Available online at www.elixirpublishers.com (Elixir International Journal)

Social Studies

Elixir Social Studies 92 (2016) 39234-39244

ISSN: 2229-712X

Ensuring fairness of high-stakes tests' score-based consequences using an argument-based approach

Forough Rahimi, Mohammad Sadegh Bagheri, Firooz Sadighi and Lotfollah Yarmohammadi

Department of Foreign Languages, Shiraz Branch, Islamic Azad University, Shiraz, Iran.

ARTICLE INFO

Article history:

Received: 26 January 2013;

Received in revised form:

21 March 2016;

Accepted: 26 March 2016;

Keywords

Test fairness,

Validity,

An argument-based framework,

IELTS test.

ABSTRACT

The study of test fairness is of crucial importance in the context of standardized high-stakes tests used to make selection decisions at different levels, for different purposes and across different groups of test takers. Fairness must be ensured in every stage of test development: from test design stage through the research stage (Kunnan, 2008). Researchers are particularly concerned about the possibility that standardized high-stakes tests may be used inappropriately to make score-based interpretations and decisions, which may introduce unfair consequences to different groups of test takers. These concerns have increased social and educational demands, subjecting these tests to rigorous validation research. This paper seeks to expand research in the area of test fairness. To this end, an argument-based approach to test fairness, developed by Xi (2010), is adopted to investigate the degree of fairness of standardized high-stakes tests of IELTS based on the inference of *Domain definition.* This article is derived from a larger study and reports the second phase of that investigation.

Introduction

Concerns about fairness among all test stakeholders are paramount in the milieu of high-stakes decision-making involving achievement, aptitude, admission, certification, and licensure tests. These high-stakes tests serve as gatekeepers to include or exclude individuals into academic communities and professional associations (Shohamy, 2001). This practice highlights the important function of tests as a source of power and educational control. Tests designed properly and used fairly can facilitate positive educational, social, and economic goals, while tests biased and misused violate social equity and its impact could ripple to the whole society. Large-scale high-stakes testing is increasingly being used for decision-making; thus, ensuring maximum fairness would be a central issue 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).

Test fairness has been a point of interest for scholars and teachers due to the importance of score-based interpretations, decisions and their social consequences. There are different approaches to the investigation of test fairness. One approach views fairness as *an independent test quality*. In this view, fairness is characterized as a test quality that is separate from validity, although some tenuous and inconsistent references may be made to validity. Xi (2010) believes that this approach "does not provide a mechanism for prioritizing them and for weighing one piece of fairness evidence against another" (p. xxx).The second sees fairness as *an all-encompassing test quality*, and gives primacy to test fairness and defines it as a test quality which subsumes and goes beyond validity. Implicit in this view is the argument that a test has to be fair to be valid. About this standpoint she maintains that "current validation frameworks … have provided means to address all the fairness qualities proposed in Kunnan (2004) in a coherent way within the framework of a validity or assessment use argument. It does not seem necessary to treat them as separate facets of fairness" (p. xxx). The third approach views fairness as *linked directly to validity*. This connection between fairness and validity suggests a strong possibility for linking fairness back to validity in a principled way. This kind of linkage would allow fairness research and practice to take advantage of a well-defined framework for validity.

The development and implementation of these frameworks to the investigation of fair testing practice highlights the importance of the topic. Xi (2010) believes that all of the empirical studies have looked at the stability of score interpretations across groups in different ways but almost none have addressed the consistency of score-based decisions (Zeidner, 1987) or the comparability of the broader effects of testing for different groups.

She further asserts that although there has been a substantial amount of work on the consequences of large-scale language tests, none of the studies have really looked at the differential impact a language test might have on different groups of test takers. No research has analyzed in depth how different manifestations of unfairness may impact the ultimate score interpretation and score-based decisions for a particular assessment. This is the main focus of this study. The researcher wants to analyze samples of high-stakes tests based on an argument based framework to estimate their fairness and the possible impact of unfair practices on final score-based interpretations, decisions and consequences.

Objectives

The main objective of this study is to apply Xian fairness framework (2010), which is established in a validity argument, to samples of standardized high-stakes tests of IELTS. The target objective is to delineate potential sources and degrees of

Tele:

E-mail address: rahimi.forough@yahoo.com

unfair testing and contribute to further fair testing practice that will benefit all groups of test takers.

To this end, we start a validity argument to provide an overall evaluation of the evidence for and against the proposed interpretation/use (or for and against the interpretive argument). The plausibility of the interpretive argument is evaluated by analyzing its overall clarity and coherence and by assessing the plausibility of its inferences and assumptions. This evaluation, or validation, of the interpretive argument generally requires many different kinds of analysis and is used to evaluate the different parts of the interpretive argument. The proposed interpretations and evidence, which uses of tests rely on all of this evidence, will support the validity argument. The second phase of validation focuses on a more objective, critical appraisal of the proposed interpretation and use of scores for a particular purpose, in particular contexts, and with a particular population, in order to evaluate the potential challenges to the proposed interpretive argument. When an inference is drawn (e.g. from an observed performance to a score) in a particular case, the warrant (e.g. the scoring rule) and its backing (judgments by panels of content experts who developed the scoring rule) may not be explicitly mentioned, especially if the inference is fairly routine in a particular context and the audience is friendly, but the warrant is invoked every time a score is interpreted.

Theoretical framework: A validity argument

The theoretical framework of the present study is based on Messick's (1989) unified validity framework and Chappelle's (1994) validity table. Messick expanded the scholarly realization of validity and the ways through which it can be investigated. He argued that validity is a unitary concept, which is based on construct-referenced evidence. Messick (1980) argues 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 evidences are by no means enough in validation. Messick argues that content-referenced evidence has an element of subjectivity since it is mainly a function of expert judgments, and leaves out the psychological processes of test takers, internal structures of the test, and differences in performance across test takers (Messick, 1988, p. 8); and that criterion-referenced evidence's correlation of test scores with future performance on a criterion may compound confusion, because the criterion will need to be validated like the test itself (Messick, 1988, p. 9).

