한국심리학회지 : 일반
Korean Journal of Psychology: General
2009, Vol. 28, No. 1, 263-281

Gender Differences on Multidimensional Personality Questionnaire(MPQ): A study of Differential Item Functioning(DIF) with Mantel-Haenszel Statistic

Jung Lee Chungnam University

Soon Mook Lee† Sungkyunkwan University

In the present study, we explored whether there are items that function differently across women and men on Tellegen's Multidimensional Personality Questionnaire(MPQ; Tellegen, 1982; Tellegen & Waller, In Press) by using Mantel-Haenszel(MH) statistic. For the purpose of the study, we applied MHITER program to a sample of responses to the 300-item MPQ personality test. The subjects were 300 individuals(137 men and 163 women, with a mean age of 39.72). The findings of the present study are as following. Firstly, some of the MPQ scales showed significant mean differences across gender; these were Social Potency, Social Closeness, Stress Reaction, Aggression, Harm Avoidance, and Absorption scales. Secondly, six MPQ scales showed more than three items functioning differently. Among others, the Traditionalism scale had the most DIF items, which is followed by Stress Reaction, Aggression, Social Potency, Achievement, and Harm Avoidance. In conclusion, there was a non-negligible number of DIF items across gender in MPQ scales, implying that there may be many DIF items in other personality tests that are currently used. The fact that men and women differently responded to some items reflect a socio-cultural impact in their perspectives on the personality items/tests.

Key words : DIF, personality test, gender difference, Mantel-Haenszel, MPQ

† Corresponding Author : Soon Mook Lee, Sungkyunkwan University, 53 Myeongnyun-dong 3-ga, Jongno-gu, Seoul 110-745, Korea
Tel: 02-2123-2445, E-mail: leex0694@hotmail.com, smlyhl@chol.com

Historically referred to as “Item Bias”, and now called “Differential Item Functioning(DIF)”, DIF refers to psychometric difference in how an item functions across two groups. In the present study, the more neutral term, differential item functioning is preferred to item bias since in many examples of items that exhibit DIF, the term “bias” does not accurately describe the situation(Ackerman, Gierl, & Walker, 2003; Holland & Thayer, 1988). DIF indicates a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by a test. The term “bias” would be justifiably used when the situation involves unfair results to one of the matched groups. If the test is free of DIF, there is no need to be concerned about the fairness of the test; however, if we detect DIF, it appears to be troublesome with its fairness. Therefore, we need to thoroughly examine the item, find out the specific possible reasons of bias, and eliminate the item to make the test more fair.

From a psycho-social perspective, it is important to find out whether a test is biased for or against certain groups such as women, African-Americans, disabled people, gay men and lesbians, and other social minority groups. Since in most cases, the groups that are the targets of DIF studies tend to be socially disadvantaged (i.e., non-English speakers, people in low SES, or women), test or item bias can mean unfairness; it could represent discrimination, prejudice, and inequality(Cole, 1993). Considering the fairness of testing as an essential element in all kinds of tests such as personality, intelligence, and other interest and aptitude tests, its importance and impact on the test results can not be over emphasized.

The present study, in this regard, tried to explore DIF in Multidimensional Personality Questionnaire(MPQ; Tellegen, 1982; Tellegen & Waller, In Press), one of the most well-known personality tests especially for its use in Minnesota twin studies(Bouchard, 1994; Tellegen et al., 1988). For its diverse use and continued work for validation over the past 20 years, it is surprising that the MPQ items were not yet fully subjected to an inquiry of the potentially differential functions over different genders. This study, in particular, approached the subject of evaluating DIF with more interest in gender differences than other factors since gender characterizes one of the most salient categorizations of human subjects.

DIF Studies in Personality Tests

Compared with achievement and scholastic aptitude tests, there have not been many DIF studies in personality tests. It is said that most DIF applications in the personality area have involved measures of attitudes such as job satisfaction, modernity, and individualism-collectivism rather than broad-band personality

inventories(Huang, Church, & Katigbak, 1997; Schnohr et al., 2008). Although the impact of personality tests is less salient than that of achievement tests in their influence on a person's choice of occupation or school, it would be difficult to understand human nature accurately without a thorough examination of gender, ethical, and socio-economic differences that frequently show up in various personality tests.

