Gender Differential Item Functioning Analysis of an English Listening Comprehension Test

Mingming Yu · Sae Il Choi[*]
(Chonnam National University · Chosun University)

Yu, Mingming & Choi, Sae il. (2013). Gender Differential Item Functioning Analysis of an English Listening Comprehension Test. *The Linguistic Association of Korea Journal, 21*(4), 79-98. The purpose of this study was to track down gender differential item functioning (DIF) in an English listening comprehension test. The study employed multiple DIF detection methods to avoid inconsistent results from a single method. The results of the study showed that out of the total 100 items in the test, 9 items were found with significant uniform DIF—4 items favoring males while the other 5 in favor of females. However, subsequent analyses of test characteristics function and test information function revealed that despite the existence of the DIF items in favor of either females or males, the test as a whole was not gender biased. The authors further provided possible explanations for the items displaying gender DIF with support from the relevant research in the literature.

Key Words: listening proficiency test, fairness, test validity, differential item functioning, item type, item content

I. Introduction

Test fairness has been a big concern among test developers. For a test to be fair, all the test takers who are of the same language proficiency level should have the same probability of getting the item correct (Camilli & Shepard, 1994). One way to achieve fairness is to make sure that test items are not biased toward a particular group. Test bias occurs when items contain sources of

[*] Mingming Yu is the first author and Sae Il Choi is the corresponding author.

difficulty that are irrelevant or extraneous to the construct or ability being measured and these extraneous or irrelevant factors affect performance (Zumbo, 1999). Biased items can not only result in systematic errors that distort the inferences made from test scores, but also reduce the validity of the measuring instruments. For example, instruments containing such items may have reduced validity for between group comparisons, since their scores may be indicative of a variety of attributes other than those the scale is intended to measure (Thissen, Steinberg, & Wainer, 1993).

Differential item functioning (DIF) analysis has become the standard procedure to investigate item bias (Zumbo, 1999). DIF is present in a test item when, despite controls for overall test performance, examinees from different groups have a different probability of answering the item correctly or when examinees from two subpopulations with the same trait level have different expected scores on the same item (Camilli & Shepard, 1994; Kamata & Vaughn, 2004). DIF analysis has been a crucial step to examine item or test bias in high-stakes test contexts (Pae & Park, 2006) and has been applied to numerous educational and psychological tests. As shown in Karami and Nodoushan (2011), DIF also has been employed in language tests to investigate potential item bias by gender (e.g., Karami, 2011; Pae, 2012; Pae & Park, 2006; Park, 2008; Takala & Kaftendjieva, 2000), by language background (e.g., Ryan & Bachman, 1992; Kim, 2001), and academic background or content knowledge (Alderson & Urquhart, 1985; Pae, 2004).

Performance difference observed between males and females on various standardized tests have been the subject of much research. Of all researched fields of DIF, however, the focus on gender DIF in large scale high-stakes English as a second or foreign language tests is not often to see (Barati & Ahmadi, 2010). Further, DIF studies on English listening comprehension tests are very rare, compared with those on the reading comprehension. Recently in Korean context, Park (2008) attempted to identify DIF across gender in the English listening part of the 2003 Korea College Scholastic Ability Test (KCSAT) and sought the sources of DIF. The study reported that out of the 17 items in the listening part, 6 items favored males while 7 items were significantly easier for females.

The finding of such heavy differential functioning was rather shocking and a partial motivation for the current study for several reasons. First of all, since KCSAT has arguably been the highest-stakes exam in Korea, such differential effects could be totally unacceptable and a serious threat to the validity of the test. If the finding is true, the same phenomenon may be observed in other listening tests implemented in Korea, imposing huge damage on the validity of such tests. Thus, a systematic investigation of DIF in listening tests is in order. Second, results of DIF analysis could be different depending on its methods. For example, results of a DIF study relying on item response theory (IRT) based methods could be different from those on non-IRT methods. Even a single method chosen can bring about different results depending on its model specifications: significance level, purification, anchor items, and iteration details, to name a few. Third, the interpretation of the results including the causes of DIF could be very subjective. Thus, for a DIF study about a subject area to be valid, consensus by multiple replications of DIF studies would be required.

The present study examined the possible presence of gender based DIF in a large scale listening comprehension test. Specifically, the study addressed the following research questions: whether the items in a listening comprehension test exhibit significant gender DIF, what are the types of DIF and possible causes of the DIF items, and whether the existence of the DIF items lead to test bias.