Messick (1989) emphasized fundamental points related to test use such as misuse of the test, social consequences of tests, and test fairness. One important reason for assigning Messick's validity argument as the underlying theoretical framework of this study was the consideration of the consequences of test use in his framework. He argued that adverse social consequences as results of invalid test interpretations, raise vital social and political issues which stand against validity issues. If the adverse social consequences are attributed to the invalidity of the test, 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.

Conceptual framework: A fairness argument in a validity argument

The conceptual 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. This conceptual approach allows the extent of fairness explorations to be expanded and clarified, taking advantage of the well-defined framework for validity.

She maintains that:

"…an assessment that is unfair, in the sense that it systematically misrepresents the standing of some individuals or some groups of individuals on the construct being measured or that tends make inappropriate decisions for individuals or groups is, to that extent, not valid for that interpretation or use. Similarly, an assessment that is not valid in the sense that it tends to generate misleading conclusions or inappropriate decisions for some individuals or groups will also be unfair" (p. 155).

Fairness is characterized as comparable validity for relevant groups that can be identified. The fairness argument consists of a series of rebuttals that may challenge the comparability of score-based decisions and consequences for sub‐groups. This framework organizes different fairness investigations into a coherent framework and offers a principled approach to evaluating the soundness of the overall fairness argument and setting research priorities. This conceptual approach allows the extent of fairness explorations to be expanded and clarified, taking advantage of the well-defined framework for validity.

This characterization of fairness as a facet of validity also augments the traditional interpretations of validity by demanding additional support for the comparability of assessment results, interpretations, decisions and consequences for relevant sub-groups. This approach draws on current argument-based methods of test validation to systematize fairness investigations. Within this framework, a fairness argument can be used to systematically generate rebuttals to the validity argument that would compromise the comparability of assessment results, interpretations, decisions and consequences for relevant sub-groups. These rebuttals are in contrast to those that would potentially weaken the validity for the whole test-taking population. This argument-based structure allows us to track how fairness issues permeate the inferential steps and become prominent in score-based decisions, actions and consequences.

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 IELTS 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 IELTS 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.

Research question

As mentioned earlier, this study aims at investigating the degree of fairness of standardized high-stakes tests of IELTS. To this end, the main objectives of the study have been formulated into a research questions as follows:

What does the argument-based approach to fairness indicate about fairness degree of IELTS in terms of domain definition inference?

Literature review

Research on validity of IELTS

Test validation should be done on an on-going basis. It involves several studies to keep the test current, complete and useful so that the decisions made about test takers on the basis of inferences from their test scores can be fully justified. There have been, however, very few published studies on IELTS predictive validity, except some Masters theses (Bellingham, 1995; Broadstock, 1995; Gibson & Rusek, 1992)) and IELTS validation reports (Fiocco, 1992). The reason might be that predictive validation must be done after the test has been released when the pressure is off the test constructors (McNamara, 1996).

Huong (2001) investigated the relationship between the IELTS scores and subsequent academic performance for Vietnamese students sponsored. Based on the results of his research he pointed to the need to carry out IELTS predictive study at specific context so that in his words "the true picture of the relationship between English language ability and academic performance at English medium universities of students from non-English speaking background can be obtained" (p. 21).

Studies on IELTS test fairness

IELTS organization released a detailed booklet in 2006 with the corporation of British Council, IELTS Australia idp and University of Cambridge ESOL examinations. "IELTS: Ensuring quality and fairness in international testing" (2006), is the only source available for research use in the field of test fairness on IELTS examinations. The manuscript has described some of the main features of IELTS and how these contribute to language assessment that is reliable, fair and relevant. It is claimed that IELTS tests meet the requirements of validity, reliability, impact and practicality, which are the four essential criteria which underpin all language assessments developed by Cambridge ESOL, in contexts that are relevant to the ways in which test takers will need to use English in their studies and working lives.

With regard to test impact, the IELTS organization has stimulated a research program called "The IELTS Impact Study Project" with an emphasis on the effects of large-scale tests on educational processes, and on society. This research has confirmed that both teachers and learners believe that IELTS has a positive influence on the classroom and on the learning experience (Hawkey, 2006). Yet, the researcher found the content unsatisfying and marked a serious gap in the literature to promote future fairness studies on IELTS tests. There is no sign of the application of an accredited test-fairness framework which can assess the fairness of these standardized high-stakes tests, on the basis of score-based interpretations, decision making and consequences.

IELTS-related research activities are guided and funded by Cambridge ESOL's Research and Validation Group. Since 1995, more than 50 funded research projects and 70 individual studies have been funded and published by the group. All the studies have examined the validity of IELTS tests and their impact. The studies were motivated by the third approach stated in the introductory chapter which views fairness as linked directly to validity. This connection between fairness and validity suggests a strong possibility for linking fairness back to validity, allowing fairness research and practice to take advantage of a well-defined framework for validity.

Fairness was not treated as a separate facet from validity. Implicit in their approach to test fairness was the standpoint that validity ensures fairness and the lack of validity marks an unfair practice of testing. Besides, the validation studies reported by IELTS or Cambridge ESOL have looked either at ensuring validation procedures at test development and/or test administration phases and the stability of score interpretations across groups in different ways but almost none have addressed the consistency of score-based decisions the broader effects of testing for different groups. None of these studies have examined the differential impact an IELTS test might have on different groups of test takers, and how different manifestations of unfairness may impact the ultimate score interpretation and score-based decisions for a particular assessment.

Methodology

Participants

The participants of this study were comprised of 125 members from three main groups including: IELTS candidates, teachers and raters. Among the total number of participants in various roles, 100 members were candidates who had taken the tests, 20 teachers, and 5 raters. More detailed description of the participants is as follows:

Group 1) A cluster-randomized sample of 100 candidates who participated in students' questionnaire survey. The sample included candidates who took IELTS at Sazmane Sanjesh (National Organization of Educational Testing).

