THE JOURNAL OF ASIA TEFL Vol. 8, No. 4, pp.175-203, Winter 2011

Beyond the Mean Differences of the SILL by Gender: Differential Item Functioning

Gi-Pyo Park
Soonchunhyang University

Brian F. French
Washington State University at Pullman

This study investigated the mean differences and differential item functioning (DIF) of the strategy inventory for language learning (SILL) by gender for university students in an English as a foreign language (EFL) context, using analysis of variance (ANOVA) and ordinal logistic regression (OLR), respectively. Only memory strategies out of the six strategy categories in the SILL showed a significant gender difference, with males using more strategies than females. After matching on score levels for DIF detection using each of the scale scores of six strategy categories as a total score, a total of 13 items showed DIF across gender with 12 items in favor of males and one item in favor of females. These 13 DIF items belonged to memory strategies (5 items), cognitive strategies (4 items), metacognitive strategies (2 items), compensation strategies (1 item), and affective strategies (1 item). This study concludes with the implications of the findings followed by future study areas.

Key words: learning strategies, strategy inventory for language learning (SILL), differential item functioning, gender, ordinal logistic regression

INTRODUCTION

Language learning strategies defined as any behaviors or thought processes used to facilitate language learning have been the subject of a growing body of research in the domain of L2 acquisition and teaching (Brown, 2000; Ellis, 1994).[1] This is mainly because good language learners use different learning strategies than less effective language learners and because learning strategies used by good language learners can be taught to less effective language learners (Cohen & Macaro, 2007; O'Malley & Chamot, 1990; Oxford, 1990). However, in order to engage in such instruction less effective learners have to be identified through an assessment process.

Oxford (1990) developed the Strategy Inventory for Language Learning (SILL) that consists of six strategy categories: memory, cognitive, compensation, metacognitive, affective, and social strategies. It has been used worldwide to investigate language learners' learning strategy use, variables affecting learning strategy use, and the relation of strategy use to L2 proficiency (Chen, 2009; Green & Oxford, 1995; Griffiths, 2003; Hong-Nam & Leavell, 2006, 2007; Kavasoğlu, 2009; Khalil, 2005; Magogwe & Oliver, 2007; McMullen, 2009; Nisbet, Tindall, & Arroyo, 2005; Oxford & Ehrman, 1995; Park, 1997; Tercanlioglu, 2004; Wharton, 2000; Yang, 1999). Even though the reliability and validity evidence to support the inferences of the SILL scores has been reported to be strong, no studies to date have explored, as a validation process, whether the items of the SILL show differential item functioning (DIF) across members of different subgroups such as gender. Investigating DIF is essential to ensure that the constructs across subgroups have the same meaning and to allow for accurate and fair mean comparisons (Cole, Maxwell, Avery, & Salas, 1993; Elder, 1997; Ferne & Rupp, 2007; Thissen et al., 1986). It should be noted that DIF investigations within the larger realm of measurement invariance studies are in accord with the

[1] This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2009-013-A00052).

guidelines in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) for providing the necessary validity evidence for all instruments.

Studies investigating gender differences in learning strategy research is crucial because men and women are considered to be different in educational and occupational outcomes in general and because they use learning strategies differently from each other to facilitate L2 acquisition in particular (Ehrman & Oxford, 1989; Radford, 1998). For instance, Ehrman and Oxford reported that women used more learning strategies compared to men and more women than men preferred intuition and feeling to sensing and thinking, respectively.

The purpose of this study is to investigate gender differences at the scale and item level of the SILL (Version 7.0) before and after matching on score/ability levels by the university students learning English in an English as a foreign language (EFL) context, using analysis of variance (ANOVA) and ordinal logistic regression (OLR). OLR has advantages over other DIF identification methods such as item response theory and the Mantel-Haenszel procedure with regard to sample size and polytomous data manipulation (Su & Wang, 2005; Zumbo, 1999).

For this purpose, the following two research questions were examined:

1. Are there mean differences in the six strategy categories of the SILL by gender?
2. Are there items in the six strategy categories of the SILL that exhibit DIF across gender?

BACKGROUND

Since the pioneering "good language learner" studies (Rubin, 1975; Stern, 1975), research on language learning strategies has burgeoned around the globe, with several books featuring a crucial chapter in the domain of L2

acquisition and teaching (Brown, 2000; Ellis, 1994). The primary reason for this surge of research on learning strategies is due to the findings that good language learners use learning strategies differently from less effective language learners in quantity as well as in quality (Chen, 2009; Gan et al., 2004; Green & Oxford, 1995; Griffiths, 2003; Magogwe & Oliver, 2007; McMullen, 2009; Nisbet et al., 2005; O'Malley et al., 1989; Park, 1997; Phakiti, 2003; Wharton, 2000). In these findings, the strategy inventory for language learning (SILL) developed by Oxford (1990) has been widely used around the world to determine L2 learners' learning strategy use.

The SILL measures foreign language learning strategies by English speakers (Version 5.0) and English learning strategies by speakers of other languages (Version 7.0, see Appendix). It consists of six different strategy categories: memory strategies for storing and retrieving new information; cognitive strategies for understanding and producing new language; compensation strategies for overcoming gaps in knowledge; metacognitive strategies for controlling or coordinating the learning process by functions such as planning and evaluating; affective strategies for regulating emotions and motivations; and social strategies for learning through interaction with others. In the administration of the SILL, the respondents are asked to complete each item on a five point Likert-scale ranging from 1 (never or almost never true of me) to 5 (always or almost always true of me). Evidence exists attesting to the reliability and validity of the SILL scores, often using Cronbach's alpha and factor analysis (Green & Oxford, 1995; Nisbet et al., 2005).

