Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 7, No. 2, October 2017
ISSN 2476-5880

On the Development and Validation of a Scale of Test Impact on Test Takers (TITT)

Mahmood Samaie[1], SaeedehMohammadi[2]

Received: 03 April 2017 Accepted: 20 September 2017

Abstract

Test impact, widely recognized as the influence of testing on learning and teaching, affects a set of stakeholders including test takers. This study defines the construct of test impact on test takers and describes the construction and validation of the scale of test impact on test takers (TITT). 410 participants having passed a language test in University Entrance Examination (UEE) were asked to answer the questionnaire containing 64 items. Exploratory factor analysis was applied in the study and yielded evidence for the expected five factor structure of the TITT scale, including the components of test results, test awareness, test experience, test importance, and test socio-cognitive effects. The final TITT Scale and its subscales consisting of 56 items demonstrated an acceptable internal consistency and expected levels of stability of the responses across time. Cronbach's alpha was .93 for the global TITT scale and between .78 and .88 for the five subscales. Implications are discussed and suggestions are provided for possible utilization and improvement of the scale, and future validity testing.

Keywords: *scale development, test impact, test impact on test takers (TITT), TITT scale*

1. Introduction

Washback or impact is generally understood as the influence of language tests on learning and teaching in relation to factors such as the individual learners and teacher's attitudes and behavior, the classroom environment, and the choice and use of teaching/learning materials (Alderson & Wall, 1993). Washback has been extensively discussed as an important aspect of consequential validity. As Messick (1996) argues, the consequential aspect as a dimension of construct validity appraises "the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to

[1] English Department, Ilam University, Iran

[2] English Department, Ilam University, Iran *(corresponding author)*, Email: saeedehm63@gmail.com

issues of bias, fairness and distributive justice (Messick 1980; 1989), as well as to washback" (p. 249).Different stakeholders (e.g. test takers, teachers, administrators, course designers, and materials developers) can be affected by the significant effects of the tests. According to Bachman and Palmer (1996), the impact of test use operates at two levels: at the micro level, in terms of the individuals including teachers and test takers who are influenced by a particular test use, and at the macro level, in terms of the educational system or society at large.

We adopt the micro level to analyze whether the test influences the perceptions, feedbacks and practices of learner. However, it's necessary to interpret test impacts upon individual learners considering the specified purpose, construct definition, test takers' characteristics, and values and goals of the society and educational program in which the test is used. By the same token, washback can vary from negative through neutral to beneficial. This also concerns the context bounded-ness of analyzing test impact as mentioned by Bachman and Palmer (1996). In other words, what is considered to be positive depends on the position adopted by those making the judgment and the educational goals he or she espouses (Hamp-Lyons, 1997).

Some prominent scholars have accounted for the paramount significance of test impact (e.g., Alderson & Wall, 1993; Bailey, 1996). It is also considered as one of the qualities of test usefulness (Bachman & Palmer, 1996) and has been recognized as one of the main criteria for developing and evaluating language tests needed to be taken into account during the ongoing iterative process of test development including purpose and design, operationalization, and administration (Bachman & Palmer, 1996; Read & Chappelle, 2001). In addition, test developers are accountable for the uses (decisions and consequences) that stakeholders make based upon their tests (Bachman & Palmer, 2010).

Given this, despite the recognized importance of washback, the exact nature of impact upon different stakeholders and involving mechanisms are not empirically documented. In a study on the key studies conducted with different sections of Messick's validity framework, Kunnan (1998), maintains that the list of studies in the consequential basis section and systematic attempts to understand the washback effect compared to the list in evidential basis section is smaller and more recent; "it is here that yawning gaps lie" (p. 6). This idea of little empirical evidence for the existence of washback is also held by Andrews (1994). Thus, we know little about students' perceptions of tests and even less about how new tests influence what students know and can do. On the same lines, related to the vagueness of test influence on learners and test takers, Bailey (1999) insists that "much more research is needed…to see whether and how these washback effects play out in the attitudes and behavior of language learners" (p. 13).

However, although the focus of impact studies has most often been on teachers and classroom practices, studies of learners have also begun to appear recently (e.g., Ross, 2005;

Green, 2006; Green, 2007; Xie, 2008; Xie& Andrews, 2013; Zhan & Wan, 2014). Green (2006), for instance, distributing 24 questions related to writing instruction among learners, aimed to examine impact of IELTS test on learning. The results indicated that learner perceptions of course outcomes are affected by the course focus reported by teachers, but that the relationship is not deterministic. Green (2007) investigated whether dedicated test preparation classes gave learners an advantage in improving their writing test scores. Findings indicated no clear advantage for focused test preparation. Zhan and Wan (2014) also examined how the revised College English Test Band 4 influenced Chinese non-English major undergraduates' out-of-class English learning practices over time through diary entries and interview. They identified the dynamic nature of washback on individual learners. Now, with reference to the focus of the present study, the techniques and instruments which have been used to investigate the test impact on different stakeholders are reviewed in the following section.

1.1 Measuring test impact

Concerning instruments to measure impact, some studies have utilized classroom observations and case studies to examine test impact (e.g., Zhan, 2003; Watanabe, 2004; Green, 2007).

