Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 10, No. 2, October 2020
ISSN 2476-5880

Construction and Validation of Educational, Social and Psychological Consequences Questionnaires of EPT as a High-Stakes Test

Mahbubeh Rezaeian [1], Seyyed Hassan Seyyedrezaei[1*],
Ghasem Barani[1], Zari Sadat Seyyedrezaei[1]

Received: 28 July 2020 Accepted: 23 September 2020

Abstract

Individuals are controlled by tests in every advanced society when they want to be admitted in educational courses, to proceed from one stage to the next, or to be given a certificate (Shohamy, 2001b). Accordingly, the present study was carried out to construct and validate educational, social, and psychological consequences questionnaires of English Language Proficiency (EPT) as a high-stakes test in Iran. To achieve the goals, after initial piloting of the item pool, a total number of 252 non-English PhD students completed the final researcher-made questionnaires developed using a comprehensive review of the related literature, experts' opinions, documents, and interviews. A number of statistical procedures were taken to validate the current questionnaires including Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA was utilized to determine the underlying factors of the scale that affect the variables in a data structure without setting any predefined structure to the outcome and to verify the number of factors; subsequently, these results were confirmed in the CFA phase. Ultimately, the results were discussed and implications of the questionnaires were presented as follows.

Keywords: Educational Consequences; High-Stakes Test; Psychological Consequences; Social Consequences

1. Introduction

According to Shohamy (1998), tests are the most powerful tools because they are often viewed as the single indicator for determining the future of individuals. Furthermore, it has been noted by Menken (2017) that a test is considered high-stakes as soon as a sole test score is utilized as the chief factor in determining substantial educational decisions. Moreover, tests are not isolated events, rather they are connected to a whole set of social, psychological, and political variables that extremely influence curriculum, ethicality, social classes, government,

[1] Department of English Language Teaching, Aliabad Katoul Branch, Islamic Azad University, Aliabad Katoul, Iran.
* (corresponding author: srezaei.sh@gmail.com)

ISSN 2476-5880

political affairs, and language knowledge (Messick, 1981). Nowadays, loads of higher education institutes over the world employ standardized English proficiency tests to evaluate learning outcomes, teaching effectiveness and achieve required educational changes (Hung, Huang, 2019).

2. Review of Literature

Shohamy (2001a) believes that the main focus of traditional testing is on designing and developing tests to correctly estimate the knowledge of testers and it entirely fails to see the test use as it is one-off phenomenon separated from test-takers, educational system, and society and strictly associated with fixed timing and stable procedures (H. D. Brown, 2004). Standard setting or traditional testing can be defined as a decision-making process with considering the classification of the test or exam results in a restricted number of levels of success (Kaftandjieva, 2010) and it aims at utilizing full-grown abilities (Alemi, Miri, Mozafarnezhad, 2019). While, in use-oriented testing, testing is seen as a rooted event in social, educational, and political contexts in which a great number of issues such as test-takers' activities for the test, the teachers' methods of teaching, materials designed for the test, and the influences of the test results on the stakeholders are all taken into consideration.

As far as the history is concerned, the traditional testing has undergone main changes after the emergence of critical pedagogy in which the test takers' viewpoints, experiences along with their expectations taken into consideration due to the fact that simply multiple-choice format cannot be considered as an proper way to evaluate different learners with dissimilar styles or individual differences (Mohammad Salehi & Tarjoman, 2017). They assert that critical language testing is viewed as a central concept in the world of language testing as making decision about individuals can be done devoid of their having control over the procedure in addition to the content of language tests. In fact, critical language testing suggests a paradigm shift in language testing in which a new criterion for the validity of language tests including the consequential validity have been introduced and language testing is responsible for questioning the employment of tests as powerful tools and scrutinizing their uses not only in educational level but also in societal level which is associated to the concern about the consequences of tests in macro level (Messick, 1989; Messick, 1996; Shohamy, 1998).

As noted by Pan (2009a), educational consequences allude to possible washback, both positive which is generally intended and negative which is normally unintended, that occurs in the educational context; for example, tests are capable to govern textbook as well as a curriculum as educational devices (Shohamy, 1998). The washback effect of a language test on learning and teaching appears to be unquestionable; however, the washback effect does not need to be always negative and unfair. When tests are designed with awareness and understanding of some factors such as the learning contexts, students, and the contents, positive washback is more likely to appear (Xerri & Vella Briffa, 2018). Apart from the educational consequences, if it is acknowledged that L2 learning is chiefly a social-psychological event, it is merely natural that social-psychological variables should be given central attention (Au, 1988). Standardized tests have been utilized as measurements of

ISSN 2476-5880

language proficiency for the purposes of graduation requirements, university admissions, hiring, and promotional decisions as some important social decisions.

According to Shohamy (2001b), tests are in charge of turning enjoyment of learning into nervousness, pain, and a feeling of inequality; moreover, they are frequently the basis and source of irritation, frustration, rivalry, pressure, and humiliation as some important affective or psychological factors (Shohamy, 2007b). Ahmadjavaheri and Zeraatpishe (2020) also state that test performance can be effected via the differences in individual features such as gender, background knowledge, age, cultural background, cognitive characteristics, and test anxiety which make the test score interpretation invalid. Furthermore, motivation as one of the main psychological factor in the present study is accountable for people' decisions, their eagerness to go on, in addition to their determination to follow the action (Dörnyei & Ushioda, 2011). Intrinsic motivation refers to the willingness to accomplish an activity since it is attractive and enjoyable (Khajavy, Ghonsooly, Hosseini Fatemi & Choi, 2014). While, extrinsic motivation drives from exterior factors such as receiving reward or evading punishment (Khajavy et al., 2014). Ryan & Deci (2000) affirm that instrumental or extrinsic motivation is widely viewed as a non-autonomous factor that encourages people to do activities to reach some goals such as job promotion.

Given the fact that high-stakes language tests generate incredible consequences for the main stakeholders (Im, Shin & Cheng, 2019), it is vital to uncover out whether suitable decisions have been made based on EPT scores' interpretations which is conducted by the Ministry of Science, Research, and Technology as an exit test for the PhD candidates of Islamic Azad universities who have chosen to continue their studies in Iran. More specifically, EPT comprises vocabulary section (25 questions in multiple-choice format), grammar section (40 questions in multiple-choice and error correction formats), and reading comprehension section (35 questions in multiple-choice and cloze test formats) in which test-takers are given 140 minutes to answer the questions.

Although applying high-stakes tests is widespread across the majority of developing societies, very little is identified about how these types of tests are designed, what criteria direct test construction, and what sort of factors manipulate or influence this process (Ali, Hamid & Hardy, 2020). Despite the roles of educational, social and psychological consequences of language testing in shaping individuals' future, to date it seems that considerable literature has grown up around the theme of micro level (test's impact on individuals), but the issue of language testing consequences in macro level (test's impact on society) is under research. It also appears that the unintended or intended consequences of language testing at educational, social, and psychological levels are still not yet fully understood and some questions have remained unanswered in this field. Furthermore, it appears that the perceptions and the voice of test-takers in high-stakes testing have not been addressed and cannot be heard in the published literature in much detail instead there are several studies based on teachers' perceptions as main stakeholders such as Gunn, Al-Bataineh and Abu Al-Rub (2016) as well as Thibodeaux (2014). Some evidence also suggests that the world of research method suffers from the lack of well-constructed questionnaires to explore educational, social and psychological consequences of English Proficiency Test

ISSN 2476-5880

(EPT) as a high-stakes test and it seems to be rather left out in critical language testing research in both micro and macro level. That is, the existing literature lacks clarity regarding the educational, social, and psychological factors among Iranian non- English PhD students in EPT as a high-stakes test. Lack of triangulation as a process of confirming evidence from diverse individuals, sorts of data, or different methods of data collection including documents and interviews in qualitative research is another major problem nested in the body of current literature.

The study tries to fill a gap in the existing literature, add to the existing knowledge base and solve the problems created by washback as an outcome of the strong authority of external testing and scrutinize its main effect on the lives of test-takers and its influences in various directions (Shohamy, 1992; Shohamy, 2007a). Data triangulation ensures that the study will be accurate because the information draws on multiple sources of information to gain multiple perspectives and validation of data. In this way, it encourages the researchers to develop a report that is both more accurate and credible. The present study informs practice and it leads to the identification of new scales by examining educational, social, and psychological factors. Furthermore, it looks as if the current study is one of the first investigations to utilize exploratory design to explore the educational, social, and psychological consequences of EPT as a high-stakes test based on students' perceptions as one of the main stakeholders at least in the Iranian context. In brief, the core objectives of the research are to discover the educational, social, and psychological consequences of EPT as a high-stakes test among Iranian non-English PhD students. Additionally, it aims to construct and validate instruments that best fit the sample under study and tries to specify latent variables that need to go into a follow-up quantitative study for further research.

3. Method

3.1. Participants at Qualitative Phase

With regard to the exploratory nature of the study, two various groups of participants were selected based on purposeful sampling strategies including extreme case sampling, in addition to convenience sampling at the qualitative phase and convenience sampling at the quantitative phase. Having taken the above-mentioned issues into consideration, sixteen Iranian non- English PhD graduates and students who passed or were engaged in EPT invited to take part in this study in a semi-structured interview at the qualitative phase to achieve data saturation after running a pilot study. They included 6 females and 10 males (M = 40, SD = 5.75) from diverse Islamic Azad universities, with various socio-economic status, employment status, and majors.

Table 1.
Interviewees' background information.

Number	Pseudonym	Gender	Age	Academic Major
1	Bita	Female	29	Fishery
2	Sara	Female	31	Fishery
3	Ali	Male	47	Fishery
4	Reza	Male	46	Political Science
5	Mohammad	Male	48	Political Science
7	Soroush	Male	41	Computer Science
8	Saeed	Male	41	Computer Science
9	Zohre	Female	38	Management
10	Sanaz	Female	35	Management
11	Amir	Male	45	Management
12	Amir Hossein	Male	37	Management
13	Roya	Female	34	Accounting
14	Saman	Male	45	Accounting
15	Armin	Male	42	Accounting
16	Ramin	Male	44	Accounting

3.2. Participants at Quantitative Phase

A total number of 252 Iranian non-English PhD graduates and students participated at the quantitative phase after a pilot study based on convenience sampling. They included 113 females (44.8) and 139 males (55.2) from diverse Islamic Azad universities in which they were studying or graduated in non- English majors. The distribution of participants by age has been shown in Table 2.