The present study aims to explore whether there are items that function differently across gender in Tellegen's Multidimensional Personality Questionnaire(MPQ), and if so, to investigate how they function in a different manner on each subdimension. The MPQ has been developed through an 'exploratory' approach in which deductive and empirical approaches are well incorporated through bidirectional inquiry "moving from ideas to data and vice versa"(Tellegen, 1985, p.262). MPQ is a self-report instrument that covers both the lower order trait and broader structural levels. Hence it has been extensively used to measure normal personality in recent years. It consists of eleven scales representing first-order personality dimensions and three second-order dimensions (Tellegen, 1982). The big three traits are positive affectivity, negative affectivity, and constraint. The eleven first-order dimensions are well-being(24 items), social potency(26 items), achievement(21 items), social closeness(22 items), stress reaction(26 items), alienation(20 items), aggression(20 items), control(24 items), harm avoidance(28 items), traditionalism(27 items), and absorption(34 items). It also has six validity scales and has been acclaimed for its high reliability and validity(Tellegen, 1982; Tellegen & Waller, In Press).

Among these eleven personality scales, some are consistently involved in a dispute of gender differences (Becker, 2006; Smith, 2002; Smith & Reise, 1998). Especially, males show higher scores on social potency, aggression, and traditionalism, and females show higher scores on social closeness, stress reaction, and harm avoidance scales(Tellegen, 1982).

Smith and Reise(1998), using an Item Response Theory(IRT) method, demonstrated a strong evidence of DIF in the Stress Reaction scale of the MPQ. Their results indicated that women were more likely to endorse items describing emotional vulnerability and sensitivity; whereas men were more likely to endorse items describing tension, irritability, and being easily upset. They used an item factor analysis program, TESTFACT(Wilson, Wood, & Gibbons, 1991), to derive five facets in the Stress Reaction scale. They argued that comparison of group mean differences can be affected by multidimensionality defined by item clusters that share similar content although they make up a unidimensional scale as a whole essentially. The five facets they came up with were: (1) *tends to worry*, (2) *nervous and tense,* (3) *sensitive and vulnerable*, (4) *easily upset and irritable,* and(5) *unaccountable mood changes*. These facets almost

exactly correspond to those reported by Tellegen(1982).

There is another DIF study supporting the potential interaction between gender and personality scales. Investigating the Eysenck Personality Questionnaire(EPQ), Francis(1993) insisted that its neuroticism scale is composed of two facets that would cause gender differences as a whole: sex-related and sex-free. Therefore, women tend to score higher on sex-related items, but they score no higher than men on sex-free items. While literature reports gender mean differences and DIF on "neuroticism" or "negative affectivity" scale, very few studies have been conducted at the item-level regarding other personality traits such as positive affectivity or openness to experience in MPQ.

In the present study, we are going to investigate DIF on all eleven scales of MPQ by using real data. In order to detect DIF between gender, we used a software MHITER(Kwak, 1997) adopting Mantel Haenszel statistical method. The aims of the present study are as following: (1) to review the Mantel-Haenszel statistic as a tool for psychologists who aim to assess DIF; (2) to detect differentially functioning items and to determine whether the items are for or against women; (3) to investigate how to interpret DIF on these personality scales, and to study the factors affecting DIF.

Mantel-Haenszel Statistical Method for DIF Studies

Research in the measurement community over the past decades has made available a variety of indices based on both classical and modern test theory for the detection of DIF. The choice of a statistic for a particular study should take into account two points: (1) the sample sizes and computing facilities available; (2) the importance of the decisions to be made, and hence, the precision required(Scheuneman & Bleistein, 1989). The literature on methods of revealing DIF generally suggests that the item response theory procedures are to be preferred, but problems with the implementation of IRT methods in psychology remain troublesome. The assumption that an IRT model fits the data cannot always be met(problem of unmet assumption) and parameters cannot always be equally well estimated for groups with different levels in a trait of interest(problem of nonequivalent precision). At the same time, a fairly large number of items and large number of testees(at least several hundreds) are necessary to do the item calibrations which is not a situation that psychologists entertain normally. The methods are also expensive and conceptually difficult to explain to a naive audience(Scheuneman & Bleistein, 1989). Compared with IRT methods, most of the classical methods are inexpensive to perform and require smaller sample sizes, which is relevant in the context of assessing non-

cognitive traits such as personality. Among other classical methods, the present study makes use of the Mantel-Haenszel(MH) statistic that has been widely applied in recent years.