Moreover, plethora of research has incorporated interviews and similar qualitative designs to examine washback (e.g., Hamp-Lyons, 1996; Kiani, Alibakhshi, & Akbari, 2009; Ramezaney, 2014; Zhan and Wan, 2014). Hamp-Lyons (1996), for example, in a study of TOEFL preparation courses in the United States, interviewed students in groups of 3 to 12 people at three different institutions. Similarly, Ramezaney (2014), through interviews, examined the nature and scope of the university entrance exam's impact on the EFL teachers' curricular planning and instruction techniques. The findings indicated that from the teachers' perspective, Iranian UEE has a significant influence on teachers' curricular planning and instruction techniques. Alderson and Özmen (2011) employing a qualitative study investigated washback effects of inter-university foreign language examination (ÜDS) on candidate academics and reported for its negative washback effect.

Yet, questionnaires and quantitative designs have been lately employed in the investigation of washback on different stakeholders (e.g. Green, 2006; Moore, Stroupe, &Mahony, 2009; Pizzaro, 2010; Akpinar&Cakildere, 2013; Green, 2014). For instance, Pizzaro (2010), exploring washback effects of a high-stakes English test, employed a teacher's questionnaire consisting of four main sections which comprised overall twenty-five, mostly closed-ended questions. It was found that the test was clearly affecting curriculum and materials and influenced teachers' methodology. However, there is no description regarding the psychometric properties of the questionnaire. Akpinar and Cakildere (2013) also, applying a 26 item questionnaire, investigated the impact of two high-stakes tests on productive and receptive

skills addressing the academic personnel. They found significant differences between different skills with respect to the impact of the two tests. As discussed, the same approach was applied by Green (2006). Green (2014) also examined the impact of the test of English for academic purposes (TEAP) through a questionnaire administered in a private university in Japan. The questionnaire showed that although the TEAP is not yet well-known, the changes it would bring to the entrance exam system are generally well regarded by high school students and teachers. These studies employed questionnaires involving open-ended and closed items for exploring test impact about the validity and reliability of which nothing has been mentioned.

In sum, although washback has been extensively discussed in recent years, research on washback incorporates a few empirical studies which explore the potential consequences and implications of language tests on teaching and learning. It's worth emphasizing that, in most of the washback studies, the methods used for data collection include interviews, testing measures, classroom observations or a combination of these involving content analysis, document review, and case studies (Alderson & Wall, 1993; Wall, 2000) which result in threats to validity. Qualitative designs have been extensively utilized in washback studies and the few existing scales incorporating mixed method approaches have mostly focused upon some specific tests such as IELTS or TOEFL or other high stakes tests. According to Green (2013), a deal of washback research into participants has been descriptive and exploratory. Although very rare efforts have been made to generate quantitative scales for examining washback, no comprehensive questionnaire has been developed and validated so as to be used in different test settings and for any kind of test. As Ozmen (2011) believes, to understand the nature of impact, it is important to refer to quantitative data where and when necessary. Indeed, to have a better understating of the nature of impact on test takers, a combination of both quantitative and qualitative designs is needed. Yet, the lack of reliable quantitative instruments to investigate test impact reveals the need to construct such tools.

Besides, no specific quantitative instrument has been developed to investigate test impact on test takers. Moreover, data collection instruments have not been developed and established on a well-grounded and organized theoretical basis. So, the lack of such basis necessitates the conceptual development of the structure of a potential instrument of test impact on test takers.

1.2 Theoretical foundation for the scale of test impact on test takers (TITT)

To conceptually guide the current study, we propose a taxonomy which involves the significant elements based upon an analysis of the fundamental literature and underlying theoretical views toward test impact on test takers (e.g., Hughes, 1993; Alderson & Wall, 1993; Bachman & Palmer, 1996; Bailey, 1996; Bailey, 1999; Bachman & Palmer, 2010). Indeed, the lack of a systematically organized and coherent theoretical framework for developing a scale for exploring

 ISSN 2476-5880

impact on test takers hastens the need for such taxonomy. Fig. 1 is thus the basis for developing the questionnaire in the current study and correspondingly decreases the threats to construct validity of the upcoming questionnaire.The following taxonomy represents our proposed taxonomy concerning the factor structure of the TITT scale needing to be confirmed in the exploratory stage of the present study. Considering whether the impact on test taker occurs before taking the test or after that, each element is designated by the time of test impact.