Group 2) A stratified-randomized sample of 20 EFL teachers who participated in teachers' questionnaire survey. The sample included teachers preparing candidates for IELTS at different language institutes in Shiraz and Tehran.

Group 3) Five volunteer raters selected though convenient random sampling who were asked to participate in rater's interview stage of the study. The sample included raters of IELTS at Sazmane Sanjesh (National Organization of Educational Testing)

Materials

The materials used for this study consisted of real samples of IELTS tests released by various organizations in charge of planning, developing, administering, and interpreting the tests and the associated results. To this end, samples of previously administered tests by Sazman-e-Sanjesh Iran and ETS TOEFL were utilized. The corpus was large enough to allow the researcher to apply an argument-based fairness framework on, and to extract regularities.

Design of the study

Due to the complex nature of research, a need was felt to imply a variety of techniques to enrich the result. Therefore, mixed-mode method was selected as the basic design of this study. The design entailed the use of both quantitative and qualitative methods. According to Creswell (2008), mixed methods designs are procedures used to collect, analyze, and integrate quantitative and qualitative data in a study. There were a number of reasons for selecting this design. The most important reason was that a mixed-method design combines qualitative and quantitative modes of research and thereby increases the strengths and eliminates the weaknesses of these approaches. The mixed methods approach is used when the researcher is unsure that one type of approach will adequately address the research problem (Creswell & Plano-Clark, 2007). In most cases, mixing quantitative and qualitative data will yield the most precise and complete picture of the research problem (Teddlie & Tashakkori, 2009).

Other merits of mixed-method designs are the improved validity and generalizability of the piece of research developed. Besides, due to the complexity of some issues in social sciences, including the concept of fairness, multi-level analyses were needed to expand the understanding of complex issues.

The typological organization of this mixed-method research was the concurrent combinations of qualitative and quantitative research (visually represented as QUAL/qual+ QUAN/quan by Johnson & Christnsen, 2004). This type of mixed-method research was a variety of concurrent designs in which we used two methods in a separate and parallel manner and the results were integrated in the interpretation phase. The main purpose of this design was to broaden the research perspective and thus provide a general picture or to test how different findings complement or corroborate each other.

Procedure

Data collection

The investigation of test fairness using an argument-based approach was conducted on some officially released test samples of IELTS. This involved a highly systematic tests content analysis, plus triangulated multi-level analyses at different levels of test development, administration, scoring and interpretation, by using various triangulation techniques and procedures. The procedures encompassed *overlapping methods, stepwise replication, inquiry audits, source triangulation, investigator triangulation* and *location triangulation.*

Tests content analysis

A careful test content analysis was conducted on two sequential phases. In phase one, the content of test samples available to the researcher was examined and analyzed for the purpose of *descriptive note taking*. This way, the researcher endeavored to search for any potential source of bias prior to the second phase. This could provide the researcher with valuable preliminary data that could be interpreted and justified later in the *reflective note talking* process (descriptive and reflective note taking processes are illustrated in detail in the instrument section). The rationale behind this was the fact that this study is a QUAL-QUAN continuum mode research. As such a need was felt to specify a room for any potential observed fact which emerged throughout the research at all phases including the content analysis phase. The second phase of content analysis was conducted based on Xi's (2010) six inferences model. In her model of analysis there is a search to link an observation in a test to any interpretations and uses which are made based on test scores. Therefore, at this phase the researcher tried to analyze the materials to regulate and enlist the observations. The data collected this way was kept in research portfolio for further use in data analysis procedure. This was a complement to phase one and was an attempt to cross-validate the quality of collected data for increasing the credibility of the research. The comprehensive elaboration is followed in data analysis section.

Triangulation

There was a number of concurrent triangulation techniques used for data collection. Overlapping methods which encompassed carefully planned *methodological triangulation*, or multiple ways of data collection (e.g., observations, interviews, and questionnaires), was applied in order to create overlapping (and therefore cross-validating) data. Stepwise replications which involved *time triangulation* or gathering data on multiple occasions (e.g., at the beginning, middle, and end of a semester) was utilized to examine the consistency of the data and interpretations over time. Inquiry audits which involved enlisting an outside expert "auditor" to verify the consistency of agreement among data, research methods, interpretations, conclusions, etc. were also applied. Other forms of triangulation were also used to enhance credibility of the qualitative procedure. Source triangulation which involved gathering data from multiple sources (e.g., people in different roles, like students, teachers, and raters) was used to minimize and understand any differences/biases held by people in various roles. Investigator triangulation which involved using multiple researchers to interpret the data was employed to minimize and understand any differences/biases the researchers may have. Location triangulation which involved gathering data at multiple sites (e.g., some different institutes/organizations) was applied to minimize and understand any differences/biases that might be introduced by the participants in each of the institutions. Where appropriate, quantitative analyses like intercoder/interrater agreement coefficients or other reliability estimates were also used to ensure maximum reliability and validity at data collection phase.

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, raters' semi-structured interviews, field notes and some focus group discussions by various parties with various roles involved in the process of testing.

Questionnaires

For the purpose of this study, two questionnaires (a candidates' questionnaire, and a teachers' questionnaire) were developed, validated, and utilized. The method through which questionnaires were developed was in line with Dörnyei (2007) ***Standard Components***. These components consisted of the followings:

1) T**itle:** This component identified the domain of the investigation and provided the respondent with initial orientation and activated relevant background knowledge and content expectations

2) General introduction: This component described the purpose of the study, describing to the participants that there are no right or wrong answers, explaining confidentiality or anonymity, asking for honest answers; and appreciating their cooperation.

3) Specific instructions: This component clarified the whole process and explained to the participants the way of going through the procedure of answering the questions in the questionnaire.

4) Q**uestionnaires items:** This component consisted of some closed-ended items and some open-ended questions which required the respondents to produce some free writing.

