On the Effect of Perceived Test Impact on Test Anxiety and Attainment of PhD Candidates: Studying English Proficiency Test in Islamic Azad University

Hamid Ashraf[1]* and Mahmood Khosravani[2]

[1]*Assistant professor, Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-e Heydarieh, Iran*
[2]*PhD candidate, Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-e Heydarieh, Iran*
Corresponding Author: Mahmood Khosravani, E-mail: mdkhosravani@yahoo.com

ARTICLE INFORMATION	ABSTRACT
	Research findings has shown that high stakes test bring about unfortunate consequences inside and beyond educational context which are known as washback and impact respectively. English proficiency test (EPT) of Islamic Azad University (IAU) in Iran, as an obligatory part of the PhD graduation program, is not an exception. Thus, considering possible adverse consequences of the test, the current research was an attempt to investigate the effect of perceived test impacts on test anxiety and attainment of the examinees. To this aim, a mixed method design was employed in which quantitative and qualitative data were collected from 81 PhD candidates in different disciplines of the humanities. In the first phase, adopting a framework for test impact perception, a structured interview was conducted on the basis of which participants were assigned to NPTI (non-perceived test impact), MPTI (moderately perceived test impact) and SPTI (severely perceived test impact). Then performance of these was compared in an adopted test anxiety scale (five-point Likert scale) and EPT of IAU. Kruskal-Wallis and one-way ANOVA were used to analyze the data. The findings on test anxiety scale indicated a significant difference between groups ($\chi2$ (2) = 52.618, p = 0.000.). However, the group means were not significantly different in EPT (F (2, 77) = 1.975, p = 0.146). Findings obtained from test anxiety supported previous studies carried out in the literature while those from EPT stood in sharp contrast. An experimental design is required to control other variables that interact with these findings. Yet, these findings have implications for applied linguistics, researchers and educational decision makers.

1. Introduction

Bachman and Palmer (1996) state that a number of phenomena work on the interests of stake holders and those who are directly involved in tests. These phenomena are categorized in two major consequences known as wash back and test impact. These terms are employed in different fields of research, and encompass different dimensions of the research in testing discipline. In this conceptualization, washback is more frequently used to refer to the effects of tests on teaching and learning at the classroom level while impact entails the effects that a test may have on individuals, policies, or practices, within the classroom, the school, the educational system, or the society as a whole. A brief review of the literature in the last three decades shows that these terms are seen as being inherently one (mostly in traditional views), as some believe that washback is part of impact or a dimension of it, however in more recent literature the effects of testing on teaching and learning have been traditionally associated with test validity (consequential validity) where washback is considered as only one form of testing consequences that need to be weighted in evaluating validity (Messick, 1989).

In fact, Messick (1989, pp. 20-29) had already postulated that consequential validity placed under the more general term of construct validity is a unified concept which deals with the impacts of tests on teaching, learning, individual test takers, teachers, society, and educational system within a country. Kiani, et al (2009) illuminated by Bachman and Palmer's taxonomy (1996) stated that, for a practical employment of these conceptualizations and terminologies in research and education, the

consequences of tests on teaching and learning are viewed as washback effects while the consequences on individual stake holders such as learners, teachers, parents, test takers' family members, and society are considered as impact. Concerning the term impact, Taylor (2005) states that language tests can have consequences beyond just the classroom and exert a significant impact on the career or life chances of individual test takers such as educational or employment opportunities or exert influence on educational systems and society at large such as using test results to make decisions about school curriculum planning, immigration policy, or professional registration for doctors.

English Proficiency Test (EPT) of Islamic Azad University is a large-scale and high-stake test that exerts impacts of different kinds on PhD Candidates and educational components involved. In spite of these impacts, a large gap in the literature exists and different dimensions of these impacts are not examined into through a profound research. In line with Chen and Garbe (2011) who hold that evaluation of a program requires both goal attainments (if the program can achieve goals established within it) and system integration issues (if the program functions in the context in which it is implemented), the current research was attempt to shed light on two variables of these impacts i.e. test anxiety and attainment of PhD candidates who are required to pass the test as a part of their educational fulfillment. Thus, adopting a theoretical framework, the present study was intended to examine the effect of the perceived tests impacts on test anxiety and attainment of Iranian PhD candidates in IAU. Therefore, the following research question was formulated:

Does perceived impact of EPT affect test anxiety and attainment of Iranian PhD candidates in IAU?