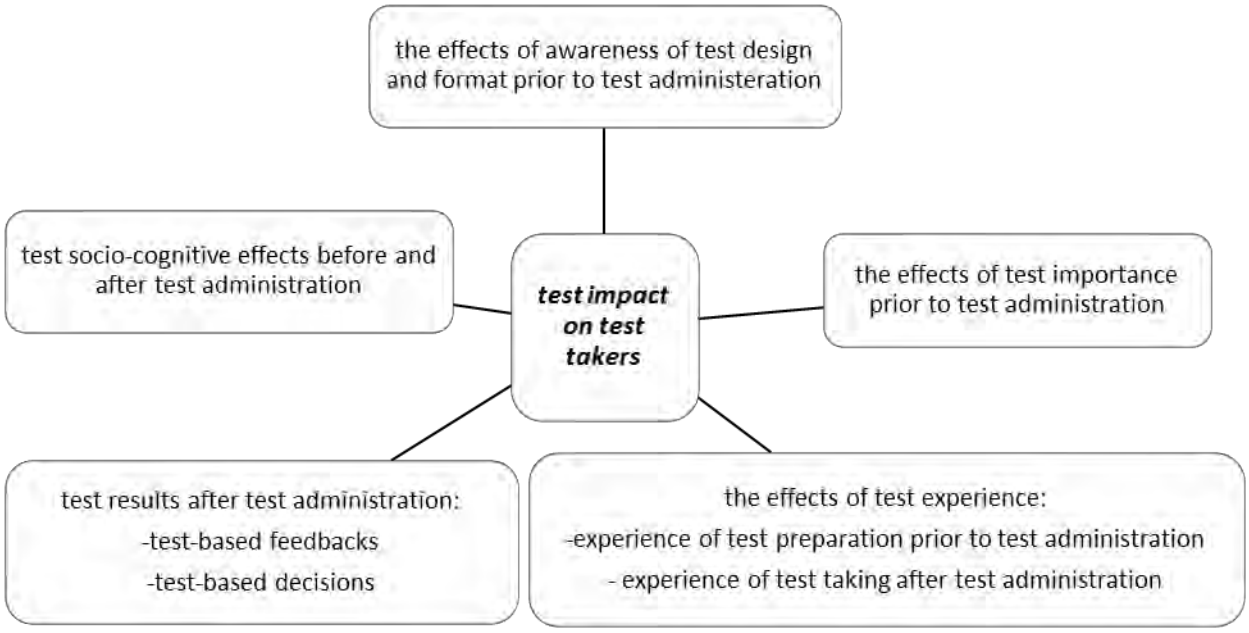

Figure 1. Taxonomy of test impact on test takers

Among different stakeholders in testing events, test takers have the highest stake of all (Hamp-Lyons, 2000). Alderson and Wall (1993, p. 120-21) pose the following hypotheses about washback on learning:

1. A test will influence learning.

2. A test will influence what learners learn.

3. A test will influence how learners learn.

4. A test will influence the rate and sequence of learning.

5. A test will influence the degree and depth of learning.

6. A test will influence attitudes to the content, method, etc., of learning.

We now provide a description of each element in the initial proposed framework of TITT.

- **The effects of awareness of test design and format prior to test administration**

This component in the hypothesized framework involves the degree to which test takers' awareness of test design and format prior to test administration affects test takers' attitudes and behaviors. Hughes (1993) emphasizes that learners' awareness of the test and its importance can influence the degree and depth, method, sequence, and rate of learners' learning and studying for the test.

- **The effects of test importance prior to test administration**

This component incorporates the effects which might arise in the test takers due to the recognized importance of the test. The degree of a test's importance can influence test takers in various aspects before they take the test. As held by Bailey (1996, p. 264-265), faced with an important test, students may participate in (but are not limited to) any of the following processes:

1) Practicing items similar in format to those on the test.

2) Studying vocabulary and grammar rules.

3) Participating in interactive language practice (e.g., target languageconversations).

4) Reading widely in the target language.

5) Listening to non-interactive language (radio, television, etc.).

6) Applying test-taking strategies.

7) Enrolling in test-preparation courses.

8) Requesting guidance in their studying and feedback on theirperformance.

9) Enrolling in, requesting or demanding additional (unscheduled)test-preparation classes or

10) Skipping language classes to study for the test.

- **The effects of test experience**

This component involves two types of test takers' experiences which might affect their characteristics and actions. They involve the experience of test preparation prior to test administration and the experience of test taking after test administration. Test takers' experience of test taking and preparation is influenced by the test which affect their perceptions, attitudes and amount of knowledge (Bachman & Palmer, 1996).

- **Test results after test administration**

Several sources of feedback following the administration of the test may impact upon the test takers' attitudes, perceptions, and behaviors, including the actual test scores provided by the scoring service, and feedback from the teachers, test-takers, and proctors (Bailey, 1999).

According to Bachman and Palmer (1996) two aspects of test takers which are influenced by the test include (a) the feedback they receive about their performance based on the test as well as its completeness, relevance and meaningfulness, and (b) the decisions that may be made about them based on the test scores and their relevance, appropriateness and fairness. Bachman and Palmer (2010) also believe that interpretations based on the test takers' language ability need to be meaningful, impartial, generalizable, relevant, and sufficient.

By the same token, this component in the taxonomy includes the effects of test-based feedbacks based on test takers' scores in the test on their perceived meaningfulness, relevance, appropriateness, and fairness of the test. Moreover, this element involves the influences of decisions made based on test takers' performance in the test on the test takers' attitudes and behaviors.

- **Test socio-cognitive effects before and after test administration**

Kiani, Alibakhshi, and Akbari (2009) have indicated that ESP tests can have several psychological (e.g., anxiety, disappointment, self-confidence), social (e.g., deprivation from education, ethical issues, acceptance of nonqualified candidates), financial (e.g., future job and income), and family consequences on the learners. The society is also influenced as the result of ESP tests scores. Besides, as Baily (1999) asserts, the language learners affected by washback may be cognitively influenced by official information about a test prior to its administration like advertisements and existing test preparation booklets, etc., or by reports from students who have taken earlier versions of the test. Thus, the last component includes the various socio-cognitive influences exerted on test takers. In other words, this concerns factors affecting test takers' cognitive and social characteristics after the test is administered.

Overall, these five elements in the taxonomy of TITT make up the theoretical foundation for the scale development in the present study. It's worth noting that the choice of these five

dimensions has been based on the fundamental aspects of impact as reviewed in the relevant theoretical and empirical literature. However, the very scarce number of designed questionnaires in the area do not account for all the factors as discussed in the above-mentioned taxonomy. As a result, the lack of such a thorough questionnaire for test impact on test takers results in subjectivity, bias and accordingly threats to validity. Furthermore, this type of study, most often, demands the strategy of triangulation or obtaining multiple data sources. This, however, increases the difficulty of conducting research concerning test impact. Hence, for the above mentioned reasons, there is an urgent need to develop and validate such an instrument in the current impact studies which can be applied in any testing setting no matter what the type of the test is with minor modifications based on the test purpose. Also, how the findings of such an instrument are interpreted is totally purpose and context specific. Accordingly, this study is intended to fill a gap in literature by developing a questionnaire to investigate the impact at the micro level of the test takers. However, it is of paramount importance to underscore that the implications and considerations resulting from the application of this questionnaire will depend on the very context in which and the specific purpose for which the test is used. As elaborated above, the development of the questionnaire in this research study is fundamentally based on the findings of the past studies and improves on their works correspondingly. Therefore, the main purpose of this study is to establish an internally consistent scale of TITT and to begin validation. Our study is, therefore, to answer the following research questions:

1. What factors underlie the test impact on test takers' questionnaire?
2. Does this questionnaire have enough evidence for construct validity?
3. Does the new scale have an acceptable reliability?