Test fairness is one of the key concerns for all test developers and DIF study has been an important statistical procedure in the examination of test fairness. However, interpretation of DIF results is usually not straightforward because of uncertainty about the complex sources of item difficulty. Statistical hypotheses about the sources of DIF would be compelling if items with the same property repeatedly show significant DIF for different groups. The results of the current DIF study are expected to contribute to identifying such potential sources of gender differences, especially among Korean English learners.

2. Previous Research

It was from the 1980s that the unfair treatment on test takers concerning gender began to be studied. Since then there have been a number of studies investigating the gender differences by a variety of DIF analyses on various language tests (Rezaee & Shabani, 2009).

Takala and Kaftandjieva (2000) conducted a study to investigate the presence of DIF in the vocabulary subtest of the Finnish Foreign Language Certificate Examination. To detect DIF, Takala and Kaftandjieva applied the one parameter logistic model (OPLM) to the data from a total of 475 examinees, 182 males and 293 females. Over 25 percent of the items showed DIF in favor of either males or females. However, the total test was not biased because, as the authors pointed out, there were equal numbers of DIF items favoring males and females in the test.

Karami (2011) made use of the Rasch model to investigate the presence of DIF between male and female examinees taking the University of Tehran English Proficiency Test (UTEPT). The results of the study indicated that 19 items functioned differentially for the two groups. Only 3 items, however, displayed DIF with practical significance. A close inspection of the items indicated that the presence of DIF may be interpreted as impact rather than bias. Thus, the research concluded that the presence of DIF may not render the test unfair and the fairness of the test under question may be due to other factors.

Applying the one parameter IRT model to a response data from a sample of 36,000 students, Barati and Ahmadi (2010) investigated DIF in the special English Test of the Iranian National University Entrance Exam (INUEE). The effect of gender and subject area was taken into account. The study confirmed the presence of gender DIF in the test and concluded it is the interaction of the subject area and item format that determines the degree and direction of DIF.

Research on gender DIF in English language tests in Korean context is very limited (e.g., Pae, 2004, 2012; Pae & Park, 2006; Park, 2004, 2008). Among these studies, there is only one DIF study (Park, 2008) on a listening comprehension test. Park (2008), employing the Mantel-Haenszel procedure, identified DIF across gender in the English listening part of the 2003 KCSAT. The participants were 20,000 males and 20,000 females who took the 2003 KCSAT. Half of the participants of either group were in liberal arts and the other half in sciences. After matching the two groups with total scores, 13 out of the total 17 items in the test showed DIF, with 6 items in favor of males and 7 items differentially easier for females. The almost equal number of DIF items for males and females might cancel out each other in the test level analysis. The study also calculated the item difficulty before matching ability level and the result was quite different from DIF analysis after matching in that 2 out of 17 item significantly easier for males while 13 favoring females. These findings suggested that item difficulty statistics should be interpreted with caution because DIF could be present beyond the item difficulty indices (Thissen et al., 1986). In order to locate the source of DIF, the study further analyzed items in terms of language type (dialogue and monolog), question type (local, global and expression), picture presence, and content following the approaches found in Nunan (1991) and Shohamy and Inbar (1991). The result revealed that the four variables in developing the test were all associated with DIF to different degrees, which provided significant implications to items developers. The study solicited replications using different samples or instruments to warrant the consistency of the findings and to check if the results reflected the unique characteristic of the Korean sample.

Pae (2004) investigated gender DIF on English reading comprehension for Korean EFL learners. The study examined gender DIF for a sample of 14,000 Korean students, 7,000 males and 7,000 females, who took the English subtest of the 1998 Korean National Entrance Exam for Colleges and Universities. The study indicated that items measuring mood, impression, and tone tended to be easier for females; however, passages with logical inferences were easier for males. Pae (2012) further tracked down gender DIF in reading subtest of three KCSAT test forms over nine years using both MH and IRT-LR procedures. The results indicated reading strategy and perceived interests could explain a part of the variance in the magnitude of gender DIF and that item type rather than item content could have a systematic relationship with gender DIF.

Causes of gender DIF, an important subject of DIF analysis, have also received considerable attention from researchers. Some studies tried to find explanations from the content or the topic of test items. For example, Donlon (1973) showed that males outperformed females on items related to practical affaires and science. Items involving visualization or eliciting information about the real life are to males' advantage (Hamilton & Snow, 1998). Some studies supported that males and females use different L2 learning strategies (e.g., Ehrman & Oxford, 1988; Bacon & Finemann, 1990; Young & Oxford, 1997).