Table 2.
Distribution of participants by age.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	less than 30	16	6.3	6.3	6.3
	30-35	64	25.4	25.4	31.7
	36-40	76	30.2	30.2	61.9
	41-45	47	18.7	18.7	80.6
	46-50	40	15.9	15.9	96.4
	more than 50	9	3.6	3.6	100.0
	Total	252	100.0	100.0	

4. Instrumentation

4.1. Semi-Structured Interview

Although the semi-structured interview is a set of pre-prepared guiding questions and prompts, the format is open-ended and the interviewees are encouraged to elaborate on the issues raised in an exploratory manner. Interviews were conducted in Persian so that the

ISSN 2476-5880

participants felt more at ease and more willingly express their feelings, emotions, and thoughts. Considering the semi-structure approach of the study, fourteen one-on-one interviews and two telephone interviews were conducted with a pre-determined set of open questions with the opportunity for the interviewer to explore particular themes or responses further and were recorded for further analysis such as: Could you please tell me about your general understanding and experience of EPT exam? as a content question. How has EPT helped you educationally? How EPT has changed and shaped your future at educational dimension? Which skill did you strengthen the most? Some educational dimension questions were: How EPT has changed and shaped your future at social dimension? To what extent, are employers' hiring decisions influenced by the exam score? Does passing the exam cause you to have to take supplementary coaching, resulting in additional expense for you? And in the case of psychological dimension, some questions were addressed such as: How EPT as a high-stakes test has changed and shaped your future at psychological dimension? How EPT has influenced your amount of stress and anxiety? How EPT has changed and shaped your motivation?

4.2. Documents

To enrich the qualitative data and extract some themes, some of social media texts available in an EPT Campaign Channel in Telegram App were analyzed with considering ethical issues; such as, "privacy, confidentiality, anonymity, data storage and disclosure" (Hammersley & Traianou, 2012, p. 99; Mazandarani, 2014, p. 143). The aim of applying documents was to obtain "language and words of the participants" (Creswell, 2009, p. 180). Namely, the subsequent quotes are illustrative of Iranian non-English PhD graduates and students' words regarding EPT in various dimentions:

Student 1: "Who is responsible for all the stress on us who have other problems in addition to the exam?"

Student 2: "I did not learn anything from this test, and I think it is just a source of income"

Student 3: "I have been undecided for a long time because I did not pass this exam"

Student 4: "No organization hears our voice. We have to interview with the national media and the press"

Student 5: "I wasted a year on this test and I have no motivation to learn English anymore"….

4.3. Researcher-Made Questionnaires

Three attitudinal researcher-developed questionnaires developed and validated (in Persian) at exploring educational, social, and psychological consequences of EPT as an exit exam that Iranian non- English PhD students must pass to graduate from Islamic Azad Universities. The developed and validated questionnaires consist of two chief sections: a closed-ended section at a five-point Likert scale arranging from 1 (strongly disagree) to 5 (strongly agree) in addition to the demographic information section to be used for further descriptive statistics. The questionnaires contain the following main constructs including:

ISSN 2476-5880

4.3.1. Educational Consequences Questionnaire (ECQ) with five subscales, 23 items at a five-point Likert-scale format scaling from 1-5 (strongly disagree to strongly agree).
Subscale 1: Learning Environment (questions 1- 4)
Subscale 2: Surface Learning (questions 5-6)
Subscale 3: Deep Learning (questions 7-16)
Subscale 4: Learning Outcomes (questions 17-19)
Subscale 5: Testing Preferences (questions 20-23)

4.3.2. Social Consequences Questionnaire (SCQ) with three subscales, 12 items at a five-point Likert-scale format scaling from 1-5 (strongly disagree to strongly agree).
Subscale 1: Job Promotion (questions 1- 3)
Subscale 2: Financial Expenses (questions 4-6)
Subscale 3: Social Justice (questions 7- 12)

4.3.3. Psychological Consequences Questionnaire (PCQ) with four subscales, 15 items at a five-point Likert-scale format scaling from 1-5 (strongly disagree to strongly agree).
Subscale 1: Self-Efficacy Beliefs (questions 1- 3)
Subscale 2: Feelings (questions 4- 9)
Subscale 3: Intrinsic Motivation (questions 10- 13)
Subscale 4: Extrinsic Motivation (questions 14- 15)

5. Procedure

5.1. Instrument Development

Initially, the related literature and documents were reviewed carefully following the standard procedure to develop a reliable and valid instrument (Dornyei, 2003; Dornyei & Taguchi, 2010), in order to extract themes and draw up an item pool on educational (85 items included in 11 subscales), social (26 items included in 6 subscales) and psychological (22 items included in 2 subscales) consequences of EPT as a high-stakes test among Iranian non-English PhD students. Then, the extracted items were checked and revised considering the face and content validity by four experts in the field of language testing and assessment. In this stage, some ambiguous words, sentences, and items, negative constructions, double-barreled questions, loaded words and sentences, item sequence (Dornyei, 2003; D rnyei & Taguchi, 2010) were modified; consequently, around 50 items were omitted from the item pool.

Following that, to find out whether all of the related themes are included or not, and to check whether any new theme can be added or not, a semi-unstructured interview was piloted and conducted but prior to commencing the study, participants received an explanation of the project before data collection and then ethical permission was sought from participants and they were asked to sign the informed consent forms without putting them under the pressure. Piloting the interview helped us to make sure everything had been covered, nothing had been missed out unintentionally in the interview, question wordings, as well as a list of probing

ISSN 2476-5880

questions were appropriate and unambiguous; moreover, some comments by interviewees were taken into consideration for constructing and conducting the final scale.

The following ways were used to ensure the quality of qualitative research: Firstly, triangulation of data sources was considered as the first strategy to enhance the credibility through the comprehensive review of related literature, semi-structured interview, and documents. As Creswell (2014) puts, triangulation is a process of confirming evidence from diverse sources to ensure that the study will be accurate because the information draws on multiple sources of information, individuals, or processes.

Member checking was utilized as the second strategy to increase the credibility of qualitative research via interviewees' consent and feedback. In this strategy, the researchers find out the accuracy of the qualitative findings by taking the ultimate report, major findings, particular descriptions or themes back to participants under study to identifying whether they feel that the qualitative findings are accurate or not. This process can be done by conducting a follow-up interview and providing an opportunity for participants to discuss and comment on the findings (Creswell, 2014). In the current study, the texts were checked to ensure that they did not include apparent mistakes made during transcription. Researchers were ensured that there was not a shift in the definition of codes, or a change in the meaning of the codes in the process of coding. This was achieved by continuously comparing the information with the codes and through writing notes about the codes and their definitions. Additionally, the communication was organized among the researchers by means of systematic documented meetings and through sharing the analysis.

After accomplishing the above-mentioned stages, researcher-developed questionnaires were developed based on the collected data in the qualitative phase. To ensure the validity and reliability as two central concepts and provide the researcher with insights of the feasibility, two types of piloting proposed by Dörnyei (2003, pp.66-67) including "initial piloting of the item pool" and "final piloting" were conducted as significant parts of research (Bryman, 2012). According to D rnyei and Taguchi (2010), preparing an initial item pool is the first occasion in the process of questionnaire construction in which a number of external feedback is essential and some of the questions might be reduced to the intended ultimate number. In the current study, after gathering the pertinent information from the initial piloting, the number of questions were reduced, the possible drawbacks of the researcher-developed questionnaire were identified, some "ambiguous", "negative constructions" including “not,” “doesn't," or “don’t”, and "double-barreled items" in which two or more questions are asked in one when a single answer is expected (Dörnyei, 2003, pp.54-55) were omitted and refined and then a "near-final version" of the questionnaire was developed (Dörnyei, 2003, p.67).

Subsequently, final piloting was carried out with 60 pilot samples. The usual sample size at this step is roughly 100 (± 20), but due to some statistical rationales, it should not be less than 50 (Dörnyei & Taguchi, 2010). This number allowed us to perform several meaningful item analyses through reducing the number of questions based on their low internal consistency reliability which was measured by the Cronbach Alpha coefficient before conducting further analysis; the negatively worded items were reversed to prevent response

ISSN 2476-5880

bias (Pallant, 2016). In this stage, considering if item deleted, Cronbach's alpha would be .79 from .72 if item 12 (I prefer to take English language courses during the doctoral course instead of taking part in EPT) was removed for the scale. Using the item analysis information, removing item 71(I blame myself for failing in this test) resulted in an increase in Cronbach's alpha from .77 to .86.

The total internal consistency reliabilities for the whole educational, social, and psychological consequences questionnaires were estimated .93, .81, and .89 in turn which were perfectly acceptable for further analysis. After more revision, three other items were totally removed due to the low reliability (0.4), and two items were transferred to a psychological questionnaire based on the participants' feedbacks. Additionally, the possible drawbacks of the researcher-developed questionnaires were identified, "ambiguous", "negative constructions" and "double-barreled items" (Dörnyei, 2003, pp.54-55) were omitted or refined and then "near-final version" of the questionnaires with 84 items were developed (Dörnyei, 2003, p.67). Afterward, online administration was utilized due to some reasons: One benefit of using online surveys was that we could get easy access to target populations who would otherwise be intricate to achieve. In this way, we were able to include scattered individuals living at a substantial geographical distance. Moreover, it helped us to save not only our time but also research costs. More importantly, a web-based survey was so effective to prevent encountering certain problems; for instance, missing data. However, despite the striking features, response rates in online administration are likely lower than return rates in traditional surveys (Dörnyei & Taguchi, 2010). The response rate was over 50 percent since 252 questionnaires out of 500 were completed and returned to us which were an acceptable rate for further analysis (Gillham, 2000).

5.2. Instruments Validation

In the current research, the validation process has been divided into two macro phases including Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) in which a number of organized micro processes have been included.