All the questionnaires were developed in English. All questionnaires 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 argument-based approach to fairness. The last part presented some open-ended question through which participants could produce some free writing and reflect their opinions. The objective was to delineate the degree of fairness enjoyed by tests based on these underlying criteria, and to detect any potential source which could support or weaken each argument in this framework. The closed-ended items were designed on a 5-point Likert scale of agreement and frequency. A combination of methods was used to increase the validity and reliability of these questionnaires. The validation process had both qualitative and quantitative essence. Firstly, the application of multi-level triangulation techniques eliminated any potential threat to accountability and credibility of the data. Besides, Cronbach's alpha was employed to ensure the construct validity of questionnaires. The reliability of the questionnaires of the study was examined through conducting a pilot study with a sample of 30 questionnaires completed by candidates, teachers and raters at some available language institutes in Shiraz.

A) Candidates' questionnaire for those who took IELTS

The questionnaire included four parts. The first part was about the candidates themselves. The purpose was to collect some preliminary information about their age, gender, educational status, and their corresponding E-mail address. The second part included some items about the test takers performance on the test. Such factors as the candidates' ability levels, difficulty of test items, familiarity or unfamiliarity with the presented topics, timing, test method, and the results were the locus of this part. The information collected from this part was used to investigate the relationship between factors perceived as affecting the candidates' performance and their scores on IELTS. The third part had ten items and was designed to evaluate the candidates' attitudes toward the test. This included their perceived opinion their scores and their language knowledge, language proficiency, the international credibility of the test, the degree of fairness, some extraneous factors such as time limit and stress, the correspondence between test performance, test scores, and the results interpretation, uses and consequences and the like. The information obtained from this part was a part of the analyses of test impact and fairness. At the end of the questionnaire, candidates were asked to answer three open-ended questions on (1) the type of knowledge or skills needed for a good IELTS score; (2) their viewpoints about the relationship between their performance on the test and the associated interpretation and use of the results; and (3) their own understanding of fair and unfair testing and the potential factors representative of unfairness involved in these tests.

B) Teachers' questionnaire for those candidates preparing to take IELTS

This questionnaire again consisted of four parts. The first part was about the teachers themselves. The purpose was to collect some preliminary information about their level of education, teaching experience, training and the like. The second part consisted of some items about their teaching activities in preparation classes such as test format, previous test samples, test taking techniques, language ability and knowledge, language skills, communication and the like. The third part included items with regard to the teachers' attitudes toward the test. This part was designed to collect some information about teaching and learning, the international credibility of the test, the correspondence between test performance, test scores, and the results interpretation, uses and consequences, fair and unfair test representations and the like. The information obtained from this part was the second part of the analyses of test impact and fairness. At the end of the questionnaire teacher participants were asked to write their comments about (1) their IELTS preparation courses highlighting the most important teaching activities and the reasons of working on them; (2) their viewpoints about the relationship between the students' performance on the test and the associated interpretation and use of the results; and (3) their own understanding of fair and unfair testing and the potential factors representative of unfairness involved in these tests

C) Raters' Interview

A series of one-session semi-structured interviews with IELTS raters were held. The main objective of these interviews was to delineate raters' attitudes and professional opinions with regard to tests, their structures, the degree of representativeness of language ability, scoring procedures, results interpretations, uses, and consequences and also their professional understanding and/or detection of any source of bias or unfairness enjoyed by these tests.

The raters' interview consisted of three parts. First of all there was an appreciation statement followed by a brief introduction about the study, its purpose and the confidentiality of the content of the interview. The main content questions focused on raters' attitudes toward the format, methodology, content, students' success and failure, their understating of fair and unfair testing, sources of bias that may lead to unfair testing at multiple phases of test planning, development, administration, scoring, interpretations, uses and consequences in each of these tests, and the degree of appropriateness of the students score-based interpretations, decisions and consequences. The concluding section was a room for the raters' final remarks with regard to fair testing. The questions of the interview were inspired by all six inferential components of the argument-based framework which underpins the conceptual framework of the present study. Each interview lasted 10-15 minutes and was

recorded. The content of the interviews was transcribed and codified for further analysis procedure.

D) Focus group's interview

At another effort to collect triangulated data, the researcher held one-session focus group interview with all three parties involved in the study, namely, test candidates, teachers, and raters.

Dörnyei (2007) defined the rationale behind focus group interviews as being based on the collective experience of group brainstorming. This involved the emergence of data inspired by participants' thinking, cooperating, challenging, and commenting together on the issue. This within-group interaction is a way of yielding high-quality data because it can create a synergistic environment that results in a deep and insightful discussion (Dörnyei, 2007).

The semi-structured focus group interview (with both open- and closed-ended questions) was conducted with three candidates, three teachers and two raters. The main objective of the focus group discussion is not different with that of semi-structure interviews stated above. It was to delineate people's (at various roles) attitudes and opinions with regard to tests, their structures, the degree of representativeness of language ability, scoring procedures, results interpretations, uses, and consequences, their perception of any source of bias or unfairness of these tests and the degree of appropriateness of score-based interpretations, decisions and consequences. The focus group discussion lasted about 30 minutes and was recorded. The content of the interview was transcribed and codified for further analysis procedure.

E) Field notes

Field notes are the observations written by a researcher throughout the data collection process. Collecting field notes is one of the most influential methods of data collection in ethnographic studies. Besides, social actors endeavor to make informed decisions based on their background knowledge. Garfinkel (1967) has defined background knowledge as knowledge which is 'seen but unnoticed' (p. 118). This includes ethnographic note-taking. Realm of testing is obviously one important territory of social sciences in which life-long consequences will be introduced to future lives of people. Therefore it seemed to the researcher that a short ethnographic attempt is needed to ensure that all aspects are taken into consideration and any sort of drawn conclusion is highly informed.

The process of taking field notes for this study was divided to two parts:

1)Observational (or descriptive) field notes which refer to the notes collected in the research field.

2)Reflective (or analytical) filed notes which refer to the notes taken after leaving the research field.