Known as the most effective tool for detecting DIF, the Mantel-Haenszel statistical procedure (Mantel & Haenszel, 1959) was applied to DIF by Holland and Thayer(1986, 1988) as a tool for studying the functioning of test items in different groups of testees. This procedure is computationally simple and easy to implement for psychologists who are not motivated to challenge using IRT models. It is by far the most popular alternative to IRT methods for detecting item-level measurement bias despite its shortcoming of not detecting "non-uniform" DIF(Rogers & Swaminathan, 1993; Zwick, Thayer, & Wingersky, 1994). Uniform DIF refers to a case where one population group consistently has a better chance of answering an item correctly, regardless of their total score. In contrast , the chance is not the same across all score levels in a case of non-uniform DIF.

The MH statistic is an extension of the traditional chi-square approaches developed by Scheuneman(1979), and Marascuilo and Slaughter (1981). In IRT methods, it is said that items function differentially or there is a lack of measurement or metric equivalence, when two groups' item characteristic curves for a given item differ by more than sampling error (Hambleton et al., 1991; Thissen, Steinberg, & Wainer, 1988). However, the Mantel- Haenszel procedure compares the performance of two groups of subjects-the reference and focal groups-on all the items in a given test, one item at a time. The group designated as the *focal group(F)* is the group that is believed to be disadvantaged by the presence of DIF in the test. The group referred to as the *reference group(R)* is taken as a standard against which we will compare the performance of the focal group for the purpose of DIF detection. For instance, the focal group may be all male testees, while the reference group may consist of the female testees.

Since MH procedure is an outgrowth of previous X^2 procedures of analyzing contingency tables (Marascuilo & Slaughter, 1981; Mellenberg, 1982; Scheuneman, 1979), conting ency tables are prepared based on response data of subjects. However, before construction of contingency tables, test scores are divided into several intervals(3-5) to control for differences in subjects' levels of attribute measured(Scheuneman, 1979). Subjects whose scores are classified into a certain interval are assumed to be at an equivalent level of the attribute. And a contingency table of two dimensions is constructed for each interval(See Tale 1): one dimension for individual scores(1, 0) on a studied item, the other for reference and focal groups. These groups are matched in terms of test scores or level of the attribute measured. Thus, at each score level *k*, individual item data from the two groups of subjects can be arranged

Table 1. Mantel-Haenszel contingency

Group	Score on studied item		
	1	0	Total
Reference	f 1rk	f 0fk	n rk
Focal	f 1fk	f 0rk	n rk
Total	n 1k	n 0k	n k

as a 2 x 2 table(see Table 1).

Since all subjects in the reference and focal groups are matched in the level of attribute measured, difference of the proportion correct (scored 1 in Table) between two groups is attributed to the differential functioning of the studied item, that is, DIF. This inference is applied to all the test score intervals and integrated to a MH X^2 statistic finally. If there are s levels or intervals for subjects' test scores, we will have an s x 2 x 2 table for any given item. Hence, the following statistics are computed (Holland & Thayer, 1988).

$$MH\ X^2 = \frac{[|\sum_j A_j - \sum_j E(A_j)| - \frac{1}{2}]^2}{\sum_j var(A_j)}$$

$$E(A_j) = \frac{N_{rj} M_{1j}}{T_j}$$

$$var\ A_j = \frac{N_{rj} N_{fj} M_{1j} M_{0j}}{T^2_j (T_j - 1)}$$

where j is from 1,2, referring to an interval of test scores, A_j is an actual frequency, and $E(A_j)$ is an expected frequency of reference group members who scored 1 on the item. $E(A_j)$ is computed as a product of marginal frequencies on the corresponding row(N) and column(M) divided by total subjects(T_j) in a given interval j.

