KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: November 26, 2016
Accepted: February 5, 2018
OnlineFirst: March 5, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.1.0606 ♦ February 2018 ♦ 18(1) ♦ 47–60

Research Article

Detection of Gender related DIF in the Foreign Language Classroom Anxiety Scale

Jongmin Ra[1]
Kookmin University

Ki Jong Rhee[2]
Kookmin University

Abstract

A fundamental challenge to understanding effects of foreign language anxiety on the foreign language learning lies in implementing reliable and valid measures. Considering importance of measurement bias and widespread usage of the foreign language classroom anxiety scale (FLCAS) in education, the aim of the current study was to detect differential item functioning (DIF) in FLCAS due to male and female students, which threatens the validity of FLCAS. Results showed that gender related DIF appeared in FLCAS. Out of 25 items of FLCAS, two items were found to exhibit gender related DIF after correcting inflated Type I error. Such results implied that what previous studies called gender differences in the mean levels of foreign language classroom anxiety might have just been response differences to FLCAS due to some of gender related DIFs on FLCAS.

Keywords

Differential item functioning • Foreign anxiety • Gender difference • Graded response model • IRTLRDIF

[1] Department of Education, Kookmin University, Seoul, 136-702, South Korea. Email: rems2002@kookmin.ac.kr

[2] **Correspondence to:** Ki-Jong Rhee, Department of Education, Kookmin University, Seoul, 136-702, South Korea. Email: rhee0408@kookmin.ac.kr

Citation: Ra, J., & Rhee, K. J. (2018). Detection of gender related DIF in the Foreign Language Classroom Anxiety Scale. *Educational Sciences: Theory & Practice, 18,* 47–60. http://dx.doi.org/10.12738/estp.2018.1.0606

Increasing attention has been paid in recent years to the importance of language anxiety in foreign language teaching and/or learning (Aida, 1994; Al-Saraj, 2014; Horwitz, 2001; Horwitz, Horwitz, & Cope, 1986; Trang, 2012). Since theoretically and empirically substantial efforts investigating foreign language anxieties on language learning have been involved (Aida, 1994; Horwitz et al., 1986), a classroom–related foreign language anxiety called the Foreign Language Classroom Anxiety Scale (FLCAS) has been developed by Horwitz et al. (1986). FLCAS measuring specific types of language anxiety (communication apprehension, test anxiety, and fear of negative evaluation) is one of the most important measures in that it conceptualizes language anxiety occurred in classroom systematically. A considerable body of previous research endorses the usefulness of FLCAS quantifying the severity of language anxiety and summarizing overall foreign language anxiety in different settings (Aida, 1994; Horwitz et al., 1986; Matsuda & Gobel, 2004; Tóth, 2008). In a review of previous studies pertaining to FLCAS, results related to the effect of gender are still inconsistent (Aida, 1994; Baker & Maclntyre, 2000; Maclntyre, Baker, Clément, & Donovan, 2002). Aida (1994) and Maclntyre et al. (2002) showed male students were less anxious than female students in the classroom setting and vice versa in Awan, Azher, Anwar, & Naz's study (2010). In addition, no statistically significant difference, in general, between male and female students regarding foreign language anxiety existed (Matsuda & Gobel, 2004).

In interpreting these inconclusive findings, other factors such as psychometric properties (e.g., differential item functioning) must be taken into account. It is possible to assume that such items are conceptually and psychometrically equivalent among groups differing in characteristics such as education, ethnicity, and gender. In this connection, it should be noted that previous ample studies investigating gender effect on the foreign language anxiety have been conducted with summary statistical techniques focused at the levels of sub-scales rather than item level (Aida, 1994; Matsuda & Gobel, 2004; Maclntyre et al., 2002).

Despite the fact that items on a test should be considered as crucial evidence for validity and reliability because items are the basic building block of a test (Lissitz & Samuelsen, 2007), there are limited studies (Koh & Ra, 2011; Panayides & Walker, 2013; Ra & Kim, 2013) examining psychometric properties of FLCAS. Among them, two studies (Koh & Ra, 2011; Panayides & Walker, 2013) simply apply IRT-related models to verify psychometric properties of FLCAS: Koh and Ra's (2011) study simply applies Samejima's (1969) Graded Response model (GRM) and Panayides and Walker (2013) uses Rasch rating scale model. Only, Ra and Kim's (2013) study investigates and shows differential item functioning (DIF) in the Korean version of FLCAS by using the multiple indicators, multiple causes (MIMIC) method. Such results indicate that simply using summed scores could result in inaccurate information (e.g., different number of factors, different foreign language anxiety between male and female students). This type of item bias could result in differences in test validity across groups (Millsap & Everson, 1993). Millsap & Everson's (1993) study shows that DIF in a test reflects measurement bias yielding a potential threat to the validity of the test. DIF occurs when individuals of different groups (e.g., male and female students) at similar levels of foreign language respond to foreign language related items differently. These differences can result in a type of bias called DIF.