Descriptive filed notes were taken regularly from all instruments specifically through individual and group interviews and also tests texts analysis. Shortly after each descriptive observation, analytical filed notes were taken. This encompassed substantial notes which reflected the opinion of the researcher(s) based on the marked factors in the descriptive phase. All entries were dated, numbered and organized in order to keep track of them. Field notes for this study contained audio, video, transcriptions, sensory detail, texts, images, descriptions of events and observations, their potential interplay and questions which emerged. It is worth mentioning here that the entire process of note taking was based on six inferences of the underlying the argument-based framework of the study.

Data analysis

The collected, transcribed and codified data was launched by qualitative data analysis software NVivo. This was an attempt to increase the accountability and credibility of this study and to manage the analyses more systematically and rigorously.

There are some stages that were followed in this procedure. First of all the researcher had to *establish a project*. The established project contained all of the documents, coding information, and associated files needed for the analysis. NVivo created a allowed addition of various types of files to the project over time.

The next step was to prepare the files for importing. The main challenge was to ensure that the data is in *text-based, electronic* format. This meant that the recorded interviews, for example, should be digital sound files that are electronic, but still needed to be transcribed in a word processing application in order to make them text-based. On the other hand, since a content analysis of paper-based test texts was performed, there was a need to use a scanner with optical character recognition (OCR) software to transform the documents to electronic format.

After these preparations, the files had to be imported into the created project. This was done by *document explorer*. The document explorer had links to many functions and could be used as a "home base" when coding various documents was done.

After importing the files of the data, some nodes were constructed. There are two type of nodes in this software: free nodes and tree nodes. Free nodes are some free and independent nodes which are not anchored to an established framework. For the purpose of this study tree nodes were constructed. The nodes construction was done under the argument-based approach and its six inferences were used as basic coding structure. Each node was named in line with one inference by the *coder* option of the software, and was provided with a short description.

The search tool is one of NVivo's most powerful functions. This allowed doing numerous types of searches through the data to find excerpts and quotes that match the specified criteria. This really helped the researcher to find related evidence which either supported to weakened each inference.

IELTS constructs and intended uses

The IELTS test was established in 1989 but has been subjected to many changes to date. IELTS assesses the language ability of people who need to study or work where English is the language used in communication. Test takers can opt for either Academic or General Training versions of the tests, according to their personal reasons for taking IELTS (Read & Hayes 2003). All candidates of IELTS test need to complete four modules of Listening, Reading, Speaking and Writing to obtain an IELTS Test Report. Listening and Speaking modules are the same for all candidates but students may choose one of the two versions of Reading and Writing Modules (Academic or General Training).

The intended uses of IELTS vary according to its version. The Academic IELTS is for people planning to study in higher education or seeking professional registration. This option assesses whether a test taker is ready to study or train in the medium of English and is a test of general academic English. Making effective use of written texts in academic work is a skill to be learnt at college or university, not one that students at all

levels should be expected to possess on entry. For this reason, the IELTS tests reflect some features of academic language but do not aim to simulate academic study tasks in their entirety. General Training IELTS is suitable for test takers planning to go to English speaking countries to undertake non-academic training or work experience, or for immigration purposes. This option emphasizes survival skills in a broad social and educational context.

All results are reported on a clear nine-band scale that is easily understood by test users.

The validity argument for IELTS tests

The standpoint adopted in this study is linking fairness to validity. When fairness is linked to validity the way we define, investigate, explain, support or weaken fairness all depends on our conceptualization of the concept of validity. The view of validity which underpins the present study, consider the validation process as building an argument which is organized around a series of inferences that lead the researcher to appropriate score-based interpretations and uses (Kane, 1992, 2004, 2006). The validation process consists of two stages. In the first stage, an interpretive argument which consists of a chain of inferences is constructed. These inferences link test performance to a decision, the warrant supporting each inference and the assumptions upon which the warrant rests. The quality of the interpretive argument in assessed in the second stage of a validity argument.

This approach to validation has been extended and applied widely in language testing realm (Bachman, 2005; Fulcher & Davidson, 2007; Chapelle et al., 2008). The framework had been applied to provide an extensive narrative of the interpretive argument for the IELTS test and an evaluation of the strength of the interpretive argument in the context of a validity argument. In the framework, six types of inferences are essential in linking performance on the IELTS test to the intended score interpretations and uses: *Domain Definition, Evaluation, Generalization, Explanation, Extrapolation,* and *Utilization.*

The illustrations of these six inferential steps and their underlying mechanisms which link observation in a test to score-based interpretation and uses have been provided in the previous chapter in detail. If these six inferences are supported, they add meaning and value to the elicited test performance and as a result score-based decisions are strongly supported. The reverse is also possible.

The fairness argument in a validity argument

The previous section provides an account of the typical inferences underlying the interpretation and use of the tests scores and the warrants supporting the inferences. Every inference addresses a different aspect of the validity argument. The next step is the establishment of a fairness argument which is embedded within a validity argument.

Each inference has a warrant for which there is a set of assumptions needed for backing it. The inferences, warrants, assumptions, and backing are key elements in most interpretive arguments. For investigating test fairness using an argument-based framework, there is a need to articulate a fairness argument by specifying the series of rebuttals which can dispute and challenge the comparability of test scores and score-base interpretations decisions and consequences for different groups of test-takers.

To demonstrate this argument, evidence is needed. This evidence has to reduce the weaken the rebuttals that the score-based interpretations and uses are not comparable across groups due to factors such as construct-irrelevant factors, construct under-representation, inappropriate score reporting practices or decision-making procedures, or unintended uses of the test scores. Failures to rebut any of these counter evidences may weaken the fairness argument and thus compromise the validity of the test.

Results and discussion

This section provides the mechanisms, results and the subsequent discussion of the application of such an argument-based framework to IELTS tests. The following argument describes the most important faces of fairness that is related to each inference.

