In The Name of God

Yazd University
Faculty of Languages and Literature
English Department

A Thesis submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Arts in Teaching English as a Foreign Language

(TEFL)

Title:

A validity Study on the PhD Entrance Exam for the TEFL Students in Iran: An Assessment Use Argument Model

Supervisor:

Dr. Mohammad Javad Rezai

Advisor:

Dr. Ali Mohammad Fazilatfar

By:

Mojdeh Dehbozorgi

October 2014

To My Guardian Angels; My Parents

My Supportive Sisters

And My Only True Friend

Acknowledgments

First and above all, I praise God, the Almighty for providing me this opportunity and granting me the capability to proceed in my postgraduate studies successfully.

My cordial thanks go to Dr. Rezai, for accepting the supervision of my M.A. thesis. This thesis could not have been written without his thoughtful guidance, critical comments, patience, and immense knowledge. I deeply appreciate him for countless hours of reflecting and reading the thesis. He not only served as my supervisor but also encouraged and challenged me throughout my academic program. He guided me through the thesis process, never accepting less than my best efforts

I am also grateful to Dr. Fazilatfar, my advisor, for his warm encouragement, thoughtful guidance, constructive comments through my academic studies, as well as enlightening me the first glance of conducting research. His guidance helped me throughout the process of conducting this research.

I would like to acknowledge and extend my heartfelt gratitude to my beloved parents and sisters whose patience and love enabled me to complete this study. Words alone cannot express what I owe them for their encouragement, material and spiritual support in all aspects of my life. They have consistently helped me keep perspective on what is important in life and shown me how to deal with reality.

Last but not the least , I need to express my gratitude to my true friend, Hossein Tehranifard, whose friendship, hospitality, knowledge, and wisdom have supported and enlightened me over the many years of our friendship.

Abstract

Developing a standardized test is one of the main objectives of educational planners at different levels. The primary concern in test development and use is demonstrating that not only the test scores are reliable, but the interpretations and uses made of the test scores are valid. PhD entrance exam for the TEFL students in Iran is considered a high-stake test which has important consequences for different stakeholders. Hence, its validity is an important factor which should be taken into consideration. The aim of this study was to examine the validity of the PhD entrance exam for the TEFL students in Iran, based on the Assessment Use Argument (AUA) model proposed by Bachman and Palmer (2010). To this end, the validity of each of the four stages, including: assessment record, interpretation, decision, and consequences according to their claims, warrants, and rebuttals was investigated through interview and opinionnaire from those participants who took part in the PhD entrance exam. The findings of this qualitative study indicated that the current status of the PhD entrance exam enjoys a moderate level of validity in terms of the claims pointed out in the AUA model. However, the assessment should be modified in particular areas such as the justification of the interpretation warrants so as to gain a more convincing justification of the assessment use and its results. This, in turn, will make our interpretations useful, appropriate and meaningful lending support to a higher validity of the assessment. The findings of the study would be applicable by future test developers in focusing more on the complexities of developing a high-stake test. Additionally, it would lead to more accurate admission of PhD students leaving more positive consequences for the higher education quality.

Key words: Validity, Assessment Use Argument, High-stake Tests, Ph.D. Language Assessment