When DIF exists, the probability of answering a specific item correctly could differ from group to group (males to females) after considering equal latent ability (Millsap & Everson, 1993). In other words, respondents with similar degrees of foreign language anxiety have different probability of responding to an item according to their population membership (e.g., male or female). Within the realm of DIF, two types of DIF, uniform and non-uniform can be detected. Uniform DIF occurs when the probability of response is in the same direction across the cognitive function continuum. In other words, DIF is in the same direction across the entire spectrum of ability (item response curves for two groups do not cross) because DIF involves the location parameter (β). Namely, threshold parameter is only investigated. Meantime, non-uniform DIF is evident when DIF is in different directions at different parts of the cognitive function continuum. An item favors one group at certain ability levels, and other groups at other levels (or the probability of item endorsement is higher for group 1 at lower ability and higher for group 2 at higher ability). The present study could improve the stability of FLCAS and could provide the plausible answers why previous gender related studies have inconsistent results. The aim of this study is to find if DIFs are present in the FLCAS and how DIF influence differences existed between male and female students.

Methods

Participants

581 Korean college students in different departments such as medical school, social work, child-education, and so on enrolled in an introductory TOEIC class four hours a week participated in this study. Of these students, 282 (48.53%) were male students and 299 (51.46%) were female students. Students in the programs generally ranged in age from 19 to 21 years. The average age of the participants is 20.11 years old; 15.7% of them have experience studying abroad. The focus group is the male group and the reference group is the female one.

Instrument

Despite the inconclusive results regarding a number of constructs of FLCAS: some studies (Horwitz et al., 1986; MacIntyre & Gardner, 1991; Tóth, 2008) suggest that FLCAS is constructed of three constructs while others (Aida, 1994) suggest more than three constructs, three constructs in FLCAS was considered in the study as Horwitz et al. (1986) suggested in their study. The FLCAS (Horwitz et al., 1986) has been widely used to measure foreign language learning anxiety within a classroom context: communication apprehension, test anxiety, and fear of negative evaluation. Communication apprehension refers to the uncomfortable feeling an individual experiences when expressing himself/herself in from of others. Inability to communicate correctly or to understand what another person says can easily result in frustration and apprehension given that the apprehensive communicator knows that total communication is not possible and he/she may be troubled by this reality (Williams & Andrade, 2008). Communication apprehension could, thus, be described as learners' shyness resulting from anxiety while using a foreign language to communicate. Test anxiety refers to a type of performance anxiety springing from a fear of failure. In general, language learners' fear of failure or poor performance leads to test anxiety. Fear of negative evaluation is likely to be manifested in a student's excessive worry about academic and personal evaluations of his or her performance and competence in the target language (MacIntyre & Gardner, 1991). Namely, fear of negative evaluation could be referred to as apprehension, avoidance, and expectation of a detrimental evaluation by others. Respondents are required to circle a number on a 5-point Likert scale that best represented their current situation to 33 items in the Korean version of FLCAS pertaining to the foreign language classroom anxiety. An answer of 5 would indicate high level of foreign language classroom anxiety while an answer of 1 indicates the other opposite end. It has been designated that the reliability coefficients of FLCAS used in the study are high based on previous studies (Aida, 1994; Horwitz et al., 1986; Koh & Ra, 2011; Tóth, 2008).

Statistical Procedures

1) Graded response model. Among existed polytomous IRT-models, Samejima's (1969) graded response model (GRM) is implemented for the current study. It models probability of responding in each category. GRM computes threshold parameters (β), with a common slope (α), for a given item and also allows for different spacing of categories across items. Since the restriction of α is the same for all categories for each item, the category order will always be the same. The item slope parameter indicates how well an item is able to discriminate between continuous trait levels near the inflection point and can either be fixed or free. The high value of α indicates the item response categories differentiate among the ability levels of those who choose adjacent response categories (Baker & Kim, 2004). In addition, β reflects the minimum level of the θ (ability parameter) needed to respond above that location with a probability of .50. Thus, $p_{ik}(\theta_j)$, the probability of an individual j's ability (θ) choosing response category *k* in item *i*, is defined as the difference between successive boundary curves as the following:

$$p_{ik}(\theta_j) = p_{ik}(\theta_j) - p_{ik+1}(\theta_j)$$
$$= \frac{1}{1 + \exp\,[-1.7\alpha_i(\theta_j - \beta_{ik})} - \frac{1}{1 + \exp\,[-1.7\alpha_i(\theta_j - \beta_{ik+1})}$$