A) Domain definition: Here the fairness issue is determining if all tasks in the IELTS tests are similarly and equally relevant to, suitable for and representative of all the sub domains for various groups of test takers. For instance, in providing the intended interpretations and uses of the IELTS test, it was stated that the intended uses of IELTS vary according to its version. The Academic IELTS is for people planning to study in higher education or seeking professional registration. This option assesses whether a test taker is ready to study or train in the medium of English and is a test of general academic English. Making effective use of written texts in academic work is a skill to be learnt at college or university, not one that students at all levels should be expected to possess on entry. For this reason, the IELTS tests reflect some features of academic language but do not aim to simulate academic study tasks in their entirety. This claim can raise fairness issues due to the fact that an academic version should assess required academic skills including Writing skill. Additionally, if this claim is proven to be true, then why the format of the test has included the Writing section? Furthermore, there is no clear definition or justifications as what exact academic skills are included or excluded and why. General Training IELTS is suitable for test takers planning to go to English speaking countries to undertake non-academic training or work experience, or for immigration purposes. This option emphasizes survival skills in a broad social and educational context. The fairness issue raised here is the fact that all the tasks in this test may not evaluate some language skills required for living in an English speaking country. Which aspect of living in an English speaking country has been brought into spotlight in the test?

B) Evaluation: As far as the evaluation inference is concerned, the test takers' individual or group differences come to spotlight. These differences among various sub-groups can pose serious fairness issues which can be affected by a variety of factors. These factors can be inconsistent test administration procedures, inappropriate item/task response format, irrelevant factors in the test delivery system, lack of or inappropriate test accommodations for test takers with disabilities, inappropriate test content, test content that under-represents the construct, rubrics that fail to represent the critical skills required in the domain or that represent irrelevant skills and rater bias against certain groups associated with the scoring of the Writing section. Besides, IELTS test format can raise further fairness issues. In IELTS test Listening and Speaking modules are the same for all candidates but students may choose one of the two versions of Reading and Writing Modules based on the version of the test uses (Academic or General Training). The fairness issue posed here is that different groups of test takers with different individual or group differences may face construct underrepresentation based on the version of the test used.

C) Generalization: This inference may raise fairness issues if differences in the generalizability of scores across different subgroups caused by construct-irrelevant factors. When the generalizability of test scores varies across different subgroups of test takers, additional investigations are needed to clarify whether the factors causing the difference are construct-irrelevant. For instance, if test takers' scores vary greatly in two administrations due to a construct irrelevant factor such as washback effect, a serious fairness issue is highlighted.

D) Explanation: The inference of explanation is directly relevant to any fairness issue which is related to differences in factorial structures or in relationships between scores on the IELTS test and other relevant test-based measures for different subgroups of test takers. These issues are caused by construct-irrelevant factors. If an evidence of irrelevant knowledge and/or processes and strategies engaged by some test taker groups to complete the tasks is found, then the explanatory power of the test scores is weakened or rebutted. Washback effect of IELTS test preparation classes is a potential evidence that can endanger the explanatory power of the test.

E) Extrapolation: The extrapolation inference is an attempt to declare that candidates' scores on the IELTS test and on a relevant criterion measure do not differ significantly across different subgroups of test takers. If evidence shows that the relationships between candidates' scores on the IELTS test and on a relevant criterion measure differ across subgroups of test takers, then the inference may be weakened or rebutted and therefore fairness issues may emerge. This potential difference may be found due to construct-irrelevant factors or under-representation of the target domain.

F) Utilization: This inference is with regard to fairness issues which involve comparability of the relevance and usefulness of the assessment results (or test takers' scores) for making decisions for different groups, and also the appropriateness of the decision-making procedures for certain groups. In addition, the impact and consequences of score-based decisions on test taker groups need to be investigated to see if the use of the test causes any non-comparable consequences on different test taker groups.

This argument incorporates that any improper interpretation or decision which is made about the assessment procedure may affect fairness issues relevant to inferences. An inappropriate decision could gain power and strength going through all inferences and as a result it can have consequent impact on the score-based decisions and consequences.

Based on this discussion and the elaborated mechanism the application of the argument-based framework to score-based decisions, interpretation and consequences of the IELTS test is presented below. The discussion illustrates that there is a fairness argument embedded in a validity argument. There are a set of inferences for each of which there are some warrants that support or back inferences. For a warrant to support an inference, some assumptions are needed. There are also some rebuttals that may challenge the validity argument and thus weaken the fairness argument. Based on the nature of evidence found, its power, degree of impact on each inference, the subsequent fairness issue and the extent to which it backs or weakens each inference the degree of fairness of the IELTS test is estimated.

Inferences, warrants and assumptions in the validity argument and counter-arguments in the fairness argument in the IELTS test

Inference 1: *Domain definition*

Warrant supporting inference 1:

Observations of performance on IELTS reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use either in English-medium institutions of higher education or for living in an English speaking country.

Assumptions underlying the warrant:

1. Assessment tasks representing the academic domain can be identified.
2. Critical English language skills, knowledge, and processes needed for study in English medium colleges and universities can be identified.
3. English language skills, knowledge, and processes needed for living in an English speaking country can be identified.
4. Assessment tasks requiring important communicative skills and representing the academic domain can be simulated.

Rebuttals that would weaken the fairness argument:

1. Assessment tasks are not equally representative of the academic domain for different groups.
2. Critical English language skills, knowledge, and processes required for some sub domains are not assessed.
3. English language skills, knowledge and processes required for living in the country with the target language domain are not assessed.
4. Varieties of English included in the test are not representative of these domains.

Inference 2: *Evaluation*

Warrant supporting inference 2:

Observations of performance on IELTS tasks are obtained and evaluated to provide observed scores reflective of targeted academic language abilities and communicative abilities.

Assumptions underlying the warrant:

1. Rubrics for scoring responses are appropriate for providing evidence of targeted language abilities.
2. Rubrics for scoring responses are appropriate for providing evidence of communicative language abilities.
3. The test provides equal opportunities for test takers to demonstrate intended knowledge, skills and abilities.
4. Task administration conditions are appropriate for providing evidence of targeted language abilities.
5. The test delivery system is appropriate for supporting the assessment of targeted language abilities.
6. The statistical characteristics of items, measures and test forms are appropriate for norm-referenced decisions.
7. Appropriate and reasonable accommodations are provided to test takers with disabilities.
8. Raters are well trained and monitored to ensure trustworthy scores.