The MH X^2 follows a X^2 distribution with 1 degree of freedom. The null hypothesis is "No DIF", that is, proportions correct(assigned "1" in Table 1) of the reference group and the focal group are equivalent across all the intervals of test scores. The alternative hypothesis states that there is a difference of proportion correct between the reference group and the focal group in at least one of the s intervals. The difference between MH X^2 and the previous X^2 procedures are three folds: (1) MH X^2 is the most powerful unbiased test of H_0 versus H_1, (2) The procedure is not iterative, rendering simplicity of computation, and (3) An estimate of the amount of DIF is provided.

In regards to the comparison between MH X^2 and IRT procedures, Holland and Thayer(1986) criticize the conventional belief that IRT based approaches to DIF would be theoretically preferred over X^2 based procedures. Holland and Thayer state that the view is "not a very precise way of describing the situation"(p.19). Although IRT based procedures would be more powerful and efficient, it is normally possible in simulation studies when the items exactly follow the hypothesized models. In real situations, however, where data or items would not exactly

follow the IRT models, IRT approaches would not be the optimal. Also, we need at least several hundreds as sample size before we attempt to apply IRT approaches to DIF. Also we need enough number of items(e.g. more than 20 items) for each scale. These two conditions are not usually met in psychological data. So it is safe to say that MH X^2 procedure is preferred to previous X^2 procedures, but it has the power and efficiency of IRT approaches as the upper bound.

Method

The present study applied the Mantel-Haenszel statistic to detect DIF in a sample of responses to the MPQ. The main goal of this investigation is to determine whether there are items that function differentially across gender and to see if the patterns of any identified gender difference are substantively interpretable. The present study is meaningful in the sense that it applies the Mantel-Haenszel DIF technique to personality item responses and also includes the analysis of all 11 scales of the MPQ.

Participants

Participants of the present study were 300 individuals(137 males and 163 females) who were administered the entire MPQ. Participants were drawn from the Minnesota Twin Registry (Lykken, Bouchard, McGue, & Tellegen, 2000). After completing informed consent, all the participants submitted the MPQ. There was no missing data.

Instrument

The 300-item MPQ is a self-report inventory that measures 11 domains and three higher-order super factors; alpha coefficients range from .76 to .89, with a median of .85(Tellegen, 1982, 1985). Besides 11 personality scales, MPQ includes six validity scales. MPQ was developed through an exploratory process that resulted in the construction of 11 primary(lower-order) scales measuring well-being, social potency, achievement, social closeness, stress reaction, alienation, aggression, control, harm avoidance, traditionalism, and absorption(Tellegen, 1982). Three higher-order factors emerged from factor analyses of the 11 primary scales, which are termed positive emotionality, negative emotionality, and constraint. The scales of well-being, social potency, social closeness, and achievement represent a higher-order factor "positive emotionality", which has clear features of a trait, extraversion. High scorers on these scales represent themselves as being engaged in active, pleasurable, and efficacious transactions with their environment and as being ready to experience the positive emotions congruent with these involvements. Primary scales such as stress reaction, alienation, and aggression are associated with negative emotionality. High scorers on

these scales describe themselves as being unpleasurably engaged, stressed, and prone to experiencing strong negative emotions such as anxiety and anger. The scales of control, harm avoidance, and traditionalism represent a higher-order constraint scale, in which high scorers represent being restrained, cautious, deferential, conventional, avoiding dangerous kinds of excitement and thrills.

Procedures

All the items in the MPQ test are dichotomous(true/false). The true/false responses were recoded into 1/0 so that we can recognize the high scores on each scale indicating a strong tendency of each trait. Computer program MHITER(Kwak, 1997) implementing Mantel-Haenszel statistic was run on each scale separately since the numbers of items are different from one another except for the scales of well-being, control, social potency, and stress reaction. For the analyses of the present study, we set the type 1 error level at the conventional level .05.