More detailed information about estimation of dichotomous and polytomous IRT models can be found other studies (Baker & Kim, 2004). It is wise to implement GRM after considering optimal number of sample sizes since IRT-based model require many sample sizes. Previous studies (Craig, Palus, & Rogolsky, 2000; Lautenschlager, Meade, & Kim, 2006; Reise & Yu, 1990) investigate optimal number of sample sizes under the GRM framework. Reise and Yu (1990) suggest at least 500 sample sizes for adequate calibration of items. Nonetheless, their work on parameter recovery for the graded response model used only limited conditions with the fixed number of items (25 items). Furthermore, they suggestions obtained from the earlier version of MULTILOG 7.03 program (Thissen, Chen, & Bock, 2003) should be revisited because of less precise algorithm (Thissen, 2001). While Craig et al. (2000) uses sample sizes that varies from 59 to 278 for 54 items with GRM, pretty good parameter recovery has been reported with sample sizes as low as 300. Furthermore, Lautenschlager et al. (2006) suggest possibility of implementing GRM with sample size as low as 300.

2) Data analysis. Analytic methods for detecting DIF are potentially important and useful in evaluating the validity of cognitive functioning of FLCAS. To perform the DIF analyses, the current study uses the IRTLRDIF program (Thissen, 2001) using the likelihood-ratio test which estimates from different groups on the same ability scale, reference group (female student). It does not require separate calibration runs but use anchor items to establish the common scale. In other words, it does automatically solve scaling issues between different groups unlike other computer programs such as the MULTILOG 7.03 program (Thissen et al., 2003) in which several runs should be performed to be placed individuals' latent ability (e.g., individuals' attitude toward foreign language anxiety) obtained from different groups into the same scale. The hypothesis of absence, which is built while analyzing DIF determining with likelihood ratio, is as there is no significant difference between the item parameters that are calculated from the focused and referenced groups. In the IRTLRDIF program, results of the compact model for the test of absence hypothesis and the augmented model are compared. In the compact model, the parameters of all items in the focused and referenced groups are supposed to be equal, in other words, none of the items are assumed as DIF. In the augmented model, it is supposed that parameters of item *i*, for the focused and referenced groups can differ, and those of the other items are supposed to be equal as happens in the augmented model. While a likelihood function could be obtained from compact model, as many likelihood functions as the number of items could be obtained from the augmented model.

The IRTLRDIF program performs a series of comparisons of compact and augmented models. Likelihood ratio tests are used for comparison resulting in goodness of fit statistics

G^2 distributed as a X^2. Four steps in general are followed to perform the DIF analysis. At the first step, no anchor items define. The first comparison is between a model with all parameters constrained to be equal for the two groups (male and female students), including the studied item, with a model with separate estimation of all parameters for the studied item. G^2 value is obtained by taking the logarithms of the likelihood function of the compact model and the augmented model (Thissen, 2001). The IRTLRDIF program is designed using stringent criteria for DIF detection, so that if any model comparison results in a X^2 value greater than 3.84 ($d.f = 1$), indicating that at least one parameter differs between the two groups at the .05 level, the item is assumed to have DIF. The quantitative value of G^2 appoints the effect degree of DIF. Taking into account Cohen's G^2 statistics, the classification made for the degree of effect is as seen below (Greer, 2004). Greer (2004) suggests three magnitude level of DIF: A level if $3.84 < G^2 < 9.4$ which indicates a negligible level of DIF, B level, if $9.4 < G^2 < 41.9$ which indicates a medium level of DIF, and C level, if $G^2 > 41.9$ which indicates large level of DIF. Once any DIF is found, further model comparisons are performed.

Second step is to set anchor item. For all models, all items are constrained to be equal within the anchor set. Anchor items are defined as those with the G^2 cutoff value of 3.84 or less for the overall test of all parameters equal versus all parameters free for the studied item. This may resulted in the selection of a very small anchor set for some comparisons. Therefore, these criteria may be relaxed somewhat, and the results of the individuals parameter estimates examine rather than the overall result. If significant DIF is observed for the α‘s or β‘s using appropriate degrees of freedom, the item will be excluded from the anchor set. Thirdly, final test for DIF is followed. After the anchor item is defined, all of the remaining (non-anchor) items are evaluated for DIF against this anchor item. Some items that have been identified as having DIF in earlier stages of the analyses, could convert to non-DIF with the use of a purified anchor set. Lastly, Adjustment for multiple comparisons should be followed. Items with values of G^2 indicative of DIF in this last stage are subject to adjustment or p values for multiple comparisons used in order to reduce over-identification of items with DIF. Bonferroni, Benjamini-Hochberg and/or other comparable method to control for false discovery could be used.