Rebuttals that would weaken the fairness argument:

1. Rubrics emphasize linguistic features not relevant to the domain or do not include some highly relevant features, biasing toward or against certain groups.
2. Rubrics emphasize communicative features not relevant to the domain or do not include some highly relevant features, biasing toward or against certain groups.
3. Inappropriate test content or construct-irrelevant knowledge and skills engaged by some test items or under-representation of the domain lead to group differences in item/test scores.
4. Inconsistent test administration practices lead to group differences in test scores.

5. Factors in the test delivery system introduce construct-irrelevant differences in test scores across groups.
6. Item/task response format introduces construct-irrelevant differences in test scores across groups.
7. Test takers with physical or learning disabilities are not provided with appropriate accommodations to help demonstrate their relevant abilities.

Inference 3: *Generalization*

Warrant supporting inference 3:

Observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters and occasions.

Assumptions underlying the warrant:

1. A sufficient number of tasks are included on the test to provide stable estimates of test takers' performances.
2. The configuration of tasks on measures is appropriate for the intended interpretation.
3. Appropriate scaling and equating procedures for test scores are used.
4. Task and test specifications are well-defined so that parallel tasks and test forms are created.

Rebuttals that would weaken the fairness argument:

1. Construct-irrelevant factors lead to differences in the generalizability of scores for different groups.

Inference 4: *Explanation*

Warrant supporting inference 4:

Expected scores are attributed to a construct of academic language proficiency or communicative language skills.

Assumptions underlying the warrant:

1. The linguistic knowledge, processes, and strategies required to successfully complete tasks are consistent with theoretical expectations.
2. The communicative knowledge, processes, and strategies required to successfully complete tasks are consistent with theoretical expectations
3. Performance on the test measures relates to performance in other test-based measures of language proficiency as expected theoretically.
4. The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components.
5. Test performance varies according to amount and quality of experience in learning English.

Rebuttals that would weaken the fairness argument:

1. Some assessment tasks engage irrelevant processes and strategies from some test taker groups.
2. Construct-irrelevant factors lead to different factor structures for different groups.
3. Construct-irrelevant factors lead to differences in the relationships between the test of interest and other relevant test-related measures for different groups.

Inference 5: *Extrapolation*

Warrant supporting inference 5:

The construct of academic language proficiency as assessed by the IELTS accounts for the quality of linguistic performance in English-Medium institutions of higher education and English language communicative skills in English speaking countries.

Assumptions underlying the warrant:

Performance on the test is related to other criteria of language proficiency in the academic or communicative context

Rebuttals that would weaken the fairness argument:

1. Inappropriate test content or construct underrepresentation lead to differences in predicting performances on relevant criterion measures for different groups.

Inference 6: *Utilization*

Warrant supporting inference 6:

The test scores and other related information provided to users are relevant and useful for making decisions about admissions, appropriate ESL coursework needed, and the selection of international teaching assistants**.**

Assumptions underlying the warrant:

1. The score reports and other related information provided users support appropriate decision-making.
2. The meaning of test scores is clearly interpreted by admissions officers, and teachers to aid relevant decision-making.
3. Reasonable admissions standards are used to ensure students can cope with the communication demands.
4. The test will have a positive influence on how English is learned and taught around the world.

Rebuttals that would weaken the fairness argument:

1. Inappropriate score aggregation and reporting practices lead to biased decisions for members in some groups.
2. Information about group differences is inappropriately used in decision-making, leading to biased decisions for members in some groups.
3. Factors in the decision-making process such as inappropriate cut score models used lead to biased decisions for some groups.
4. Construct-irrelevant factors, construct underrepresentation, or inappropriate decision making processes cause negative impact on some groups.
5. Different groups of test takers have differential access to test preparation materials, thus impacting the equity of the testing practice.
6. Inappropriate use of test results causes negative impact on some groups.

Interpretation

The argument structure was applied on the first inference called the domain definition. The inference rests on the ground that all tasks in the IELTS test are relevant, suitable and representative of the sub domain for all groups of test takers. This incorporates that all tasks in the IELTS tests are genuine indications of English language skills for use in institutes of higher education for graduate and undergraduate test takers.

The warrant supporting this inference is that observations of performance on the IELTS reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use either in English-medium institutions of higher education or for living in an English speaking country. Certain assumptions are needed for this warrant to hold true.

First of all test tasks should represent the academic domain of language use. If evidence is found that shows all test tasks are not equally representative of the academic domain of language use for different groups of test takers, then the inference will lose power and its validity may be compromised. The data collected and analyzed from the questionnaires, interviews and also tests' content analyses reveled that assessment tasks are not completely representative of the targeted academic domain. In providing the intended interpretations and uses of the IELTS test, it was stated that the intended uses of IELTS vary according to its version. The Academic IELTS is for people planning to study in higher education or seeking professional registration.

This option assesses whether a test taker is ready to study or train in the medium of English and is a test of general academic English. Making effective use of written texts in academic work is a skill to be learnt at college or university, not one that students at all levels should be expected to possess on entry. For this reason, the IELTS tests reflect some features of academic language but do not aim to simulate academic study tasks in their entirety. This claim can raise fairness issues due to the fact that an academic version should assess required academic skills including Writing skill. General Training IELTS is suitable for test takers planning to go to English speaking countries to undertake non-academic training or work experience, or for immigration purposes. This option emphasizes survival skills in a broad social and educational context. The fairness issue raised here is the fact that all the tasks in this test may not evaluate some language skills required for living in an English speaking country. Which aspect of living in an English speaking country has been brought into spotlight in the test?