2. Literature Review
The history of investigating the effects of tests started from 1950s and 1960s when practitioners and scholars in general education started asserting that exams (high-stakes tests) may exert influence of some kinds on teachers, learners, and other stake-holders. For instance, Vernon (1956) reported that teachers, through distorting and changing the curriculum, show their inclination toward ignoring the content that enjoys little chance of appearance in the exam. A reverse effect was also reported by Davies (1968) discussed that tests and test materials were likely to be used as teaching devices to narrow down the teaching content, emphasizing the effects of test on teaching and curriculum. This phenomenon was later studied by Davies (1968), Swain (1985), and Hughes (1989), Bailey (1999), Taylor (2005) and therefore attracted attention of practitioners and researchers in the last decades.

In particular, in impact studies, according to McNamara (2006), Messick's (1989) validity model was the most influential model in language assessment research especially in the field of validation. From the perspective of validity, impact is referred to as a quality which needs to be designed into the test. In addition to that, Bachman and Palmer (1996) place "impact" under the notion of test usefulness to sit beside qualities such as reliability, construct validity, authenticity and interactiveness, arguing that the overall usefulness of a test is a function of these qualities. Later Bachman and Palmer (2010) proposed an assessment-use-argument framework and linked consequences of test use to construct validity.

Later, Cizek (2012) claimed that unintended adverse social consequences should be incorporated into a code of ethics or a code of fair testing practice in issue already established by Bachman (2005). The ethics perspective to "impact" however, focuses on the sociopolitical dimension of language testing. Yet, in this line of impact research too, the critical ethical and sociopolitical perspective presents methodological challenges by giving prominence to value implications underlying test interpretation and test use.

In a study, Hammond and Forrester (2005) investigated the consequences of student testing for teaching and teacher quality. The findings of the study showed that uses of student testing for accountability purposes has some influences on curriculum, classroom practices, student opportunities, and the decision making processes of schools and teachers. They concluded that the kinds of assessments used, consequences attached to these tests, affect both teaching and teacher quality.

Also, Kiani, Alibakhshi and Akbari (2009) investigated the test consequences of an ESP test on Iranian language learners and found that the test imposes enormously negative psychological and social consequences on the examinees as perceived by them. In adition, in a study by Pan (2009), the appropriateness and consequences of test use were investigated. It was concluded that, to evaluate the appropriateness of test use, it is essential to investigate the consequences of test use under test, educational, and social considerations and since those who make decisions of test-use are not experts at test validation, developing a local model of test impacts and consequences serves as guideline for decision makers to evaluate the appropriateness of test use.

 In addition, a good bulk of studies have inquired into the effects of testing on language learning and teaching. Mahmoudi (2014) studied the washback effect of Iranian national university entrance exam on English teaching and learning in pre-university. The findings of the research showed that many factors affect the process of English language learning and teaching in

pre-university schools. Furthermore, Sadighi, et al. (2018) examined the possible washback effect of the university entrance exam and concluded that the perceptions of EFL fourth-graders' teachers compel them to use supplementary materials geared toward university entrance exams' demand, contrasting with the main textbook's educational objectives.

3. Method
In the current study, a mixed method approach which involves different combinations of qualitative and quantitative techniques, either at the data collection or at the analysis levels, was employed. A mixed-methods approach can be employed when a researcher combines both quantitative and qualitative data to understand and investigate a research problem (Creswell, 2008).

3.1 Participants
In order to answer the research question, 112 PhD candidates who participated in IAU-EPT were selected through volunteer sampling. All participants who aged between 29 to 63 years old were PhD candidates in different disciplines of humanities that were selected in a testing site in IAU in different months. In the first month in the testing site of female candidates, 46 examinees participated in the study while in the later, in the testing site of male candidates, 66 examinees participated in the study. To complete the data collection process, all the participants were asked to mail their EPT score but only 39 female candidate and 42 male candidates replied our emails and sent their scores. Finally, the data collected from 81 examinees were selected for further analysis.

3.2 Instruments
In order to answer the questions of the research, two main instruments were used:
A) The first instrument used in this test was a structured interview guided by the theoretical findings of Kiani, et al (2009), which encompassed two major dimensions of test impacts i.e. psychological dimension and social dimension. Each of these dimensions included 9 different components. This instrument was used to discover whether the participants perceive test impacts or not.
B) The second instrument used in this study was a test anxiety scale adopted from In'nami (2006). This scale was a modified version of Sarason (1975). This five-point Likert scale encompasses 37 different items, evaluating test relevant and test-irrelevant factors. This instrument was used to investigate whether perception of test impacts affects anxiety of EPT examinees.