The SILL has been primarily used in the following two subareas of learning strategy research: strategy use affected by variables such as gender and nationality and the relation of strategy use to L2 proficiency and achievement (El-Dib, 2004; Green & Oxford, 1995; Hong-Nam & Leavell, 2006, 2007; Khalil, 2005; McMullen, 2009; Nisbet et al., 2005; Oxford & Ehrman, 1995; Park, 1997; Tercanlioglu, 2004). In these subareas, gender as an underlying variable of strategy use has attracted specific attention among researchers because the findings show that males and females use different

behaviors and thought processes to facilitate language learning, which in turn influences variables underlying strategy use and the relation of strategy use to L2 proficiency. Thus, the findings of learning strategy research without considering the intervening variable of gender can be superficial and misleading.

Nevertheless, gender had been ignored among L2 acquisition researchers until Oxford et al. (1988) raised a serious concern regarding the lack of gender studies in learning strategy research. After reviewing the previous limited number of studies on gender, they found that females in general used more learning strategies compared to males because of females' dominant use of socially oriented learning strategies. Since then, gender has been the subject of learning strategy research around the world, especially using the SILL (Ehrman & Oxford, 1989; El-Dib, 2004; Green & Oxford, 1995; Hong-Nam & Leavell, 2006; Kavasoğlu, 2009; Khalil, 2005; McMullen, 2009; Mochizuki, 1999, Nisbet et al., 2005; Oxford & Nyikos, 1989; Oxford & Ehrman, 1995; Tercanlioglu, 2004; Wharton, 2000).

Confirming the early gender studies reviewed by Oxford et al. (1988), many researchers have shown that females used significantly more learning strategies than males in the total strategy use of the SILL by the foreign language learners and teachers in the USA (Ehrman & Oxford, 1989), the university students in Puerto Rico (Green & Oxford, 1995), the foreign language learners in FSI (Oxford & Ehrman, 1995), the EFL university students in Japan (Mochizuki, 1999), the EFL high school and university students in Israel (Khalil, 2005), and the EFL pre-service teachers in Turkey (Kavasoğlu, 2009). However, these findings are not always consistent with other studies. For instance, the male students at the university level in Kuwait used significantly more learning strategies than their counterparts (Tercanlioglu, 2004). In addition, no significant gender differences were observed elsewhere by the foreign language learners at the university level in Singapore (Wharton, 2000), the English as a second language (ESL) university students in the USA (Hong-Nam & Leavell (2006), and the EFL university students in Kuwait (El-Dib, 2004), China (Nisbet et al., 2005), and

Saudi (McMullen, 2009).

More specifically, in the six strategy categories of the SILL, females used significantly more learning strategies compared to males in memory strategies (Green & Oxford, 1995; Khalil, 2005; Kavasoğlu, 2009), cognitive strategies (Mochizuki, 1999), compensation strategies (Oxford & Ehrman, 1995; Mochizuki, 1999), metacognitive strategies (Green & Oxford, 1995; Mochizuki, 1999; Khalil, 2005; Kavasoğlu, 2009), affective strategies (Green & Oxford, 1995; Hong-Nam & Leavell, 2006; Mochizuki, 1999), and social strategies (Green & Oxford, 1995; Mochizuki, 1999; Kavasoğlu, 2009). Conversely, males used significantly more learning strategies than their counterparts in cognitive and metacognitive strategies (Tercanlioglu, 2004).

Unfortunately, most of these studies have not shown item-based analysis of the SILL concerning gender differences, with a few exceptions (Green & Oxford, 1995; Wharton, 2000). Green and Oxford reported that females used 14 strategy items out of the 50 SILL items significantly more than males, whereas males used only one strategy item significantly more than their female counterparts. Wharton found that 13 and 10 items out of the 80 SILL items showed significant variation by gender in favor of males and females, respectively. It should be noted that this similar number of items might balance out each other in the SILL as a whole and produce no significant gender differences. The point is that items that are endorsed by females or males at a different rate due to gender and not ability could influence the inferences and conclusions we make at the group level about strategy use.

Thus, previous studies to date have shown that in general females used more language learning strategies compared to males to facilitate L2 acquisition. However, uncertainty still remains as to whether this result was because females used truly more learning strategies than their counterparts or whether this difference was because the items of the SILL showed differential item functioning (DIF) for females. DIF is present when different group members have differential probabilities of getting an item correct in a test or when they endorse a certain level of an item in a questionnaire after matching on ability/score levels (Clauser & Mazor, 1998; Zumbo, 1999). That

is, if men and women have the same level of strategy use, yet on a given item they have a different probability of endorsing the same point (e.g. a 4) on the scale, the item is said to exhibit DIF. Such items can threaten the meaning of the ability assessed and lead to inaccurate conclusions about groups.

DIF investigations are common in both educational and psychological measurements. Most relevant to this work are scales that contain item types that are self-report rating scales (e.g., 1-5 scale) for traits and not typical achievement items (e.g. correct/incorrect). Gender DIF studies are quite common on such measures ranging from the MMPI (Waller et al., 2000) and the NEO-PI-R (Reise, Smith, & Furr, 2001) to university course evaluation forms (Finch & French, 2010) and integration to college scales (Breidenbach, & French, 2011). All such investigations are concerned with gathering the needed evidence to support measurement invariance across groups in order to support accurate group comparisons (e.g., mean differences).

