THE JOURNAL OF ASIA TEFL
Vol. 19, No. 3, Fall, 2022, 835-851
http://dx.doi.org/10.18823/asiatefl.2022.19.3.6.835

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511 © 2004 AsiaTEFL.org. All rights reserved.

Exploring the Role of Emotioncy in Reading Comprehension Test Bias

Elahe Moradi
Department of English, Ferdowsi University of Mashhad, Iran

Zargham Ghabanchi
Department of English, Ferdowsi University of Mashhad, Iran

Reza Pishghadam
Department of English, Ferdowsi University of Mashhad, Iran

Given the significance of psychological factors in forming biases, this study intends to introduce emotioncy, emotions evoked by senses, as a potential source of reading comprehension test bias. To this end, 514 English as a Foreign Language (EFL) learners were asked to take a 30-item multiple-choice reading comprehension test along with the emotioncy scale. Based on the emotioncy scores, participants were classified into two groups of Low and High-emotioncy. Subsequently, Rasch model-based Differential Item and Test Functioning (DIF/DTF) analyses were employed across the two target groups. The results showed that the reading comprehension test functioned differentially both at the level of the individual items and the whole test as a set of items, favoring the examinees with higher levels of emotioncy. Thus, the study provides evidence for emotioncy as a potential psychological source of reading comprehension test bias and discusses implications for educators and test developers.

Keywords: Differential Item Functioning, Differential Test Functioning, emotioncy, reading comprehension test, test bias

Introduction

Decisions made on the basis of test results cannot be appropriate, if the test results are not the true indicator of the abilities and knowledge of the test takers. Results sometimes depend on the test itself as well as on the population being tested. This dependability reduces the possibility of objective measurement. For a test to act as a fair measuring device, it must be valid for its recipients. It must be constructed in such a way as to minimize the extraneous factors and maximize the effects of the abilities being measured. Characteristics of the test takers can be considered among the factors which can impact the objectivity of the test and its results. The test or some of its items may perform differently for different test takers, known as Differential Item Functioning (DIF). Since items that indicate DIF can act as a threat to the validity of a test, DIF analysis has taken into account as a necessary step in the validation of tests. The results of any assessment can be jeopardized by test bias (Shohamy, 1997). Thus, potential sources of bias in various fields, including language education, must be discovered.

Regarding the concept of test bias, previous studies have generally focused on specific characteristics of the examinees such as gender or race (McNamara & Roever, 2006). However, any characteristic that test takers have in different levels can also lead to bias and systematically influences test performance (McNamara & Roever, 2006). However, there are few studies on these factors, including psychological characteristics. Consequently, more research is required to fill this gap.

For this purpose, the current study aims to scrutinize the potential of one of these characteristics, emotioncy, to bias EFL learners test performance. The term emotioncy, a combination of emotion and frequency, is a new phenomenon first proposed by Pishghadam and Adamson et al. (2013). A clear definition of emotioncy was given by Pishghadam, Jajarmi, and Shayesteh (2016) as the emotions evoked by the senses from which he receives input. Each individual has a degree of emotioncy towards different language entities (Pishghadam et al., 2013). For example, there are words that carry emotioncy for some people due to the fact that they have heard it, seen it, touched it, or experienced it in some way. Such words are obtained and retained faster and easier compared to those with less or no emotioncy (Pishghadam & Shayesteh, 2016).

Research on the concept of emotioncy has shown some evidence that it can perform a role in facilitating language learning for students (e.g., Pishghadam & Shayesteh, 2016). Given this evidence and regarding the fact that emotions vary for different individuals, which means that different students have different sensory experiences, it can be assumed that test takers' level of emotioncy can influence their test scores and may also lead to test bias. To be precise, test takers of the same level of subject knowledge will act differently on the tests due to different emotioncy levels.

As the relevant literature demonstrates, there is little research on the relationship between test bias and emotioncy. Pishghadam, Baghaei et al. (2016) took an implicit view to describe emotioncy as a potential source of test bias. Their results showed that emotioncy could predict an individual's better test achievement. However, their study only focused on word forms and did not address emotioncy for the meaning of words. Along the same lines, Karami, Pishghadam, et al. (2019) examined the role of emotioncy as a probable source of bias through examining test takers' emotioncy for both word form and meaning. They concluded that learners who had higher levels of emotioncy for the form and meaning of words outperformed those with lower levels of emotioncy.

Given the fact that vocabulary knowledge is essential to the development of reading and vocabulary knowledge and understanding of reading materials are strongly related to each other, the current study aims to extend previous research studies and investigate how emotioncy toward words can assist understanding of reading texts and, in turn, can serve as a resource of possible test bias. Since no previous research study has investigated this hypothesis, there appears to be a need to investigate this issue. Hence, the results of this study can enrich the theoretical underpinnings of the testing domain in general and the bias of language testing in particular. In other words, the current research study may contribute valuable insights to the development of valid and unbiased tests by revealing a potential psychological source of bias. For this purpose, using IRT-based differential item and test functioning (DIF/DTF), the present study aims to probe into the role of emotioncy (sense-induced emotions) in EFL learners' reading comprehension test performance in an attempt to close the existing gap.

Literature Review

Differential Item/Test Functioning

Contemporary validity theory has expanded procedures to support the rationality of decisions based on tests, thus addressing issues of test fairness (McNamara & Roever, 2006). After Messick, Bachman introduced validity as a unitary concept; requiring evidence to support the conclusions we reach on the basis of test scores (McNamara & Roever, 2006). Correct and accurate testing is an important issue in social research. Many important decisions are made based on the results of tests taken under different conditions in the fields of educational and psychological testing. Inaccurate conclusions are often made if the property of measurement stability is not evaluated across these conditions (Makransky & Glas, 2013).

Recently, a major concern of test developers in the field of second language assessment has been the question of test fairness. The issue of fairness has become increasingly controversial in recent years, and assessment practitioners have developed processes to reduce unfairness (Nisbet & Shaw, 2022). Some testing specialists view fairness as being treated within concepts such as bias, justice, and equality (Moghadam & Nasirzadeh, 2020). According to Wallace (2018), fairness is related to the evidential basis for interpretation and use of test scores, with construct validity, procedural equality, and psychometric properties being major sources of evidence for validity.

