, and the result of which is presented in Table 7, TOEFL iBT tests.

The results of ANOVA show that differences among groups are statistically significant at Sig = 0 < α = 0.05. This can be indicative of the fact that observed test scores cannot be generalized to similar language tasks in the universe, test forms and occasions.

Table 4 Descriptive statistics for observed performance on the test and observed test scores in the TOEFL iBT test

	Mean	Max	Std.	*N*
OPerf.	33.15	55	9.40	140
OTS	24.51	40	7.28	140

Table 5 Correlation between observed performance on the test and observed test scores in the TOEFL iBT test

		OPerf.	ALA
OPerf.	Pearson correlation	1	.697**
	Sig. (two-tailed)		.000
	N	140	140
OTS	Pearson correlation	.697**	1
	Sig. (two-tailed)	.000	
	N	140	140

Sig = .0 < .05
**Correlation is significant at the 0.01 level (two-tailed)

In order to detect exactly where differences fall, a post hoc analysis using Tamhane test was used. Table 8 summarizes the results of the post hoc analysis for the TOEFL iBT.

As we can see, the results indicate means which are statistically significantly different among groups. This incorporates that observed test scores can be generalized to other language tasks in the universe, occasions, and test forms.

Quantitative results for "explanation" inference

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 9 for the TOEFL iBT tests.

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 10 for the TOEFL iBT test.

As can be seen in Table 10, the correlation between the two variables of expected test scores and academic language abilities is reported to be .436 for the TOEFL iBT test. The results are statistically significant at a 0.05 level. This incorporates that there is a strong relationship between expected test scores and underlying language abilities in academic environments.

Table 6 Descriptive statistics for the inference of generalizability in the TOEFL iBT test

	N	Mean	Std.	Minimum	Maximum
Strongly disagree	18	6.14	3.83	3.00	13.50
Disagree	36	7.53	3.19	3.00	15.00
Undecided	24	9.11	2.09	6.00	13.50
Agree	37	10.66	2.72	4.50	15.00
Strongly agree	25	12.82	3.07	6.00	15.00
Total	140	9.42	3.60	3.00	15.00

Table 7 ANOVA for the TOEFL iBT test

	Sum of squares	df	Mean square	*F*	Sig.
Between groups	543.503	4	135.876	15.561	.000
Within groups	1004.145	115	8.732		
Total	1547.648	119			

Sig = 0 < α = 0.05

Quantitative results for "extrapolation" inference

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 11 for the TOEFL iBT test.

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 the test. In order to find the relationship between these two variables, Pearson correlation was run. The results are reported in Table 12.

As shown in the table, the correlation between variables of academic language ability and the quality of language performance in target domains is .459 for the TOEFL iBT

Table 8 Results of post hoc analysis using Tamhane test for multiple comparisons among groups in the TOEFL iBT test

(I) obs.test.score	(J) obs.test.score	Mean difference (I–J)	Sig.
Strongly disagree	Disagree	− 1.39048	.944
	Undecided	− 2.97078	.149
	Agree	− 4.51891*	.008
	Strongly agree	− 6.68214*	.000
Disagree	Strongly disagree	1.39048	.944
	Undecided	− 1.58030	.308
	Agree	− 3.12843*	.001
	Strongly agree	− 5.29167*	.000
Decidedun	Strongly disagree	2.97078	.149
	Disagree	1.58030	.308
	Agree	− 1.54813	.185
	Strongly agree	− 3.71136*	.001
Agree	Strongly disagree	4.51891*	.008
	Disagree	3.12843*	.001
	Decidedun	1.54813	.185
	Strongly agree	− 2.16324	.126
Strongly agree	Strongly disagree	6.68214*	.000
	Disagree	5.29167*	.000
	Undecided	3.71136*	.001
	Agree	2.16324	.126

Sig = 0 < α = 0.05
*The mean difference is significant at the 0.05 level

Table 9 Descriptive statistics for expected scores and underlying language abilities in academic environment in the TOEFL iBT test

	Mean	Max	Std.	*N*
ETS	11.88	20	3.27	140
TSI	32.25	40	8.42	140

test. The results are statistically significant to claim that there is a strong relationship between these variables.

Conclusions

After Messick (1988) proposed the notion of the consequential aspect of validity, the issue of test fairness gained prominence in language testing. Messick (1989) believed that there are some test properties which can affect language teaching and learning outcomes on an evidential basis. High-stakes tests have a crucial role in language teaching and testing domain and serve as a point of emphasis by pioneers in this field. Some scholars maintain that the production, execution, and successive use of language test score can significantly affect individuals, institutions, and the society (Bachman 1990; Messick 1980, 1988, 1989, 1996). This issue stresses the importance of validity and fairness as a test quality which should be taken into serious consideration by test developers. As such, this study was conducted to quantitatively evaluate and determine the relationship between a series of inferences in an interpretive argument-based framework as potential threats to the validity of one of the most prevalent tests of English language proficiency.