Two types of DIF can be assessed: Uniform DIF, which is concerned with item difficulty/endorsement and non-uniform DIF which is concerned with item discrimination. Item discrimination refers to how well the item differentiates between those participants who have higher proficiency on the trait being measured by the instrument from those who have lower proficiency. Studies on DIF are important because the mean score difference of a construct between group members may not be due to the true difference of the construct, but be due to construct irrelevant variance or construct underrepresentation (Maller 2003; Messick, 1989).

Several studies on DIF have been published in the domain of L2 acquisition and teaching (Elder, 1997; Ferne & Rupp, 2007; Geranpayeh & Kunnan, 2007; Kim, 2001; Kunnan, 1990; Pae & Park, 2006; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000; Zumbo, 2003). These studies investigated whether the items of a test displayed DIF or if the mean differences of a test were present across members of different subgroups such as gender or nationality.

One of the earliest but insightful studies regarding DIF across subgroups was undertaken by Kunnan (1990) who investigated by the Rash model from

IRT whether the ESL Place Examination (ESLPE) at the University of California at Los Angeles (UCLA) displayed DIF across native languages and gender. Kunnan reported that 20 items were differentially easier for the males, whereas only three items were differentially easier for their female counterparts.

Ryan and Bachman (1992) detected DIF across native language background and gender in the Test of English as a Foreign Language (TOEFL) and in the First Certificate of English (FCE) by the Mantel-Haenszel procedure. The result revealed that six items out of the 140 TOEFL items and two items out of the 38 FCE items showed DIF across gender. In terms of DIF direction, four items were in favor of males and two items were in favor of females in the TOEFL, whereas one item was a non-uniform item and thus both males and females were favored depending on ability level, respectively.

Pae and Park (2006) investigated DIF across gender in the Korea College Scholastic Aptitude Test (KCSAT) and found that six items were in favor of males and seven items were in favor of females. More specifically, items related to sports, advising, and cleaning the road were differentially easier for males, whereas items related to shopping, watching TV, and dining out were differentially easier for females.

More recently, Geranpayeh and Kunnan (2007) identified six DIF items across age groups in the listening section of the Cambridge Certificate in Advanced English examination by the Marginal Maximum Likelihood Ratio Test. However, in the follow-up content analysis by review panels, seven DIF items were detected. These findings by statistical analysis and review panels indicate that there are limitations in identifying the sources of DIF on test items.

As seen in these examples, research on DIF to date has focused on high stakes tests. Regardless of many studies conducted utilizing a self-report questionnaire such as the SILL in the domain of L2 acquisition and teaching, no study to date has investigated whether the items of the SILL show DIF across different subgroups. Before the field goes too far down the path of

relying on the SILL for comparing mean differences, evidence is needed to support that the items of the SILL function the same across males and females and support that strategy use is measured equally well in both groups.

METHODOLOGY

Participants

A total of 948 students (N_{males}=368, $N_{females}$=580) who took an English conversation course in the spring semester of 2009 as a required course at a university in Korea participated in this study. Most of the participants were sophomores with the average age of 21, studying in different colleges such as humanities (177), social sciences (269), natural sciences (127), engineering (179), medical sciences (174), and medicine (22). There were additional 23 participants who were excluded from data analysis because they did not respond to the SILL items sincerely and/or failed to indicate gender.

Two assumptions were made in this study. First, taking into consideration that the participants had studied English with a focus on four skills from seven to ten years in public schools at the time of data collection, their overall English proficiency was assumed to be intermediate, showing huge individual differences in English proficiency depending on various factors such as study hours outside of the classroom, learning strategy use, and motivation (Skehan, 1989). Second, among the 113 students who reported the scores of the Test of English for International Communication (TOEIC), males (*m*=653 out of 990, SD=183) outscored females (*m*=627, SD=203). Thus, the overall English proficiency of male participants was assumed to be higher than that of female participants.

Data Collection Procedure

The first author contacted eight native English professors who taught

English conversation at a university in Korea, as per the recommendation of the staff who was in charge of assisting and managing native English professors in the Department of Practical English. After assuring their assistance in collecting the data in class, the first author met each professor individually and explained how to collect the data with the SILL: First, briefly explain the purpose of this study to the participants and ask for their participation in this study. Second, tell the participants to decide whether or not they will participate in this study, followed by the notice that participation does not affect course grades. Third, explain to the participants what the SILL assesses and how to respond to each item of the SILL on a five-point Likert scale. Fourth, ask the participants to respond sincerely and honestly to each item of the SILL which does not have correct or wrong answers. And finally, tell the participants to continue to think about various learning strategies which will facilitate their English learning.

Data collection required approximately 50 minutes. Most participants in class decided to cooperate with this study by responding to the SILL. However, those who did not reveal their gender or failed to respond to more than five items out of the 50 SILL items were excluded in the data analysis, leading to a total of 948 data which were included in analyses.