As Camille and Sheppard (1994) have argued, all test-takers who have the same level of knowledge should have the same chance of success in endorsing test items; consequently, the first step in discovering item bias is to examine scores for trace of DIF, which happens when two groups of test-takers who have the same level of knowledge, have different performance on the test (Thissen et al., 1993). Test fairness may be at risk by bias on the item, group of items, or level of the test. According to Zumbo (2003), item bias occurs when one group of examinees is less likely to answer one item correctly due to the test or characteristics of its items that are irrelevant to the purpose of the test.

As an initial step for determining item bias, differential item functioning (DIF) has been well studied (Yu et al., 2006). Brinbaum (1968) emphasized that the Item Response Theory DIF detection method is a robust technique for investigating item and test bias. According to Stark and Dragosow (2004), at the item level, bias refers to the differences in the probability of correctly endorsing an item between individuals who have the same level of ability, but belong to different subgroups. Over the entire test level as a set of items, bias refers to differences in the total expected scores of these individuals. The accumulation of small differences in items can become very large at the test level, thus biasing the entire test in favor of one group over another. Differential Test Functioning (DTF) can also occur in test cases where DIF analyses indicate that no single item appears to display a significant amount of DIF (Chalmers et al., 2015). Thus, assessing DTF in isolation and in combination with DIF analysis can be useful for test developers. DIF/DTF studies are performed with two groups called the focal and reference group. The first group can refer to a minority group of test takers, and the latter group concerns those potentially favored by the test (Geramipour & Shahmirzadi, 2019).

Several DIF studies have been done to distinguish group differences in examinees performance and to scrutinize if the test items are consistent across members of different subgroups in the context of DIF (e.g., Bay, 2004; Carlton and Harris, 1992; Chen and Henning, 1985; Elder, 1996; Gaffney, 1991; Lawrence and Corley, 1989; Mahler, 2001; Ryan and Bachman, 1992; Scheunemann and Geritz, 1990; Sehmitt & Dorans, 1990; Thissen, Steinberg, & Wainer, 1988, 1993). Grouping has been done in terms of gender (e.g., Aryadoust et al, 2011; Breland, et al, 2004; Maller, 2001), ethnicity (e.g., Sehmitt & Dorans, 1990), and academic backgrounds (e.g., Pae, 2004), linguistic backgrounds (e.g., Chen & Henning, 1985; Ryan & Bachman, 1992), and disability status (e.g., Maller, 1997).

Although many research studies have probed into different sources of test bias such as age, ethnic, gender, and language background, psychological variables that may lead to assessment bias have not received much attention (McNamara & Roever, 2006; Pishghadam, Baghaei, et al., 2016).

Emotion as a psychological factor can lead to DIF in test performance. Research results have shown that negative achievement emotions are associated with low academic success and conversely, positive achievement emotions are related to high academic success (e.g., Frenzel et al., 2007). Indeed, research by Pekrun et al (2002) has indicated that the emotions experienced by learners in academia, in general, and while taking a test, in particular, are central to their educational attainment and may be a determinant of their performance. Emotions associated with academic settings are referred to as achievement emotions and are defined as "emotions tied directly to achievement activities or achievement outcomes" (Pekrun, 2006, p. 317). Pishghadam, Baghaei et al. (2016) stated that because test takers experience different emotions, their emotional relationships can be considered as an indicator of achievement and, therefore, as a source of bias. For this reason, the current study aspires to address the potential of a psychological variable, namely emotioncy to bias EFL learners' test performance.

Emotioncy

The concept of emotioncy proposed by Pishghadam and Tabatabaian et al. (2013) was adopted primarily from Greenspan and his developmental model, Individual Differences, Relationship-Based (DIR) for language acquisition (Greenspan & Shanker, 2004). The notion of emotioncy maintains that each individual has a certain level of emotion towards each concept in the language. The emotion felt has varying degrees based on the individual's experience with the entity whether it is heard, seen, smelled, touched, or research has been done related to the concept. Pishghadam and colleagues argue that creating emotional connections with words and concepts promotes deeper and more effective second language learning through emotions. According to Pishghadam, Jajarmi et al. (2016), individuals' understanding of reality is shaped based on the sensory input they receive through different senses. In fact, sense-provoked emotions deal with the combination of sensing (senses), feeling (emotion), and doing (frequency), which are supposed to have an effect on our judgments and decisions.

To have a closer look, based on the emotioncy literature, individuals may be Avolved (null emotioncy), Exvolved (auditory, visual, and kinesthetic emotioncies) or Involved (inner and arch emotioncies) toward a particular concept (Pishghadam, 2015). This classification displays the degree of involvement of individuals in a particular area, which greatly affects the way they perceive reality and understand the world. Table 1 presents the types of emotioncy and their definitions.

TABLE 1
Emotioncy Types

Type	Kind	Experience
Avolved	Null emotioncy	When an individual has not seen, heard about or experienced an object or concept
Exvolved	Auditory emotioncy	When an individual has merely heard about an object or concept
	Visual emotioncy	When an individual has both heard about and seen the object
	Kinesthetic emotioncy	When an individual has heard about, seen, or touched the real object
Involved	Inner emotioncy	When an individual has directly experienced the word/concept
	Arch emotioncy	When an individual has deeply done research to get additional information

Adapted from Pishghadam, Jajarmi et al. (2016)

To extend the notion of emotioncy, Pishghadam (2015) constructed a continuum and assigned scores to each type of emotioncy. In his sequence, 0 was equal to Null (no emotion), 1 to Auditory emotioncy, 2 to Visual emotioncy, 3 to Kinesthetic emotioncy, 4 to Inner emotioncy, and 5 to arch emotioncy.