Results

Descriptive Statistics

First of all, we computed the basic descriptive statistics of the sample. The total number was 300(137 men and 163 women), and the mean age was 39.72. As can be seen from Table 2, the reliabilities (coefficient α) of each scale are relatively high, indicating that items are highly correlated with one another. Comparing the means of males and females, it was found that the discrepancies between men and women were very close to those reported in Tellegen(1982). Consistent with Tellegen's report, several scales showed significant mean differences across gender; these are social potency, social closeness, stress reaction, aggression, harm avoidance, and absorption scales. The scales of achievement, alienation, control, and traditionalism showed some mean differences between males and females, although the differences were not statistically significant. Of note, though, these group mean differences exhibited in Table 2 do not indicate whether these differences are real or caused by DIF at item-level.

DIF analyses

As Table 3 manifests, six scales out of 11 showed more than three DIF items. Among others, the scale traditionalism needs more attention in that it has six items that are functioning differentially across gender. Following traditionalism, stress reaction and aggression scales are revealed to have four DIF items each, and three scales such as social potency, achievement, and harm avoidance manifested three differentially functioning items. The rest of the scales, well-being, social closeness, alienation, control, and absorption are shown to have few DIF items, if any.

Table 2. MPQ Scale means, standard deviations, and reliabilities

Scale	Males(N=137)			Females(N=163)		
	Mean	SD	α	Mean	SD	α
Well-being	18.81	5.08	.89	18.63	4.67	.86
Social Potency *	10.98	6.70	.91	7.70	6.08	.90
Achievement	13.09	4.80	.84	12.15	4.56	.82
Social Closeness *	13.61	5.03	.86	15.50	4.37	.82
Stress Reaction *	9.90	6.11	.88	12.07	6.46	.89
Alienation	3.33	4.24	.90	2.53	3.23	.84
Aggression	5.15	3.60	.77	3.35	2.79	.72
Control	16.27	4.23	.77	16.75	4.86	.84
Harm Avoidance *	17.94	5.89	.86	22.30	4.53	.82
Traditionalism	19.48	4.20	.75	19.23	4.99	.82
Absorption *	15.01	6.48	.85	17.72	7.17	.88

* $p < .05$

Table 3. Number of DIF and non-DIF items for each scale of MPQ

MPQ Scales	Number of non-DIF items	Number of DIF items
Well-being	23	1
Social Potency	23	3
Achievement	18	3
Social Closeness	21	1
Stress Reaction	22	4
Alienation	20	0
Aggression	16	4
Control	22	2
Harm Avoidance	25	3
Traditionalism	21	6
Absorption	32	2

The DIF items of each scale are listed in Table 4. On the rightmost column of the Table 4, we put the log-odds-ratios of each item. When the log-odds-ratio is below zero, the item favors women. On the other hand, when the ratio is above zero, the item favors men. We see a mixture of items favoring women or men in the same scale as in the Table 4.

For instance, in the social potency scale, the item "I often monopolize conversations" was more frequently responded by females, on the other hand, the item "I would enjoy being a powerful executive or a politician" by males. Likewise, the scale of stress reaction showed a mixture of items favoring women or men. Women endorsed the items such as "My feelings