5.2.1. Exploratory factor analysis of ECQ.

EFA was applied to prove or decrease the number of factors (N=42). Because of cross-loadings of some items on more than one factor, 19 items were removed; thus, the number of items was reduced to 23. Then, Kaiser–Meyer–Olkin test (KMO) and Bartlett's Test of Sphericity were conducted to measure sampling adequacy and confirm the validity and appropriateness of the responses gathered to the problems being addressed in the study (Rokhsari, 2017). As illustrated in Table 3, the KMO value is estimated .93 (which is above 0.7). High value (close to 1.0) indicate that factor analysis may be useful with the data; furthermore, Bartlett's Test of Sphericity is less than 0.05 (sig <0.05), thus, the collected data are adequate to be examined via factor analysis.

 ISSN 2476-5880

Table 3.
KMO and Bartlett's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.930
Bartlett's Test of Sphericity	Approx. Chi-Square	4137.848
	Df	253
	Sig.	.000

The next output is a table of communalities which gives researchers information how much of the variance in each item can be explained and low values less than 0.3 specify that the item with low value does not fit well with other items in its components and should be removed to refine and improve the scale (Pallant, 2016). According to Table 4, all of the extraction communalities are above 0.3 which depict the appropriacy of all items in the factor analysis process.

Table 4.
Communalities.

	Initial	Extraction
LE1	1.000	.668
LE2	1.000	.725
LE3	1.000	.703
LE4	1.000	.740
SL1	1.000	.729
SL2	1.000	.673
DL1	1.000	.619
DL2	1.000	.652
DL3	1.000	.708
DL4	1.000	.716
DL5	1.000	.681
DL6	1.000	.794
DL7	1.000	.779
DL8	1.000	.684
DL9	1.000	.749
DL10	1.000	.755
LO1	1.000	.867
LO2	1.000	.873
LO3	1.000	.791
TP1	1.000	.600
TP2	1.000	.780
TP3	1.000	.788
TP4	1.000	.792

Note. LE= learning Environment; SL= Surface Learning; DL= Deep Learning; LO= Learning outcome; TP= Testing Preferences

ISSN 2476-5880

As Pallant (2016) puts, the next way of determining the number of factors to maintain is called parallel analysis. For this process, the list of eigenvalues presented in the Total Variance Explained table and some extra information can be applied. Using Kaiser's criterion or the eigenvalues rule, only factors with an eigenvalue of 1.0 or more should be retained for further research. As shown in Table 5, only the first five components recorded above 1 in the values provided in initial eigenvalue (10.306, 2.259, 1.890, 1.251, and 1.158). These five components explain a total 73.32 percent of the variance (above 55%) which manifest no difference from the initial solution (73.32). Thus, no variation explained by the initial solution is lost and variability simply can be explained by the factor model (IBM Knowledge Center, n.d.).

Table 5.
Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	10.306	44.811	44.811	10.306	44.811	44.811	9.240
2	2.259	9.823	54.634	2.259	9.823	54.634	4.625
3	1.890	8.219	62.853	1.890	8.219	62.853	6.112
4	1.251	5.439	68.292	1.251	5.439	68.292	6.711
5	1.158	5.037	73.328	1.158	5.037	73.328	3.784
6	.636	2.764	76.093				
7	.621	2.700	78.793				
8	.580	2.521	81.314				
9	.507	2.205	83.519				
10	.420	1.824	85.343				
11	.386	1.680	87.022				
12	.375	1.629	88.651				
13	.336	1.461	90.112				
14	.318	1.384	91.496				
15	.291	1.265	92.761				
16	.277	1.203	93.964				
17	.262	1.139	95.103				
18	.234	1.019	96.123				
19	.224	.972	97.095				
20	.212	.920	98.015				
21	.174	.756	98.771				
22	.153	.663	99.434				
23	.130	.566	100.000				

Component Correlation Matrix in the next Table shows the strength of the relationship among the majority of factors (above .3). It provides useful information to decide whether it is rational to the employ of matrix rotation or whether it is required to use, and

 ISSN 2476-5880

report the oblique rotation solution. Since, the strength of the relationship is above .3, oblique rotation is preferred (Pallant, 2016).

Table 6.
Component correlation matrix.

Component	1	2	3	4	5
1	1.000	.394	.532	.620	.395
2	.394	1.000	.323	.287	.223
3	.532	.323	1.000	.523	.423
4	.620	.287	.523	1.000	.405
5	.395	.223	.423	.405	1.000

Given the above points, it would be concluded that the Promax rotation would recover this structure better than the orthogonal Varimax since the underlying latent traits are more greatly correlated (Finch, 2006). The results demonstrate that factors 1, 2, 3, 4, and 5 consist of 4, 2, 10, 3, and 4 items respectively (Table 7).

Table 7.
Pattern matrix[a].

	Component				
	1	2	3	4	5
LE1			.880		
LE2			.892		
LE3			.798		
LE4			.752		
SL1					.907
SL2					.807
DL1	.448				
DL2	.785				
DL3	.869				
DL4	.659				
DL5	.801				
DL6	.961				
DL7	.982				
DL8	.652				
DL9	.952				
DL10	.759				
LO1				.962	
LO2				.942	
LO3				.839	
TP1		.708			
TP2		.871			
TP3		.897			
TP4		.894			

ISSN 2476-5880

5.2.2. Exploratory factor analysis of SCQ.

No item deletion was found based on EFA in this part. KMO value is estimated .82 and the Bartlett's Test of Sphericity is less than 0.05 (sig <0.05), so the collected data are sufficient to be examined using factor analysis.

Table 8.
KMO and Bartlett's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.825
Bartlett's Test of Sphericity	Approx. Chi-Square	1838.516
	df	66
	Sig.	.000

According to Table 9, all of extraction communalities are above 0.3 which portray all items are appropriate for the factor analysis process. Moreover,the communalities are all high, which specifies that the extracted components signify the variables well. If communalities for a variable are less than 0.3, that variable possibly will struggle to load considerably on any factor (IBM Knowledge Center, n.d.).

Table 9.
Communalities.

	Initial	Extraction
JP 1	1.000	.782
JP 2	1.000	.886
JP 3	1.000	.861
FE 1	1.000	.675
FE 2	1.000	.814
FE 3	1.000	.777
SJ 1	1.000	.623
SJ 2	1.000	.806
SJ 3	1.000	.726
SJ 4	1.000	.633
SJ 5	1.000	.709
SJ 6	1.000	.624

Note. JP= Job Promotion; FE= Financial Expenses; SJ= Social Justice

As illustrated in Table 10, these three components clarify a total 74.29 percent of the variance (above 55%). Thus, the complexity of the data set can considerably be reduced by using these components, with nearly 26% loss of information.

ISSN 2476-5880

Table 10.
Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	4.392	36.602	36.602	4.392	36.602	36.602	4.243
2	2.956	24.636	61.239	2.956	24.636	61.239	2.929
3	1.567	13.060	74.298	1.567	13.060	74.298	2.486
4	.582	4.850	79.148				
5	.503	4.189	83.337				
6	.483	4.027	87.364				
7	.354	2.951	90.315				
8	.332	2.767	93.082				
9	.270	2.249	95.331				
10	.225	1.878	97.209				
11	.185	1.541	98.750				
12	.150	1.250	100.000				

Table 11 also confirms that the strength of the relationship among the factors is very low (below 0.3). Therefore, the varimax and oblimin rotation will present the similar results. As asserted by Pallant (2016), numerous researchers conduct both varimax and oblimin rotation and subsequently report the most obvious and easiest one to interpret. Following this, Table 12 displays that factors 1, 2, and 3 consist of 3, 3 and 6 items in turn which are labeled as job promotion, financial expenses and social justice.

Table 11.
Component correlation matrix.

Component	1	2	3
1	1.000	.227	-.093
2	.227	1.000	.267
3	-.093	.267	1.000

 ISSN 2476-5880

Table 12.
Pattern matrix[a].

	Components		
	1	2	3
JP 1		.874	
JP 2		.931	
JP 3		.945	
FE 1			.839
FE 2			.911
FE 3			.846
SJ 1	.787		
SJ2	.913		
SJ 3	.839		
SJ 4	.767		
SJ 5	.847		
SJ 6	.785		

5.2.3. Exploratory factor analysis of PCQ.

Conducting the first EFA led to the deletion of 4 items due to items crosssloading, afterwards, 16 items remained for further analysis (Table, 13). Additionally, all of extraction communalities are above 0.3 so all items are suitable for the factor analysis procedure (Table. 14).

Table 13.
KMO and Bartlett's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.889
Bartlett's Test of Sphericity	Approx. Chi-Square	2172.771
	df	120
	Sig.	.000

Table 14.
Communalities.

	Initial	Extraction
SEB 1	1.000	.714
SEB 2	1.000	.779
SEB 3	1.000	.674
F1	1.000	.510
F2	1.000	.719
F3	1.000	.699
F4	1.000	.699
F5	1.000	.488
F6	1.000	.598
IM 1	1.000	.753
IM 2	1.000	.888
IM 3	1.000	.819
IM4	1.000	.807
EM 1	1.000	.651
EM 2	1.000	.645
EM 3	1.000	.546

Note. SEB=Self- Efficacy Belief; F=Feeling; IM= Intrinsic Motivation; EM= Extrinsic Motivation

ISSN 2476-5880

As clarified in Table 15, four components enlighten a total 68.67 percent of the variance (above 55%). Following this, Table 16 verifies that most of the factors are correlated (above .3) expect item 4 but Pattern Matrix[a] (Table 17) is preferable since it provides simple structure (J.D. Brown, 2009).

Table 15.
Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	6.512	40.703	40.703	6.512	40.703	40.703	5.225
2	1.970	12.314	53.017	1.970	12.314	53.017	5.287
3	1.427	8.919	61.935	1.427	8.919	61.935	4.075
4	1.079	6.741	68.676	1.079	6.741	68.676	1.892
5	.783	4.895	73.572				
6	.618	3.861	77.433				
7	.583	3.646	81.079				
8	.532	3.328	84.406				
9	.486	3.037	87.443				
10	.450	2.815	90.258				
11	.378	2.362	92.620				
12	.320	2.001	94.621				
13	.304	1.900	96.521				
14	.230	1.437	97.958				
15	.198	1.238	99.196				
16	.129	.804	100.000				

Table 16.
Component correlation matrix.

Component	1	2	3	4
1	1.000	.552	.470	.034
2	.552	1.000	.547	.248
3	.470	.547	1.000	.079
4	.034	.248	.079	1.000

ISSN 2476-5880

Table 17.
Pattern Matrix[a].