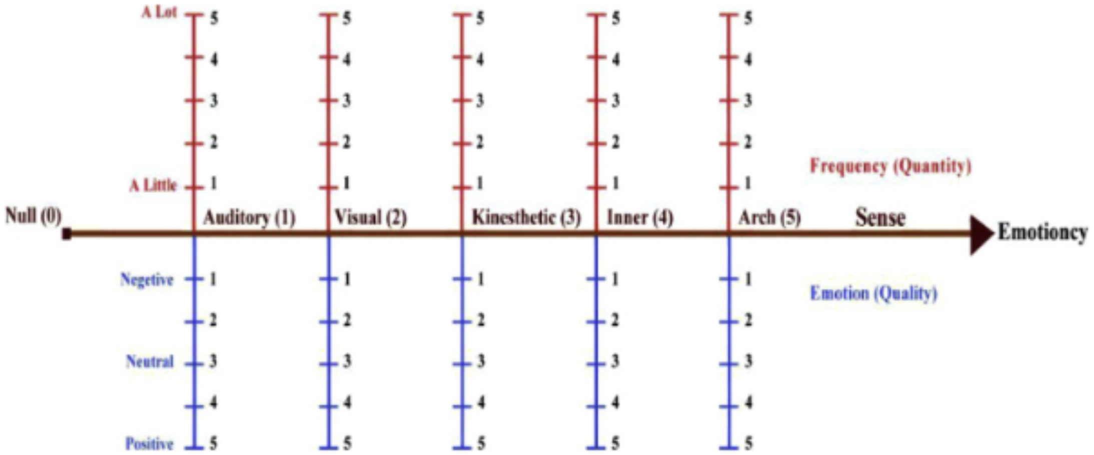


Figure 1. A metric for measuring emotioncy (Adapted from Pishghadam, 2016a).

> Based on the premise that cognition and emotion are interrelated, each degree of the emotioncy model indicates the level and depth of acquiring language entities (Pishghadam, 2015). Individuals can have higher emotioncy for some specific words in a language because they have sensed those entities. They can have heard, seen, smelled, touched, or experienced them; however, as they don't have any kind of experience related to a specific language entity, they may have low or no emotioncy.

Along the same lines, the researchers of this study believe that the idea of reading comprehension can be considered related to the concept of emotioncy. Studies by Borsipour et al. (2019) and Shahiyan et al. (2017) limited to examining emotioncy and its relationship to willingness to read and reading topics respectively; However, the current study investigates emotioncy and its relationship to reading comprehension in order to accurately test students' reading skill in the light of emotioncy towards the basic concepts of texts. In addition, several studies (Bower, 1992; Schutz & Lanehart, 2002; Schutz & Pekrun, 2007) have been conducted to show the importance of emotion in education; however, none of them studied the concept of emotioncy(sense-induced emotions) and its relationship to reading comprehension performance to examine the role of emotioncy as a probable source of test bias. Thus, the current study theorized that sensory emotions towards key concepts of texts may perform a role in assessing reading comprehension. Thus, it aims to find answers to the following research questions:

1. Does emotioncy toward key concepts in reading comprehension texts bias EFL learners' reading comprehension test scores at the level of item?
2. Does emotioncy toward key concepts in reading comprehension texts bias EFL learners' reading comprehension test scores at the level of the whole test as a set of items?

Method

Participants

This study was conducted among 100 males and 414 intermediate and upper intermediate EFL learners at three universities in Mashhad (Iran). The participants aged between 18 and 45 years (Mean = 25.63, SD = 6.43). They spoke Persian as their first language. Convenience or opportunity sampling technique was employed to select the participants for this study.

Instruments

First, the participants' language proficiency level was distinguished by administering the Oxford Quick Placement Test and those who were at the intermediate and upper-intermediate levels of proficiency were chosen. Next, two instruments were employed to collect the required data: the first instrument was a 20-item emotioncy scale. Based on the emotioncy scale proposed by Pishghadam (2016a), each item includes three subcategories. The first subcategory, measuring the sense aspect of the emotioncy, includes six points (null, auditory, visual, kinesthetic, inner, or arch). The second and third subcategories are a five-point Likert- type scale for exploring the levels of emotions towards each concept and the frequency of exposures (See Appendix A).

Emotioncy = sense (frequency + emotion)

According to this formula, the emotioncy scores can be variable from 0 to 50 (The score of zero = Avolved, 1 to 30 = Exvolved, and 30 to 50 = Involved).

For example, a student who expressed his/her feelings about the word “fever” as follows: I have experienced it (feeling score: 4) several times (frequency score: 5) and feel very negative about the fever (emotion score: 1). His total emotional score would be 4 (5 + 1) = 24, which indicates that the student is completely Exvolved in this concept.

To validate the scale, a Rasch rating scale model was utilized. Rasch analysis was carried out employing WINSTEPS 3.74.0 software (Linacre, 2012). The results indicated that the scale was one-dimensional and all the items were within the acceptable limit. Moreover, to ensure the reliability of the scale, Cronbach's alpha (α = 0.92) was estimated, and the scale showed high reliability.

Second, to test the examinees' reading comprehension, three texts, each containing 10 questions, were selected from the reading component of the Longman Complete TOEFL course (Philips, 2005). The rationale behind the selection of these texts was to ensure that the desired range of emotioncy was achieved among the examinees. The first reading comprehension text was about "Thunderstorm", the second text was about "Aspirin and its origin", and included questions 11-20. The third text with questions 20-30 was related to a rare fish called “Coelacanth”.

Procedure

First, to measure participants’ sensory emotions towards key concepts of the reading comprehension texts, the emotioncy scale was answered by the participants. Next, the multiple-choice reading comprehension test was administered to estimate the participants’ reading test scores. Finally, based on the participants’ emotioncy levels, they were divided into two groups of Low and High –emotioncy. To analyze the collected data, DIF and DTF analyses were conducted through Winsteps 3.74.0 Software and MIRT package in R Statistical Software, respectively.

Results

Descriptive statistics, including mean scores and the standard deviations for each item of the emotioncy scale were computed to organize and summarize the characteristics of the data set, and the results are presented in Table2.