Data Analysis

All the data were included in statistical analyses performed by analysis of variance (ANOVA) and ordinal logistic regression (OLR) unless the data included substantial missing information. That is, those participants who did not respond to more than five items in the SILL or who did not indicate their gender were eliminated from data analysis. ANOVA was performed to investigate gender differences in scale means in learning strategy use before matching on score levels and OLR was used to detect DIF after matching on score levels. OLR has not been commonly used in the domain of L2 acquisition and teaching and is worth detailed description. In OLR, predictors are used to model the probability of observing a given level of an item

response. For DIF detection, predictors typically include the total score as an ability measure, a grouping variable (e.g. gender), and the interaction between ability and group. An item is identified as a DIF item if the latter two variables show a statistically significant improvement in the data-to-model fit beyond a model that includes only ability. A statistically significant group or interaction term signals uniform or non-uniform DIF, respectively. OLR DIF was employed because of practical advantages over other DIF methods for polytomous items. Specifically, OLR DIF (a) is less expensive to implement than IRT methods (e.g. sample size, software), and (b) requires less data manipulation compared to the polytomous variants of the Mantel-Haenszel procedure (Su & Wang, 2005). OLR DIF detection in this study follows the 6 subscales of the SILL and was applied separately to items related to each of the subscales. Purification of the matching criterion (i.e. total score) improves DIF detection (Su & Wang, 2005; Zumbo, 1999). As recommended by Holland and Thayer (1988), a two step purification process was employed (e.g., French & Maller, 2007). This included an iterative process that analyzed items for DIF to create a matching score that did not contain DIF items except for the studied item. Once this purified matching score was created the final analysis for a given item was conducted. See French and Maller for procedure details.

To classify a DIF item, the chi-square difference test was used to compare models. Variables were entered in the order suggested by Zumbo (1999): (a) total score (the conditioning variable); (b) gender (the grouping variable); and (c) the interaction term. Additionally, an ordinal R^2 value (McKelvey & Zavoina, 1975) associated with each step was used as the effect size measure. The sequential nature of this DIF process was used where comparisons were made between the models at step 3 vs. step 1, using a two-degree-of-freedom chi-square difference test simultaneously tests for uniform and non-uniform DIF. This simultaneous test allows the analyst to quickly determine which items exhibit DIF. As suggested by Zumbo (1999), the criteria of a significant 2-df $\chi^2_{difference}$ test between models ($p<0.01$) was employed. The $R^2_{difference}$ criterion was reported to aid interpretation. As there is no agreed

upon guideline in the literature for what is the most accurate level for this criterion (e.g., French & Maller, 2007; Hidalgo & Lopez-Pina, 2004), especially for polytomous items, it was not used to classify DIF items.

RESULTS

In order to investigate gender differences in the six learning strategy categories of the SILL (Version 7.0, see Appendix), one-way ANOVA was performed. As shown in Table 1, only memory strategies showed a statistically significant gender difference with males using more strategies compared to females [$F(1, 931)=4.600$, $p<.05$]. However, the effect size of the difference was small (d=0.144). No significant gender differences were found in other strategy categories with mean scores almost identical between the two groups and effect sizes small.

Table 1.
Gender Difference by ANOVA

Strategy Category	*Gender*	*Mean*	*Mean Difference*	*F*	*d*
Memory	Male	2.938	0.084	4.600*	0.144
	Female	2.854			
Cognitive	Male	2.793	0.057	2.381	0.103
	Female	2.736			
Compensation	Male	3.390	-0.004	0.009	-0.010
	Female	3.394			
Metacognitive	Male	2.895	0.036	0.651	0.056
	Female	2.859			
Affective	Male	2.768	-0.010	0.057	-0.016
	Female	2.778			
Social	Male	2.923	0.023	0.231	0.033
	Female	2.900			

Note: The probability level was set at .05.

Since ANOVA was calculated before matching on score levels, we were

not sure yet whether the group differences or non-differences of each strategy category in Table 1 were due to true learning strategy use or if the items that exhibit DIF across males and females assisted to create such situations. Thus, DIF analyses were performed to identify DIF on the SILL items across gender after matching on score levels by OLR using each of the scale scores of six strategy categories as a total score.

In order to detect DIF items in memory strategies, OLR was performed with the scale score of memory strategies as a total score to match on score levels. As shown in Table 2, four items (item 1, item 2, item 8 and item 9) out of the nine items in memory strategies showed DIF across gender with all the four items in favor of males. More specifically, two items such as item 8 "reviewing English lessons" and item 9 "remembering English words or phrases by remembering their location" showed uniform DIF, whereas the other two items such as item 1 "thinking of relationships between words" and item 2 "using new English words in a sentence" showed non-uniform DIF.

In the same way, OLR was performed to detect DIF items in the cognitive strategies with the scale score of cognitive strategies as a total score to match on score levels. Table 3 shows that five items out of the 11 items in cognitive strategies displayed DIF across gender. More specifically, item 10 "saying or writing new English words several times," item 13 "using the English words in different ways," item 14 "starting conversations in English," and item 23 "making summaries of information" displayed uniform DIF for males, whereas item 11 "trying to talk like native speakers" displayed non-uniform DIF for females.

Table 2.
DIF Across Gender by the Scale of Memory Strategies

Item	$\chi^2_{difference}$	$R^2_{difference}$	*Favors*	*DIF*
i1	***21.744***	***0.016***	***Male***	***Nonuniform***
i2	***25.503***	***0.018***	***Male***	***Nonuniform***
i3	3.288	0.002		
i4	2.250	0.001		
i5	2.866	0.001		

i6	6.404	0.004		
i7	1.549	0.001		
i8	***11.158***	***0.009***	***Male***	***Uniform***
i9	***14.49***	***0.014***	***Male***	***Uniform***

Note. Italicized and bold rows have significant $\chi^2_{difference}$ (2-*df*, $p < 0.01$). None of the items have effect size ($R^2_{difference}$) greater than or equal to 0.130.