	Components			
	1	2	3	4
SEB 1			.781	
SEB 2			.928	
SEB 3			.844	
F1	.807			
F2	.800			
F3	.701			
F4	.781			
F5	.747			
F6	.762			
IM 1		.900		
IM 2		.892		
IM 3		.948		
IM 4		.879		
EM 1				.819
EM 2				.756
EM 3				.595

5.3. Confirmatory factor analysis (CFA).

At the second phase of conducting factor analysis, CFA was applied to confirm factor structure attained in the EFA. As Howitt and Cramer (2000) assert CFA proves that the factor structure attained in the EFA is robust and not only the outcome of accidental variability in data. To achieve the goal, Smart PLS 3 was utilized to do required statistical analyses.

5.3.1. Confirmatory factor analysis of ECQ.

Outer model loadings in Figure 1 show that all of items with measurement loading are above .40; therefore, they should be retained in the study. On the whole, the loading fluctuates from 0 to 1 and the larger the loadings, the more reliable and stronger the measurement model would be (Hair, Hult, Ringle, & Sarstedt, 2014).

 ISSN 2476-5880

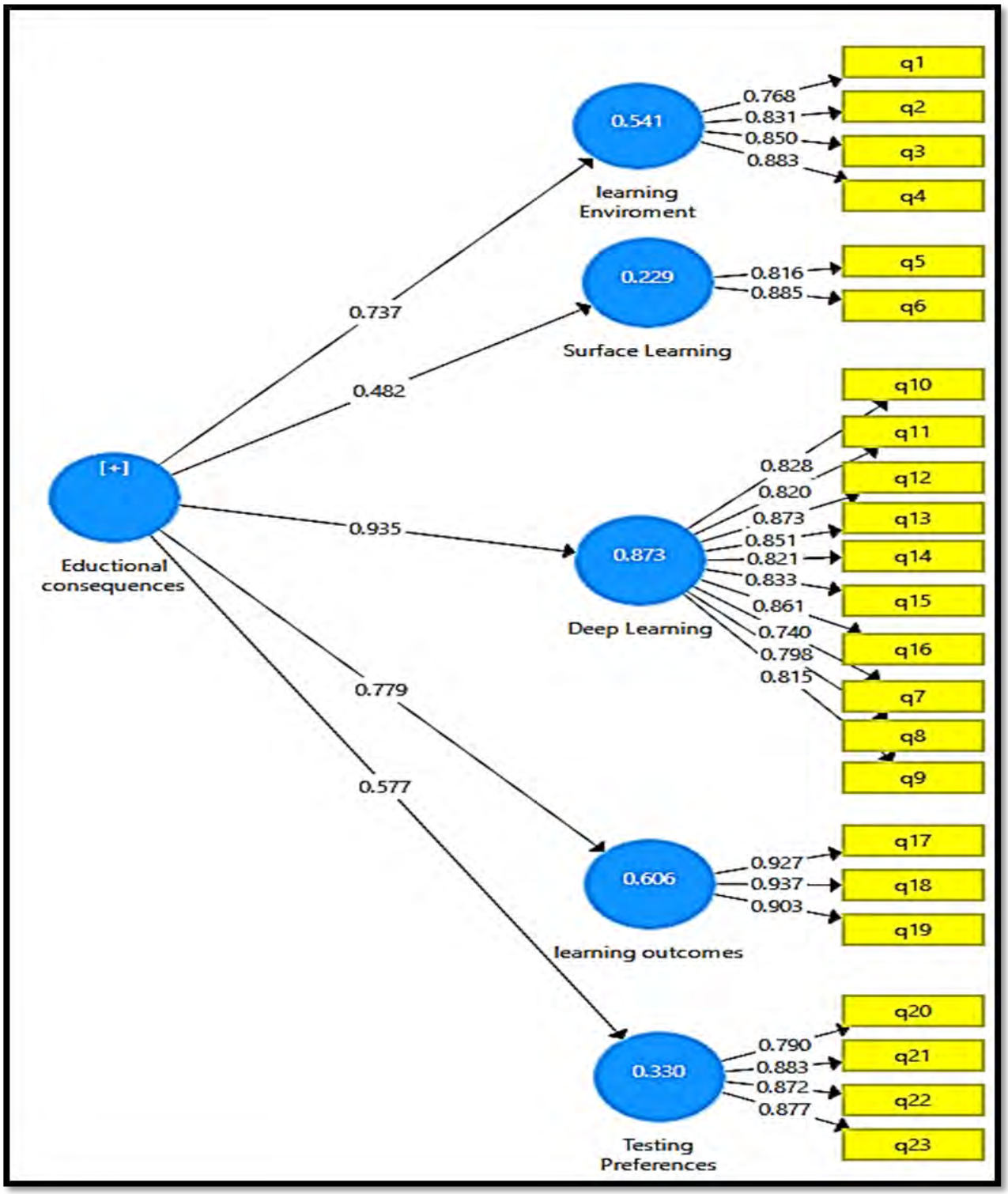

Figure 1. Confirmatory factor analysis of educational consequences.

5.3.1.1. Goodness of fit for measurement models.

As Garson (2016) states, global goodness of fit measure is not available in PLS-SEM. For that reason, the SmartPLS report presents different coefficient to model fit instantly after the listing of the input data including Composite Reliability, Cronbach's Alpha, Average Variance Extracted (AVE), and Heterotrait-Monotrait Ratio table to estimate discriminant

ISSN 2476-5880

validity criterion. As Shown in Table 18, the reliability values above 0.60 to 0.70 are acceptable in exploratory research (Hair, Risher, Sarstedt, & Ringle , 2018). Composite Reliability is preferable among PLS-based study since Cronbach's Alpha may under or overestimate scale reliability. Composite Reliability varies from 0 to 1. For an adequate model, it should be equal to or higher than.70. Lastly, AVE which is applied to test both convergent and divergent validity should be above 0.50 (Hair et al., 2018; Hock & Ringle, 2010). Heterotrait-Monotrait Ratio (HTMT) table is the last step to check discriminant validity in which threshold values should be less than 0.90 (Table 19).

Table 18.
Construct reliability and validity.

Subscales	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
TP	0.878	0.878	0.917	0.733
DL	0.947	0.948	0.955	0.680
SL	0.624	0.643	0.840	0.725
LE	0.854	0.865	0.901	0.696
LO	0.912	0.912	0.945	0.851

Note. TP= Testing Preferences; DL= Deep Learning; SL= Surface Learning; LE= Learning Environment; LO= Learning Outcomes

Table 19.
Heterotrait-monotrait ratio (HTMT).

Subscales	Testing Preferences	Deep Learning	Surface Learning	learning Environment	learning Outcomes
TP					
DL	0.458				
SL	0.271	0.462			
LE	0.405	0.612	0.517		
LO	0.364	0.698	0.463	0.581	

Note. TP= Testing Preferences; DL= Deep Learning; SL= Surface Learning; LE= Learning Environment; LO= Learning Outcomes

5.3.1.2. Goodness of fit for structural models.

Subsequent to measurement fit is confirmed to be acceptable, the results of Structural Models are assessed though Smart PLS. R^2 varies from 0 to 1 and higher values indicate a greater explanatory power. The results above 0.67, 0.33 and 0.19 are considered as substantial, moderate and weak in turn in line with Hock and Ringle (2010).

ISSN 2476-5880

Table 20.
R-square (R2).

Subscales	R Square	R Square Adjusted
TP	0.333	0.330
DL	0.873	0.873
SL	0.232	0.229
LE	0.543	0.541
LO	0.607	0.606

Note. TP= Testing Preferences; DL= Deep Learning; SL= Surface Learning; LE= Learning Environment; LO= Learning Outcomes

Predictive relevance (Q^2) evaluates the predictive validity of the a large complex model using PLS and it can be obtained through two types of prediction techniques; that is, Construct Crossvalidated Redundancy and Construct Crossvalidated Communality (Akter, D'Ambra, & Ray, 2011). Garson (2016) puts, Q2 greater than 0 is regarded as a predictive model. While a Q2 with 0 or negative value shows the model is irrelevant to prediction of the given endogenous factors. As a rule, 0.02, 0.15, and 0.35 values present small, medium and high effect size (Cohen, 1988). On the basis of the following tables, it can be concluded that the model has a high degree of predictive relevance.

Table 21.
Construct crossvalidated redundancy.

	SSO	SSE	Q^2 (=1-SSE/SSO)
TP	1,008.000	773.267	0.233
DL	2,520.000	1,084.533	0.570
SL	504.000	423.519	0.160
LE	1,008.000	642.640	0.362
LO	756.000	378.974	0.499

Note. TP= Testing Preferences; DL= Deep Learning; SL= Surface Learning; LE= Learning Environment; LO= Learning Outcomes

Table 22.
Construct crossvalidated communality.

	SSO	SSE	Q^2 (=1-SSE/SSO)
TP	1,008.000	584.764	0.420
DL	2,520.000	1,396.517	0.446
SL	504.000	400.602	0.205
LE	1,008.000	595.852	0.409
LO	756.000	388.293	0.486

Note. TP= Testing Preferences; DL= Deep Learning; SL= Surface Learning; LE= Learning Environment; LO= Learning Outcomes

5.4. Confirmatory factor analysis of SCQ.

 ISSN 2476-5880

Outer model loadings in figure 2 confirm that all of items with measurement loading are above .40; consequently, the items should be retained in this investigation. As noted, the loading fluctuates from 0 to 1 and the larger the loadings, the more reliable the measurement model would be (Hair et al., 2014).