2. Methodology

2.1 Participants

The participants recruited in the study include 410 Iranian students from different universities at BA level in two provinces of Ilam (a city in the west part of Iran) and Zanjan (a city in the northwest of Iran). They were 227 female and 168 male students and the age range was 18 to 22 years (M= 20.16, SD= 1.47). The universities were chosen on the basis of credibility and feasibility.

University Entrance Examination (UEE) called Konkoor in Iran served as a nation-wide competitive selection test is applied for making important educational decisions and placing students in different university courses of study. Konkoor is a comprehensive, 4.5-hour multiple-choice standardized test that covers all subjects taught in Iranian high schools—from math and science to Islamic studies and foreign language. The students usually spend a year to get prepared for the exam, and if anyone fails, they are allowed to take the exam in the following

years. General English test is included in all levels of the university entrance examinations in Iran.

The participants who have passed UEE were selected to identify the validity of the TITT scale which addressed the impact of these high stakes tests and more specifically their General English test, on the test takers.

2.2 Procedure

After designing a potential questionnaire based on the proposed taxonomy in the current study, we asked a number of experts in the field of applied linguistics to judge both the positive and negative aspects of each subscale and to give feedback about the items in terms of their comprehensibility and relevance to the topic. This was done to enhance content and face validity. The set of potential scale items was modified and expanded, so that a final version of the questionnaire was generated. The scale was then translated into Persian.

The phase for content validity was undertaken to ascertain whether the content of the questionnaire was appropriate and relevant to the study purpose. To ensure the content validity of the TITT, the researchers clearly defined the conceptual framework of test impact by undertaking a thorough literature review and seeking expert opinion. Once the conceptual framework was established, three experts in applied linguistics were asked through formal interviews to judge the relevance of the items to the elaborated conceptual framework. Collected data in this stage showed that the items were relevant to our theoretical model in fig. 1 and the underlying theoretical foundation concerning impact on test takers. Regarding the face validity, which indicates the appropriateness of the questionnaire on the face of it, the same experts were asked to evaluate the questionnaire in terms of feasibility, readability, clarity of wording, and consistency of style, layout and formatting.

Moreover, in order to make sure of the comprehensibility of the items among our target individuals, a second phase of pilot testing involved administering the translated potential items to an additional group of participants. Pilot testing was conducted among 100 test takers who have passed the entrance exams. In this stage, the Cronbach alpha coefficient was found to be .91. Moreover, all the five hypothesized subscales of test awareness (TA hereafter), test importance (TI hereafter), test experience (TE hereafter), test results (TR hereafter), and test socio-cognitive effects (TSCE hereafter) yielded high reliability estimates of .869, .777, .930, .835, and .780 respectively.

Participants answered potential items and a series of open-ended questions about hypothesized components relevant to test impact on learners, tailored to explore each of the main factors of the construct. They were also asked to add any other point which they felt was missing in the questionnaire concerning each subscale. This phase helped us to generate relevant items which can be easily understood by the average person. These test takers were also asked to give feedback about the items in terms of their comprehensibility and check any items that seemed unclear or confusing, and items checked more than once were subsequently deleted from the

pool. In result, the set of potential scale items was modified and expanded, so that by the end of testing, a pool of potential scale of 64 items had been emerged.

In the next phase of scale construction, we administered the pool of potential items to a larger group of participants, so that we could choose our final scale items based on their loadings on hypothesized subscales as well as their reliability.

2.3 Analysis

In the current research, exploratory factor analysis were utilized to ascertain construct validity. *Exploratory factor analysis* (EFA) attempts to discover the constructs affecting a set of data. In this study, responses to items assessing the components of the questionnaire were analyzed separately using EFA. Items with loadings lower than 0.3 were omitted from final versions of the subscales.

Furthermore, it was essential to estimate the reliability of any developing questionnaire. Reliability refers to the repeatability, stability or internal consistency of a questionnaire (Jack & Clarke, 1998). The present study incorporated Cronbach's $\propto$ statistic. This statistic uses inter-item correlations to determine whether constituent items are measuring the same domain. If the items show good internal consistency, Cronbach's $\propto$ should exceed 0.70 for a developing questionnaire or 0.80 for a more established questionnaire (Bryman & Cramer, 1997). Cronbach's $\propto$ statistic is reported for the separate domains and subscales within the questionnaire as well as the entire questionnaire.

3. Results

3.1 Scale internal consistency

We had 410 participants in EFA stage. Fourteen of these participants were removed as multivariate outliers with endorsement patterns that could be considered markedly atypical, such as endorsing every item with either a 0 or a 4. Having removed the outliers from 410 participants, finally we analyzed data obtained from 396 participants.