Using the IRTLRDIF program instead of MULTILOG-MG 7.03 has several advantages. Besides its flexibility (possibility of manipulating missing data), its performance (e.g., statistical power) is superior to non-parametric methods with the small sample size (Bolt, 2002). It could examine uniform and non-uniform DIF simultaneously. Furthermore, it does not require equating because of simultaneous estimation of group parameters.

Despite those advantages of implementing IRTLRDIF programs over MULTILOG-MG 7.03 mentioned above, possible disadvantages of the IRTLRDIF program are that assumption before implementing IRT models should be met and no formal magnitude summary measure or guidelines are available. Last but not least concern is the type I error inflation which occur when the suggested model does not fit into data (Bolt, 2002). DIF is examined to determine whether or not the FLCAS has valid measurement properties that are invariant across groups (males and females students). In order to detect the gender related DIF, 581 students are divided into two groups based on their gender. Once the probability of correct response is calculated for each item in each group by using Samejima's (1969) GRM, item response curves (IRC) of DIF obtained from two groups are separately presented with the IRTLRDIF program (Thissen, 2001).

After performing explanatory factor analysis, all the statistical analyses (descriptive statistics, unidimentionality assumption) are followed. In addition, prior to performing GRM in the IRTLRDIF program, individuals' responses should be placed from the lowest response to 0 and increase incrementally from there. Thus, original 5-point scale starting from 1 to 5 has been transformed as 0, 1, 2, 3, and 4. Once obtaining the initially flagged DIF items at = .05 level for 33 items, inflated Type I error is corrected using the Benjamini and Hochberg's (1995) approach since the same items are used several times to be detected as the DIF related items. Furthermore, graphs describing DIF related items between two groups were drawn by R program (R Development Core Team, 2007).

Findings

A major purpose of this investigation was to determine if any DIF was present in FLCAS. In order to investigate DIF, the unidimensionality assumption, single underlying trait exclusively determining the probability of item responses, was examined with the 33 items before performing DIF analysis. Since no one accepted method for determining unidimensioanlity prerequisite assumption before performing IRT-based models, the current study implemented exploratory factor analysis (Funk & Rogge, 2007). More specifically, principal axis factoring (PAF) with varimax was conducted to find the number of factors of FLCAS. Furthermore, eigenvalues greater than one, scree plot, and minimum average partial (MAP) test were used to determine the optimal number of factors in PAF analysis.

Not surprisingly, results from the factor analysis revealed different number of optimal factors: seven factors from the eigenvalues greater than one criterion, two factors from the scree plot, and three factors from MAP. Based on the interpretability of analyses and suggestions of Horwitz et al. (1986), three-factor solution was optionally chosen. This questionnaire consisted of 33 statements, of which 9 items were for communication apprehension (1, 2, 3, 9, 13, 14, 20, 24, and 33), 9 items for test anxiety (4, 10, 15, 16, 19, 21, 25, 26, and 29), and 7 items for fear of negative evaluation (5, 8, 11, 22, 28, 30, and 32). In addition, 4 items (12, 18, 27, and 31) having double cross-loadings were eliminated and 4 items (6, 7, 17, and 23) having

Table 1
Descriptive Statistics and Results of t-Test for the FLCAS

Anxiety	Male students (n = 282)		Female students (n = 299)		t	Coefficient's
	M	SD	M	SD		
Total foreign language anxiety	96.96	19.66	106.06	15.96	-4.05**	.92
Communication apprehension	27.15	6.71	31.23	5.93	-5.04**	.89
Test anxiety	24.89	6.89	25.86	5.29	-1.29	.84
Fear of negative Evaluation	21.12	4.11	23.60	3.57	-5.06**	.72

* statistically significant at < .05 ** statistically significant at < .01.

low factor loading (< .40) were excluded (Matsunaga, 2010). In this research, results obtained from factor analysis had been used as data in order to inspect the DIF determining techniques. The results showed that the percentage of variance accounted by the first factor was about 32%, satisfying the unidimensionality assumption based on the guideline (20%) suggested by previous studies (Hattie, 1985). Results from the study showed that female students rated their own language anxiety significantly higher than male students (t (579) = -4.05, p < .01). Mean scores of female students