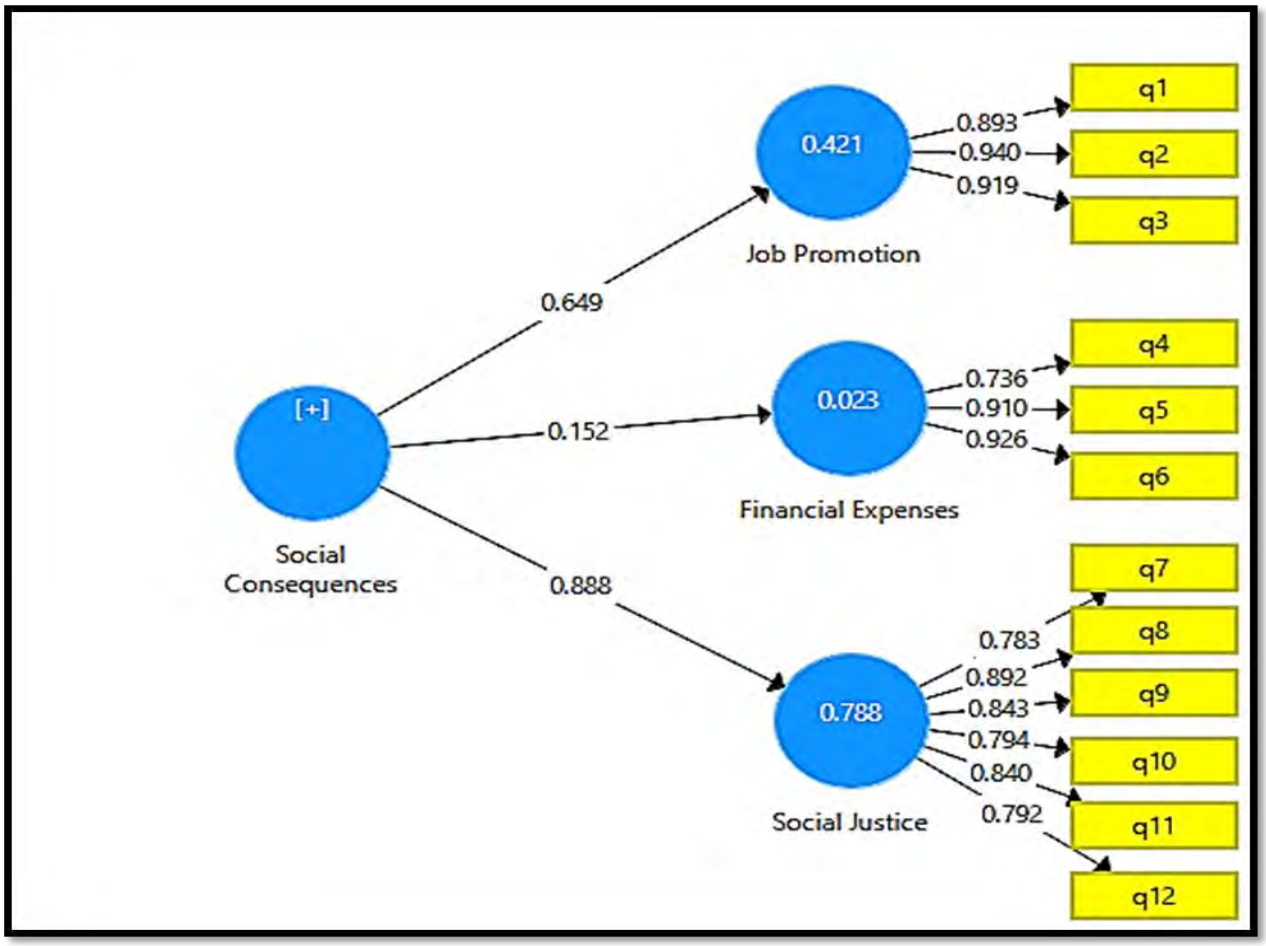

Figure 2. Confirmatory factor analysis of social consequences.

5.4.1. Goodness of fit for measurement models.

As illustrated in Table 23, the reliability values are higher than 0.70. Composite Reliability values are greater than.70. Finally, AVE values are above 0.50. Heterotrait-Monotrait Ratio (HTMT) in Table 24 shows that discriminant validity threshold values are less than 0.90.

Table 23.
Construct reliability and validity.

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
FX	0.833	0.933	0.896	0.743
JP	0.906	0.908	0.941	0.842
SJ	0.905	0.907	0.927	0.680

Note. FX= Financial Expenses; JP= Job Promotion; SJ=Social Justice

 ISSN 2476-5880

Table 24.
Heterotrait-monotrait ratio (HTMT).

	FX	JP	SJ
FX			
JP	0.300		
SJ	0.111	0.260	

Note. FX= Financial Expenses; JP= Job Promotion; SJ=Social Justice

5.4.2. Goodness of fit for structural models.

In conformity with Cohen (1988), the results of R^2 above 0.26, 0.13 and 0.02 are considered as large, medium and small effect sizes respectively. Given to Tables 26 and 27, the model has an acceptable degree of predictive relevance.

Table 25.
R-square (R2).

	R Square	R Square Adjusted
FX	0.023	0.019
JP	0.421	0.418
SJ	0.788	0.787

Note. FX= Financial Expenses; JP= Job Promotion; SJ=Social Justice

Table 26.
Construct crossvalidated redundancy.

SSO	SSE	Q^2 (=1-SSE/SSO)	
FX	756.000	748.735	0.010
JP	756.000	501.495	0.337
SJ	1,512.000	734.001	0.515

Note. FX= Financial Expenses; JP= Job Promotion; SJ=Social Justice

Table 27.
Construct crossvalidated communality.

SSO	SSE	Q^2 (=1-SSE/SSO)	
FX	756.000	456.084	0.397
JP	756.000	383.634	0.493
SJ	1,512.000	862.864	0.429

Note. FX= Financial Expenses; JP= Job Promotion; SJ=Social Justice

5.5. Confirmatory factor analysis of PCQ.

As depicted in figure 3, all of factor loading values are above 0.40 except item 15 which must be omitted (0.329).

 ISSN 2476-5880

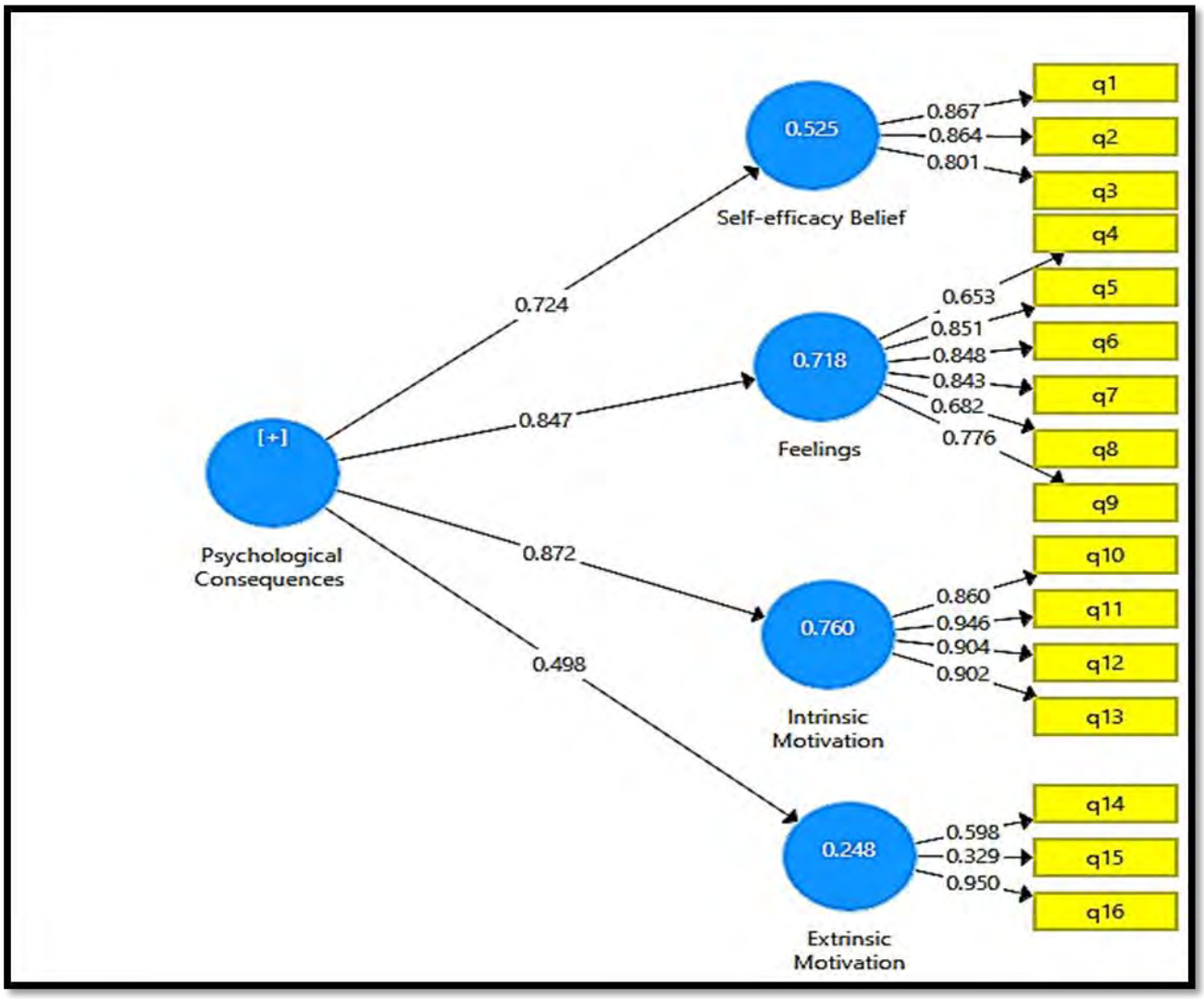

Figure 3. Confirmatory factor analysis of psychological consequences.

5.5.1. Goodness of fit for measurement models.

According to Table 28, most of the reliability values are higher than 0.70 except extrinsic motivation but since its Composite Reliability value is 0.76, it can be acceptable in this stage. Moreover, AVE values area above 0.50. On the basis of Heterotrait-Monotrait Ratio (HTMT) table, discriminant validity threshold values are smaller than 0.90 (Table. 29).

Table 28.
Construct reliability and validity.

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
EM	0.484	0.757	0.766	0.632
F	0.868	0.884	0.902	0.608
IM	0.925	0.928	0.947	0.816
SEB	0.799	0.805	0.882	0.713

Note. EM=Extrinsic Motivation; F=Feelings; IM=Intrinsic Motivation; SEB= Self-Efficacy Belief

Table 29.
Heterotrait-monotrait ratio (HTMT).