The second and third assumptions underlying this inference is that the test should assess certain critical language skills and knowledge which are required either for studying at English-medium institutes of higher education or for living in an English speaking country. Academic critical language skills are defined to be the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to manage test tasks. Bloom (1956) has provided a taxonomy that illustrates critical skills. This taxonomy has six levels each of which incorporates certain skills. The first level is memorization in which the candidate produces a solid knowledge base via recognizing, recalling, reciting, naming, defining and describing. The second level is comprehension and manifests an understanding of facts and knowledge with an active application of strategies such as restating, explaining, interpreting, discussing, summarizing and defending. The third level is application in which the candidate extends his/her acquired facts by getting involved in classifying, applying, producing, discovering, modifying and preparing. The next level is analysis in which the test taker breaks the ideas apart and tries to relate to other ideas. This encompasses comparing, contrasting, connecting, relating, categorizing and analyzing. The nest level is synthesis in which the candidate creates new organization of ideas and involves designing, organizing, constructing, composing, revising and developing. The final level is evaluation and encompasses making well-reasoned judgments and decisions via recommending, judging, criticizing, deciding, evaluating and supporting. The analyses of tests' content revealed some limited and inconsistent reference to critical language skills such as referencing, deduction and logical implication in some sub skills, but the wider range of critical language skills that correspond to further targeted domain were neglected. Referring back to Bloom's (1956) taxonomy, we recognize that some critical skills are covered minimally or even ignored in either versions of the test. For example, test takers, teachers and raters reported application of skills such as recognizing, recalling, naming, defining, describing, explaining and interpreting, but not defending, classifying, applying, producing, modifying, comparing, contrasting, relating, categorizing, analyzing, designing, constructing, composing, revising, criticizing, deciding, evaluating and supporting.

The last assumption upon which this inference rests is that the test tasks should encompass important skills which represent the academic domain of language use. These skills go beyond linguistic or critical skills and should incorporate communicative language skills. The rebuttal that weakens this assumption and the associating inference is that the varieties of English included in the test be not representative of the targeted domain. The IELTS tests merely use the British version of language, while test takers taking the test may aim at using the results for admitting or living in an English-medium institution or English-speaking country rather than the UK. This rebuttal weakens the force of the inference and compromises the fairness argument.

Conclusion

This study was an attempt to delineate potential sources of unfair testing which can lead to subsequent unfair score-based decisions and consequences. In this argument-based framework, the inferences address various aspects of validity. As a result, it can be said that various inferences highlight various aspects of fairness. This study provided the mechanisms and the results of the application of such an argument-based framework to IELTS tests. The interpretation part highlighted the findings that compromised the fairness argument in the inference of *Domain definition*.

The results indicated that the degree of fairness enjoyed by the IELTS is compromised based on the counter evidences found against the assumptions that trigger the domain definition inference. This study has focused only of the initial inference. The rest of inferences can be followed in the same way to enrich the results.

References

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

Bellingham, L. (1995) Navigating choppy seas: IELTS as a support for success in higher education. *The TESOLANZ Journal, 3,* 21-28.

Broadstock, H. J. (1995). *The predictive validity of the IELTS and TOEFL: A comparison.* Unpublished master's thesis, University of Melbourne.

Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research*? Second Language Research, 10, 157-187.*

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (Eds.).(2008). *Building a validity argument for the Test of English as a Foreign Language*™. Mahwah, NJ: Lawrence Erlbaum.

Creswell, J.W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

Fiocco, M. (1992). *English proficiency levels of students from a non-English Speaking background: A study of IELTS as an indicator of tertiary success.* Unpublished paper, Centre for International English, Curtin University of Technology.

Gibson, C., & Rusek, W. (1992*). The validity of an overall band score of 6.0 on the IELTS as a predictor of adequate English language level appropriate for study.* Unpublished MA in Applied Linguistics Dissertation, Macquarie.

Huong, T.T.T. (2001).The predictive validity of the International English Language Testing System (IELTS) Test. *Post-Script, 2* (1), 66-96. Faculty of Education at the University of Melbourne. Australia.

Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000.* Cambridge University Press, Cambridge.

Hayes, B. & Read, J. (2004).IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In Cheng, L., Watanabe, Y. &Curtis, A. (Eds.), *Washback in language testing: Research contexts and methods*, Lawrence Erlbaum Associates, Mahwah, NJ, 97–112.

Johnson, R. B. & Christnsen, L. (2004).*Education research: Quantitative, qualitative, and mixed approaches*. Boston: Allyn and Bacon.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement,38*, 319–342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issuesand Practice, 21*, 31–41.

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

Kunnan, A. J. (2004). Test fairness. In Milanovic, M. & Weir C., (Eds.), European language testingin a global context: Proceedings of the ALTE Barcelona Conference (pp. 27–48). Cambridge, UK: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

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 & H. I. Braun (Eds.), *Test validity* (pp. 33 - 45). Hillsdale, N.J. :: Lawrence Erlbaum Associates.

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

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

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences.* Thousand Oaks: SAGE.

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

Zeidner, M. (1987). A comparison of ethnic, sex and age biases in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing, 4*, 55–71.

Bio data:

Forough Rahimi is a PhD candidate in Applied Linguistics at Islamic Azad University, Shiraz Branch, where she presently teaches. She is also a member of 'Young Researchers' Club'. She has published some books and articles and presented at some national and international conferences. Her main areas of interest include critical applied linguistics, teacher education, and specifically language assessment

Mohammad Sadegh Bagheri holds a PhD in TEFL and is currently the Humanities Faculty Dean at Islamic Azad University, Shiraz, Iran. He has published many books and articles and delivered lectures at local, national and international conferences. His main areas of interest and research are international exams, learning strategies, multiple intelligences, e-learning and assessment.

Firooz Sadighi is a professor of TEFL at Islamic Azad University, Shiraz Branch. He has published many books and articles and presented at local, national and international conferences. His main areas of interest and research are first language acquisition and linguistics.

Lotfollah Yarmohammadi is a professor of TEFL at Islamic Azad University, Shiraz Branch. He has published many books and articles. His main areas of interest are critical discourse analysis and socio-linguistics.