Table 2
Item Parameters for Focal (Males) and Reference (Females) Groups

		Focal group (Males)					Reference group (Females)				
Item	G^2	α	β_1	β_2	β_3	β_4	α	β_1	β_2	β_3	β_4
1	5.70	1.35	-2.49	-2.32	-1.15	0.10	1.12	-3.65	-2.67	-1.15	0.22
2	22.40	1.59	-2.48	-1.97	-0.77	0.14	1.15	-4.53	-2.50	-0.64	0.79
3*	**24.10**	**1.72**	**-1.56**	**-1.35**	**-0.86**	**0.34**	**2.48**	**-2.60**	**-1.60**	**-0.79**	**0.08**
4	16.60	0.84	-4.22	-3.37	-0.04	2.33	1.70	-10.06	-1.80	-0.01	1.15
5	1.30	0.81	-3.86	-2.73	-0.7	1.42	0.65	-4.12	-2.85	-0.55	1.74
8	13.30	0.92	-2.43	-2.31	-1.49	0.38	0.69	-4.20	-2.89	-1.30	1.10
9	10.30	1.34	-1.68	-1.60	-1.13	0.11	1.33	-2.52	-2.19	-1.07	0.34
10	8.20	0.79	-2.18	-2.08	-1.23	0.61	1.01	-2.37	-1.93	-0.87	0.34
11	6.50	0.65	-4.22	-3.65	-2.4	0.56	0.52	-4.44	-3.78	-1.75	1.91
13	4.70	1.68	-1.75	-1.55	-0.71	0.41	1.56	-2.41	-1.89	-0.78	0.41
14	6.60	1.37	-2.07	-1.64	-0.86	0.31	1.18	-3.18	-1.91	-0.80	0.31
15	2.20	0.77	-4.86	-3.68	-1.13	1.34	1.07	-3.21	-2.42	-0.87	1.10
16	3.70	1.29	-3.00	-1.97	-0.13	1.21	1.67	-2.59	-1.59	-0.37	0.77
19	10.40	1.51	-2.82	-1.74	0.21	1.60	1.52	-3.26	-1.61	-0.32	1.13
20	12.50	2.35	-1.62	-1.23	-0.48	0.45	2.38	-2.61	-1.34	-0.36	0.41
21	14.50	1.13	-3.30	-1.79	0.16	1.51	2.22	-2.05	-1.15	-0.26	1.00
22	11.50	0.52	-5.87	-4.72	-1.66	1.63	0.89	-5.50	-2.54	-0.88	1.45
24	5.80	1.87	-1.91	-1.58	-0.67	0.38	1.81	-2.40	-1.55	-0.43	0.45
25	7.30	0.90	-4.26	-2.68	0.00	2.01	1.59	-2.67	-1.56	-0.27	1.30
26	2.00	1.82	-2.50	-1.31	0.34	1.29	1.88	-2.45	-1.20	0.13	1.17
28	9.30	1.51	-2.33	-2.09	-1.18	0.45	1.06	-3.77	-2.71	-1.00	1.02
29	15.40	0.77	-4.60	-3.62	-0.22	2.69	1.84	-2.94	-1.83	-0.35	0.84
30	16.60	0.75	-4.20	-4.04	-2.72	1.10	1.00	-4.30	-2.85	-1.18	1.20
32	17.00	0.84	-2.14	-2.00	-1.53	0.37	0.91	-3.37	-2.57	-1.25	0.75
33*	**24.60**	**1.42**	**-1.86**	**-1.72**	**-1.23**	**0.40**	**1.64**	**-3.10**	**-1.75**	**-0.80**	**0.64**

were higher than those of male students. It was, however, interesting to note that statistically significant differences between male and female students occurred only in communication and negative evaluation anxiety sub-scales but not in test anxiety.

Information regarding item discrimination and difficulty parameters was shown in Table 2. Results from the study first showed that 21 of 25 items or about 84 % of the total items displayed significant DIF before correcting inflated Type I error: only four items (5, 15, 16, and 26) were not DIF-related items out of 25 items by the guidelines Greer's (2004) suggested. Correcting inflated Type I error with the Benjamini and Hochberg's (1995) method, however, yielded only two items (3 and 33) with DIF. Those items were nested in the communicative anxiety domain.

Table 3 showed descriptive statistics (e.g., mean, min, max, and SD) about item and ability parameters after correcting the inflated Type I error. The mean values of ability parameter in the reference group (female students) were higher than those in the focal group (male students). This result implied that female students had more classroom-related foreign language anxiety compared to male students. For instance, mean values of females' attitude toward foreign language classroom anxiety was higher than that of males' attitude toward foreign language classroom anxiety. In other words, female students were more vulnerable than male students in terms of the foreign language classroom anxiety.