	Extrinsic Motivation	Feelings	Intrinsic Motivation	Self-Efficacy Belief
EM				
F	0.366			
IM	0.573	0.618		
SEB	0.350	0.545	0.618	

Note. EM=Extrinsic Motivation; F=Feelings; IM=Intrinsic Motivation; SEB= Self-Efficacy Belief

5.5.2. Goodness of fit for structural models.

The results of R^2 Table prove that values are adequate for the present study. Construct Crossvalidated Redundancy and Construct Crossvalidated Communality tables manifest that Q2 are greater than 0; thus, they are regarded appropriate for a predictive model.

Table 30.
R square(R2).

	R Square	R Square Adjusted
EM	0.240	0.237
F	0.717	0.716
IM	0.761	0.760
SEB	0.525	0.523

Note. EM=Extrinsic Motivation; F=Feelings; IM=Intrinsic Motivation; SEB= Self-Efficacy Belief

Table 31.
Construct crossvalidated redundancy.

SSO	SSE	Q² (=1-SSE/SSO)
EM	504.000	436.162	0.135
F	1,512.000	880.148	0.418
IM	1,008.000	405.606	0.598
SEB	756.000	484.236	0.359

Note. EM=Extrinsic Motivation; F=Feelings; IM=Intrinsic Motivation; SEB= Self-Efficacy Belief

Table 32.
Construct crossvalidated communality.

SSO	SSE	Q² (=1-SSE/SSO)
EM	504.000	458.004	0.091
F	1,512.000	952.564	0.370
IM	1,008.000	496.278	0.508
SEB	756.000	500.219	0.338

Note. EM=Extrinsic Motivation; F=Feelings; IM=Intrinsic Motivation; SEB= Self-Efficacy Belief

ISSN 2476-5880

6. Discussion and Conclusion

As mentioned earlier, although educational, social and psychological issues of language testing play vital roles in determining individuals' future, it appears that sizeable literature has pained attention to the small effects of high-stakes testing on individuals but the issue of language testing consequences in educational, social and psychological scope in broader sense is under research. The principal objective of this study was to construct and validate educational, social and psychological consequences questionnaires of EPT through exploring its underlying educational, social and psychological constructs among Iranian PhD non-English students and give voice to students as one of the most important stakeholders in language testing. In order to achieve the goals, the researcher-made questionnaires validated in the following three steps: (1) reviewing the related literature and document (2) conducting EFA in order to determine the number of underlying factors, (3) performing CFA to confirm the underlying factors of the questionnaire. The statistical analyses of EFA and EFA were supported five components in educational consequences questionnaire (learning environment, surface learning, deep learning, learning outcomes and testing preferences; three social consequences components (job promotion, financial expenses and social justice); in addition to four psychological consequences components (self-efficacy beliefs, feelings, intrinsic motivation, extrinsic motivation).

In educational consequences questionnaire, item 1 (Instructors play a supportive role for students participating in test preparation classes/ courses), (and item 2 (Instructors provide appropriate answers to students' concerns about this test in preparation classes / courses) refer to the supportive role that instructors play in front of students who take classes or preparation courses for this exam. Teacher support is considered as one of the three components of the language classroom environment which is an ecological standpoint about individual development (Peng & Woodrow, 2010) in which they support, assist, trust, befriend, and are paying attention to the students (Dorman, Fisher, & Waldrip, 2006). Moreover, items 3 and 4 (Assignments designed for EPT preparation classes or courses are clear) in addition to (Assignments designed for EPT preparation classes or courses are useful) focus on task orientation that implies the degree to which it is essential to finish activities and elucidate problems (Dorman et al., 2006). As Kubanyiova (2007) notes, constructive and interesting tasks can lead to learner's engagement, in fact, tasks can increase performance when they are pertinent and meaningful with a sensible extent of difficulty. Falout, Elwood, and Hood (2009) also confirm that learning environments that need high-stakes testing and provide unsuitable level of courses possibly will cause some motivational troubles.

Items 5 and 6 (EPT makes students memorize test-related contents and this test increases superficial learning among students) are related to uncover the participants' views on surface learning. High-stakes testing may persuade students as passive learners to focus on surface or superficial approaches to learning rather than deep or meaningful learning (Newstead & Findlay, 1997). Items 7-16 (Appendix I) have been included to understand about deep learning (This test introduces students to more useful learning methods in English; makes students analyze English; makes students produce English; helps students to better understand English; the questions of this test are related to the actual use of the

ISSN 2476-5880

language (writing, speaking, reading, etc.) English; this test makes students use English in a real environment in the future and lastly this test improves students' speaking, reading, listening and writing skills) in which further quantitative analyses can reveal participants' perceptions about the amount, types, quality of learning, as well as the authenticity of the test items or their usage in the real world. The next subscale which is named learning outcomes, items 17-19, refers to what EPT can bring for its participants; that is, whether it can improve their vocabulary knowledge; grammar knowledge or their translation skills. Lastly, item 20 (I prefer all language skills, including speaking and writing, to be assessed in this test), item 21 (I prefer questions that require more or deeper thinking), item 22 (I prefer questions that require creativity), and item 23 (I prefer questions that need further analysis) are going to estimate the participants' EPT testing preferences in which they can express their opinions on enhancing their critical thinking as their preference via EPT since critical thinking is an dynamic process of thinking and analyzing what students obtain rather than simply achieving and accepting information (Fisher, 2001).Unluckily, tests can not improve the growth of critical thinking skills as The test-takers may possibly memorize the materials to take the test (Bell, 2000). Critical thinking can be linked to higher order thinking in Bloom's taxonomy in which a number of levels of learning may cause diverse levels of critical thinking which lead to deep learning (Huffman, Carson & Simonds, 2000).

Additionally, social consequences questionnaire with three subscales demonstrates that items 1, 2 and 3 (The result of this test helps me to find my job easier / faster; the result of this test will help me get a job promotion; the result of this test improves my salary or income status) refer to participants' point of views on job promotion to see whether they believe the result of EPT helps them to get job promotion as well as salary increase as social consequences of the test or not? Items 4-6 (This test makes me pay for the purchase of books / educational materials related to the test; this test makes me pay for a private class; this test makes me pay for traveling to another city .)are labeled as financial expenses that are brought via EPT such as books industry; educational materials expenses along with tutoring fees. As noted by Minarechová (2012), high-stakes testing is concerned with the issue of money as well as finance. For example, parents or families need to assist test-takers financially to afford tutoring or test preparation materials expenses (Pan, 2000). Questions **7**-12(Appendix II) are designed to evaluate social justice in EPT; that is, whether the test is the same for all participants, regardless of some issues such as social status, religion, ethnicity / race just to name a few. The focus of this subscale is on justice with regards to "social equity" (Mcnamara & Ryan, 2011, p. 165). As asserted by Kunnan (2000), concept of societal equity concentrates on the social outcomes or consequences of testing with reference to whether testing agendas are contributed to social equity or whether there are any destructive effects because of them or not. As a result, similar to educational consequences, intended or unintended social consequences may possibly be brought through the use of the test (Pan, 2009b) as they are influential tools in control social order (Hamp-Lyons, 2000).

The last questionnaire is titled as psychological consequences questionnaire with 4 subscales. Self-efficacy belief as the first construct consists of 3 items. In this questionnaire, item 1 (I believe I have the ability to pass this test), item 2 (I'm sure I can improve my

ISSN 2476-5880

English with more effort), and item 3(I think learning English is very easy) have been designed to specify participants' judgment of their own capabilities. Self- efficacy beliefs should be taken into consideration since they have an indirect outcomes or impacts on the participants' English language test scores (Zimmerman & Kitsantas, 2005).

Questions 4-9(Appendix III) have been included and named as feelings. As note by Shohamy (1982), affective considerations such as stress, anxiety, nervous and aggressiveness should not be ignored or underestimated in language testing as they have taken a noteworthy place in education. The third one is labeled as intrinsic motivation in which participants demonstrate their perspectives on their inner force to achieve their goal. This subscale consists of question 10 (Preparing for the test increases my interest in learning English), question 11(Preparing for this test makes the learning process enjoyable for me), question 12 (Preparing for this test makes me aware of the value of learning English), and question 13 (Preparing for this test gives me inner satisfaction). Item 14 (I want to pass this exam to get a job promotion) in addition to item 15(Making a distinction among the participants based on the result of this test will increase my motivation to study) are labeled as extrinsic motivation which are derived from external some factors.

The findings of the present research accord with our earlier studies, which showed that how deeply rooted in social and education systems the current language testing process is. As an example, A survey study in this area conducted by Al Amin and Greenwood (2018) demonstrated that due to washback effects, teachers diminished classroom teaching and learning to a curriculum focused almost totally on what was expected in the test. Additionally, testing was likely to decrease academic curiosity, applied huge pressure on students' lives, and supported an education industry regarding "coaching centers and commercial publishers, and advantage those who can afford to pay for extra tuition" (p.15). Interestingly, the chance to earn further income via after-school coaching also persuaded low-paid teachers to alter their attention from making effort to meet national curriculums objectives to focusing on the limited framework of examinable materials (Al Amin & Greenwood, 2018).

A recent study by Castro and Vega (2017) confirmed that washback effect could change the students' perceptions and attitudes toward language testing in various ways. In another research, Simpson (2016, p.14) found that "the effects of standardized testing on students' physical and emotional well-being are worrisome". He concluded that these effects were various from occurrence of "vomiting to anxiety attacks" (p.14). Accordance with the present results, a range of the collected data confirmed that students were influenced harmfully from high-stakes testing. For instance, they were anxious and frustrated about testing (Fitzgerald, 2015). According to Pan (2009b), some students endeavored greatly to study English regarding the exit requirement since they were afraid of not being able to graduate. Nevertheless, they complained it was unfair to prevent them from graduation due to a failure to pass the English certification test. Meanwhile, the financial burden because of taking lessons in order to pass the test was an extra negative washback emerged by the exit requirement (Shen as cited in Pan, 2009b). In a fascinating study conducted by Wheelock,

ISSN 2476-5880

Bebell, and Haney (2000), students drew themselves as irritated, nervous, tired, and negative when asked to illustrate a self-portrait in testing situations.

By similarity, some research conducted in Iranian context also manifested that the effects of language testing cannot be overlooked both at micro and macro levels. The findings of the recent study done by Mohammad Salehi and Tarjoman (2017) indicated that washback considerably influenced both the learning and teaching methodologies. They found that MA Entrance Exam divided the curriculum into relevant and irrelevant or important and unimportant sections since teachers got interested in improving the scores of the students; thus, it created fear in students in the process of learning. Furthermore, the results of their study specified that every university professors and the most of the students wanted to have control over some matters such as the content of the test and the time of the administration. Students also demanded that MA Entrance Exam should be applied as an indicator of their language ability or knowledge instead of test-taking skills.