Table 4. MPQ scales, DIF items of each scale and their log-odds-ratio

MPQ Scales	*DIF Items*	*Log-odds-ratio*	
Well-being	1. I am usually light-hearted.	-.98059	**F**
Social Potency	1. I often monopolize conversations.	-1.0183	**F**
	2. I would enjoy being a powerful executive or politician.	1.5817	**M**
	3. On social occasions, I don't particularly care to "run the show." (R)	-1.4864	**F**
Achievement	1. I see no point in sticking with a problem if success is unlikely. (R)	-.87301	**F**
	2. I like to try difficult things.	.77339	**M**
	3. I don't like to do more than is really necessary in my work. (R)	-1.5284	**F**
Social Closeness	1. Often I go a whole morning without wanting to speak to anyone. (R)	.88701	**M**
Stress Reaction	1. My feelings are hurt rather easily.	-.82816	**F**
	2. I am easily startled by things that happen unexpectedly.	-.86672	**F**
	3. Minor setbacks sometimes irritate me too much.	.88491	**M**
	4. If I have a humiliating experience, I get over it very quickly. (R)	-.65113	**F**
Aggression	1. I enjoy violent movies.	2.1058	**M**
	2. When I have to stand in line, I never try to get ahead of others. (R)	-1.0516	**F**
	3. I like to watch a good, vicious fight.	2.7624	**M**
	4. Sometimes I just like to hit someone.	-1.5740	**F**
Control	1. I almost never do anything reckless.	-.90032	**F**
	2. I am a cautious person.	-1.2214	**F**
Harm Avoidance	1. I might enjoy riding in an open elevator to the top of a tall building under construction. (R)	-1.0895	**F**
	2. I would not enjoy fighting a forest fire.	-.93817	**F**
	3. Of the following situations I would like least: (a) Being in a flood, (b) Carrying a ton of coal from the backyard into the basement.	.93713	**M**
Traditionalism	1. I would be very embarrassed to tell people that I had spent my vacation at a nudist camp.	-1.1548	**F**
	2. More censorship of books and movies is a violation of free speech and should be abolished. (R)	-1.2651	**F**
	3. I am disgusted by foul language.	-1.3815	**F**
	4. Of the following statements I agree more with: (a) If a boy 6- or 7-years old lies or steals, he should be punished severely, (b) Lying and stealing aren't very serious in boys aged 6 or 7.	.79772	**M**
	5. Of the following two statements I agree more with: (a) Parents should ignore it when small children use naughty words, (b) Parents should punish small children when they use naughty words. (R)	1.0918	**M**
	6. High moral standards are the most important thing parents can teach their children.	1.1508	**M**
Absorption	1. Textures-such as wool, sand, wood - sometimes remind me of colors or music.	-.73563	**F**
	2. My thoughts often don't occur as words but as visual images.	1.0885	**M**

Note. (R) indicates that the reverse score is true. F stands for the item that favors females, and M stands for the item that favors males.

are hurt rather easily", "I am easily startled by things that happen unexpectedly", and "If I have a humiliating experience, I get over it very quickly(R)"; men more frequently endorsed the item "Minor setbacks sometimes irritate me too much." This finding strongly confirms the results of DIF analyses of Smith and Reise(1998) where they used the IRT method. In their study on a stress reaction scale, they reported DIF items across gender: women were more likely to endorse items describing emotional vulnerability and sensitivity; men were more likely to favor items indicating tension, irritability, and being easily upset.

The scales, aggression and traditionalism are also interesting in Table 4. The items such as "I enjoy violent movies" and "I like to watch a good, vicious fight" which traditionally represent typical masculine characteristics were endorsed by men. On the other hand, women tended to endorse the items, "Sometimes I just like to hit someone" and "When I have to stand in line, I never try to get ahead of others(R)." There is a similar finding in the traditionalism scale, that is, women were inclined to endorse the items such as "disgusted by foul language", opinion supporting "censorship of books and movies" and "very embarrassed to tell people that I had spent vacation at a nudist camp." These are also traditionally feminine characteristics; shy, soft, not violent, warm, nurturing and so on. In contrast, men greatly favored the items suggesting "high moral standards of parents' way of bringing up their children" and "strict punishment towards their children."

Table 5. Comparison of DIF items in Stress Reaction Scale between MH method and IRT method

Items	MH log-odds-ratio		IRT Adjusted *b* difference	
Feeings hurt easily	-.82816	**F**	-.65	**F**
Easily startled	-.86672	**F**	-.56	**F**
Easily rattled	NS		-.54	**F**
Gets over humiliation easily (R)	-.65113	**F**	-.53	**F**
Minor setbacks irritate	.88491	**M**	.35	**M**
Often nervous	NS		.35	**M**
Often irritated	NS		.36	**M**
Moody	NS		.36	**M**
Is tense	NS		.39	**M**

Note. (R) indicates that the reverse score is true. F stands for the item that favors females, and M stands for the item that favors males.