Table 3
Summary Statistics of the FLCAS

		Item parameters Estimates (*I* = 25)					Ability (*J*= 581)
		α	β_1	β_2	β_3	β_4	
Focal group (Male)	Min	0.52	-5.87	-4.72	-2.72	0.10	-4.83
	Max	2.35	-1.56	-1.23	0.34	2.69	1.84
	Mean	1.22	-2.96	-2.34	-0.86	0.92	-0.20
	SD	0.46	1.20	0.95	0.76	0.73	1.33
Reference group (Female)	Min	0.52	-10.06	-3.78	-1.75	0.08	-3.94
	Max	2.48	-2.05	-1.15	0.13	1.91	3.97
	Mean	1.39	-3.53	-2.10	-0.71	0.86	0.44
	SD	0.53	1.61	0.64	0.44	0.47	1.10

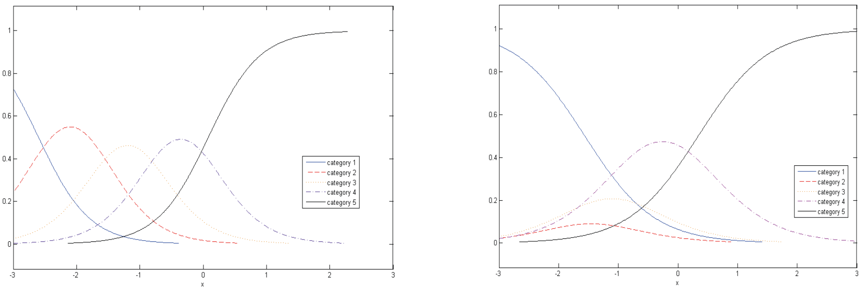

Figure 1. Item response curve of item 3 (DIF item) between females (left) and males (right).

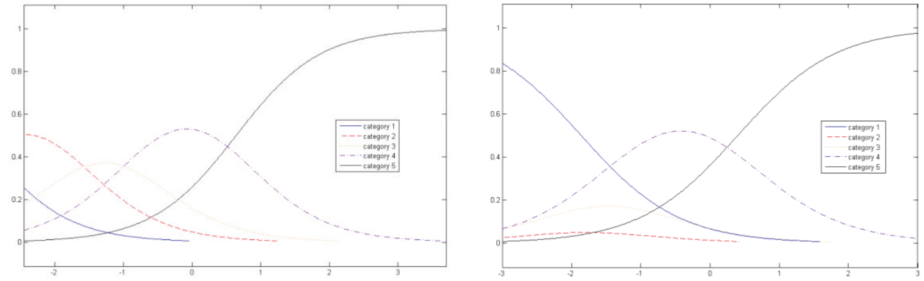

Figure 2. Item response curve of item 33 (DIF item) between females (left) and males (right).

In addition, Figure 1 and 2, describing the item response curve for the DIF items (3 and 33), showed that there were clearly different response patterns between males and females. IRC of female students had four thresholds while male students had two thresholds. Such results could imply that the probability of being at a severe level of language anxiety in the reference group is higher than that in the focal group. Similar response patterns occurred in item 33.

Furthermore, Table 4 also showed how DIF items affected levels of raw score. Mean scores of communicative anxiety in the reference group (female students) were higher than those in the focal group (male students) with and without DIF items. Difference between the reference (female students) and the focal group (male students) were statistically significant, t (579) = 5.12, p < .01 for the situation including DIF items and t(579) = 4.69, p < .01 for the situation excluding DIF items. Even though two separate situations were still statistically significant, it was worthwhile noting the shifted mean scores from high to low scores.

Table 4
Descriptive Statistics of Communicative Anxiety Domain

	Reference Group (Female)				Focal group (Male)			
	DIF included		DIF excluded		DIF included		DIF excluded	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Communicative apprehension	27.86	5.25	20.76	4.02	24.21	5.85	18.21	4.43

With respect to those DIF items, mean scores of the focal group (male students) were smaller than those of the reference group (female students). It implied that male students showed less communicative classroom foreign language anxiety than female students did.

Discussions

The numerous previous studies about foreign language classroom anxiety focus on discoveries of personal characteristics (e.g., gender) that are associated with different levels of foreign language classroom anxiety. Despite the tremendous efforts for developing cognitive test items related foreign language classroom anxiety, much research investigating psychometric properties of FLCAS is limited to factor analysis, in which a large number of factors are confirmed (Ra & Kim, 2013; Tóth, 2008). One of the drawbacks associated with studies based on traditional approach is that most of them look at the differences between females and males rather than the causes of the difference. More importantly, instruments containing items which are unfairly advantages one group over the other increase risks of jeopardizing validity for between-group comparisons because their scores are influenced by a variety of attributes other than those intended. The current study, thus, brings the necessity of examining DIF to reveal the true difference between male and female students in the EFL context and to increase reliability and validity of FLCAS. Although FLCAS is one of the most widely used foreign language measure, helping researchers investigate gender differences regarding the foreign language anxiety (Horwitz et al., 1986), it suffers from the lack of psychometric studies (Ra & Kim, 2013). After controlling for the effects of DIF on the perception of foreign language anxiety between male and female students, the true difference of foreign language anxiety between two groups on FLCAS could be revealed.