3.3 Procedures
In the very beginning of the research, 112 male and female PhD candidates in different disciplines of humanities were selected through volunteer sampling. Using the first instrument of the research (perception of test impacts), the participants underwent a face to face structured interview just after taking the EPT exam in an IAU site. Their interviews were recorded. Also, they were administered the second instrument (test anxiety scale) to be filled by them. In order to investigate the effect on attainment of examinees, considering ethical considerations of the research and informed consent, emails of participants as well as their demographic information were collected. They were asked to mail their EPT score, replying to an email sent by research group. In the second round of the research, qualitative data obtained from the interview were analyzed through content analysis, on the basis of which respondent were categorized into three groups of Not Perceiving Test Impacts (NPTI), Moderately Perceived Test Impacts (MPTI), and Severely Perceiving Test Impacts (SPTI). Since only 81 participants emailed their EPT score, quantitative manipulations were done only for this candidate and 31 participants who did replied the email were removed from the groups.

For test anxiety, considering the non-parametric nature of the data obtained from the Likert scale, the data were entered into SPSS (Version 21), changed into mean rank and analyzed through Kruskal-Wallis test to investigate any possible significant difference between the mean ranks of the three groups. Also, the EPT scores of participants were collected, entered into SPSS and, considering the nature of the data and assumptions involved, analyzed through one way analysis of variance (One-Way ANOVA). A schematic representation of the procedures of the study is shown table 1.

Table 1: *The schematic representation of the procedure of the study*

Step 1		Step 2	Step 3			Step 4	
Sampling		Structured face to face interview (Audio-Recorded)	Grouping (based on Interview results)			Comparison	
Early Sample	Final Sample	Content Analysis	NPTI	MPTI	SPTI	Test Anxiety	Attainment
112	81		18	27	36	Kruskal-Wallis	One-Way ANOVA

4. Results and Discussion

In different steps of the study, different kinds of results were obtained. In step two of the study, as shown in table.1), content analysis was performed and on the basis of the frequency of the propositions found for components of perceived test impact (Kiani, et al, 2009), the participants were placed in one of NPTI, MPTI and SPTI groups. Then, their performances on two variables (Test Anxiety and EPT) were compared, aimed at identifying in statistically significant difference. The results of these statistical comparisons are reported below:

A)   *Test Anxiety*

Since the scale used for evaluating test anxiety of PhD candidates was a Likert one, for ease of statistical manipulations, the collected five-point Likert data were rendered into mean ranks in SPSS and analyzed through Kruskal-Wallis. The findings are reported in following tables:
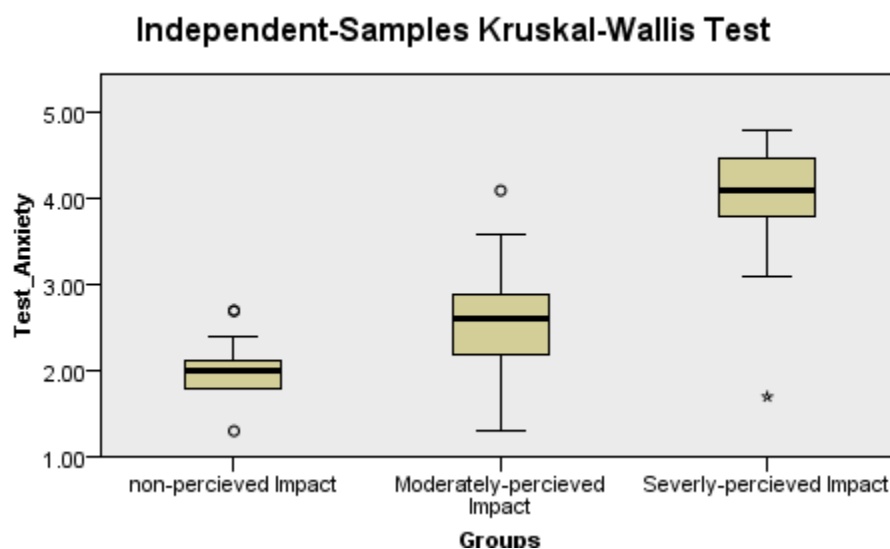
Table 2: *Test Statistics[a,b] for Test Anxiety*

	Test Anxiety
Chi-Square	52.618
df	2
Asymp. Sig.	**0.000**

This table shows that the p values for test anxiety is more than 0.05 indicating that these three groups perform differently on the scale. In sum, a Kruskal-Wallis H test indicated a statistically significant difference in mean scores between NPTI, MPTI and SPTI groups , $\chi 2$ (2) = 52.618, p = 0.000. In order to find out where the difference lies, a post-hoc multiple comparison was performed, the results of which are shown below:

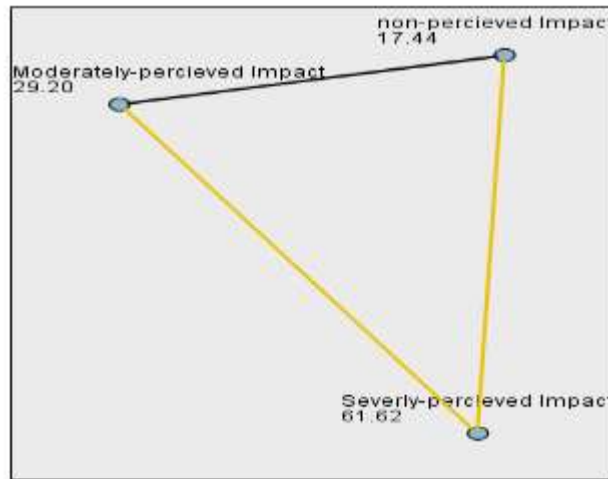Table 3: *Hypothesis Test Summary of Kruskal-Wallis (Asymptotic significances for test anxiety)*

Null Hypotheses	Test	Sig	Decision
The distribution of Test Anxiety is the same across categories of Groups	Kruskal-Wallis	**0.000**	Reject the null hypothesis

Figure 1: *Visualization of Independent- Samples Kruskal-Wallis test for Test Anxiety*

As it is shown in this figure, the mean rank for NPTI, MPTI and SPTI groups stands at different points of a Likert scale, submitting proofs on the difference in mean rank of the three groups.

Figure 2: *Pair-wise comparison visualization of groups' performance in test anxiety scale*

As it is shown in the picture, each node represents the average rank of groups. Furthermore, the lines drawn between nodes show the significance of the differences between them. Accordingly, there is a significant difference between NPTI and SPTI. Also there is a significant difference between MPTI and SPTI. However, the difference between MPTI and NPTI is not significant as the line drawn between is colored differently. Pair-wise comparisons statistics are also shown in the following table:

Table 4: *Pair-wise comparison statistics of groups*

Samples	Test Statistics	Std. E	Std. Test Statistics	Sig	Adj. Sig
NPTI-MPTI	-11.759	7.151	- 1.644	0.100	**0.300**
NPTI-SPTI	- 44.181	6.784	-6.513	0.000	**0.000**
MPTI-SPTI	- 32.421	5.983	-5.419	0.000	**0.000**

The data presented in this table also shows that the difference between NPTI and MPTI is not significant (p=0.300) but the differences between these two groups and SPTI is significant (p=0.000). This means that different perception of test impacts has brought different levels of test anxiety among the examinees.

B) Attainment

EPT administered by IAU, brings about a continuous score between 0 and 100. The scored collected from 81 participants of the study were entered into SPSS and analyzed through ANOVA, the findings of which are reported in the following tables:

Table 5: *ANOVA Output for EPT*

		Sum of Squares	df	Mean Square	F	**Sig.**
EPT	Between Groups	633.890	2	316.945	1.975	**0.146**
	Within Groups	12354.310	77	160.446		
	Total	12988.200	79			

As it is indicated in Table 5, the significance value for EPT is *0.146* (i.e., $p = 0.146$) which is above 0.05 and, therefore, there is not a statistically significant difference in the mean of scores between the different groups. The results of the post-hoc in table 6, submitted more detailed evidence of the difference.

Table 6: *Post-Hoc Multiple Comparison for EPT*

Tukey HSD

Dependent Variable	(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.
EPT Score	NPTI	**MPTI**	-7.611	3.854	**0.125**
		SPTI	-5.214	3.674	**0.336**
	MPTI	**NPTI**	7.611	3.854	**0.125**
		SPTI	2.397	3.244	**0.741**
	SPTI	**NPTI**	5.214	3.674	**0.336**
		MPTI	-2.397	3.244	**0.741**

In sum, there was not a statistically significant difference between groups as determined by one-way ANOVA (F (2, 77) = 1.975, p = 0.146). This shows that different perception of test impacts might not be a significantly effective variable in EPT scores of the PhD candidates.