The quantitative analyses suggested that:

1. There is a direct relationship between the performance on the TOEFL iBT and the target domain of English language use in English-medium institutions.
2. There is a direct relationship between performances on the TOEFL iBT and observed test scores, and these scores are reflective of intended academic language abilities.
3. TOEFL iBT observed test scores can be generalized to similar language tasks in the universe, test forms, and occasions.
4. TOEFL iBT expected test scores can be accounted for by underlying language abilities in an academic environment.

Table 10 Correlation between expected scores and underlying language abilities in academic environment for the TOEFL iBT test

		ex.test.score	acad.envir
ETS	Pearson correlation	1	.436**
	Sig. (two-tailed)		.000
	N	140	140
TSI	Pearson correlation	.436**	1
	Sig. (two-tailed)	.000	
	N	140	140

Sig = .0 < .05
**Correlation is significant at the 0.01 level (two-tailed)

Table 11 Descriptive statistics for academic language abilities and language performance in target domain for the TOEFL iBT test

	Mean	Max	Std.	*N*
ALA	26.89	30	7.11	140
QLP	27.02	45	6.70	140

5. The theoretical construct of academic language abilities accounts for the quality of language performance in English-medium institutions.

This study puts emphasis on fair testing practice at different phases of test planning, development, administration, and specifically interpretation and use. High-stakes tests are too important in educational settings and can introduce consequences for test takers. This also leads test takers to test preparation classes that can open it to questions on washback, a practice that affects English language teaching and learning. The findings of this study can have the pedagogical implication of promoting positive washback effect on the way English language is taught and learned in different teaching and testing contexts (Rahimi et al. 2014).

There should be equilibrium in language classes between enhancing academic and general language skills and knowledge and also between the communicative approaches to teaching English and the design of the standardized high-stakes tests. Language teachers, institutions, and also learners should equally pay attention to improving communicative skills and gaining desirable test results at the same time (Rahimi et al. 2014).

In the same line, utilizing the results of such tests for fair classroom testing and teaching is of utmost importance in language education. While high-stakes tests certainly introduce more significant impacts on test takers, the role of classroom assessment should not be ignored in enhancing language learning also (Rahimi et al. 2014).

To sum up this part, it can be concluded that the pedagogical implication of this study can be classified into three categories. Firstly, the study can feed influential implications on the impact associated with language tests and positive and negative washback effect. Secondly, it can contribute to promoting professional and developmental awareness and wisdom of language teachers, institution, learners, and raters on the ways through which maximum fairness can be enhanced. And finally, it can help promoting fairer classroom assessment which can promote language learning via testing.

Table 12 Correlation between academic language abilities and language performance in target domain for the TOEFL iBT test

		ac.lan.ability	qual.lan.perform
ALA	Pearson correlation	1	.459**
	Sig. (two-tailed)		.000
	N	140	140
QLP	Pearson correlation	.459**	1
	Sig. (two-tailed)	.000	
	N	140	140

Sig = .0 < .05
**Correlation is significant at the 0.01 level (2-tailed)

Limitation of the study

The inference of utilization which aims to determine if test scores have beneficial consequences for the teaching and learning of English is based on the warrant that test scores and other information provided to users are relevant and useful for evaluating the students' English proficiency for studying at English-medium institutions and have beneficial consequences for the teaching and learning of English. This inference is supported by the evidence provided by other inferences in this chain and proved that score-based interpretations are relevant, useful, and sufficient for evaluating the adequacy of test takers' English language proficiency. However, one limitation of this study was that the inference of utilization could not be assessed quantitatively. Further studies can address this inference using qualitative techniques.

Additional files

Additional file 1: Candidates' questionnaire for those who took TOEFL iBT. (DOCX 23 kb)
Additional file 2: Teachers' questionnaire for those candidates preparing to take TOEFL iBT. (DOCX 24 kb)
Additional file 3; Raters' questionnaire. (DOCX 20 kb)

Abbreviations

ANOVA: Analysis of variance; TOEFL iBT: Test of English as a Foreign Language Internet-based Test

Acknowledgements

We would like to express our gratitude to the journal team and the study participants.

Availability of data and materials

All data and material used for this study are available.

Authors' contributions

All authors contributed to conducting this study at various phases of data collection, reading, and approving the final manuscript.