The current study verified the unidimensionality as other previous studies (Koh & Ra, 2011; Panayides & Walker, 2013). It is, however, noteworthy that the original three constructs that Horwitz et al. (1986) suggested in developing the measure does not emerge as factors of the FLCAS in three subsequent studies (Aida, 1994; Cheng, Horwitz, & Schallert, 1999; Matsuda & Gobel, 2004; Tóth, 2008). For example, Tóth (2008) verifies three components suggested in Horwitz et al. (1986) with the Hungarian version of FLCAS. On the other hand, Aida (1994) found four factors for the FLCAS in a sample of 96 American students learning Japanese. A few years later, Cheng et al. (1999) extracted two factors as did Matsuda and Gobel (2004).

After verifying the unidimensiaonlity, results obtained from the current study show that a significant number of DIF in FLCAS are revealed before controlling the inflated type I error across the level of difficulty indicating gender-related DIF occurs at all levels of foreign language–related attitude. However, after controlling gender-related DIF using Benjamini and Hochberg's (1995) approach, only two items (item 3 and 33) are present in the domain of communicative anxiety in the FLCAS. In order to better understand those two items, it is reasonable to find out what items cause DIF. Items showing DIF are "I tremble when I know that I'm going to be called on in language class (item 3)" and "I get nervous when the language teacher asks questions which I haven't prepared in advance (item 33)." In light of previous findings that those items are closely related to the typical gender characteristics, it might be possible to imply that male students did not fear of speaking out in from of class compared to female students. Compare to Ra & Kim's (2013) study examining non-equivalent item using MIMIC approach, only item 33 was identically included in both studies; item 3 in the present study while item 4, and 6 were only included in Ra & Kim's (2013) study. Theoretically, results from Ra & Kim's (2013) study and the present study should be consistent. However, considering the different results about DIF in FLCAS, it seems that more research in this areas is needed.

Although in the present study there is statistically significant difference between male and female students in mean scores of communication apprehension, the fact that the existence of DIF affects somewhat decrease of levels of communicative apprehension is noteworthy. These results from the current study indicate necessity of examining not only mean level differences in a construct, but also the deeper structure of the construct because the traditional mean level difference across genders premises conceptual equivalence across two different groups. Thus, gender differences shown in the previous studies using mean scores of FLCAS could be spurious. DIF in FLCAS shows that students having the same level of language anxiety, but of different gender, do not have the same probability of endorsing items. In other words, it is likely that the differences in the means of anxiety levels between male and female students are unclear due to true differences or due to differential functioning of some items.

One of significant contribution of this study is to use relatively easy computer program. A fundamental important issue as practitioners and researchers in assessing DIF is choosing accessible computer programs among extant statistical methods and software programs. The IRTLRDIF program is relatively easy to be used for the DIF analysis, which does not required several separate runs (e.g., MULTILOG-MG 7.03) to placed latent variables on the same scale. This program is efficient in terms of time and cost. It should be noted that the present study has some limitations with regard to its sample sizes, unbalance sample sizes and utilization of only one software program. First and foremost is the question of sample sizes. As recommended in previous studies (Craig et al., 2000; Lautenschlager et al., 2006; Reise & Yu, 1990), further analysis with a different number of sample sizes should be followed. The other one is that further study should be conducted to include unbalanced sample sizes. Since unequal same sizes could possible yield incorrect information about the gender-related DIF. In addition, impacts of deleting DIF should be considered. It is might reasonable to replace items showing DIF with an item measuring similar threshold/discrimination parameter if there are a large item pools.

Nonetheless, it should be cautious because dropping items might adversely affect the content validity of the instrument. The reliability of FLCAS in the study decreased from .96 to .92 if two items (3 and 33 items) were not excluded for the analysis. Constructs of FLCAS with three factors were shrunken into two factors. Furthermore, it is possible to use an instrument that is not comparable to other research using that instrument. Last notable psychometric property of FLCAS is its relatively high reliability. Previous studies (Adia, 1994; Koh & Ra, 2011; Panayides & Walker, 2013) including the present study show high internal consistency of FLCAS, which could be indication of item redundancy and narrowness of the scale (Boyle, 1985). Instead of trying to achieve high internal consistency, it is advisable to consider guidelines Nunnally (1978) recommended. He suggests that about .70 of reliability be enough for instrument in basic research and does not need to increase its reliability beyond .80. Furthermore, Nunnally (1978) also suggests necessity of increasing reliability until .95 if critical decision is made on the selection.