Taylor (2005), believes that language tests result in consequences which are sometimes beyond the educational context to include career or life chances of individual test takers (educational/employment opportunities). Bachman (1990) had already noticed these consequences and therefore, used the term impact to describe these consequences of tests. Besides, Bachman and Palmer (1996) believe that high stake tests result in macro effects within society. Exit EPT test of IAU is a very high stakes test which has affected PhD candidates to a great extent, imposing on them financial or other social consequences. The findings of the present study are in line with Kiani, et al (2009) who reported that stress, depressing and anxiety are among the negative impacts of high-stakes test on Iranian language learners. They also reported that the financial aspects of the test negatively affect test takers. The findings of the current research supported Cheraghian, et al (2008) that observed no meaningful relationship between test anxiety and academic performance. In addition, these finding are in harmony with Aydin (2006) which stated that negative attitudes of learners towards language learning, test invalidity, fear of negative evaluation, bad experiences on tests, time limitation pressure, difficulty of course contents and parental expectations brings about test anxiety. Though SPTI group had higher levels of test anxiety, the attainment part of the study stood in sharp contrast with Ellis (1994) and Gardner (1985) who had asserted that language anxiety is an effective factor that affects achievement negatively. Also, unlike Cakici (2016) that demonstrated a significant reverse correlation between students' foreign language anxiety and their language achievement, the perception of negative impacts of EPT test did not significantly affected the attainment of PhD candidates.

5. Conclusion

The current study was conducted to investigate whether PhD candidates with different perception of negative consequences of test impacts have different test anxiety and EPT scores or not. The findings show that the more sever perception of negative social and psychological impacts of test result in greater test anxiety while their performance on EPT of IUA is not significantly affected by the classification of different perception groups (MPTI, NPTI and SPTI). Further research is required to investigate the correlation between test anxiety and EPT attainment. Also, in order to control variables involved in perceiving test consequences, test anxiety and attainment, an experimental or quasi-experimental research design is required. Therefore, the generalization and conclusion derived from the current study is required to be down with caution. Yet, these findings have significant implications for language researchers, educational policy makers, educational consolers and PhD candidates of IAU.

References

[1] Aydin, S. (2006). Test anxiety in foreign language learning. *Journal of graduate school of social sciences of Balikesir University, 9,* 145-160.

[2] Bachman, L, & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

[3] Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice: Developing and using language assessments in the real world*. Oxford University Press.

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

[5] Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

[6] Cakici, D.(2016). The correlation among EFL learners' test anxiety, foreign language anxiety and language achievement. *English Language Teaching, 9(8),*109-203.

[7]   Cheraghian, B., Fereidooni Moghadam, M., Baraz-Pardjani, S. H., & Bavarsad, N. (2008). Test anxiety and its relationship with academic performance among nursing students. *Journal of Knowledge, and Health, 3,*25-29.

[8]   Darling-Hammond, L & Rustique-forrester, E. (2005). *The consequences of student testing for teaching and teacher quality*. Yearbook of the National Society for the Study of Education. 104. 289 - 319. 10.1111/j.1744-7984.2005.00034.x.

[9]   Cizek, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31–43.

[10]   Creswell, J.W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson Education.

[11]   Chen, H. T., & Garbe, P. (2011). Assessing program outcomes from the bottom-up approach: An innovative perspective to outcome evaluation. In H. T. Chen, S. L. Donaldson, & M. M. Mark (Eds.), Advancing validity in outcome evaluation: Theory and practice (No. 130, summer, pp. 93-106). Jossey-Bass.

[12]   Davies, A. (1968). *Language testing symposium: A psycholinguistic approach*. Oxford University Press.

[13]   Ellis, R.(1994). *The study of second language acquisition*. Oxford University Press.

[14]   Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. Edward Arnold.

[15]   Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

[16]   In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System*, *34*(3), 317-40.

[17]   Kiani, G.R; Alibakhshi, G & Akbari, R. (2009). On the consequential validity of ESP tests: a qualitative study in Iran. *The journal of applied linguistics, 2*(1), 103-126.

[18]   Mahmoudi, L. (2014). The washback effect of Iranian national university entrance exam (INUEE) on pre-university English teaching and learning (Unpublished doctoral dissertation). University of Malaya, Malaysia.

[19]   Messick, S. (1989). Validity. In: Linn, R.L. (Ed.), Educational Measurement (3rd Edition), (pp.13-103). American Council on Education/Macmillan, New York.

[20]   McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51.

[21]   Pan, Y.C. (2009). Evaluating the appropriateness and consequences of test use. *Colombian applied linguistics journal,11,* 95-103.

[22]   Sarason, I.G. (1975). The Test Anxiety Scale: Concept and research. In I.G. Sarason & C.D. Spielberger (Eds). *Stress and Anxiety* (pp. 193-217). Hemisphere.

[23]   Sadighi, S. et al. (2018), Wash-back effect of Iranian students' pre-university English textbook and university entrance examinations: Teachers-based perspectives. *African Educational Research Journal, 6*(4), 303-316.

[24]   Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), Input and Second Language Acquisition (pp. 235-253). Rowley, MS: Newbury House.

[25]   Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 1-2.

[26]   Vernon, P. E. (1956). The Measurement of Abilities. (2nd. Ed.), University of London Press