TABLE 2
Descriptive Statistics for the Items of the Emotioncy Scale

Item	Avolved	Exvolved	Involved	Emotioncy mean Score	SD
1.Coelacanth	68.5%	29%	2.57%	3.54	7.81
2. Fossil		60.7%	39.3%	25.99	14.15
3. Paleontologist	40.3%	54.7%	5.1%	7.86	10.75
4. Extinct	27.2%	60.1%	12.6%	13.53	13.08
5. Prehistoric	50.4%	43.6%	6%	8.24	11.51
6. Carnivore	43.8%	48.6%	7.6%	8.78	12.01
7.Living specimen	28%	63.2%	8.8%	10.39	11.96
8. Aspirin	1.6%	49.4%	49%	29.78	14.42
9. Fever	3.1%	62.1%	34.8%	25.86	13.76
10. Pain	1.6%	62.3%	36.2%	28.59	12.63
11. Chemical	2.7%	57.8%	39.5%	25.97	14.08
12. Medicinal value	9.3%	66.7%	23.9%	19.82	14.55
13. Relieving ache	36.4%	50.6%	13%	11.69	14.13
14. Lightning	4.5%	61.9%	33.7%	24.34	14.88
15. Air temperature	1.9%	43.6%	54.5%	30.29	14.05
16. Altitude	27%	57.4%	15.6%	14.21	14.90
17. Thunderstorm	12.3%	65.8%	22%	18.75	14.43
18. Collision	31.7%	60.1%	8.2%	10.04	12.09
19. Cumulus cloud	14%	61.9%	24.1%	18.93	15.05
20. Tornado	8.8%	70.4%	20.8%	18.69	13.87

The emotioncy score of each participant for each item of the scale was estimated by the emotioncy formula and they were classified based on their emotioncy scores into three groups of Avolved, Exvolved, and Involved. Furthermore, the emotioncy mean score and the standard deviation for each item were calculated. Items one to seven of the emotioncy scale are the key concepts of the third text of the reading comprehension test. As the descriptive statistics in Table 2 suggests, on the average, the lowest emotioncy scores are related to the third text. A large number of examinees were Avolved toward the related concepts of this text. The emotioncy mean score for the key concepts of this text was 11.19. Items eight to thirteen in this scale are related to the second reading comprehension text. As the table displays, the highest levels of emotioncy are related to the key concepts of this text. The number of participants who were Involved towards the items of this text was higher compared to the other texts. The emotioncy mean score for the key concepts of the second text was 23.62. Items fourteen to twenty encompass the key concepts of the first reading comprehension text. Most of the participants were Exvolved towards the key concepts of this text. The emotioncy mean score for the items of this text was reported 19.32.

As the emotioncy mean sore for each individual item indicates, the mean for item 15 (Air temperature) is the highest of all, which means that the EFL learners had the highest level of emotioncy for this word. 54.5 percent of participants were Involved towards this concept. Also, it is shown that the mean for item 1 (Coelacanth) is the lowest of all, meaning that the EFL learners had the lowest level of emotioncy for this word. 68.5 percent of individuals were Avolved towards this concept meaning that they had no information about it.

TABLE 3

Item Measures and Fit Statistics for the Items of Reading Comprehension Test

Item	Measure	Error	Infit MNSQ	Outfit MNSQ
24	1.29	0.12	1.02	1.16
20	1.06	0.11	1.22	1.36
21	0.79	0.11	1.32	1.46
29	0.70	0.11	0.94	0.95
17	0.62	0.10	1.20	1.23
5	0.43	0.10	1.08	1.15
7	0.41	0.12	0.98	1.03
30	0.41	0.10	0.97	0.97
6	0.38	0.10	0.97	0.98
16	0.36	0.10	0.87	0.83
23	0.31	0.10	0.87	0.84
22	0.29	0.10	0.86	0.82
10	0.24	0.10	0.94	0.92
12	0.21	0.10	0.95	0.92
15	0.18	0.10	1.15	1.15
11	0.15	0.10	1.12	1.17
8	-0.1	0.10	0.94	0.92
14	-0.4	0.10	1.06	1.05
26	-0.15	0.9	0.89	0.86
19	-0.27	0.9	1.06	1.08
28	-0.31	0.9	0.96	0.94
25	-0.35	0.9	0.88	0.86
9	-0.39	0.9	1.07	1.11
4	-0.50	0.9	1.02	1.01
2	-0.52	0.9	1.03	1.03
3	-0.91	0.10	1.04	1.01
27	-0.94	0.10	0.89	0.86
18	-0.96	0.10	0.92	0.87
1	-1.12	0.10	0.96	0.90
13	-1.36	0.10	0.85	0.77

The Infit and Outfit MNSQ (mean-square) statistics are shown in Table 3. According to Linacre (2012), the expected value for MNSQ is 1; however, the range of 0.7 to 1.3 or 1.4 is suggested (Linacre, 1999; Bond & Fox, 2015). As the table indicates, the infit and outfit indices for all the items are within the acceptable range: hence, it can be concluded that all the items fit the Rasch model and the test is unidimensional. In this table, the items are presented in a descending order of difficulty, indicating that item 24 with the difficulty measure of 1.29 logits and the standard error of 0.12 was the most difficult one. This item was related to the third text of the reading comprehension test towards which a large number of participants showed lower levels of emotioncy and they gained the lowest emotioncy scores for this text. Item 13 with the difficulty measure of -1.36 logits and the standard error of 0.10 was the easiest one. This item belonged to the second reading comprehension text towards which most of the participants showed higher levels of emotioncy and on the average, the emotioncy scores for this text were reported the highest of all.

According to Lord (1980), differential item functioning (DIF) is referred to as lack of invariance of item parameters across different subsamples. DIF is considered as an evidence of item bias. The difficulty of each item of the scale for each group of respondent is shown by DIF measures. Thus, in order to examine whether reading comprehension test items functioned differentially for the examinees, the sample was categorized into two groups. Test takers whose emotioncy scores ranged from 0 to 30 in each item (Avolved and Exvolved) were referred to as the Low-Group and those who scored from 30 to 50 (Involved) were assigned the High-Group label (presented in Table 4). Next, for each item, Rasch model-based DIF analysis was conducted. The findings are shown in Table 5.

TABLE 4
Number of the Participants and the Emotioncy Mean Scores in the Two Groups

Item	Number of Participants (Low Group)	Emotioncy Mean (Low Group)	Number of Participants (High Group)	Emotioncy Mean (High Group)
1. Coelacanth	499	2.52	15	37.60
2. Fossil	296	15.66	218	40.02
3.Palentologist	482	5.82	32	38.59
4.Extinct	429	9	85	36.35
5.Prehistoric	476	5.88	38	37.74
6.Carnivore	467	5.87	47	37.74
7.Living Specimen	466	7.43	48	39.13
8.Aspirin	240	16.68	274	41.27
9.Fever	282	15.55	232	38.40
10.Pain	259	18.54	255	38.80
11.Chemical	291	15.51	223	39.61
12.Medicinal Value	380	12.84	134	39.62
13.Relieving Ache	443	7.06	71	40.59
14.Lightning	323	14.75	191	40.54
15.Air Temperature	221	16.41	293	40.77
16.Altitude	426	8.70	88	40.88
17.Thunderstorm	388	12.02	126	39.47
18.Collision	468	7.12	46	39.65
19.Cumulus Cloud	376	11.35	138	39.59
20.Tornado	392	12.39	122	38.96

The item difficulties for each emotioncy group and their standard errors are shown in Table 5. The higher the DIF index, the more difficult the item is. The next column shows the effect size (in logits) which is the variation of the difficulty between the two groups. The DIF must be large enough to be noticeable. A difference of 0.05 logits is usually recommended (Linacre, 2012). Also, the table shows the probability of Welch t and Welch. For a DIF that is statistically significant on an item, a probability less than 0.05 is usually required.