The purpose of measurement is to discriminate, evaluate, and predict individuals' abilities (Kirshner & Guyatt, 1985). Given the extensive use of FLCAS within the foreign and/or second language learning environment, it is of the utmost importance to uncover reasons of systematic different performances of FLCAS and to explain the possible difference. Work on the psychometric properties of FLCAS could shed light on the construct validation of FLCAS which is currently underway to establish foreign language anxiety (Horwitz et al., 1986). The inclusion of educational background, cultural and historical characteristics in DIF analyses will yield even relevant insights into the stability issues of FLCAS across different circumstances.

References

Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *The Modern Language Journal, 78*, 155–168.

Al-Saraj, T. M. (2014). Revisiting the Foreign Language Classroom Anxiety Scale (FLCAS): The anxiety of female english language learners in Saudi Arabia. *L2 Journal, 6*, 50–76.

Awan, R., Azher, M., Anwar, M. N., & Naz, A. (2010). An investigation of foreign language classroom anxiety and its relationship with students' achievement. *Journal of College Teacher & Learning, 7*(11), 33–40.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation technique* (2nd ed.). New York, NY: Marcel Dekker, Inc.

Baker, S. C., & MacIntyre, P. D. (2000).The role of gender and immersion in communication and second language orientations. *Language Learning*, *50*(2), 311–341.

Benjamini, Y., & Hochberg, Y. (1995).Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B*(57), 289–300.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 123–141.

Boyle, G. J. (1985). Self-report measures of depression: Some psychometric considerations. *The British Journal of Clinical Psychology*, *24*, 45–59. http://dx.doi.org/10.1111/j.2044-8260.1985.tb01312.x

Cheng, Y. S., Horwitz, E. K., & Schallert, D. L. (1999). Language writing anxiety: Differentiating writing and speaking components. *Language Learning*, *49*, 417–446.

Craig, S. B., Palus, C. J., & Rogolsky, S. (2000, April). *Measuring change retrospectively: An examination based on item response theory*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology. New Orleans, LA.

Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology*, *21*, 572–583.

Greer, T. G. (2004). Detection of differential item functioning (DIF) on the SATV: A comparison of four methods: Mantel-Haenszel, logistic regression, simultaneous item bias and likelihood ratio test (Doctoral dissertation, University of Houston).

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139–164.

Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*, 112–126.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*, 125–132.

Kirshner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Clinical Epidemiology*, *38,* 27–36.

Koh, B.-R., & Ra, J.-M. (2011). The validity of FLCAS based on item response theory. *Modern British and American Language and Literature*, *29*(3), 21–40.

Lautenschlager, G. J., Meade, A. W., & Kim, S.-H. (2006, April). *Cautions regarding sample characteristics when using the graded response model*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.

Maclntyre, P. D., Baker, S. C., Clément, R., & Donovan, L. A. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Langue Learning, 52*, 537–564.

MacIntyre, P. D., & Gardner, R. C. (1991). Investigating language class anxiety using the focused essay technique. *Modern Language Journal*, *75*, 296–304.

Matsuda, S., & Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System, 32*, 21–36.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*(1), 91–110.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297–334.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Panayides, P., & Walker, M. J. (2013). Evaluating the psychometric properties of the Foreign Language Classroom Anxiety Scale for Cypriot Senior High School EFL students: The Rasch measurement approach. *Europe's Journal of Psychology*, *9*(3), 493–516. http://dx.doi/org/10.5964/ejop.v9i3.611

R Development Core Team. (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Ra, J.-M., & Kim, J.-C. (2013). Examining measurement invariance using the MIMIC model: Gender difference in FLCAS. *The Journal of Educational Research*, *11*(2), 1–20.

Reise, S. P., & Yu, J. (1990).Parameter recover in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144.

Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Thissen, D. (2001). *IRTLRDIF v.2.0b*: *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill, NC: University of North Carolina at Chapel Hill.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). *MULTILOG 7.03* [computer software]. Lincolnwood, IL: Scientific Software International.

Tóth, Z. (2008). A foreign language anxiety scale for Hungarian learners of English. *Working Paper Language Pedagogy,* 2, 55–78.

Trang, T. T. T., (2012). A review of Horwitz, Horwitz and Cope's theory of foreign language anxiety and the challenges to the theory. *English Language Teaching, 5*(1), 69–75.

Williams, K. E., & Andrade, M. R. (2008). Foreign language learning anxiety in Japanese EFL university classes: Causes, coping and locus of control. *Electronic Journal of Foreign Language Teaching*, *5*(2), 181–191.