Using a two group mixture IRT model analysis, Cohen and Bolt (2002) first showed that the manifest characteristics assumed to be related to gender DIF often has a very weak relationship with the latent groups that are actually advantaged or disadvantaged by the items. Then, they proposed a two stage alternative to DIF: the first stage define the primary dimensions that contributed to DIF using an exploratory IRT mixture model analysis and the second stage studies examinee characteristics associated with those dimensions to understand the causes of DIF. They applied the alternative to a college level English placement test and argued that the model could successfully explained the causes of gender DIF.

The relationship between gender DIF and test format is also one of the active research fields. For example, evidence indicates that males generally perform better than females on multiple choice items but females perform better on essay type (constructed) items (Bolger & Kellaghan, 1990; Linn, De Benedictis, Delucchi, Harris, & Stage, 1987; Mazzeo, Schmitt, & Bleistein, 1993). In measures of quantitative abilities (vocabulary, grammar, etc.), females tend to perform better than males when constructed response items are included (Lane, Wang, & Magon, 1996). This difference has been attributed to the stronger writing skills and neater and more comprehensive answers that are provided by females (Lane et al., 1996; Mazzeo et al., 1993; Willingham & Cole, 1997). It has also been suggested that girls are more reluctant to guess on multiple choice questions than boys while boys overestimate their likelihood of success and hence take risks unknowingly (Linn et al., 1987). However, there are also studies against the findings above. For example, Barati and Ahmadi (2010) reported that the reading comprehension section in their study favored males and females equally and the item format, multiple choice questions, alone could not explain DIF. Lin and Wu (2003) performed DIF and DBF (differential bundle functioning) analyses to study differential performance by gender on the English Proficiency Test for a sample of 4459 adult Chinese EFL learners. Using the SIBTEST as a main analytical tool, they provided empirical evidence that the bundle of listening comprehension systematically favored females, whereas the bundles of grammar/vocabulary and cloze favored males.

These previous studies above contributed significantly to identifying gender DIF effects in language tests. However, they suffered from some methodological deficiencies. Most of all, these studies relied on just one or two DIF detection methods, which questioned the validity of the studies since they could not consider inconsistencies of the results. Another problem with the majority of the previous studies, especially studies in Korean context, is that they heavily relied on the Rasch model or, equivalently, the one-parameter logistic model. Those models could consider item difficulties only; thus, the scope of the studies was quite limited because non-uniform DIF could not be detected. An additional motivation of the present study was to fill in this methodological deficiencies in DIF study by encompassing multiple detection methods and adopting more realistic IRT models.

3. Method

3.1 Instrument and subjects

The test used in this study was a TOEIC practice test developed by a private testing company under the same test specifications of the official TOEIC and administered to college students across the country in March, 2009. The test was a two-hour multiple choice exam that consisted of two sections: listening and reading sections, each section with 100 multiple choice questions, respectively. The listening section had four parts: Photographs (10 items), Question Response (30 items), Conversations (30 items), and Short Talks (30 items). Test takers listen to various statements, questions, conversations, and talks recorded in English and then answer questions based on the listening segments. The reading section also includes four parts: Incomplete Sentence (40 items), Text Completion (12 questions), Single Passages (28 items), and Double Passages (20 items). For the current study, however, only the listening section was used.

A matrix of responses to the test was obtained from a sample of 592 students, 249 boys and 333 girls, at a local university. The majority of participants were students from various academic fields in the university and a few graduate students also participated. Since the test takers covered almost all academic disciplines on the campus and no academic field dominated the examinee group, interaction between the examinee’s gender and academic field were not expected. In this study, the boys belonged to the focal group and the girls were in the reference group.