Table 3
DIF Across Gender by the Scale of Cognitive Strategies

Item	$\chi^2_{difference}$	$R^2_{difference}$	*Favors*	*DIF*
i11	***10.322***	***0.007***	***Female***	***Nonuniform***
i12	2.260	0.001		
i13	***25.832***	***0.019***	***Male***	***Uniform***
i14	***16.797***	***0.013***	***Male***	***Uniform***
i15	1.991	0.001		
i16	3.652	0.002		
i17	1.802	0.001		
i18	3.920	0.002		
i19	2.761	0.002		
i20	4.811	0.002		
i21	8.845	0.007		
i22	7.745	0.008		
i23	***20.167***	***0.067***	***Male***	***Uniform***

Note. Italicized and bold rows have significant $\chi^2_{difference}$ (2-*df*, $p < 0.01$). None of the items have effect size ($R^2_{difference}$) greater than or equal to 0.130.

Table 4 and Table 5 show the results of DIF on the items of compensation strategies and metacognitive strategies across gender by OLR which used the scale scores of compensation strategies and metacognitive strategies as a total score, respectively. In Table 4, only item 27 “reading English without looking up every new word” exhibited uniform DIF favoring males out of the six items in compensation strategies. In Table 5, two items such as item 30 “trying to find many ways to use English” and item 31 “using mistakes to help learn better” exhibited uniform DIF favoring males out of the nine items in metacognitive strategies.

Table 4
DIF Across Gender by the Scale of Compensation Strategies

Item	$\chi^2_{difference}$	$R^2_{difference}$	*Favors*	*DIF*
i24	9.030	0.006		
i25	3.256	0.002		
i26	2.653	0.002		
i27	***12.552***	***0.011***	***Male***	***Uniform***
i28	0.595	0.001		
i29	0.557	0.001		

Note. Italicized and bold rows have significant $\chi^2_{difference}$ (2-*df*, $p < 0.01$). None of the items have effect size ($R^2_{difference}$) greater than or equal to 0.130.

Table 5
DIF Across Gender by the Scale of Metacognitive Strategies

Item	$\chi^2_{difference}$	$R^2_{difference}$	*Favors*	*DIF*
i30	***20.192***	***0.013***	***Male***	***Uniform***
i31	***16.317***	***0.012***	***Male***	***Uniform***
i32	1.764	0.001		
i33	7.970	0.004		
i34	0.045	0.001		
i35	3.917	0.002		
i36	4.297	0.001		
i37	0.092	0.001		
i38	2.788	0.002		

Note. Italicized and bold rows have significant $\chi^2_{difference}$ (2-*df*, $p < 0.01$). None of the items have effect size ($R^2_{difference}$) greater than or equal to 0.130.

Table 6 and Table 7 show DIF statistics across gender in affective strategies and social strategies by OLR which used the scale score of affective strategies and social strategies as a total score to match on score levels, respectively. In affective strategies, item 40 “encouraging oneself to speak English” showed uniform DIF in favor of males. In social strategies, no items showed significant χ^2 variance and no DIF in OLR after matching on score levels.

Table 6
DIF Across Gender by the Scale of Affective Strategies

Item	$\chi^2_{difference}$	$R^2_{difference}$	*Favors*	*DIF*
i39	1.168	0.001		
i40	***31.120***	***0.022***	***Male***	***Uniform***
i41	2.918	0.001		
i42	5.923	0.005		
i43	1.490	0.001		
i44	7.116	0.004		

Note. Italicized and bold rows have significant $\chi^2_{difference}$ (2-*df*, $p < 0.01$). None of the items have effect size ($R^2_{difference}$) greater than or equal to 0.130.

Table 7
DIF Across Gender by the Scale of Social Strategies

Item	$\chi^2_{difference}$	$R^2_{difference}$	*Favors*	*DIF*
i45	5.245	0.004		
i46	4.962	0.002		
i47	1.963	0.001		
i48	4.154	0.002		
i49	0.370	0.001		
i50	5.881	0.004		

Note. Italicized and bold rows have significant $\chi^2_{difference}$ (2-*df*, $p < 0.01$). None of the items have effect size ($R^2_{difference}$) greater than or equal to 0.130.

DISCUSSION

This study investigated gender differences in the six strategy categories of the SILL by performing ANOVA and OLR after matching on score levels. In general, the study supports the validity of the score inferences of the SILL because even though several items were found to display DIF across gender, the effect sizes of these differences were small.

More specifically, one of the findings was that even though effect size was small, males used memory strategies significantly more than females, supporting one study where the males in Kuwait used more learning

strategies compared to the females (Tercanlioglu, 2004). However, this study is in contrast with many other studies where the females in Puerto Rico (Green & Oxford, 1995), Japan (Mochizuki, 1999), Israel (Khalil, 2005), and Turkey (Kavasoğlu, 2009) reported using memory, cognitive, metacognitive, and social strategies more frequently than their counterparts.

Two reasons can be attributed to the mean differences in ANOVA in favor of males. First, males might truly use more learning strategies compared to females because there could be many variables affecting strategy use, including social context, proficiency level, and motivation for and attitude toward English learning (Oxford & Nyikos, 1989; Wharton, 2000). In terms of learning contexts, the university which the participants of the current study attended was different from other universities in EFL settings. Because there were many native English teachers and international students at the university, say about 6%, which introduced various programs such as English Village, Chinese Village, and English Zone designed to help Korean students to interact with native English teachers and international students. In a somewhat conservative society like Korea, males might have an advantage over females in inviting input, interacting with native English teachers and international students, and being acculturated to various cultures, leading to more learning strategy use by males (El-Dib, 2004; Schumann, 1986; Tercanlioglu, 2004). In proficiency level, taking into consideration that males were assumed to be more proficient than females and that learning strategy use was related to proficiency level, the former might have used more learning strategies than the latter at the time of data collection (Green & Oxford, 1995; Griffiths, 2003; Magogwe & Oliver, 2007; Park, 1997). In this regard, we point out that investigating the proficiency level of males and females must occur in the studies on the relation of learning strategy use to gender, a point not examined in most previous studies. Otherwise, the issue of gender in learning strategy use can be misleading and biased.