As the table demonstrates, items (1, 3, 11, 12, 13, 15, 16, 20, 29, and 30) showed DIF. Although the p-values for 5 of the 10 difficulty contrasts were less than 0.05 (items 3, 12, 15, 20 and 29), in 5 other items (1, 11, 13, 16, 30) p-value is greater than 0.05, but the effect size meets the requirement of 0.5 logits, which makes the contrast undeniable. The value of the difficulty contrast is the most important number in DIF analysis (Linacre, 2012). Statistical significance can depend on many factors such as sample size; consequently, a large p-value does not necessarily indicate that the result is ignorable or has no value for decision-making. 6 items (11, 12, 13, 15, 16 and 20) out of these 10 DIF items were related to the second text of the reading comprehension test. Of these (items 1 and 3) were related to the first text, and the other two were related to the third text (items 29 and 30). Content analysis of the ten DIF items showed that items that were related to “making inferences and drawing conclusions,” “locating references within the text,” “identifying sources of inferred information,” and “understanding exceptions” were shown to have DIF, which means that in such items there were more differences between the performance of the two groups of Low and High-emotioncy examinees.

TABLE 5

DIF Statistics for the Reading Comprehension Test

Item	Low Group Difficulty (S.E.)	High Group Difficulty (S. E.)	Difficulty Contrast (S. E.)	Weltch *t*	Weltch Prob
1	-1.06 (.10)	-1.82 (.45)	.77 (.47)	1.65	.10
2	-.44 (.10)	-.40 (.36)	-.05 (.38)	-.12	.90
3	-.91 (.10)	-.14 (.36)	-.77 (.37)	-2.07	.04
4	-.45 (.10)	-.40 (.36)	-.6 (.38)	-.16	.87
5	.45 (.11)	.92 (.38)	-.47 (.39)	-1.20	.23
6	.41 (.10)	.38 (.36)	.04 (.38)	.10	.92
7	.48 (.10)	.64 (.37)	-.16 (.38)	-.41	.68
8	.3 (.10)	-.3 (.36)	.00 (.37)	.00	1.00
9	.34 (.10)	-.14 (.36)	-.20 (.37)	-.55	.58
10	.29 (.10)	-.10 (.36)	.39 (.38)	1.04	.30
11	.24 (.10)	-.27 (.36)	.50 (.37)	1.34	.18
12	.31 (.10)	-.70 (.38)	1.01 (.40)	2.54	.01
13	-1.32 (.11)	-1.83 (.46)	.50 (.47)	1.07	.29
14	.01 (.10)	-.27 (.38)	.28 (.37)	.75	.45
15	.07 (.10)	1.15 (.40)	-1.08 (.41)	-2.64	.01
16	.43 (.11)	-.14 (.36)	.56 (.37)	1.51	.13
17	.64 (.11)	.92 (.38)	-.28 (.39)	-.71	.47
18	-.94 (.10)	-1.10 (.39)	.16 (.41)	.39	.69
19	-.28 (.10)	-.31 (.37)	.03 (.38)	.08	.93
20	.93 (.12)	2.63 (.49)	-1.70 (.50)	-3.40	.001
21	.75 (.11)	1.06 (.39)	-.31 (.40)	-.77	.44
22	.28 (.10)	.38 (.36)	-.09 (.37)	-.25	.80
23	.29 (.10)	.43 (.37)	-.14 (.39)	-.37	.71
24	1.26 (.13)	1.54 (.42)	-.27 (.44)	.62	.53
25	-.40 (.10)	-.66 (.37)	.26 (.38)	.68	.50
26	-.29 (.10)	-.14 (.36)	-.15 (.37)	-.40	.69
27	-1.00 (.10)	-1.45 (.44)	.44 (.45)	.98	.32
28	-.42 (.10)	-.56 (.37)	.14 (.38)	.36	.71
29	.72 (.11)	-.40 (.36)	1.12 (.38)	2.95	.00
30	.31 (.11)	1.06 (.38)	-.75 (.40)	-1.88	.06

Figure 2 represents the DIF plot which was drawn based on the DIF measures of each individual item reported for the two classes of Low and High- emotioncy. As the plot indicates, there is a very strong main diagonal, meaning that the outlying points on each side have DIF. As DIF in one direction on one item may be balanced by DIF in other directions on other items, it is suggested to investigate DIF of the whole test as a set of items to verify if the test functions appropriately for the two target groups.

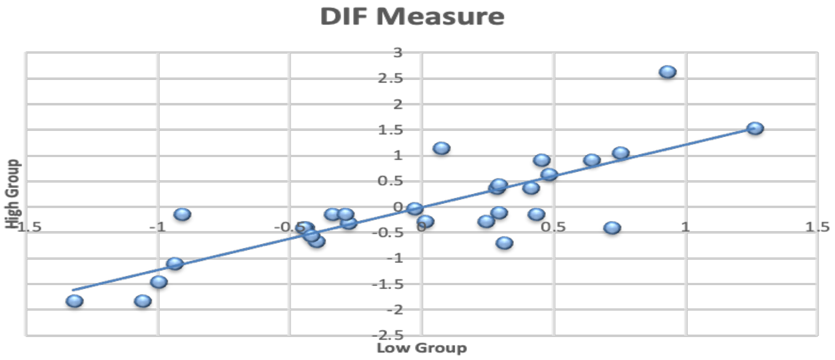
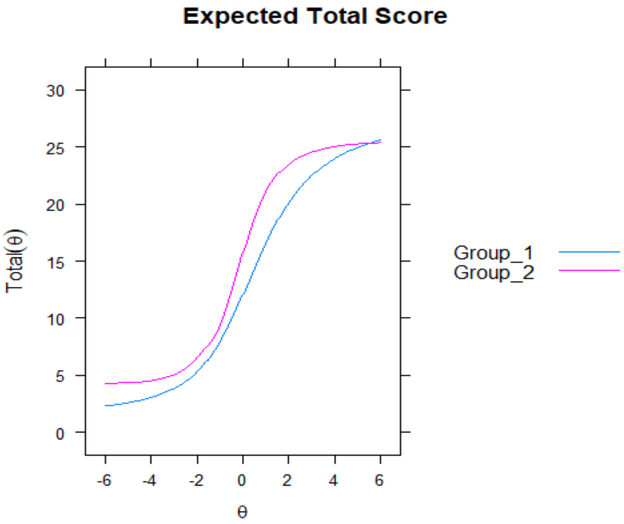