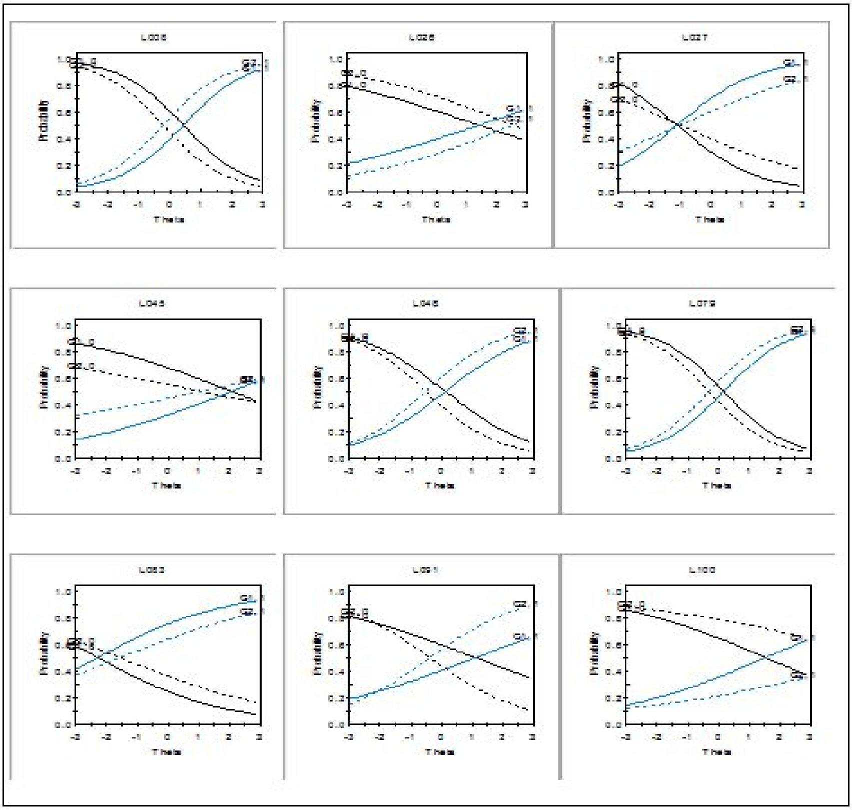
3.2 Analysis

Many detection methods have been developed to identify DIF items and usually each method comes with its own software program, which makes it hard to compare the detection results. Recently Magis, Beland, Tuerlinckx, and De Boeck (2010) developed an *R* package for DIF, called *difR*, which contains 11 traditional methods to detect DIF in dichotomously scored items. The package can compare DIF results from different methods. Both uniform and non uniform DIF effects can be detected with methods based on item response theory models. In the current study, this *difR* package was used for preliminary detection of DIF items by comparing a variety of DIF results by different methods. For non-IRT method DIF that is usually based on statistical methods for categorical data with the total test score as a matching criterion, the modified version of Mantel-Haenszel (M-H) procedure (Mazor, Clauser, & Hambletonn, 1994), Standardization procedure (Dorans and Kullick, 1986), and Logistic regression procedure (Swaminathan & Rogers, 1990) were used. For IRT-based method DIF that is based on the asymptotic properties of statistics derived from the IRT estimation results, Lord's chi-square test (Lord, 1980), Raju's area test (Raju, 1990), and Likelihood ratio test (Thissen, Steinberg, & Wainer, 1988) were used. For the IRT and non-IRT methods, both uniform and non-uniform DIF detection were allowed and item purification procedure was taken for each method. In addition, for the IRT-based method, two parameter Logistic IRT model was employed. For more details about using the *difR* package and specifications of each DIF detection method, readers are referred to Magis et al. (2001). Finally, the DIF results by these six methods from the *difR* program were then compared to the one obtained from DIF analysis using IRTPRO (Cai, 2012) program. Eventually, potential DIF items were chosen only if all the seven methods above agreed on their DIF effects.

Table 1. DIF Results

item	Totalχ^2	df	p	χ^2_a	df	p	χ^2_b	df.	p
8	9.5	2	0.008	0.00	1	0.942	9.5	1	0.002
26	7.5	2	0.023	0.10	1	0.737	7.4	1	0.006
27	6.7	2	0.035	1.40	1	0.244	5.4	1	0.020
45	9.9	2	0.007	1.00	1	0.330	9	1	0.003
48	7.5	2	0.024	0.20	1	0.625	7.3	1	0.007
79	6.4	2	0.042	0.00	1	0.914	6.3	1	0.012
83	8.4	2	0.015	0.20	1	0.643	8.2	1	0.004
91	16.1	2	0.000	2.30	1	0.131	13.8	1	0.000
100	15.7	2	0.000	0.70	1	0.419	15	1	0.000

Figure 1. Trace Lines of DIF Items



Note: G1 = male group, G2 = female group

4. Results

The preliminary screening by the *difR* program indicated that of the total 100 items in the test, 11 items have significant gender DIF (p<.05). However, the DIF analysis by the IRTPRO program was a little conservative and chose only 10 items as potential DIF items. Table 1 presents results of the DIF analyses for the 9 Items on which the two programs agreed about their DIF effects. The table shows that all the nine items have so called uniform DIF effects; all item difficulties have significant DIF effects while none of the discriminations shows any significant DIF effect. Such uniform DIF effects can also be observed in Figure 1, which shows item characteristic curves of all the potential DIF items. According to the Figure 1, items 8, 45, 46, 79, and 91 were easier for the female students whereas items 26, 27, 83, and 100 favored the male students. Further, the figure also shows that items 27 and 91 had minor non-uniform DIF effects but they were not significant (p>.244, and p>.131, respectively).