Second, the results might be due to the items in the SILL that exhibited DIF for males after matching on score levels (Cole et al., 1993; Thissen et al., 1986). That is, it was easier for males to endorse the items than females.

There were a total of 13 items that exhibited DIF with 12 items in favor of males and one item in favor of females. Among the 13 DIF items across gender, 10 items exhibited uniform DIF concerned with item difficulty/endorsement, whereas three items (item 1, item 2, and item 11) displayed non-uniform DIF concerned with item discrimination.

In the DIF items in strategy categories, cognitive strategies included five DIF items, consisting of the most DIF items out of the six strategy categories in the SILL. Since the current study is exploratory in nature and since no studies to date conducted DIF on the SILL across gender, explaining the underlying reasons for these findings and comparing these findings across studies are limited. However, as were the findings, male students in general had an unfair advantage over female students in cognitive strategies which were used to understand and produce new language by different ways. That is, given equal self-reported strategy use, males endorsed higher levels of strategy use compared to females. This was true especially with the items such as item 10 "saying or writing new English words several times," item 13 "using the English words in different ways," item 14 "starting conversations in English," and item 23 "making summaries of information." We speculate that males had an advantage over females in item 13 and item 14, which are related to language use, probably because of the specific English learning context of the university, as discussed above. For item 10 and item 23, taking into consideration that using these two learning strategies requires hard work, it might be that males are more motivated for improving English proficiency and differentially used these learning strategies compared to females (Oxford & Nyikos, 1989; Wharton, 2000). It is interesting to note that females, especially frequently strategy-using females, had an advantage over males in item 11 "trying to talk like native speakers" which exhibited non-uniform DIF. This finding is in accord with previous findings where females tended to learn a foreign language through imitation more than males (Brown, 2000).

Memory strategies included four DIF items in favor of males, consisting of the second most DIF items in the six strategy categories. This finding evidences that males had an advantage over females in storing and retrieving

new information in general and the following items in particular: item 1 "thinking of relationships between words," item 2 "using new English words in a sentence," item 8 "reviewing English lessons," and item 9 "remembering English words or phrases by remembering their location." Considering that three items, item 1, item 2, and item 9, are related to schema, males, especially frequently strategy-using males because of the two non-uniform DIF items (item 1 and item 2), took advantage of their schema in remembering words and idioms more than females. Worthy of notice is the crucial role of schema in memorizing words, L2 listening and reading comprehension, and creation (Anderson et al., 1977; Carrell & Eisterhold, 1983). For item 8, it is difficult to hypothesize the reason why it differentially functioned for males because from teaching experiences with these students females tended to review their lessons more sincerely and frequently than males to improve their grades. Further investigation is needed.

In general, the items in compensation, metacognitive, affective, and social strategies were relatively fair across gender with the limited number of DIF items in each strategy category. There was one DIF item in favor of males in compensation strategies which was related to using the language in spite of knowledge gaps: item 27 "reading English without looking up every new word." Considering this item, male students preferred to use guessing strategies more than female students, and any items on guessing strategies could differentially function for males.

In metacognitive strategies which are concerned with coordinating the learning process, two items exhibited DIF in favor of males: item 30 "trying to find many ways to use English" and item 31 "using mistakes to help learn better." In affective strategies which are related to regulating emotions in language learning, only item 40 "encouraging oneself to speak English" showed DIF flagging for males. Again, it is interesting to note that male students preferred to use English, taking advantage of making mistakes for their benefit and encouraging themselves to speak English. Thus, any items on language use for these students can be candidacy for DIF in favor of males.

No DIF items were found in social strategies which were used to learn

with other people. This result might be because several items in social strategies in the SILL including item 45, item 46, item 48, and item 49 are concerned with "asking for help." Considering the relatively independent nature of males, they might avoid the strategies related to "asking for help" which might offset the strategies related to "language use," leading to no DIF on the items in social strategies across gender. Caution should be exercised with item 47 "practicing with other students," which is concerned with language use but did not show DIF for males. Our best speculation is that the participants might interpret "practicing with other students" as practicing with other Korean students rather than practicing with international students. Taking into consideration that Korean students tended to avoid using English among themselves because of peer pressure, we can understand why this item did not display DIF in favor of males (Park, 2000).

CONCLUSION

This study investigated the mean differences and differential item functioning of the SILL by gender before and after matching on score/ability levels for university students in an EFL context. Several important findings were made, using ANOVA to compare mean differences and OLR to compare item functioning after matching on score levels.

One of the implications of these findings is that DIF should be investigated in order to understand the mean differences of learning strategy use by gender. Otherwise, we are not sure yet whether the mean differences between males and females are due to true differences in learning strategy use or due to DIF items. Nevertheless, most studies on learning strategy use in regards to gender to date make a conclusion on the basis of mean difference studies using ANOVA or t-tests, for instance, without in-depth assessing for DIF items. However, these conclusions are hasty and mislead the exact nature of group differences.

Another implication is that DIF items should be either revised or

eliminated because they do not function the same way across gender. However, it is worth noting that DIF items do not necessarily turn out to be biased and that item elimination can cause construct underrepresentation. Rather, it is recommended that review panels composed of content experts review DIF items identified by statistical analysis. Even though the findings by statistical analysis do not often match with the findings by review process, review process by content experts is a critical step to supplement statistical findings (AERA, APA, & NCME, 1999; Geranpayeh & Kunnan, 2007).