Discussion

DIF Items in MPQ

The present study suggests that there are many non-negligible number of DIF items in MPQ scales. Especially, the scales such as traditionalism, aggression, stress reaction, social potency, and harm avoidance showed more than three DIF items, which is very noticeable. It means that some items on the MPQ scale are relatively easier for women to endorse and some are relatively easier for men to endorse.

Among the scales containing DIF items, there are several noteworthy scales: traditionalism with the most DIF items, stress reaction and aggression, both ending up with four DIF items. The nature of the DIF items in these scales will be discussed. The traditionalism scale directs our attention mostly. As can be seen in Table 4, each gender tended to endorse a different subset of three items out of the total six. Whereas few mean differences were found in Table 2 on the scale of traditionalism, interestingly enough, Table 4 reveals that there are significant differences in item-level response patterns on the same scale. Women, in particular, expressed feelings of shame and embarrassment when talking about a nudist camp(item 1), being disgusted by use of foul language(item 3), and favorable attitudes toward censorship of books and movies(item 2). On the other hand, men showed stronger adherence to traditional values and standards. For example, items asking punishment of children regarding lying and stealing(item 4) and using naughty words(item 5) were highly endorsed by men. At the same time, men also have a tendency to prefer high moral standards(item 6) compared with women.

The results of the present study on the stress reaction scale are consistent with those of Smith and Reise ' study(1998) which was conducted with an IRT method. According to their study, "women were more likely to endorse items describing emotional vulnerability and sensitivity, whereas men were more likely to endorse items describing tension, irritability, and being easily upset."(Smith & Reise, 1998, p.1359). The results of the present study agree with them in that items such as "hurt rather easily"(item 1), "easily startled"(item 2), and "hard to get over a humiliating experience"(item 4) were strongly favored by women, and the item "minor setbacks sometimes irritate me" was highly favored by men(Refer to Table 4 & 5). As illustrated in Table 5, however, Smith and Reise' study demonstrated more DIF items on Stress Reaction Scale than the present study. At any event, though, the items which are endorsed by women have a tendency to reflect emotional vulnerability and sensitivity in situations that involve self-evaluation. On the other hand, items which are endorsed by men tend to reflect the general experience of nervous tension, unexplainable moodiness, irritation, frustration, and being on-edge.

In relation to aggression scale, four items out

of 20 manifested DIF. Among these four, two items were more endorsed by men and the other two were more endorsed by women. As indicated in Table 4, men tended to favor items such as "enjoy violent movies"(item 1) and "like to watch a good, vicious fight"(item 3), reflecting some degree of direct violent tendencies. On the other hand, women tended to favor items which indirectly show aggressiveness in the form of impatience(item 3) and uncontrollable hot temper(item 4). This finding seems to be consistent with Tellegen and Waller's(2008) argument for the five facets of aggression scale in MPQ in that the aspects of responses men and women exhibited a bit of directional difference. In other words, item-level interpretation of the aggression scale needs a caution, even though the scale aggression may be generally seen as masculine.

Factors affecting DIF

The phenomena of differential functioning by items in those three noteworthy scales can be investigated in terms of potential factors affecting DIF. We observed the particular preference towards certain items by both women and men in the aforementioned scales. This could be traced back to biological evolution in human beings(Buss, 1991). However, we are more interested in looking into potential relationship with sociocultural factors since personality scales are used more for inquiring social and adaptive behaviors in modern days(Houtman, 1990). In terms of behavioral ideology, there are several traits differentially recommended for men and women respectively. Drawing upon Horney's(1930/1967) position that the more powerful side between women and men will generate an ideology facilitating to maintain the different pattern of behaviors, Smith and Riese(1998) showed sociocultural explanations for DIF in personality scales.

Recently, Huang and his colleagues(1997) conducted a DIF study on the NEO-PI personality tests. They explored cultural explanations of DIF on the inventory. Their results demonstrated that women, as compared to men, saw themselves as more soft-hearted, warm, cheerful, and emotionally responsive, but also as more anxiety-prone, stress reactive, and self-conscious. They also reported different responses of subjects across western and oriental cultures. For instance, Filipino students, as compared to American students, saw themselves as being more easily rattled or vulnerable, as placing more value on aesthetic experiences, and as being less outgoing, talkative, and affectionate (Huang et al., 1997).