Table 2. 2PL-IRT Model Item Parameter Estimates

	male				female			
item	a	s.e.	b	s.e.	a	s.e.	b	s.e.
8	1.09	0.24	0.30	0.15	1.12	0.18	0.29	0.11
26	0.34	0.14	1.17	0.61	0.40	0.14	2.18	0.81
27	0.86	0.31	1.09	0.56	0.46	0.15	1.00	0.36
45	0.40	0.15	1.77	0.70	0.21	0.13	0.87	0.78
48	0.81	0.23	0.04	0.20	0.95	0.18	0.54	0.14
79	1.07	0.31	0.07	0.13	1.11	0.19	0.36	0.11
83	0.55	0.26	2.14	0.93	0.41	0.15	1.49	0.55
91	0.4	0.16	0.92	0.48	0.73	0.16	0.42	0.16
100	0.44	0.15	1.29	0.51	0.27	0.15	4.77	2.70

Note: a: discrimination, b: difficulty, s.e.: standard error

Table 2 provides numerical differences of item parameter estimates for the male and female groups. The table shows that all the differences of item discriminations between the two groups are within ±2·s.e; however, some differences of item difficulties between the two groups(e.g., items 26, 45, and 100) are beyond ±2·s.e. Further, for some reasons, items 26 and 100 were very challenging for the female students while item 4 was differentially difficult for the male students. Clearly, these items warrant careful examination of their properties, including item formats , content, and type.

Although about 10% of the items showed DIF effects, the test as a whole did not favor either gender. Shown in Figure 2 are the test characteristic curves for the male and female students, which are almost indistinguishable over the entire range of the student ability. The figure indicates that the effects of those DIF items were cancelled out and no serious bias toward either gender occurred.

Figure 2. Test Characteristic Curve

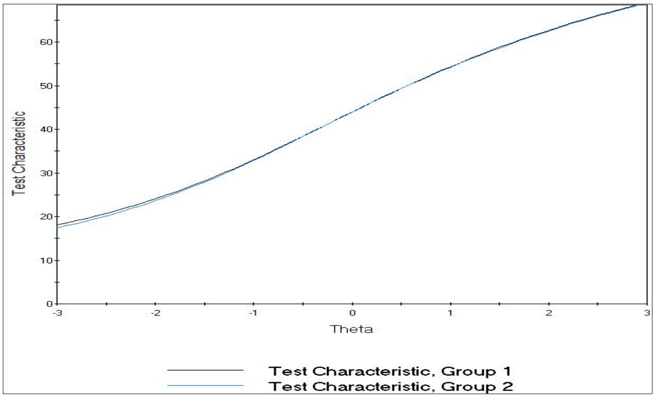
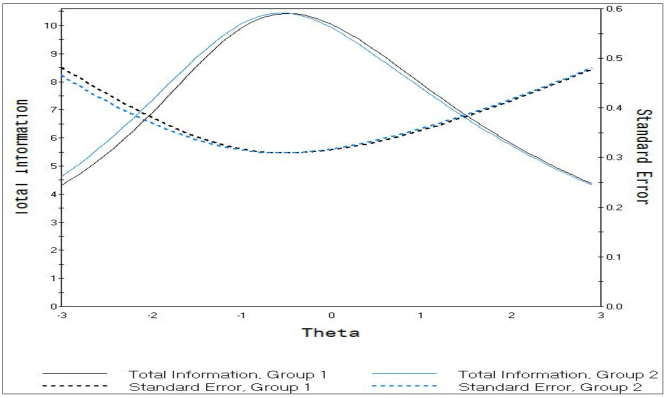


Figure 3. Test Information Curve



Perhaps almost the same number of items indicating DIF in favor of either group may counterbalance each other and DIF at the level of individual items may be canceled at the test level (Drasgow, 1987; Roznowki & Reith, 1999; Zumbo, 2003). In order to render the results above more plausible, the study examined the test information function shown in Figure 3. The information curves in Figure 3 are very similar to each other. Although the male group shows slightly larger test information beyond 0.5 ability scale and the female group has slightly larger test information over the other range of the ability scale, the difference seems to be trivial. In terms of the standard error of measurement (SEM), the difference beyond the ability scale 1.5 is almost indistinguishable. In fact, the reliability for the male and female groups are 0.90 and 0.89, respectively, which again manifests that the test as a whole was not gender biased. Based on the evidence shown above, it is reasonable to conclude that the test was not gender biased, despite the existence of a few DIF items.