Figure 2. DIF plot of the items for the two groups of low and high-emotioncy.

To examine if the test, consisting of all the items, function the same way for the Low and High-emotioncy groups, Differential Test Functioning (DTF) was investigated by doing separate analyses for the two groups. If DTF exists, then the test characteristic curves will differ. Rasch model Test Characteristic Curve (TCC) shows the relationship between total scores and Rasch measures on all the items of the test. A TCC correlates the expected total score to trait level and is derived by summing the item response functions for the respective groups (Hulin et al., 1983). Comparison of the test characteristic curves for the Low and High-emotioncy groups, shown in Figure 3, revealed difference in the learners' total scores in terms of the emotioncy group they belong.



Group 1= Low emotioncy group
Group 2= High emotioncy group

Figure 3. Test Characteristic Curves (TCC) for low and high-emotioncy groups.

TABLE 6
DTF Report

sDTF.score	sDTF(%).score	u.DTF.score	uDTF(%).score
-1.918	-6.396	1.928	6.428

To investigate overall bias in the reading comprehension test, signed (sDTF) and unsigned (uDTF) were calculated using the method described by Chalmers, Counsell, and Flora (2015). uDTF is a measure of average absolute bias, regardless of which group is advantaged, and sDTF is a measure of average directional bias. The uDTF value covers the mean area between the test curves, indicating the absolute deviations in the item properties collected over the entire test. When the area between the curves is zero, the test works the same for groups and there is no bias. The uDTF for the reading comprehension test has a potential range from 0 (no bias) to 30. As Table 6 suggests, the results represent a total score bias of approximately two raw points (or 6.42%) in favor of the High-emotioncy examinee group.

Discussion

The extensive use of tests and the great diversity of test takers have made designing, administering, and analyzing tests a challenging and complex task. The different race, gender, nationality and background information of individuals as well as many other unrelated external factors can influence test results. As a result, careful attention must be paid to the analysis of test results so that what are measured will be true abilities of the test takers. Fairness is related to validity (Kunnan, 2004). This can be achieved by reducing irrelevant factors and increasing the ability factor effect of the individuals being tested. The correct test should indicate the same results for different groups of examinees after being matched on the underlying ability. Among the various sources of test bias, psychological variables, which may play a critical role in the development of bias for or against test takers, have not received sufficient attention (Garrett & Young, 2009). Accordingly, this study aimed to investigate the idea of emotioncy as a potential source of reading comprehension test bias. To this end, the differential item and test functioning were examined by the Rasch model in order to verify whether emotioncy could be considered as the potential source of reading comprehension test bias at the item and test levels.

With respect to the first question, the results of the DIF analysis showed that there were 10 items which functioned differentially among the two groups of Low and High-emotioncy. 6 out of these 10 DIF items were related to the second reading comprehension text, and the other four DIF items were related to the first and third reading comprehension texts. Thus, it can be concluded that emotioncy can perform as a source of test bias at the item level. But when the overall DIF size of an instrument seems to be small, it may be concluded that DIF in one direction on one item would be balanced out by DIF in other directions on other items. Consequently, it is recommended to examine Differential Test Functioning (DTF) to check whether the test functions the same way for both reference (High-emotioncy) and focal (Low-emotioncy) groups.

For this reason, to answer the second question of the study and to investigate if the whole test as a set of items performs the same for the two groups of Low and High-emotioncy, DTF was examined through test characteristic curves and item response theory (IRT)-based analysis. As the results indicated, the whole test as a set of items functioned differentially for the two groups and EFL learners who had higher levels of emotioncy outperformed those who carried lower levels of this psychological variable. Therefore, it can be concluded that emotioncy can also be taken into consideration as a potential source of bias at the level of the whole test as a set of items.

In addition, as indicated by the item difficulty level, the most difficult items were related to the third text of the reading comprehension test. The results showed that the average degree of EFL learners' emotioncy for this text was the lowest and that a large number of participants were Avolved towards the key concepts of this text. On the other hand, the easiest items were related to the second and first reading text, respectively. As the results showed, most EFL learners were primarily Exvolved and Involved toward the key concepts of these texts. Thus, it can be explained that by increasing the level of emotioncy, the difficulty level of the item might be decreased. Pishghadam, Jajarmi et al. (2016) stated that learners with lower levels of emotioncy experience distal emotions that are far from reality and process input trivially, while those with higher levels of emotioncy enjoy proximal emotions and process input more deeply. Also, based on the notion of emotioncy, learners who exhibit higher levels of emotions for different concepts experience a higher degree of participation in the relevant activity. Therefore, higher levels of emotion for reading comprehension test concepts will lead to more participation and more chances of success. The results of this study also provided evidence to support the relationship between emotioncy and test performance suggested by Pishghadam and Baghaei et al. (2016), who found that individuals with higher levels of emotion toward concepts outperformed those with lower levels of emotion. They emphasized that emotioncy as a source of test bias has a dynamic nature, whereas other sources such as gender and age are static.

The results of this study are also consistent with Karami, Pishghadam et al. (2019), who examined emotioncy as a potential source of vocabulary test bias. Their findings revealed that EFL test takers' emotions for both meaning and vocabulary form is likely to lead to test bias and may alter test taker performance.