A third implication of this study is that in consideration of the previous findings that learning strategies are associated with L2 learning, teachers should encourage their students to use learning strategies inside and outside the classroom (McMullen, 2009; Park, 1997; Phakiti, 2003). What teachers should keep in mind in teaching learning strategies is that males use different learning strategies than females because several factors including social context and proficiency level underlie learning strategies. In addition, teachers should inform the students that why learning strategies, employed by learners to facilitate learning, work in L2 acquisition because informed learners will be sure to use more learning strategies than non-informed learners.

Regardless of the empirical findings and theoretical justifications of the current study, generalization of this study should be made with caution due to a few limitations. First, the sample was chosen from a population of EFL university students in Korea and generalizing the findings to other learning contexts will reveal limitation. Second, DIF analyses were carried out using an internal matching criterion, assuming the validity and fairness of the criterion without considering external matching criteria (Ferne & Rupp, 2007). This implies that any mean difference on the total scores is valid and reflects the underlying trait or ability distribution of the two groups being compared. Even with the purification process, this concern remains. However, the internal criterion is often the best choice for matching because (a) the total scores likely have high reliability, (b) validity evidence often has been provided for the scores, and (c) the scores are obtained under similar

conditions for all persons and remain the best option in DIF studies (Dorans & Holland, 1993). Third and last, the underlying reasons for the DIF items were by no means easy to explain because of the lack of data regarding review processes by content experts for DIF items, which was beyond the scope of this study.

Future studies need to be undertaken for more clear understanding of the scientific phenomena of DIF on the SILL across gender by different samples from ESL as well as EFL learning contexts, using various DIF detection methods. In addition, DIF items deserve to be reviewed by inviting review panels that consist of gender experts with an eye toward the combination of such information assisting the field to understand why these differences exist.

THE AUTHORS

Gi-Pyo Park is a Professor at Soonchunhyang University in Asan, Korea. He is interested in L2 acquisition, teaching, and testing. More specifically, his research areas of interest have included critical period in L2 acquisition, language learning strategies, anxiety, teaching listening and reading skills, corrective feedback, differential item functioning, and effective teachers.
Email address: gipyop@sch.ac.kr

Brian F. French is an Associate Professor at Washington State University in Pullman, Washington, USA. He is interested in educational and psychology test score validity issues. More specifically, his research area focuses on measurement invariance issues and advanced statistical modeling.
Email address: frenchb@wsu.edu

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D. C.: American Educational Research Association.

Anderson, R., Reynolds, R., Schallert, D., & Goetz, E. (1977). Frameworks for comprehending discourse. *Educational Research Journal, 14*, 367-381.

Breidenbach, D., & French, B. F. (2011). Ordinal logistic regression to detect differential item functioning for gender in the institution integration scale. *Journal of College Student Retention, 12,* 339-352.

Brown, H. D. (2000). Principles of language learning and teaching (5th ed.). White Plains, NY: Addison Wesley Longman.

Carrell, P., & Eisterhold, J. C. (1983). Schema theory and ESL reading pedagogy. *TESOL Quarterly*, *17*(4), 553-573.

Chen, M.-L. (2009). Influence of grade level on perceptual learning style preferences and language learning strategies of Taiwanese English as a foreign language learners. *Learning and Individual Differences, 19*, 304-308.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-47.

Cohen, A., & Macaro, E. (2007). Language learner strategies: Thirty years of research and practice. Oxford: Oxford University Press.

Cole, D. A., Maxwell, S. E., Avery, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114*, 174-184.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Ehrman, M., & Oxford, R. L. (1989). Effects of sex differences, career choice, and psychological type on adult language learning strategies. *Modern Language Journal, 73*, 1-13.

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing, 14*, 261-277.

El-Dib, M. A. (2004). Language learning strategies in Kuwait: Links to gender, language level, and culture in a hybrid context. *Foreign Language Annals, 37*, 85-95.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Finch, W. H., & French, B. F. (2010). Detecting differential item functioning of a course satisfaction instrument in the presence of multilevel data. *The Journal of the First-Year Experience and Students in Transition, 22*, 27-48.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for DIF detection. *Educational and Psychological Measurement, 67,* 373-393.

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese universities. *The Modern Language Journal, 88*, 229-243.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly, 4*, 190-222.

Green, J., & Oxford, R. (1995). A closer look at learner strategies, L2 proficiency, and gender. *TESOL Quarterly, 29*, 261-297.

Griffiths, C. (2003). Patterns of language learning strategy use. *System, 31*, 367-383.

Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64,* 903-915.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NG: Lawrence Erlbaum.

Hong-Nam, K., & Leavell, A. G. (2006). Language learning strategy use of ESL students in an intensive English learning context. *System, 34*, 399-415.

Hong-Nam, K., & Leavell, A. G. (2007). A comparative study of language learning strategy use in an EFL context: Monolingual Korean and bilingual Korean-Chinese university students. *Asia Pacific Education Review, 2007*, 71-88.

Kavasoğlu, M. (2009). Learning strategy use of pre-service teachers of English language at Mersin University. *Procedia Social and Behavioral Sciences, 1*, 993-997.

Khalil, A. (2005). Assessment of language learning strategies used by Palestinian EFL learners. *Foreign Language Annals, 38*, 108-117.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test.

Language Testing, 18, 89-114.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741-746.

Magogwe, J. M., & Oliver, R. (2007). The relationship between language learning strategies, proficiency, age and self-efficacy beliefs: A study of language learners in Botswana. *System, 35*, 338-352.