The 64 items of the TITT scale were subjected to reliability analysis. 8 items were removed from the scale due to their higher reliability alpha than the whole scale and their comparatively low item-total correlations. The remaining 56 items resulting in the final version of the developing scale (see appendix) showed high internal consistency with an alpha coefficient of .93 and an average item-total correlation of r=.63. The item content of the deleted items was general in scope, and there was no clear evidence of redundancy in meaning among the remaining items based on magnitude of inter-item correlations and face validity. According to Nunnally and Bernstein (1994), item-total correlation should be at least .3 in order for the item to be considered a meaningful contribution to the scale. In this study, inter-item correlations within each subscale were all above .3, indicating acceptable relationships between items and the

whole scale. Table 1 shows the reliability statistics for the whole scale and its underlying subscales.

Table 1

Reliability Statistics: Final Version of the Scale of TITT

Reliability statistics						
	Overall Scale	**TA subscale**	**TI subscale**	**TE subscale**	**TR subscale**	**TSCE subscale**
Cronbach's alphas	.931	.803	.785	.885	.818	.794
Number of Items	56	10	9	17	12	8

3.1.1 Test–Retest Reliability.

Test-retest reliability is estimated by administering the same tool to the same sample on two different occasions hypothesizing that there will be no substantial change in the construct over time (two-month span in our case). The duration of time between the two tests is critical and a proper time interval must be selected between the two test administrations. Test-Retest reliability of the scale of TITT was undertaken by administrating the questionnaire to 50 participants at two times.

Good test–retest reliability was obtained when participants' responses to the TITT Scale were compared across Time 1 and Time 2. Test–retest correlations were as follows: TITT scale (overall score): .90; TA subscale: .87; TI subscale: .86; TE subscale: .88; TR subscale: .89; and TSCE subscale: .82.

3.2 Scale validity

At the exploratory stage of our analysis, to ascertain about sufficiency of sampling and appropriateness of the factor model for each of our main variables, Kaiser-Meyer-Oklin (KMO) measure of sampling adequacy and Bartlett's Test of Sphericity were estimated. Inspection of the correlation matrix revealed the presence of many coefficients of .3 and above in line with what Tabachnick and Fidell (2007) have recommended. KMO value was .816, exceeding the recommended value of .6 (Kaiser, 1970) and Bartlett's test of Sphericity (Bartlett, 1954) reached statistical significance, supporting the factorability of the correlation matrix. Furthermore, confidence level of 0.00 for Bartlett's test conveyed appropriateness of factor model.

Then, the structure of the TITT Scale was examined by subjecting the 64 items to an exploratory factor analysis using methods of principal component extraction and a varimax rotation. The Kaiser's criterion was used as the technique to retain factors. The factor analysis yielded eight factors and the items selected for the final version of the scale were selected in this step, that is, the items were selected if they had loadings on the extracted factors equivalent to .30 or greater. Throughout this process, eight items were removed from the scale. The resulting 56 items were subjected to factor analysis with principal component extraction and a varimax rotation for the second time. This time, the analysis yielded five expected factors and initial and extracted communalities were obtained via Principal Axis Factoring from the fifty six items comprising the scale of TITT.

The result of Varimax with Kaiser Normalization was a rotated component matrix. Table 2 shows the rotated factor matrix of 56 items of the scale of TITT.

Table 2

Rotated Factor Matrix of 56 Items Comprising the Scale of TITT

Item	**Factors**					**Item**	**Factors**				
	1	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
1	*	.67	*	*	*	**29**	*	*	.53	*	*
2	*	.38	*	*	*	**30**	*	*	.63	*	*
3	*	.48	*	*	*	**31**	*	*	.49	*	*
4	*	.58	*	*	*	**32**	*	*	.49	*	*
5	*	.55	*	*	*	**33**	*	*	.50	*	*
6	*	.54	*	*	*	**34**	*	*	.42	*	*
7	*	.60	*	*	*	**35**	*	*	.49	*	*
8	*	.53	*	*	*	**36**	*	*	.34	*	*
9	*	.44	*	*	*	**37**	.65	*	*	*	*
10	*	.34	*	*	*	**38**	.64	*	*	*	*

 ISSN 2476-5880

11	*	*	*	.34	*	**39**	.44	*	*	*	*
12	*	*	*	.50	*	**40**	.57	*	*	*	*
13	*	*	*	.62	*	**41**	.58	*	*	*	*
14	*	*	*	.36	*	**42**	.62	*	*	*	*
15	*	*	*	.51	*	**43**	.63	*	*	*	*
16	*	*	*	.71	*	**44**	.66	*	*	*	*
17	*	*	*	.51	*	**45**	.56	*	*	*	*
18	*	*	*	.64	*	**46**	.56	*	*	*	*
19	*	*	*	.73	*	**47**	.53	*	*	*	*
20	*	*	.42	*	*	**48**	.72	*	*	*	*
21	*	*	.64	*	*	**49**	*	*	*	*	.66
22	*	*	.44	*	*	**50**	*	*	*	*	.54
23	*	*	.63	*	*	**51**	*	*	*	*	.38
24	*	*	.60	*	*	**52**	*	*	*	*	.40
25	*	*	.44	*	*	**53**	*	*	*	*	.35
26	*	*	.51	*	*	**54**	*	*	*	*	.37
27	*	*	.41	*	*	**55**	*	*	*	*	.38
28	*	*	.51	*	*	**56**	*	*	*	*	.42

The eigenvalues as well as the variance explained by the extracted five rotated factors also reveal that each of the five factors enjoys an eigenvalue higher than one and together they explain almost 51% of variance in the TITT scale. Thus, further support for the necessity of establishing factorial validity is provided. The variance explained for each factor is presented in the discussion section.