After DIF items are detected, it is crucial to investigate the causes of DIF across the two gender groups. According to Tittle (1982), a test or an exam can favor female or male testees in three possible ways: content, format, and type. Therefore, the authors discussed plausible causes of the gender DIF items in light of the three ways.

In the listening comprehension section, all the 9 items with significant gender DIF (item 8, 26, 27, 45, 48, 79, 83, 91, 100) displayed uniform DIF. Item 27 exceptionally exhibited minor non-uniform DIF but it was not significant (p> .244) and its DIF effect was mainly due to the difference of item difficulty (p> .020). Among the 9 items with uniform DIF, items 8, 45, 48, 79 and 91 are partial to females and the rest four items (item 26, 27, 83 and 100) favor male students.

Items 8, 48, 79, and 91 seemed to favor female students for a common reason: attention to details. First, item 8 requires students to listen to the four options that describe a given picture and to choose the right answer that matches some details of the picture. To get the correct answer, students should note that a couple of people in the picture are looking in the same direction. Because females tend to be more sensitive and attentive to details in visual input, they might get right to the point of a picture more accurately than males and could determine the right answer. Similarly, item 48 requires test takers' attention to specific details in a stretch of dialogue by a couple over the phone.

Most of the dialogue is about a husband's excuse for the cancellation of dinner with his wife owing to a sudden staff meeting at work but the question itself is about what his wife is going to eat for dinner. Thus, females might outperform males in this item, relying on their strong sense of details.

Item 79 is based on a long monologue about a car advertisement. The listening text is filled with a full description of a brand new car, including its interior, capacity, and installment plan. However, the test item is about a customer service for color mentioned very briefly at the end of the long advertisement. Again this item requires test takers to pay attention to a very small detail that deviates from the mainstream of the text. Item 91 also had a very similar pattern to those items mentioned above. The majority of the listening text in item 91 was based on an advertisement about a rising jazz band, including their music style and ticket selling policy for upcoming recital series. The question of item 91, however, was very confined to a small detail of the advertisement, the benefit of purchasing tickets in advance. Thus, for the same reason behind item 8, 48, and 79, females student could outperform males. These observations support Pae's (2012) findings in his DIF analysis of the reading subtest in 2003 KCAT that the specific information items flagged for DIF never favored males.

Males and females differ in their perceived interests and activities, and these differences are likely to have an influence on school activities, grades and test scores (Willingham & Cole, 1997). Specifically, females are more likely to be interested in humanities, arts, families, education and social sciences than males are. This may serve as evidence for item 45, an item to females' advantage, since its topic about shopping is in females' interest list. In his DIF analysis of the 2003 KCAT English section, Park (2008) also found that items about shopping favored females. More generally, the results of DIF analysis of item 45 is consistent with the conclusion made by Stricker and Emmerich (1999) that there is a significant relationship between gender differences in the examinee's perceived interest and the magnitude of gender DIF.

In contrast, item 26 requires test takers to infer the most reasonable and logical response after listening to a statement. In item 26, a person says he thought his interlocutor was in another city, which was not true. Given this context, test takers were required to choose the most reasonable reaction from the interlocutor using their logical deduction. Males could be comparatively better at logical analysis, so they might outscore females in the item. Similarly, Lin and Kong (2009) also found in their differential bundle analysis of the English Proficiency Test in China that males performed much better with items about scientific research, logical deduction and verifying facts than females do.

Items 83 and 100 may be explained from the perspective of content familarity. Content familiarity means that particular groups will have better performance on a text they are familiar with. Content familiarity has been known to be a significant factor in gender DIF and was found to have significantly affected students' overall language performance (Brantmeier, 2003; Bügel, & Buunk, 1996; Floyd & Carrell, 1987; Hyde & Lynn 1988). The topic of item 83 is about technical details of a latest printer model: its power consumption, warm up time, installation environment, and adapter requirement, just to name a few. Since the description does not explicitly name the model as a printer, the test takers should make an educated guess about what the item could be after hearing a long description of technical details.