A correlational study revealed that test anxiety had debilitative impacts in language learning and it is was negatively associated to foreign language test performance (Massomeh Salehi & Marefat, 2014). As asserted by Kheirkhah and Ghonsooly (2014), Iranian English university entrance examination for the Humanities (IEUEEH) negatively affected the Iranian English language teaching and learning program. In the same way, Tahmasebi and Yamini's study (2013) proved that University Entrance Examinations (IUEE) might be tools of power that use to give power to parties' policies and influence stakeholders' lives. Mohammadi (2010) uncovered that the conditions of this kind of exam made the students as well as their families anxious. Secondly, he found that if the students could not perform the exam successfully, they would possibly be demotivated and frustrated. He also found MA Entrance Examination in Iran as a high-stakes test worked powerfully in order to lead to changes. he asserted that these tests were capable to influence not only the participants, but also process and product of an educational system.

Furthermore, several studies have been published on motivation and autonomy such as Bravo, Intriago, Holguín, Garzon and Arcia (2017); Chan (2016); Kelly (2014); Lamb (2009); Ushioda (2011); Ushioda and Dörnyei (2009); just to name a few and the relationship between them are well established in these studies. To date, it seems that there is a relatively small body of literature that is concerned with motivation and autonomy in language testing. the results of the study seem to be consistent with other research conducted by Buyukkeles (2016) who found that the foreign language achievement test (FLAT) as the exit test pushed a reasonable number of students to do test-related language practice including vocabulary and grammar autonomously. On the contrary, the results of the study indicated that FLAT had no substantial washback on students' intrinsic motivation; while, the amount of washback on students extrinsic motivation was significant.

Although every precaution was taken to carefully prepare and conduct the research, the present study suffered from some shortfalls. Namely, in terms of methodology, the present research was limited in a number of ways. First, this study was based on voluntary participation of candidates. Thus, the volunteer participants may share some common features not presented in those member of population not willing to take part in the study.

ISSN 2476-5880

This could question the generalizability of the findings of the study. Because of limited access to eligible participants, some demographic factors as gender, socioeconomic status and work experience and their effects were not controlled and investigated in the current study. The response rate of the questionnaires is one of the limitations that the researchers encountered in the research. The implications of this study are the possibility that individuals especially policy makers, test developers, teachers, or learners may benefit from its practical knowledge. It possibly will lead to new policies about administering tests through enhancing their awareness of the educational, social and psychological consequences of EPT as a high-stakes test among main stakeholders. Additionally, through identifying test-takers' experiences, both classroom teachers and test designers may incorporate test processes and procedures too much better to make sure that their interpretations and use of test scores are precise and accurate (Cheng & DeLuca, 2011).

The present study might have a number of influential implications thinking about testing and assessment in a new way by listening to students voices as marginalized agents silenced, not heard, or rejected in language testing as it is essential to study how test-takers evaluate testing events, and how their experiences might be contributed to test validity (Cheng & DeLuca, 2011). A high-stakes language test can bring intended and unintended consequences for its stakeholders including students given to the fact that their test performance might have effects on their occupational and educational promotion, employment status, social and psychological health just to name a few. To develop a full picture of intended and unintended consequences of EPT, additional studies will be needed that explore the other stakeholders' voices such as families or test developers. Further studies which take demographic information into account will need to be undertaken.

Further investigations are also recommended to develop a full picture of educational, social, and psychological factors in high-stakes testing quantitatively via utilizing these questionnaires among non-English PhD students in Iran. Furthermore, supplementary research is required to explore the voice of teachers and parents as the main stakeholders regarding educational, social, and psychological consequences of EPT as a high-stakes test in Iranian context. There is also room to do further studies to explore teachers' preferences between testing and assessment and ask them to provide their reasons to support their choice. Lastly, additional studies on some demographic variables such as age, gender, and economic status of participants will be useful in this field as well.

References

Ahmadjavaheri, Z., & Zeraatpishe, M. (2020). The impact of construct-irrelevant factors on the validity of reading comprehension tests. *International Journal of Language Testing*,*10*(1).1-10.

Akter, S., D'Ambra, J., & Ray, P. (2011). An evaluation of PLS based complex models: The roles of power analysis, predictive relevance and GoF index. *Proceedings of the 17th Americas Conference on Information Systems* (pp. 1-7). Detroit, USA: Association for Information Systems.

AlAmin, M., & Greenwood, J. (2018). The examination system in Bangladesh and its impact: On curriculum, students, teachers and society. *Language Testing in Asia*, *8*(4), 1-18. doi: 10.1186/s40468-018-0060-9

Alemi, M., Miri, M., & Mozafarnezhad, A. (2019). Investigating the effects of online concurrent group dynamic assessment on enhancing grammatical accuracy of EFL learners. *International Journal of Language Testing, 9*(2), 29-43.

Ali, M. M., Hamid, M. O., & Hardy, I. (2020). Ritualisation of testing: problematising high-stakes English-language testing in Bangladesh, Compare. *A Journal of Comparative and International Education, 50*(4), 533-553.doi: 10.1080/03057925.2018.1535890

Au, S. Y. (1988). A critical appraisal of Gardner's social-psychological theory of second-language (L2) Learning. *Language Learning, 38*(1), 75-99.doi:10.1111/j.1467-1770.1988.tb00402.x

Bell, J. (2000). Framing and text interpretation across languages and cultures: A case study. *Language Awareness, 9*(1), 1-16.doi:1080/09658410008667133

Bravo, J. C., Intriago, E. A., Holguin, J. V., Garzon, G. M., & Arcia, L. O. (2017). Motivation and autonomy in learning English as foreign language: A case study of Ecuadorian college students. *English Language Teaching, 10*(2), 100-113.doi:10.5539/elt.v10n2p100

Brown, H. D. (2004). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13*(3), 20-25.

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford: Oxford University Press.

Buyukkeles, G. (2016). *The washback effect of a high-stakes exit test on students' motivation in a Turkish pre-university EFL preparatory school* (Master's dissertation). University of Reading, UK, Reading.

Castro, M. C., & Vega S. A. L. (2017). *The impact of tests in students' perceptions and attitudes towards their learning task* (Master's thesis). Universidad Distrital Francisco José de Caldas: Bogotá, Colombia.

Chan, H. W. (2016). Popular culture, English out-of-class activities, and learner autonomy among highly proficient secondary students in Hong Kong. *Universal Journal of Educational Research, 4*(8), 1918-1923. doi:10.13189/ujer.2016.040823

Cheng, L., & Deluca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104-122. doi:10.1080/10627197.2011.584042

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*(4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Dorman, J. P., Fisher, D. L., & Waldrip, B. G. (2006). Learning environments, attitudes, efficacy and perceptions of assessment: A LISREL analysis. In D. L. Fisher & M. S. Khine (Eds.), *Contemporary approaches to research on learning environments* (pp. 1–28). Singapore: World Scientific.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

D rnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research construction, administration, and processing* (2nd ed.). NewYork: Routledge

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation*. Harlow: Pearson Education.

Falout, J., Elwood, J., & Hood, M. (2009). Demotivation: Affective states and learning outcomes. *System, 37*(3), 403-417.

Finch, H. (2006). Comparison of the Performance of Varimax and Promax Rotations: Factor Structure Recovery for Dichotomous Items. *Journal of Educational Measurement, 43*(1), 39-52. doi:10.1111/j.1745-3984.2006.00003.x

Fisher, A. (2001). *Critical thinking: An introduction*. Cambridge: Cambridge University Press.

Fitzgerald, L. (2015). *Consequences of high-stake testing* (Master's thesis). Rochester, NY: St. John Fisher College, Fisher Digital Publications.

Garson, G. D. (2016). *Partial least squares: Regression and structural equation models.* Asheboro, NC: Statistical Associates Publishers Publications.

Gillham, B. (2000). *Developing questionnaire*. London: Continuum.

Gunn, J., Al-Bataineh, A., & Abu Al-Rub, M. (2016). Teachers' perceptions of high-stakes testing. *International Journal of Teaching and Education, 6*(2), 49-062.doi: 10.20472/TE.2016.4.2.003

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks, CA: Sage Publications.

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2018). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2-24. doi:10.1108/ebr-11-2018-0203

Hammersley, M., & Traianou, A. (2012). *Ethics in qualitative research: Controversies and contexts*. Los Angeles: Sage Publications.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*(4), 579-591.

Hock, M., & Ringle, C. M. (2010). Local strategic networks in the software industry: An empirical analysis of the value continuum. *International Journal of Knowledge Management Studies 4*(2), 1-15. doi: 10.1504/IJKMS.2010.030789

Howitt, D., Cramer, D. (2000). *An introduction to statistics in psychology: A complete guide for students* (2nd ed.). Prentice Hall, Hemel Hempstead.

Huffman, K.J., Carson, C.L., & Simonds, C.J. (2000). Critical thinking assessment: The link between critical thinking and student application in the basic course. *Basic Communication Course Annual, 12*(7), 60-96.

Hung, S.A., Huang, H.D. (2019). Standardized proficiency tests in a campus-wide English curriculum: *a washback study. Language Testing* in *Asia, 9*(21), 1-17. doi:10.1186/s40468-019-0096-5

IBM Knowledge Center. (n.d.). Retrieved from https://www.ibm.com/support/knowledgecenter/en/

Im, G.H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(14), 1-26, doi: 10.1186/s40468-019-0089-4

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito.

Kelly, M. (2014). *Fostering autonomy, generating motivation and shaping identities in the adolescent language classroom: An experimental research project* (Doctorial thesis), Dublin City University, Dublin.

Khajavy, G. H., Ghonsooly, B., HosseiniFatemi, A., & Choi, C. W. (2014). Willingness to communicate in English: A microsystem model in the Iranian EFL classroom context. *TESOL Quarterly, 50*(1), 154-180, doi:10.1002/tesq.204

Kheirkhah, H., & Ghonsooly, B. (2014). Qualitative study of Iranian English university entrance examination in the light of positive washback strategies. *Studies in English Language Teaching, 2*(1), 37-65. doi:10.22158/selt.v2n1p38

Kubanyiova, M. (2007). *Teacher development in action: An empirically based model of promoting conceptual change in in-service language teachers in Slovakia* (Unpublished doctoral dissertation). Nottingham, England: University of Nottingham.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, UK: Cambridge University Press.