4. Discussions

This study sought to develop a multidimensional scale of test impact on test takers, particularly appropriate for students who have passed a high stakes exam. The research was designed to examine the factor structure, internal consistency, and stability of the scale and subscales. In the

following, the structure of the final version of the questionnaire is elaborated in light of related literature.

4.1 Factor 1: Test results

According to Bailey (1999), learners may be influenced by several sources of feedback and reactions following the administration of the test. The first factor or subscale in this newly developed scale explaining the largest proportion of the total variance (12.68 %) accounts for the impacts of the test results on the test takers. These results involve feedbacks received by the test takers after the test administration about their performance as well as decisions made based on those feedbacks. This contributes to the two aspects of the test takers influenced by the test as held by Bachman and Palmer (1996).

The highest amount of variance suggests the significance of test results in investigation of washback in the field of language testing. Bachman and Palmer (2010) believe that interpretations of test takers' language ability and the intended uses of a particular assessment must be meaningful, impartial, generalizable, relevant, and sufficient for a particular group of test takers, and in a particular setting. In fact, the items assess if the test feedbacks and decisions have been fair, appropriate, relevant, meaningful, and informative for the test takers. This factor is composed of 12 items including items 37-48 with high factor loadings ranging between .53 and .72. This subscale demonstrated high internal consistency (reliability) with a Cronbach's alpha of .818, strongly suggesting that this set of items is tapping a common underlying concept. Two marker items: *"These decisions have been fair for me."* and "*The decisions and final results are appropriate and effective to me.*" loaded highest on this factor.

4.2 Factor 2: Test awareness

The items in this factor involve the influence of the test takers' cognizance of test design and format on their various test preparation activities. They also show the extent to which test awareness influences the degree and depth, method, sequence, and speed of learners' learning and studying for the test (Hughes, 1993). Bailey (1996) asserts that faced with an important test, students may participate in different processes such as practicing items similar in format to those on the test. This factor consists of ten items (1-10) which explain 10.88 % of the total variance. It involves loadings ranging between .34 and .67. This subscale has a Cronbach's alpha of .803 indicating high internal consistency. Item 1 in the questionnaire (*My awareness of the format and design of the exam affected the degree of my learning and studying.*) has the highest loading in this component.

4.3 Factor 3: Test experience

These items reflect influences the experience of test preparation and test taking exerts on learners' characteristics and activities. In fact, related to the first aspect of test takers influenced by the test (Bachman & Palmer, 1996), the items tend to reveal the degree to which the social, cultural, and topical knowledge, attitudes, views, and perceptions, and strategies and techniques in the test takers have been influenced by the test experience of taking and preparing for the test.

The items of this factor explains 10.87 % of the total variance and include items 20-36. The loadings in this factor vary between .38 and .64. Cronbach's alpha for this subscale is .885 and item 21 (*My experience of test preparation influenced my social and cultural knowledge.*) has the highest loading.

4.4 Factor 4. Test importance

Items 11-19 explaining 9.62 % of the total variance refer to the category of test importance. They have the factor loadings from .34 to .73 and the highest loading refers to item 19 (*The importance of the test made me skip some classes and set aside some of my academic and social activities to study for the test.*). The reliability statistic for this subscale is r= .785. These clusters of items depict the effects of test importance on the test takers. Hughes (1993) emphasizes the influence of test importance on the rate, sequence, depth, and methods of learning. Bailey (1996) also asserts that when, faced with an important test, learners participate in some processes and activities (e.g., applying test-taking strategies, enrolling in test-preparation courses, skipping language classes to study for the test).

4.5 Factor 5: Test socio-cognitive effects

The last factor in this scale involves items 49-56 and enjoys 7% of the explained variance. This subscale also has internally consistent items (r= .794) with factor loadings ranging between .35 and .66. Item 49 in this component, serves the highest factor loading (*I have been influenced by existing test preparation booklets and other advertising materials.*). These items refer to the psychological, financial, and cognitive effects of the test on the test takers at the individual and social level. As Bailey (1999) holds, the language learners may be influenced by official information about a test prior to its administration, including advertising materials from the test publisher, existing test preparation booklets, etc. Conducted study by Kiani, Alibakhshi, and Akbari (2009) on the impacts of ESP test, in line with the present study, reported for some psychological, social, financial consequences on learners. Furthermore, according to Bachman and Palmer (1996), at micro level, the test influences learners' perceptions of TLU (target language use) domain, goals and values. On the same lines, the items in this category tend to show whether the test has influenced learners' motivation, autonomy, expectations, attitudes, etc.

5. Conclusions

The present study was designed to address the lack of a valid and reliable measure of test impact in the context of language testing by establishing and validating a scale of test impact on test takers that demonstrated adequate psychometric properties. The construction of TITT Scale was guided by the theoretical framework of test impact as summarized in figure 1. The suggested taxonomy also provides a coherent theoretical framework for impact on test takers which can be used as the basis for developing future potential questionnaires.

Developing a multidimensional scale of test impact on test takers assessing the expected five bases of test impact on test takers yielded support for the existing literature. In result, using exploratory factor analysis, a five-factor structure was extracted, each factor having an acceptable internal consistency. This is an evidence of construct validity which suggests that a common cause underlie the covariance among the subscales and test items (Baghaei&TabatabaeeYazdi, 2016). Since test impact provides evidence for construct validity, the extracted factors can thus guide test developers in considering varied aspects of test impact throughout the iterative process of test development and validation.