The current study presents a number of implications as well. In contrast to static sources of test bias such as gender or race, emotioncy is dynamic in nature which means that individuals can move from one emotional level to another in different contexts. For this reason, it is suggested that test takers enhance their emotional levels towards concepts up to the inner or arch levels. Furthermore, the study will help introduce a new role for test developers as envolvers. This means that test designers may adopt three different approaches while designing tests: through the items, issues or concepts that they present in their tests, they have the power to decide for which items the testees are to be Avolved (i.e., with zero emotioncy), Exvolved (i.e., with auditory, visual, and kinesthetic emotioncies), or Involved (i.e., with inner and arch emotioncies). So, this study introduces the test designers as an envolver who determines what should be included or excluded on a test. Overall, the findings of the current research may be utilized to bring about consciousness-raising of test developers, test takers, teachers and material developers. They should be more aware of the concept of emotioncy caused by sensory involvement and its role in language learning and test performance. Subsequently, appropriate measures can be employed to make the best use of emotioncy in improving language teaching, learning, and testing.

Conclusion

Many studies have examined various sources of bias in language testing. However, the relevant literature on test bias has taken into account just a small number of factors that lead to test bias, such as ethnicity, gender, social class, and language background. For this reason, this study aspires to probe into the concept of a psychological variable, namely, emotioncy as a possible source of bias in the reading comprehension test. To this end, the reading comprehension test was tested for evidence of bias in terms of emotioncy level. The results of the data analysis showed that the students' emotioncy level can act as a potential source of bias at the item level as well as the entire test as a set of items.

In light of the results of this study, it is suggested that test designers choose those texts that can cover a broad range of concepts so that they can predict the emotioncy levels of all test takers. While a certain class of learners may be Avolved toward the concepts of reading comprehension texts, others may have experienced or even researched the same concepts previously. Therefore, the main recommendation for researchers is for test developers. They should be aware of the negative impact of the differential item and test functioning on test validity and test results to design and construct tests that are not beneficial to any group by considering as many factors as possible specifically psychological factors of a dynamic nature.

The results of the current study carry some limitations that could affect the results and limit the generalizability of the conclusions. First of all, the results were contextualized in the context of an Iranian sample of EFL students. It is not possible to choose a true random sample to achieve perfect results. Second, like any other questionnaire-based survey, not all questions may be answered with due care. The current study was also conducted in a single cultural context. However, since sensory emotions can be influenced by culture, it is advisable to examine the concept of emotioncy in other cultural settings so that more global generalizations can be made. Furthermore, this study investigated emotioncy with reference to a reading comprehension test. Future research could examine emotioncy as a potential source of bias in other skills including listening, speaking or writing to check whether emotioncy can bring about DIF or DTF in relation to other language skills.

Acknowledgement

The authors wish to express their sincere gratitude to the editor and anonymous reviwers of this manuscript who assisted us to fine-tune the paper.

The Authors

Elahe Moradi is a PhD candidate at Ferdowsi University of Mashhad, Iran. Her research interests include Language Testing, Psycholinguistics, and Sociolinguistics.

Department of English, Ferdowsi University of Mashhad, Iran.
Email: elahe.moradi@mail.um.ac.ir

Dr. Zargham Ghabanchi holds a PhD in Applied Linguistics from the University of Liverpool, the UK. He has a chair at Ferdowsi University of Mashhad. He has published several articles.

Department of English, Ferdowsi University of Mashhad, Iran.
Email: ghabanchi@um.ac.ir

Professor Reza Pishghadam (corresponding author) is a professor of language education and a courtesy professor of educational psychology at Ferdowsi University of Mashhad, Iran. In 2010, he was classified as the distinguished researcher of humanities in Iran. In 2014, he also received the distinguished professor award from Ferdowsi Academic Foundation, Iran.

Department of English, Ferdowsi University of Mashhad, Iran.
Email: pishghadam@um.ac.ir

References

Aryadoust, V., Goh, C. C., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8*(4), 361-385. http://dx.doi.org/10.1080/15434303.2011.628632

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Addison-Wesley.

Bond, T. G., & Fox, C.M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Lawrence Erlbaum.

Borsipour, B., Pishghadam, R., & Naji Meidani, E. (2019). The role of sensory emotions in increasing willingness to read in EFL learners. *Publicaciones, 49*(2), 169-189. http://dx.doi.org/10.30827/publica ciones.v49i2.8094

Bower G. W. (1992). How might emotions affect learning? In Christianson, S. A. (Ed.). *The handbook of emotion and memory: Research and theory* (pp. 3-31). Erlbaum.

Breland, H., Lee, Y., Najarian, M., & Muraki, E. (2004). An analysis of TOEFL-CBT writing prompt difficulty and comparability for different gender groups. *ETS Research Report Series, 1*, 1-54. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01932.x

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and majority /minority group comparisons* (ETS Research

Report, 92-64). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1992.tb01495.x

Chalmers, R. P. Counsell, A. & Flora, D. B. (2015). It might not make a big DIF: Improved. Differential Test Functioning statistics that account for sampling variability. *Educational and Psychological Measurement, 76*(1), 114-140. http://dx.doi.org/10.1177/0013164415584576

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-163.

Elder, C. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and Modern Greek. *Language Learning, 46*(2), 233-282. http://dx.doi.org/10.1111/ j.1467-1770.1996.tb01236.x

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics: A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education, 22*(4), 497-514. http://dx.doi.org/10.1007/BF03173468

Gafni, N. (1991). *Differential Item Functioning: Performance by sex on reading comprehension tests.* ERIC. https://files.eric.ed.gov/fulltext/ED331844.pdf

Garrett, P., & Young, R. F. (2009). Theorizing affect in foreign language learning: An analysis of one learner's responses to a communicative Portuguese course. *The Modern Language Learner, 93*(2), 209-226. http://dx.doi.org/10.1111/j.1540-4781.2009.00857.x

Geramipour, M., & Shahmirzad, N. (2019). A gender-related differential item functioning study of an English test. *The Journal of Asia TEFL, 16*(2), 674-682. https://dx.doi.org/10.18823/asiatefl.2019.16.2.15.674

Greenspan, S. I., & Shanker, S. G. (2004). *The first idea: How symbols, language, and intelligence evolved from our primate ancestors to modern humans.* Da Capo Press.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Irwin.