Lamb, T. (2009). Controlling learning: Learners' voices and relationships between motivation and learner autonomy. In S. Toogood, R. Pemberton & A. Barfield (Eds.), *Maintaining control: Autonomy and language learning* (pp. 67-86). Hong Kong: Hong.

Mazandarani, O. (2014). *EFL Lecturers' perceptions of teacher effectiveness and teacher evaluation in Iranian universities* (Doctorial thesis). University of Exeter, Exeter.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8(*2), 161-178. doi: 10.1080/15434303.2011.565438

Menken, K. (2017). High-stakes tests as de facto language education policies. In E. Shohamy, I.G. Or & S. May (Eds.), *Language testing and assessment. Encyclopedia of language and education* (pp. 385 -397). Cham, Switzerland: Springer. doi: 10.1007/978-3-319-02261-1_35

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20. doi:10.2307/1174731

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). London: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256. doi:10.1177/026553229601300302

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy, 3*(1), 82-100. doi: 10.2478/v10159-012-0004-x JoP

Mohammadi, M., (2010). *High-stakes testing washback: A survey on the effect of Iranian MA entrance examination on teaching*. Paper presented at 19th MELTA International Conference, Kuala Lumpur, Malaysia.

Newstead, S. E., & Findlay, K. (1997). Some problems with using examination performance as a measure of teaching ability. *Psychology Teaching Review, 6(1),* 23-30.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using SPSS*. Maidenhead: Open University Press.

Pan, Y. C. (2009a). Evaluating the appropriateness and consequences of test use. *Colombian Applied Linguistics Journal*, *11*, 93-105.

Pan, Y. C. (2009b). Test impact: English certification exit requirements in Taiwan. *TEFLIN Journal, 20*(2), 119-139.

Peng, J., & Woodrow, L. (2010). Willingness to communicate in English: A Model in the Chinese EFL classroom context. *Language Learning, 60*(4), 834-876.

Rokhsari, S. (2017). *An investigation into the relationship of emotional and spiritual intelligence with job satisfaction and organizational commitment among Iranian EFL faculty and non-faculty university instructors: A mixed-methods approach* (Doctorial thesis). Ferdowsi University, Mashhad.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54-67.

Salehi, M. [Massomeh], & Marefat, F. (2014). The effects of foreign language anxiety and test anxiety on foreign language test performance. *Theory and Practice in Language Studies, 4*(5). 931-940. doi:10.4304/tpls.4.5.931-940

Salehi, M. [Mohammad], & Tarjoman, M. (2017). An investigation of a nationwide exam from a critical language testing perspective. *Cogent Social Sciences, 3*(1). doi:10.1080/23311886.2017.1396639

Shohamy, E. (1982). Affective considerations in language testing. *The Modern Language Journal, 66*(1), 13-17. doi:10.2307/327810

Shohamy, E. (1992). New models of assessment: The connection between testing and learning. In E. Shohamy & R. Walton (Eds.), *Language assessment for feedback: Testing and other strategies* (pp. 1–28). Dubuque, IA: Kendall Hunt Publishing Company.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331–345. doi:10.1016/s0191-491x(98)00020-0

Shohamy, E. (2001a). *The power of tests: A critical perspective on the uses of language tests.* Harlow: Pearson Education.

Shohamy, E. (2001b). Democratic assessment as an alternative. *Language Testing, 18*(4), 373–391. doi:10.1177/026553220101800404

Shohamy, E. (2007a). Language tests as language policy tools. *Assessment in Education, 14*(1), 117–130. doi: 10.1080/09695940701272948

Shohamy, E. (2007b). Tests as power tools: Looking back, looking forward. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner & C. H. Doe (Eds.), *Language testing reconsidered* (pp.141–153). Ottawa: University of Ottawa Press.

Simpson, C. (2016). *Effects of standardized testing on students' well-being*. Harvard: Harvard Graduate School of Education.

Thibodeaux, A. K. (2014). *The effects of leadership and high-stakes testing on the retention of teachers* (Doctorial thesis). University of Southern Mississippi, Hattiesburg.

Ushioda, E. (2011). Motivating learners to speak as themselves. In G. Murray, X. Gao & T. Lamb (Eds.), *Identity, motivation and autonomy in language learning. Second language acquisition* (pp.141–153). Bristol: Multilingual Matters.

Ushioda, E., & Dörnyei, Z. (2009). Motivation, language identities and the L2 self: A theoretical overview. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2. Second language acquisition* (pp. 1-8). Bristol: Multilingual Matters.

Wheelock, A., Bebell, D. J., & Haney, W. (2000). *What can student drawings tell us about high-stakes testing in Massachusetts? Teachers College Record*. Retrieved from https://www.researchgate.net/publication/331346728_What_Can_Student_Drawings_Tell_Us_About_High-Stakes_Testing_in_Massachusetts

Xerri, D., & Vella Briffa, P. (Eds). (2018). *Teacher involvement in high-stakes language testing.* Springer International Publishing: Cham, Switzerland.

Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility belief. *Contemporary Educational Psychology, 30* (4), 397–417. doi: 10.1016/j.cedpsych.2005.05.003

Appendix I: Educational Consequences Questionnaire (English Version)

Dear Respondent,

We would like to ask you to help us by answering the following questions concerning EPT as a high-stakes test. This survey is conducted with the aim of looking into PhD non-English students' perceptions of the educational consequences of EPT. This is not a test so there is no "right" or "wrong" answers and you don't even have to write your name on it. We are interested in your personal opinion. The information will be kept confidential and will be used just for research purposes. Please give your answers sincerely as only this will guarantee the success of the investigation. Thank you very much for your help and cooperation.

1. Strongly agree 2. Agree 3. Undecided 4. Disagree 5. Strongly disagree

Constructs	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
Learning Environment					
1. Instructors play a supportive role for students participating in test preparation classes / courses.					
2. Instructors provide appropriate answers to students' concerns about this test in preparation classes / courses.					
3. Assignments designed for EPT preparation classes or courses are clear					
4. Assignments designed for EPT preparation classes or courses are useful.					
Surface learning	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
5. This test makes students memorize test-related contents.					
6. This test increases superficial learning among students.					
Deep Learning	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
7. This test makes students learn more useful ways to learn English.					
8. This test makes students analyze English.					
9. This test makes students produce English.					
10. This test helps students to better understand English.					
11. The questions of this test are related to the actual use of the language (writing, speaking, reading, etc.) English.					
12. This test makes students use English in a real environment in the future.					
13. This test improves students' speaking skills.					
14. This test improves students' reading skills.					

ISSN 2476-5880

1**5.** This test improves students' listening skills.					
16. This test improves students' writing skills.					
Learning Outcomes	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
17. This test improves students' vocabulary.					
18. This test improves students' grammar.					
19. This test improves students' translation skills.					
Testing Preferences	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
20. I prefer all language skills, including speaking and writing, to be assessed in this test.					
21. I prefer questions that require more or deeper thinking.					
22. I prefer questions that require creativity to be answered.					
23. I prefer questions that need further analysis.					

Appendix II: Social Consequences Questionnaire (English Version)

Dear Respondent,

We would like to ask you to help us by answering the following questions concerning EPT as a high-stakes test. This survey is conducted with the aim of looking into PhD non-English students' perceptions of the social consequences of EPT. This is not a test so there is no "right" or "wrong" answers and you don't even have to write your name on it. We are interested in your personal opinion. The information will be kept confidential and will be used just for research purposes. Please give your answers sincerely as only this will guarantee the success of the investigation. Thank you very much for your help and cooperation.

1. Strongly agree 2. Agree 3. Undecided 4. Disagree 5. Strongly disagree

Constructs	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
Job Promotion					
1. The result of this test helps me to find my job easier / faster.					
2. The result of this test will help me get a job promotion.					
3. The result of this test improves my salary or income status.					
Financial Expenses	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
4. This test makes me pay for the purchase of books / educational materials related to the test.					
5. This test makes me pay for a private class.					
6. This test makes me pay for traveling to another city.					
Social Justice	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
7. This test is the same for all participants, regardless of social status.					
8. This test is the same for all participants regardless of ethnicity / race.					
9. This test is the same for all participants, regardless of religion.					
10. This test provides equal opportunities for all participants, including the disabled ones.					
11. This test is the same for all participants, regardless of their economic status.					
12. This test is the same for all participants, regardless of their majors.					

Appendix III: Psychological Consequences Questionnaire (English Version)

Dear Respondent,

We would like to ask you to help us by answering the following questions concerning EPT as a high-stakes test. This survey is conducted with the aim of looking into PhD non-English students' perceptions of the psychological consequences of EPT. This is not a test so there is no "right" or "wrong" answers and you don't even have to write your name on it. We are interested in your personal opinion. The information will be kept confidential and will be used just for research purposes. Please give your answers sincerely as only this will guarantee the success of the investigation. Thank you very much for your help and cooperation.

1. Strongly agree 2. Agree 3. Undecided 4. Disagree 5. Strongly disagree

Constructs	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
Self-Efficacy Beliefs					
1. I believe I have the ability to pass this test.					
2. I'm sure I can improve my English with more effort.					
3. I think learning English is very easy.					
Feelings	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
4. This test increases my stress and anxiety.					
5. The result of this test makes me feel hopeless.					
6. The result of this test makes me nervous / aggressive.					
7. This test has caused stress and tension in my family.					
8. The difficulty of preparing for this test will dampen my motivation to study English in the future.					
9. This test makes students think about dropping out of university.					
Intrinsic Motivation	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
10. Preparing for the test increases my interest in learning English.					
11. Preparing for this test makes the learning process enjoyable for me.					
12. Preparing for this test makes me aware of the value of learning English.					
13. Preparing for this test gives me inner satisfaction.					
Extrinsic motivation	**Strongly agree**	**Agree**	**Undecided**	**Disagree**	**Strongly disagree**
14. I want to pass this exam to get a job promotion.					
15. Making a distinction among the participants based on the result of this test will increase my motivation to study.					

Thanks for your cooperation