Accordingly, the study offers a promising reliable and valid measure of test impact on test takers with good psychometric properties. TITT scale may be a useful instrument to evaluate the amount of impact that a test can have on the test takers in each level and help advance the knowledge about test impact. Thus, this scale can contribute to understanding the consequential basis of the validity of any potential test as an evidence for the recognition of a test's unified progressive construct validity (Messick, 1989), its usefulness (Backman& Palmer, 1996), consequential aspect of construct validity (Messick, 1996) and fairness (Kunnan, 1998). Moreover, for the purpose of research and practice, the information resulting from the scale and its subscales might enhance an awareness of test takers' various attributes after test taking, including their future academic performance, their actual socio-cognitive characteristics, etc.

In addition, although originally designed for high stakes exams, this newly devised questionnaire can be used to investigate the test impact for any high stakes or low stakes tests in different EFL/ ESL contexts of language testing across different levels of language proficiency. The study can also be replicated to see the validity findings for other high stakes tests. Besides, each subscale has the potential to be explored separately in distinct studies. Further studies can be conducted to ascertain the validation results in other language contexts. Moreover, minor modifications in the scale may be required to investigate test impact in other contexts than language settings.

References

Akpinar, K. D., &Cakildere, B. (2013). Washback effects of high-stakes language tests of Turkey (KPDS and ÜDS) on productive and receptive skills of academic personnel. *Journal of Language and Linguistic Studies, 9(2),* 81-94.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13(3),* 280-297.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14(2),* 115-129.

Andrews, S. (1994). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, R. Berry, & V. Berry, (Eds.), *Bringing about change*

inlanguage education. Proceedings of the International Language in Education Conference 1994 (pp. 67-81). Hong Kong: University of Hong Kong.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: developing language tests and justifying their use the real world.* Oxford: Oxford University Press.

Baghaei, P., &TabatabaeeYazdi, M. (2016). The Logic of Latent Variable Analysis as Validity Evidence in Psychological Measurement. *The Open Psychology Journal, 2016, 9,* 168-175.

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

Bailey, K. M. (1999). *Washback in language testing. TOEFL Monograph Series. Report Number: RM-99-04, TOEFL-MS-15.* Princeton, NJ: Educational Testing Service.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of Royal Statistical Society, 16 (Series B),* 296-298.

Bryman, A., & Cramer, D. (1997). *Quantitative Data Analysis with SPSS for Windows.* Routledge, London.

Byrne, B. M. (2001). *Structural equation modeling with AMOS. Basic concepts, applications, and programming.* Mahwah, N.J.: Lawrence Erlbaum Ass.

Green, A. (2006). Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing, 11 (2),* 113-134.

Green, A. (2007). *IELTS Washback in Context: Preparation for academic writing in higher education.* Studies in Language Testing 25. Cambridge: Cambridge University Press.

Green, A. (2013). Washback in language assessment. *International Journal of English Studies, 13 (2)*, 39-51.

Green, A. (2014). *The test of English for academic purposes (TEAP) impact study: report 1 - preliminary questionnaires to Japanese high school students and teachers.* Eiken Foundation of Japan.

Hamp-Lyons, L. (1997). Washback, impact, and validity: Ethical concerns. *Language Testing, 14(3),* 295-303.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28,* 579-591.

Hu, L., &Bentler, P. M. (1999). Cut-off criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hughes, A. (1993). *Backwash and TOEFL 2000.* Unpublished manuscript, University of Reading.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jack, B., & Clarke, A. (1998).The purpose and use of questionnaires in research. *Professional Nurse, 14*, 176–179.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35,* 401-415.

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.

Kunnan, A. J. (Ed). (1998). *Validation in language assessment: Selected papers from the 17th language testing research colloquium.* Long Beach Mahwah, New Jersey: Lawrence Erlbaum Associates.

Messick, S. (1989). Meaning and Values in Test Validation: the Science and Ethics of Assessment. *Educational Researcher, 18(2),* 5-11.

Messick, S. (1996). Validity and Washback in Language Testing. *Language Testing, 13(3),* 241-256. Retrieved from http://dx.doi.org/10.1177/026553229601300302

Moore, S., Stroupe, R., &Mahony, P. (2009). Perceptions of IELTS in Cambodia: A case study of test impact in a small developing country. *IELTS research reports, 13 (6),* 1-109.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd Ed.).* New York: McGraw-Hill.

Özmen, K. S. (2011). Washback effects of the Inter-University Foreign Language Examination on Foreign Language Competences of Candidate Academics. *Novitas-ROYAL (Research on Youth andLanguage), 5(2),* 215-228.

Pizzaro, M. A. (2010). Exploring the washback effects of a high-stakes English test on the teaching of English in Spanish upper secondary schools. *RevistaAlicantina de EstudiosIngleses, 23,* 149-170.

Ramezaney, M. (2014). The washack effects of university entrance exam on Iranian EFL teachers' curricular planning and instruction techniques. *Procedia - Social and Behavioral Sciences, 98,* 1508 -1517.

Read, J., &Chapelle, C.A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18 (1)*, 1-32.

Ross, S. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics, 26 (3),* 317-342.

Tabachnick, B. G., &Fidell, L. S. (2007). *Using multivariate statistics* (5thedn). Boston: Pearson Education.

Wall, D. (2000). The Impact of high Stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28 (4),* 499-509.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 129-146). Mahwah, NJ: Lawrence Erlbaum Associates.

Xie, Q. (2008). Students' perception of the CET4 listening and test preparation practices-implications for washback. *Research Studies in Education, 6,* 32-47.