Item 100 is very similar to item 83 in structure; however, the subject is not about an electronic appliance (printer) but now about a new mobile service provided by a consortium of telecom firms. The description of the service is technically thick since the text is mainly about which firm in the consortium (three firms with one firm as an affiliate of another) is in charge of which part of the service, say, who is going to provide the mobile service itself and who is going to provide the main infrastructure. These are all areas that males may enjoy and explore most. Therefore, it is not surprising that males performed better on the two items. Conversely, females had much difficulty handling these items; in fact, item 100 was the most difficulty item for females. Again, the results are in accordance with the conclusion of Mazzeo et al. (1993) and O'Neill and McPeek (1993) that items related to science or stereotypical male pursuits are to males' advantage.

5. Discussion and Implications

The present study was an attempt to examine the gender differential item functioning (DIF) on the listening section of a English proficiency test. Multiple DIF detection methods, including three IRT based methods (Lord, Raju, LRT) and three non-IRT methods (M-H, Standardization, Logistic), were employed to avoid method-dependent DIF results. The results of the study indicated that 9 items were functioning differentially; all 9 items displayed uniform DIF—5 items functioned differentially for females while the other 4 showed DIF in favor of males. Subsequently some possible explanations for the causes of DIF items were pursued. Despite the existence of DIF items in the test, however, the test as a whole did not demonstrate much gender difference; specifically, the test characteristic curves for both groups are almost indistinguishable over all the ability range and the test information curves shows very negligible differences.

The results of this study have some implications for item and test construction, listening pedagogy, and test fairness. First of all, potential sources of biased items in terms of question type, content, perceived interest, and stereotypical gender pursuits should be considered in the process of item development. For example, the DIF analyses in this study showed that items measuring a very small detail that deviates away from the mainstream of the talk (e.g., items 8, 48, 79, and 91) can differentially function for females. The study also observed that items assessing heavy technical content (e.g., item 83 and 100) may function in favor of males while content familiarity and perceived interest can work differentially for either gender. Such type of items could be carefully reexamined for their validity by an item selection committee or test item writers before they are operationalized.

Second, results of DIF analyses in this study provide a valuable lesson in listening pedagogy; the current listening instruction should be expanded in scope and depth for each gender. For instance, considering items measuring deviating small details can work against males, teachers can provide remedial instructions for male students on this area. In addition, taking DIF items with gender biased interest and content into consideration, teachers can help their students by intentionally training each gender with items constructed from topics of the opposite gender's interests.

Third, the present study also supports the crucial point echoed in almost all DIF studies that a test with items flagged for DIF is not necessarily unfair or invalid. In this study, 5 items favored females while 4 items worked differentially for males. The difference in the number of DIF items was minimal and their DIF effects might be cancelled out by each other. However, there could be another possible explanation for the cancellation. The test of the current study was exceptionally long; it contained 100 items covering the four subsections. It might be possible that the outnumbering non-DIF items (91 items) dominated both the test characteristics function and the test information function, masking substantial DIF effects between the two groups. Thus, the cancelling result should be accepted as an exceptional case, not as a general case with a test that contains plenty of items.

Some limitations of the current study must also be noted. As in most DIF studies, the study inevitably risked some degree of subjectivity in choosing the DIF methods, setting up delicate specifications for each method, and explaining possible causes. While most DIF studies in the literature employed just one or two detection methods, this study was extremely conservative in choosing the DIF methods to avoid method-dependent DIF results. Altogether, 7 different methods, six in the preliminary DIF study and one installed in IRTPRO program, were used to detect DIF and items all the seven methods agreed on were selected for further analyses. The main reason the DIF procedures from the IRTPRO program was chosen was that the program provides rigorous statistical testing for all aspects of DIF. However, this never means that the program is superior to any other procedures; its choice is primarily for convenience. Further it should be acknowledged that such conservative approach was taken at the cost of so called inflated Type II error in that existing true DIF effects could not be detected due to the extreme conservative procedures.

Since DIF detection methods are a sophisticated statistical procedure, they inherently entail many delicate specifications: anchor items, significance levels, purification procedures, iteration numbers, stopping rules, etc. Thus DIF results could differ not only between different methods (inter-method inconsistency) but also within a detection method (intra-method inconsistency). A simultaneous evaluation of all these methods on the same metric using simulated data sounds promising but the computation would be too heavy to implement. Although technically limited, the *difR* program was quite handy in dealing with this inter-method inconsistency. However, even the use of this handy program cannot help making subjective decisions on the choice of each specification for different detection method.