Table of Contents

Chapter One Introduction 1

1.1. Preliminaries 3

1.2 Statement of the problem 6

1.3 Purpose of the study 7

1.4 Research questions 8

1.5 Significance of the study 8

1.6 Theoretical framework 9

1.7 Definition of key terms 11

1.8 Outline of the study 12

Chapter Two Review of the Related Literature 15

2.1 Language Assessment 17

2.2 Qualities of Language Tests 19

2.3 Model of Test Usefulness 20

2.3.1 Reliability 20

2.3.2 Construct validity: 21

2.3.3 Authenticity 21

2.3.4 Interactiveness 22

2.3.5 Impact 23

2.3.6 Practicality 24

2.4 Validity 25

2.4.1 Validity vs. Reliability 27

2.4.2 Validity as a Unitary Concept 27

2.4.2.1 Content Relevance (Content Validity) 29

2.4.2.2 Criterion Relatedness (Criterion Validity) 30

2.4.2.3 Construct Validity 30

2.5 Models of Validity 31

2.5.1 Messick's Conceptualization of a Unified Validity 31

2.5.2 Interpretive Argument 35

2.5.3 Evidence-Centered Design 37

2.6 Assessment Use Argument 40

2.6.1 Practical Reasoning 40

2.6.2 Structure of Assessment Use argument 42

2.7 Studies on Validity 49

2.8 Impetus to the present study 51

Chapter Three Methodology 53

3.1. Participants 55

3.2. Instruments 56

3.2.1. Ph.D Entrance Exam 56

3.2.2. Opinionnaire 58

3.2.3. Interview 60

3.3. Procedures 60

3.3.1. Test administration 61

3.3.2. Opinionnaire 61

3.3.3. Interview 61

3.4. Data Analysis 62

Chapter Four Data Analysis 65

4.1. Analysis of the Opinionnaire Results 67

4.1.1. Beneficiality of the Testing Program 67

4.1.2. Beneficiality of the Admission Program and Decisions 69

4.1.3. Treating the Reports Confidentially 70

4.1.4. Understandability of the Assessment Reports 72
4.1.5. Timely Distribution of the Assessment reports..... 73
4.1.6. Consideration of the Educational and Societal Values 74
4.1.7. Consideration of the Educational and Societal Values in the Score Based Decisions 76
4.1.8. Providing Equal chances of Admission 78
4.1.9. Informing the Stakeholders about Decision Making Procedure 79
4.1.10. Enabling the Participants to Perform at Their Highest Level of Ability. 81
4.1.11. Assessment Records as Indicator of the Participants' Ability 82
4.1.12. Assessment Records as Indicator of the Participants' General English Proficiency 83
4.1.13. Assessment Records as Indicator of the Participants' Content Knowledge Level..... 84
4.1.14. Impartiality of the Ph.D Entrance Exam 86
4.1.15. Understandability of the Assessment Records' Procedures..... 87
4.1.16. Treating Individuals Impartially 88
4.1.17. Resemblance between the Test Tasks and TLU Domain..... 89
4.1.18. Resemblance between General English Tasks and Academic Communicative Context..... 90
4.1.19. Relevancy of the test scores 92
4.1.20. Sufficiency of the test scores..... 93
4.1.21. Consistency of the Test Score Across Different Administration of the Exam 94
4.2. Analysis of the Interview Result 95
Chapter Five Discussion and Conclusion..... 107
5.1 Restatement of the Problem 109

5.2 Discussion of the results 111

5.3 Concluding remarks 126

5.4 Implications of the study 128

5.5 Limitations of the study 129

5.6 Suggestions for Further Studies 131

Appendix I 133

Appendix II 185

References 191

List of tables

Table 2.1 Progressive Matrix View of Validity 33

Table 2.2 Layers of Evidence-Centered Design for educational assessments. (Adapted from Mislevy & Haertel 2006, p.8)..... 39

Table 3.1 Detailed information of the Participants 56

Table 4.1 (Frequency of the choices made for beneficiality of the testing program) 69

Table 4.2 Frequency of the choices made for beneficiality of the admission program and decisions..... 70

Table 4.3 Frequency of the choices made for treating the reports confidentially 71

Table 4.4 Frequency of the choices made for understandability of the assessment reports 73

Table 4.5 Frequency of the choices made for timely distribution of the assessment reports 74

Table 4.6 Frequency of the choices made for consideration of the educational and societal values 76

Table 4.7 Frequency of the choices made for consideration of the educational and societal values in the score based decisions 77

Table 4.8 Frequency of the choices made for providing equal chances of admission 79

Table 4.9 Frequency of the choices made for informing the stakeholders about decision making procedure..... 80

Table 4.10 Frequency of the choices made for enabling the participants to perform at their highest level of ability 82

Table 4.11 Frequency of the choices made for assessment records as indicator of the participants' ability 83

Table 4.12 Frequency of the choices made for assessment records as indicator of the participants' general English proficiency 84

Table 4.13 Frequency of the choices made for assessment records as indicator of the participants' content knowledge level 85
Table 4.14 Frequency of the choices made for impartiality of the Ph.D entrance exam.. 87
Table 4.15 Frequency of the choices made for understandability of the assessment records' procedures 88
Table 4.16 Frequency of the choices made for treating individuals impartially 89
Table 4.17 Frequency of the choices made for resemblance between the test tasks and TLU domain 90
Table 4.18 Frequency of the choices made for resemblance between general English tasks and academic communicative context 91
Table 4.19 Frequency of the choices made for relevancy of the test scores 92
Table 4.20 Frequency of the choices made for sufficiency of the test scores 94
Table 4.21 Frequency of the choices made for consistency of the test score across different administration of the exam 95
Table 4.22 Topics mentioned in the interview 95

List of Figures

Figure 1.1 Structure of an assessment use argument. Adapted from Language Assessment in Practice by Bachman and Palmer (2010, p.104) 11

Figure 2.1 Links in an interpretive argument, Adopted from Kane, Crooks,& Cohen (1999, p.9) 36

Figure 2.2 Toulmin's model of practical reasoning.(Adapted from Toulmin,2003, p.97) 41