Xie, Q., & S. Andrews. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30 (1),* 49-70.

Zhan, Y. (2003). *Washback on Chinese learners: An impact study of the College English Test Band 4*, Retrieved from http://www.iaea2008.cambridgeassessment.org.uk/ca/digitalAssets/ 164766_Zhan.pdf.

Zhan, Y., & Wan, Z. H. (2014). Dynamic nature of washback on individual learners: The role of possible selves. *Assessment & Evaluation in Higher Education*, 39 (7), 821-839.

Appendix

Final version of the questionnaire

Dear participants,

You are kindly asked to answer the following questionnaire about *the English test that you have taken in the university entrance exam* lastly. Please read each statement and consider to what extent each one is correct for you. Please fill the blanks before you start to answer the questionnaires.

Your age:…….

Your gender: Male ◯ Female ◯

The name of the high-stakes test you have recently taken: BA ◯ MA ◯ Ph.D.◯ Exam

Test awareness					
	1. Strongly disagree, 2. Disagree, 3. Agree, 4. Strongly agree	1	2	3	4
1	My awareness of the format and design of the exam affected the degree of my studying and effort.				
2	My awareness of the format and design of the exam made me follow my teacher's instruction.				
3	My awareness of the format and design of the exam affected the rate of my learning and studying.				
4	My awareness of the format and design of the exam affected the way and sequence of my learning and studying.				
5	My awareness of the format and design of the exam affected the time that I spent for learning, effective use of my time and timetabled studying.				
6	My awareness of the format and design of the exam made me participate in specific types of preparation classes.				
7	My awareness of the format and design of the exam made me get more help from my teachers.				
8	My awareness of the format and design of the exam made me search in reference materials and internet before the exam.				
9	My awareness of the format and design of the exam made me practice similar exercises to test items.				

10	My awareness of the format and design of the exam made me practice some test taking strategies in finding the correct answer.				
Test importance					
	1. Strongly disagree, 2. Disagree, 3. Agree, 4. Strongly agree	**1**	**2**	**3**	**4**
11	The importance of the test made me participate in different activities and take part in different test preparation classes before the exam.				
12	The importance of the test has influenced the degree and depth of my studying.				
13	The importance of the test made me be more involved in class activities.				
14	The importance of the test influenced my selection of sources and studying different sources and materials.				
15	The importance of the test influenced the amount of money that I spent for preparing for exam.				
16	The importance of the test made me get more help from my teachers.				
17	The importance of the test made me cooperate more with my friends and get their help.				
18	The importance of the test made me spend more time and have a timetable for learning and studying.				
19	The importance of the test made me skip some classes and set aside some of my academic and social activities to study for the test.				
Test experience					
	1. Strongly disagree, 2. Disagree, 3. Agree, 4. Strongly agree	**1**	**2**	**3**	**4**
20	My experience of test preparation influenced my knowledge of English language.				
21	My experience of test preparation increased my social and cultural information.				
22	My experience of test preparation made me deeply understand what I had studied in my English courses.				
23	My experience of test preparation influenced my perceptions of language use.				
24	My experience of test preparation changed my views and attitudes toward the subject matter.				
25	My experience of test preparation influenced my perceptions of test and different test tasks.				

26	My experience of test preparation made me identify my weaknesses and strengths in different parts of the exam.				
27	My experience of test preparation made me use different strategies and techniques for learning.				
28	My experience of test preparation influenced the way I answered the exam questions while taking the test.				
29	My experience of test taking and answering questions influenced my knowledge of English language.				
30	My experience of test taking and answering questions influenced my social and cultural knowledge.				
31	My experience of test taking and answering questions made me deeply understand what I had studied in my English courses.				
32	My experience of test taking and answering questions influenced my perceptions of target language use domain.				
33	My experience of test taking and answering questions changed my views and attitudes toward the subject matter.				
34	My experience of test taking and answering questions influenced my perceptions of test and different test tasks.				
35	My experience of test taking and answering questions made me feel I need to work more on the areas I had problems during exam.				
36	My test taking experience made me use different strategies and techniques for learning.				
Test results					
	1. Strongly disagree, 2. Disagree, 3. Agree, 4. Strongly agree	**1**	**2**	**3**	**4**
37	I am satisfied with the feedback given to my performance.				
38	This feedback has been complete and meaningful to me.				
39	This feedback made me aware of my weaknesses and strengths and increased my amount of studying.				
40	This feedback increased my amount of knowledge and information.				
41	This feedback has been appropriate and effective for me.				

42	I feel satisfied with the decisions made based on my final marks and results.				
43	My marks/percentages have affected the kind of decisions made for me and my overall success.				
44	These decisions have been fair for me.				
45	I know that for making fair decisions, fixed and standard criteria have been applied.				
46	I am aware of the procedures of decision making.				
47	The decisions accord with my marks and percentages.				
48	The decisions and final results are appropriate and effective to me.				
Test socio-cognitive effects					
	1. Strongly disagree, 2. Disagree, 3. Agree, 4. Strongly agree	**1**	**2**	**3**	**4**
49	I have been influenced by existing test preparation booklets and other advertising materials.				
50	This test has influenced my expectations of the English courses.				
51	This test made me recognize my own strengths and weaknesses.				
52	The test has influenced my future job and my personal and social life.				
53	The test made me understand the concepts and materials better.				
54	This test has influenced my perspectives and attitudes toward language learning.				
55	This test influenced my autonomy and independence in language learning.				
56	This test has influenced my motivation to learnand continue education.				

Scoring:All items are written in the positively keyed direction, so no reverse scoring of items is required.