Maller, S. (2003). Best practices in detecting bias in nonverbal tests. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 23-47). Kluwer Academic/Plenum Publishers.

McKelvey, R. D., & Zavoina, L. (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology, 4,* 103-120.

McMullen, M. (2009). Using language learning strategies to improve the writing skills of Saudi EFL students: Will it really work? *System, 37*, 418-433.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education & Macmillan.

Mochizuki, A. (1999). Language learning strategies used by Japanese university students. *RELC Journal, 30*, 101-113.

Nisbet, D., Tindall, E., & Arroyo, A. (2005). Language learning strategies and English proficiency of Chinese university students. *Foreign Language Annals, 38*, 100-107.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition.* Cambridge: Cambridge University Press.

O'Malley, J. M., Chamot, A. U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*, 418-437.

Oxford, R. L. (1990). Language learning strategies: What every teacher should know. Boston: Heinle & Heinle.

Oxford, R. L., & Ehrman, M. (1995). Adult's language learning strategies in an intensive foreign language program in the United States. *System 23*, 359-386.

Oxford, R. L., & Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *Modern Language Journal, 73*, 291-300.

Oxford, R. L., Nyikos, M., & Ehrman, M. (1988). Vive la Différence? Reflections on sex differences in use of language learning strategies. *Foreign Language Annals, 21*, 321-329.

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

Park, C. (2000). Peer pressure and learning to speak English: Voices from the selected learners. *English Teaching, 55*, 231-268.

Park, G.-P. (1997). Language learning strategies and English proficiency in Korean

university students. *Foreign Language Annals, 30*, 211-221.

Phakiti, A. (2003). A close look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*, 26-56.

Radford, E. (1998). Why can't a woman be more like a man, or vice versa? In J. Radford, (Ed.), *Gender and choice in education and occupation* (pp. 174-191). London: Routledge.

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R Neuroticism scale. *Multivariate Behavioral Research, 36*, 83-110.

Rubin, J. (1975). What the "good language learner" can teach us. *TESOL Quarterly, 9*, 41-51.

Ryan, K., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*, 12-29.

Schumann, J. H. (1986). Research on the acculturation model for second language acquisition. *Journal of Multilingual and Multicultural Development, 7*, 379-392.

Skehan, P. (1989). Individual differences in second-language learning. London: Edward Arnold.

Stern, H. H. (1975). What can we learn from the good language learner? *Canadian Modern Language Review, 31*, 304-318.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18,* 313-350.

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

Tercanlioglu, L. (2004). Exploring gender effect on adult foreign language learning strategies. *Issues in Educational Research, 14*, 183-193.

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

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods*, *5*, 125-146.

Wharton, G. (2000). Language learning strategy use of bilingual foreign language learners in Singapore. *Language Learning, 50*, 203-243.

Yang, N. D. (1999). The relationship between EFL learners' beliefs and learning strategy use. *System, 27*, 515-535.

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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analysis? Implications for translating language tests. *Language Testing, 20*, 136-147.

Appendix

Strategy Inventory for Language Learning (SILL, Version 7.0)

Part A

1. I think of relationships between what I already know and new things I learn in English.
2. I use new English words in a sentence so I can remember them.
3. I connect the sound of a new English word and an image or picture of the word to help me remember the word.
4. I remember a new English word by making a mental picture of a situation in which the word might be used.
5. I use rhymes to remember new English words.
6. I use flashcards to remember new English words.
7. I physically act out new English words.
8. I review English lessons often.
9. I remember new English words or phrases by remembering their location on the page, on the board, or on a street sign.

Part B

10. I say or write new English words several times.
11. I try to talk like native English speakers.
12. I practice the sounds of English.
13. I use the English words I know in different ways.
14. I start conversations in English.
15. I watch English language TV shows spoken in English or go to movies

spoken in English.
16. I read for pleasure in English.
17. I write notes, messages, letters, or reports in English.
18. I first skim an English passage (read over the passage quickly), then go back and read carefully.
19. I look for words in my own language that are similar to new words in English.
20. I try to find patterns in English.
21. I find the meaning of an English word by dividing it into parts that I understand.
22. I try not to translate word-for-word.
23. I make summaries of information that I hear or read in English.

Part C

24. To understand unfamiliar English words, I make guesses.
25. When I can't think of a word during a conversation in English, I use gestures.
26. I make up new words if I do not know the right ones in English.
27. I read English without looking up every new word.
28. I try to guess what the other person will say next in English.
29. If I can't think of an English word, I use a word or phrase that means the same thing.

Part D

30. I try to find as many ways as I can to use my English.
31. I notice my English mistakes and use that information to help me do better.
32. I pay attention when someone is speaking English.
33. I try to find out how to be a better learner of English.
34. I plan my schedule so I will have enough time to study English.
35. I look for people I can talk to in English.
36. I look for opportunities to read as much as possible in English.

37. I have clear goals for improving my English skills.
38. I think about my progress in learning English.

Part E

39. I try to relax whenever I feel afraid of using English.
40. I encourage myself to speak English even when I am afraid of making a mistake.
41. I give myself a reward or treat when I do well in English.
42. I notice if I am tense or nervous when I am studying or using English.
43. I write down my feelings in a language learning diary.
44. I talk to someone else about how I feel when I am learning English.

Part F

45. If I do not understand something in English, I ask the other person to slow down or say it again.
46. I ask English speakers to correct me when I talk.
47. I practice English with other students.
48. I ask for help from English speakers.
49. I ask questions in English.
50. I try to learn about the culture of English speakers.