Figure 2.3 Data- Claim inferential links in AUA. Adapted from Language assessment in practice, (2010), by Lyle Bachman & Adrian Palmer, p.100. 43

Figure 4.1 Beneficiality of the testing program 68

Figure 4.2 Beneficiality of the admission program and decisions 70

Figure 4.3 Treating the reports confidentially 71

Figure 4.4 Understandability of the assessment reports 72

Figure 4.5 Distribution of the assessment reports in timely manner 73

Figure 4.6 Consideration of the educational and societal values 75

Figure 4.7 Consideration of the educational and societal values in the score based decisions 77

Figure 4.8 Providing equal chances of admission 78

Figure 4.9 Informing the stakeholders about decision making procedure 80

Figure 4.10 Enabling the participants to perform at their highest level of ability 81

Figure 4.11 Assessment records as indicator of the participants' ability 83

Figure 4.12 Assessment records as indicator of the participants' general English proficiency 84

Figure 4.13 Assessment records as indicator of the participants' content knowledge level85

Figure 4.14 Impartiality of the Ph.D entrance exam 86

Figure 4.15 Understandability of the assessment records' procedures 87

Figure 4.16 Treating individuals impartially 89

Figure 4.17 Resemblance between the test tasks and TLU domain 90

Figure 4.18 Resemblance between general English tasks and academic communicative context 91

Figure 4.19 Relevancy of the test scores 92

Figure 4.20 Sufficiency of the test scores 93

Figure 4.21 Consistency of the test score across different administration of the exam 94

Figure 4.22 Ideas about sources of the exam 97

Figure 4.23 ideas about two administration of the exam 98

Figure 4.24 ideas about elimination of the GMAT 99

Figure 4.25 ideas about the quality of the questions 100

Figure 4.26 ideas about increasing the number of questions 101

Figure 4.27 ideas about changes in the type of questions 101

Figure 4.28 ideas about participants' satisfaction with the interview 102

Figure 4.29 Ideas about elimination of the interview 103

Figure 4.30 Ideas about changes in the scales 104

Figure 4.31 Ideas about inclusion of open-ended questions 105

Chapter One

Introduction

1.1. Preliminaries

In any learning domain, the most common way of confirming whether the subject matter has been acquired or not is through testing. According to Webster's seventh new collegiate dictionary (1963), a test is a critical examination, observation, or evaluation. Longman advanced American dictionary defines test as a set of questions or problems used as a means of evaluating the abilities, skills or performance of an individual or a group.

A test cannot be considered an appropriate tool for eliciting the test takers' related knowledge, and the test score cannot be a real representative of test takers' actual ability unless it has a certain set of characteristics. Developing a standardized test is one of the main objectives of educational planners at different levels. The primary concern in test development and use is demonstrating not only the test scores are reliable, but that the interpretations and uses we make of the test scores are valid. The other features of a standardized test , as mentioned by Bachman 1996, are authenticity, interactiveness, impact and practicality.

When it comes to the field of language assessment, one of the main requirements is to define language ability. Language ability, as argued by Bachman 1990, "is the ability to use language communicatively" (p.81). This ability has two components: (a) language knowledge and (b) strategic competence. Language knowledge is a combination of organizational knowledge and pragmatic knowledge. Organizational knowledge refers to how utterances, sentences and texts are organized, and divided in two branches of (a) grammatical knowledge, which is required for producing or comprehending formally accurate utterances and sentences, and (b) textual knowledge, which is involved in producing or comprehending the

sequence of units of information in text. Pragmatic knowledge pertains to how utterances or texts and sentences are related to the communicative goals of language users and to the features of language use setting. It consists of both functional and sociolinguist knowledge.

Functional knowledge relates to illocutionary competence, and enables individuals to interpret relationships between utterances, sentences or texts and the intentions of the language users. Sociolinguistic knowledge is the perceptiveness which enables individuals to create or interpret language that is appropriate to a particular language use setting.

Strategic competence refers to a set of metacognitive strategies such as goal setting, appraising, and planning that direct the ways in which individuals use their language knowledge, topical knowledge, affective schemata to communicate (Bachman & Palmer,2010).

In the field of language assessment, the purpose of giving a language test and gaining a test score is to make a link between the performance on the test and the domain of language knowledge the test taker has. In other words, something the test taker can do with language in some language use domain other than the test itself (Bachman, 2004). Mislevy (1996) believes that by using a test score we are essentially reasoning from evidence, using the test score as the evidence for the inferences or interpretations and decisions we want to make. Bachman and Palmer (2010) state that using a language assessment has four stages, including obtaining samples of individuals' language performance, recording their performance, interpreting these records as indicators of some aspect of the test takers' language ability, and making decisions on the basis of these interpretations.