Karami, M., Pishghadam, R., & Baghaei, P. (2019). A probe into EFL learners' emotioncy as a source of test bias: Insights from differential item functioning analysis. *Studies in Educational Evaluation, 60*, 170-178. https://doi.org/10.1016/j.stueduc.2019.01.003

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Wei (Eds.), *European language testing in a global context: proceedings of the ALTE Barcelona Conference* (pp. 27-48). Cambridge University Press.

Lawrence, I. M., & Curley, W.E. (1989). *Differential Item Functioning for males and females on SAT-Verbal Reading sub score items: follow-up study* (ETS Research Report 89-22). Educational Testing Service.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3,* 103-122.

Linacre, J. M. (2012). *A user guide to WINSTEPS MINISTEPS Rasch-model computer programs.* Winsteps.com. https://www.winsteps.com/winman/copyright.htm

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates. http://dx.doi.org/10.4324/9780203056615

Maclntyre, P. D., Baker, S., Clement, R. & Donovary, L. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning, 52,* 537-564. http://dx.doi.org/10.1111/1467-9922.00226

Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement, 46*(9), 3228-3237. https://doi. org/10.1016/j.measurement.2013.06.020

Maller, S. J. (1997). Deafness and WISC-III item difficulty: Invariance and fit. *Journal of School Psychology, 35*(3), 299-314. http://dx.doi.org/10.1016/S0022-4405 (97)00010-1

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817. http://dx.doi.org/10.1177/00131640121971527

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell Publishing.

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia, 10.* https://doi.org/10.1186/s40468-020-00105-2

Nisbet, I., & Shaw, S. (2022). Fair high-stakes assessment in the long shadow of Covid-19. *Assessment in Education: Principles, Policy & Practice.* http//dx.doi.org/10.1080/0969594X.2022.2067834

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21*(1), 53-73. http://dx.doi.org/10.1191/0265532204lt274oa

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315-334. http://dx.doi.org/10.1007/s10648-006-9029-9

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative- search. *Educational Psychologist, 37*(2), 91-105. http://dx.doi.org/10.4324/9781410608628-4

Pishghadam, R. (2015 October 6). *Emotioncy in language education: From exvolvement to involvement* [Conference paper]. 2nd conference of interdisciplinary approaches to anguage teaching, literature, and translation studies. Mashhad, Iran.

Pishghadam, R. (2016a May). Emotioncy, extraversion, and anxiety in willingness to communicate in English [Conference paper]. 5th International conference on language, education, and innovation.

Pishghadam, R. (2016b September). Introducing emotioncy tension as a potential source of identity crises [Conference paper]. Interdisciplinary conference on cultural identity and philosophy of self.

Pishghadam, R., Adamson, B., & Shayesteh, S. (2013). Emotion-based language instruction (EBLI) as a new perspective in bilingual education. *Multilingual Education, 3*(9), 1-16. http://dx.doi.org/10.1186/2191-5059-3-9

Pishghadam, R., Baghaei, P., & Seyednozadi, Z. (2016). Introducing emotioncy as a potential source of test bias: A mixed Rasch modeling study. *International Journal of Testing, 17*(2), 127-140. http://dx.doi.org/10.1080/15305058.2016.1183208

Pishghadam, R., Jajarmi, H., & Shayesteh, S. (2016). Conceptualizing sensory relativism in light of emotioncy: A movement beyond linguistic relativism. *International Journal of Society, Culture & Language, 4*(2), 11-21.

Pishghadam, R., & Shayesteh, S. (2016). Emotioncy: A post-linguistic approach toward vocabulary learning and retention. *Sri Lanka Journal of Social Sciences, 39*(1), 27-36. http://dx.doi.org/10.4038/sljss. v39i1.7400

Philips, D. (2005). *Longman complete course for the TOEFL test: Preparation for the computer and paper tests.* Longman.

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12-29. http://dx.doi.org/10.1177/026553229200900103

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00737.x

Schutz, P. A., & Lanehart, S. J. (2002). Introduction: Emotions in education. *Educational Psychologist, 37*(2), 67-68. http://dx.doi.org/10.4324/9781410608628-1

Schutz, P. A., & Pekrun, R. (2007). *Emotion in education*. Elsevier.

Sehmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*(1), 67-81. http://dx.doi.org/10.1111/j.1745-3984.1990.tb00735.x

Shahian, L., Pishghadam, R. & Khajavi, G. H. (2017). Flow and reading comprehension: Testing the mediating role of emotioncy. *Issues in Educational Research, 27*(3)*,* 527-549.

Shayesteh, S., Pishghadam, R., Khodaverdi, A. (2020). FN400 and LPC responses to different degree of sensory involvement: A study of sentence comprehension. *Advances in Cognitive psychology, 16*(1), 45-58.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14*(3), 340-349. http://dx.doi.org/10.1177/026553229701400310

Stark, S., & Drasgow, F. (2004). Examining the effects of Differential Item (Functioning and Differential) Test Functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*(3), 497-508. https://doi.org/10.1037/0021-9010.89.3.497

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-172). Lawrence Erlbaum Associates, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item function in using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Lawrence Erlbaum Associates.

Yu, L., Lei, P.W., & Suen, H. K. (2006*).* Using a Differential Item Functioning (DIF) procedure to detect differences in opportunity to learn (OTL) [Conference paper]. The annual meeting of the American Educational Research Association*,* San Francisco, California.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests*. Language Testing, 20*(2)*,* 136-147. http://dx.doi.org/10.1191/0265532203lt248oa

Wallace, M. P. (2018). Fairness and justice in L2 classroom assessment: Perceptions from test takers. *The Journal of Asia TEFL, 15*(4), 1051-1064. http//dx.doi.org/10.18823/asiatefl.2018.15.4.11.1051

(Received January 27, 2022; Revised August 20, 2022; Accepted Sep 18, 2022)

Appendix A

Sample Item of the Emotioncy Scale

1.Coelacanth	**My feeling about this word**	**My frequency of exposure to this word**
I don't know what it is□		
I have heard about it □	Extremely negative□	Very rarely□
I have heard about and seen it □	Negative□	Rarely□
I have heard about, seen and been in touch with someone who has used this word□	Neutral□	Occasionally□
Including the previous ones, I have used this word myself □	Positive□	Frequently□
Including the previous ones, I have conducted research on this word □	Extremely positive□	Very frequently□