Lastly, the search for the possible causes of DIF between males and females was quite subjective although every effort was made to explain the causes objectively with support of the results from previous DIF studies in the literature. More specifically, the explanation for DIF in this study was a two-stage trial in that DIF analyses was performed first and then some possible causes for the DIF effects were brought in from the literature, not from some simultaneous measurements of the subject characteristics. Simply the current data structure did not allow such modeling because only measurement data were provided. Due to this lack of the structural information about the test takers, the study could not take other potential sources into consideration such as testing strategies, cognitive skills, general language ability and academic background.

Test fairness is a broad concept which encompasses much more than a mere DIF analysis of the items (Davies, 2010; Kane, 2010; Kunnan, 2010; Xi, 2010) and content analysis alone cannot lead us far in this regard (Nandakumar, 1993; Pae, 2004; Scheuneman & Gerritz, 1990). For a better understanding of the causes of gender DIF on language tests, further research is expected to investigate both test taker's characteristics and item properties simultaneously.

References

Bacon, S. M., & Finemann, M. D. (1990). A study of the attitudes, motives, and strategies of university foreign language students and their disposition to authentic oral and written input. *Modern Language Journal, 74*(4), 459 – 473.

Barati, H., & Ahmadi, A. R. (2010). Gender based DIF across the subject area: A study of the Iranian National University Entrance Exam. *The Journal of Teaching Language Skills, 2*(3), 1-26.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender inferences in scholastic achievement. *Journal of Educational Measurement, 27*(2), 157-164.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cohen, A., & Bolt, D. (2002). *A mixture model analysis of differential item functioning*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Davies, A. (2010) Test fairness: A response. *Language Testing, 27*(2), 171-176.

Donlon, A. (1973). *Content factors in sex differences on test questions*. Research Memorandum, 73-28. NJ: Educational Testing Service.

Ehrman, M., & Oxford, R. (1988). Effects of sex differences, career choice, and psychological type on adult language learning strategies. *Modern Language Journal, 72*, 253 – 265.

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests*. CSE Technical Report 483.

Kamata, A., & Vaughn B. (2004). An Introduction to differential item functioning analysis. Learning Disabilities: *A Contemporary Journal, 2*, 49-69.

Kane, M. (2010) Validity and fairness. *Language Testing, 27*(2), 177-182.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies, 5*(2), 27-38.

Kunnan, A. J. (2010) Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183-189.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium*, Orlando, Florida (pp. 1-14). Cambridge: Cambridge University Press.

Lane, S., Wang, N., & Magon, M. (1996). Gender related differential item functioning on a middle school mathematics performance assessment. *Educational Researcher, 15*, 21-27.

Linn, M. C., De Benedictis, T., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: what does "don't know" really mean? *Journal of Research in Science Teaching, 24*(3), 267 278.

Lin, J., & Wu, F. (2003). *Differential performance by gender in foreign language testing*. Poster presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). *Sex related performance differences on constructed response and multiple choice sections of Advanced*

Placement Examinations. (College Board Report No.92-7; ETS RR-93-5). New York: College Entrance Examination Board.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

Nunan, D. (1991). *Language teaching methodology: A textbook for teachers*. New York: Prentice Hall.

Pae, T. I. (2002). *Gender differential item functioning on a national language test*. Unpublished doctoral dissertation, Purdue University, West Lafayette, Indiana, United States.

Pae, T. I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32*(2), 265–281.

Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple data analysis over nine years. *Language Testing, 29*(4), 533-554.

Pae, T., & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475-496.

Park, G. P. (2004). Comparison of L2 listening and reading comprehension by university students learning English in Korea. *Foreign Language Annals, 37*, 448–458.

Park, G. P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly, 42*(1), 115–122.

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*(13), 109-131.

Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing, 8*, 23-40.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*, 323–340.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

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

Willingham, W. W., & Cole, N. S. (1997). Fairness issues in test design and use. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 227-346). Hillsdale, NJ: Lawrence Erlbaum Associates.

Xi, X. (2010) How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.

Young, D. J., & Oxford, R. (1997). A gender related analysis of strategies used to process written input in the native language and a foreign language. *Applied Language Learning, 8(*1), 43 – 73.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert type (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Mingming Yu
Department of English Education
Chonnam National University
Yongbong-ro, Buk-gu,
Gwangju 500-757, Korea
Phone: 010-3717-9608
Email: sandrayu168@hotmail.com

Sae Il Choi
Department of English Education
Chosun University
375, Sesek-dong, Dong-gu
Gwangju 501-759, Korea
Phone: 62-431-8788
Email: csieagles@gmail.com

Received on September 30, 2013
Revised version received on November 30, 2013
Accepted on December 10, 2013