A test score is the representation of test taker ability; especially in cases of high-stake tests which has important consequences for the test taker. Such a score will affect the test taker's life; thus, the consequences of the decision which is made based on the test score should be taken into consideration.

Among all the features of the standardized test, validity is one of the most important factors which is related to the adequacy and appropriateness of the interpretations and uses which make of the test score. Validity is an essential quality of a language test and many test users believe that a valid test measures what it is supposed to measure; therefore, to them a valid test means a good test.

In order to have a justifiable interpretation of the test score, there exist different frameworks for assessment design and assessment use justification, each of which completes and dispels the lacks of the previous models. In the first framework which has been known as Mesickian conceptualization of a unified validity, Mesick (1989) defined validity as an overall evaluative judgment of the degree to which empirical evidences and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment.

The second framework known as Interpretive argument approach was developed by Kane (1990). It is an argument based approach to validity as a means of addressing the Messick's model difficulties. In this framework, the interpretive argument provides both the basis for organizing and prioritizing validity evidence and a means for gauging the progress of the validation effort.

The third model is Evidence-centered assessment design which was introduced by Mislevy, Steinberg, and Almond (1997). It is a conceptual framework which introduces a principled structure for designing, producing and

delivering educational assessments. This model explicates the relationships among the inferences the assessors want to make about the student, what needs to be observed to provide evidence for those inferences and what features of situations evoke that evidence.

The fourth and the latest model of validity, acting as a framework for the current study, is Assessment Use Argument (AUA) which is used for justifying the intended use of an assessment. It was first introduced by Bachman in 2003 but was later modified in 2010. A full description of this model is presented in section 1.6.

Since Ph.D entrance exam for the TEFL students in Iran has been administered communally for the past four years and it is considered as a high-stake exam, its validity is one of the main factors which should be taken into consideration. This study tries to have a critical analysis of the Ph.D entrance exam and discusses its validity from the AUA perspective.

1.2 Statement of the problem

In any high-stake language assessment, the validity of both the test and interpretations of the test results has been considered one of the most challenging factors because the interpretations made based on the test scores will have serious consequences for the test takers and test users. Lack of predetermined sources for the exam, existence of a subjective scoring of the interviews and inappropriate types of questions are all among the factors which, in the test takers' opinion, put the validity of this exam into question.

From all the previously mentioned models of validity, assessment use argument is the one which can provide a better basis for investigating the justifiability of an assessment use. Like many other high stake tests, the Ph.D

entrance exam for the TEFL students in Iran is considered a high-stake test which has important consequences for each test taker, so its validity is an important factor which should be taken into consideration.

The admission criteria for Ph.D studies have been modified since 2010. Previously each university had its own matriculation standards, but now test takers have to participate in a communally entrance exam. To the best knowledge of the researcher, although the system of Ph.D admission has been changed, but very few studies have been conducted on the validity of such a high stake test based on the AUA model.

The current study will examine the validity of Ph.D entrance exam for the TEFL students in Iran on the basis of AUA model. Hence, the validity of each of the four steps, including assessment record, interpretation, decision , and consequences, according to their claims , warrants and rebuttals will be investigated through interview and opinionnaires from the stakeholders perspective.

1.3 Purpose of the study

Based on the problems mentioned regarding the validity of Ph.D entrance exam for the TEFL students in Iran, this study aimed at investigating the following multiple factors.

Firstly, it sought to investigate the degree of the benefits of the consequences of Ph.D entrance exam for test takers.

The second purpose of this study pertained to the equitability of the decisions which are made based on the test score.

Thirdly, the study attempted to discover the degree of the justifiability of the interpretations, and aimed at accounting for the consistency of the assessment records.

Overall, the study tried to find out the justifiability of the Ph.D entrance exam and investigated the validity of such a test based on AUA model in accordance with the warrants and rebuttals that are needed to justify the consequences of Ph.D entrance exam.

1.4 Research questions

This study aimed at examining the validity of Ph.D entrance exam for the TEFL students in Iran, on the basis of AUA model. Based on the problems mentioned, the following research questions were addressed in the current study.

RQ1. To what extent are the consequences of the Ph.D entrance exam beneficial for the TEFL students?

RQ2. To what extent are the decisions made for Ph.D entrance exam equitable for the TEFL students?

RQ3. To what extent are the interpretations made of the Ph.D entrance exam justifiable?

RQ4. To what extent are the assessment records consistent across different administration of the Ph.D entrance exam?

RQ5. To what extent are the uses made of the Ph.D entrance exam justifiable?

1.5 Significance of the study

As it was mentioned previously, few studies have been conducted on the validity of language assessments based on AUA framework. Particularly, no studies have been reported on investigating the validity of Ph.D entrance exam in Iran.