Authors' information

Mohammad Reza Esfandiari is an assistant professor at Islamic Azad University, Shiraz Branch. His main areas of interest include curriculum development, translation competence, and language testing.
Mohammad Javad Rikanatia is an assistant professor at Islamic Azad University, Shiraz Branch. His main areas of interest include language education and assessment and teacher education.
Helia Vaezian is an assistant professor at Khatam University in Tehran. Her main areas of interest are translator training, teacher education, and language assessment.
Forough Rahimi is an assistant professor at Shahid Beheshti University of Medical Sciences in Tehran. Her main areas of interest include teacher education, ESP, and language teaching and testing.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Foreign Languages, Shiraz Branch, Islamic Azad University, Shiraz, Iran. [2]English Language Department, Khatam University, Tehran, Iran. [3]School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Received: 31 January 2018 Accepted: 25 April 2018
Published online: 28 May 2018

References

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*, 1–15.
Bachman, L.F. & Cohen, A.D. (1998). Interfaces between second language acquisition and language testing research. Cambridge: Cambridge University Press.
Bachman, LF (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Bachman, LF. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1–42.

Bachman, LF (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.
Bachman, LF, & Eignor, DR (1997). Recent advances in quantitative test analysis. In C Clapham, D Corson (Eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment*, (pp. 227–242). Dordrecht: Kluwer Academic.
Bachman, LF, & Palmer, A (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
Banerjee, J, & Luoma, S (1997). Qualitative approaches to test validation. In C Clapham, D Corson (Eds.), *Encyclopedia of language and education, volume 7: Language testing and assessment*, (pp. 275–287). Dordrecht: Kluwer Academic.
Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research*? Second Language Research*, *10*, 157–187.
Chapelle, C. (1999). Validity in language testing. *Annual Review of Applied Linguistics*, *19*, 254–274.
Chapelle, CA, Enright, MK, Jamieson, JM (Eds.) (2008). *Building a validity argument for the test of English as a foreign language*™. Mahwah: Lawrence Erlbaum.
Clark, J. (1975). Theoretical and technical considerations in oral proficiency testing. In Jones, S. & Spolsky, B. (Eds.), Language Testing Proficiency, Center for Applied Linguistics. Arlington, 10–24.
Cronbach, LJ (1988). Five perspectives on the validity argument. In H Wainer, HI Braun (Eds.), *Test validity*, (pp. 3–17). Hillsdale: Lawrence Erlbaum.
Cumming, A, & Berwick, R (Eds.) (1996). *Validation in language testing*. Clevedon: Multilingual Matters.
Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.
Green, A (1997). *Verbal protocol analysis in language testing research*. Cambridge: Cambridge University Press.
Guion, RM. (1977). Content validity-the source of my discontent. *Applied Psychological Measurement*, *1*(1), 1–10.
Kane, MT (2006). Validation. In RL Brennan (Ed.), *Educational measurement*, (4th ed., pp. 18–64). Washington, DC: American Council on Education/ Praeger.
Kunnan, AJ (1998). Validation in language assessment. In *Selected papers from the 17th Language Testing Research Colloquium*. Mahwah: Long Beach, Lawrence Erlbaum.
Kunnan, AJ (2004). Test fairness. In M Milanovic, C Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference*, (pp. 27–48). Cambridge: Cambridge University Press.
Kunnan, AJ (2008). Large-scale language assessment. In E Shohamy, N Hornberger (Eds.), *Encyclopedia of language and education, 2nd edition, volume 7: Language testing and assessment*, (pp. 135–155). Amsterdam: Springer Science.
Lazaraton, A (2002). *A qualitative approach to the validation of oral tests*. Cambridge: Cambridge University Press.
Lederman, L. M., & Burnstein, R. A. (2006). Alternative approaches to high-stakes testing. Virginia: Phi Delta Kappan.
McNamara, TF. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012–1027.
Messick, S (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H Wainer, HI Braun (Eds.), *Test validity*, (pp. 33–45). Hillsdale: Lawrence Erlbaum Associates.
Messick, S (1989). Validity. In RL Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.
Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.
Rahimi, F, Bagheri, MS, Sadighi, F, Yarmohammadi, L. (2014). Using an argument-based approach to ensure fairness of high-stakes tests' score-based consequences. *Procedia - Social and Behavioral Sciences*, *98*(2014), 1461–1468. https://doi.org/10.1016/j.sbspro.2014.03.566 Elsevier, Science Direct.
Rahimi, F, Esfandiari, MR, Amini, M. (2016). An overview of studies conducted on washback, impact and validity. *Studies in Literature and Language*, *13*(4), 6–14.
Sackett, PR, Shewach, OR, Keiser, HN. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435–1447. https://doi.org/10.1037/apl0000236.
Schmidgall, JE (2017). Articulating and evaluating validity arguments for the *TOEIC*® tests. In *ETS research report series*. https://doi.org/10.1002/ets2.12182.
Shohamy, E (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Longman.
Toulmin, SE (2003). *The use of argument (updated edition)*. Cambridge: Cambridge University Press.
Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing*, *22*(4), 463–508.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.