The tendency of different rate in endorsing particular items in traditionalism would reflect different ways in which women and men feel about traditional values and define them implicitly. When we consider the parental methods of raising children, we can easily discover the different perspectives of mothers and fathers. Generally and universally, fathers favor

stricter discipline including physical punishment and mothers tend to tolerate mischievousness and misbehavior of children. In addition to the difference of role-playing at home, women are discouraged to use foul language and unsociable behaviors at work, social organizations, and wherever they are. However, due to increased participation of women in education and employment, women become less traditional(Mc Broom, 1986), although men's tarditional values are not significantly affected(Madden, Alee, & Smith, 1989). Further research, therefore, needs to be done in analyses of traditionalism scale or its items by taking potentially moderating factors into consideration.

Observations of DIF in stress reaction scale and aggression scale may be explained in a similar sociocultural context of being endorsed or not by either women or men. Historically, women have been brought up and educated with the ideology of a patriarchal social system. In male-dominant societies, women must have good feminine characteristics such as beauty, cleanness, shyness, politeness, soft mind, relationship-oriented, sensitiveness and so on. The traits that are typically regarded as masculine such as irritable responses to stress, or aggressive, tough, violent, assertive, bold, and achievement-oriented behaviors have not been acceptable for women(Smith & Riese, 1998). Nowadays, even with the influence of the feminist movements, there are several traits highly recommended for women and men respectively. Therefore, this sociocultural trend may have influenced different item responses and mean differences in the scores of stress reaction scale and aggression scale.

How to deal with DIF Items?

Basically there are three different ways in dealing with DIF items (Smith, 2002; Smith & Riese, 1998): discarding DIF items, constructing separate tests for different subgroups, retaining DIF items. First, when there are not many DIF items, they can be discarded without suffering much loss of reliability and criterion validity. It is desirable to discard the troubling items to avoid the difficulty of locating the respondents on a common scale, when respondents are from different groups to which items are functioning differentially.

However, there is a reservation in discarding DIF items. On one hand, DIF items may be useful for theoretical purposes. In developing a test we would have a goal of learning or developing a construct through the scale construction(Tellegen & Waller, 2008). Tellegen and Waller argue that test construction can be "exploratory" in a sense that a developer may be "moving from ideas to data and vice versa"(p.262). That approach points out potential contributions of the DIF item to teaching us new aspects of the construct of interest. If these contributing items are completely discarded because of their differential functioning, we might lose sight over the complexity of the

construct that might be theoretically interesting. On the other hand, discarding DIF items raises a thorny question where to stop. We usually analyse the data again after discarding DIF items. Then there could come out another DIF items. After discarding these DIF items, we follow the same procedure. There is no clear guideline where we should stop this continuous procedure practically. So discarding does not resolve the problems, but may create more problems.

Constructing separate tests is the second option to choose in dealing with DIF items. Once all the DIF items are removed, the test seems to be most appropriate. This scale is appropriate for "purposes of prediction or case evaluation rather than comparing across groups"(Smith, 2002, p.761). Removing all DIF items leads to development of many tests that seem to be well matching with particular groups, one test for each partitioning of the testee population. However, separates tests tend to make comparisons between groups or between constructs in a group difficult, if not possible since each test covers a relatively narrow aspect of a construct. If comparisons are to be made, scale scores should be linked or equated so that substantive interpretations can be attempted across scales. Considering time and expenses for linking or similar work, developing separate tests is not ideal. However detailed scales are developed, there could still be individuals who may respond in an inconsistent pattern that is expected from the group membership. So developing separate tests does not provide a resolution to the problems. Smith(2002) argues that it "can pose more problems than it solves"(p.761).

Final option is the retention of DIF items. By this option the complexity of the construct is respected, which is an advantage. With this option we can cluster non-DIF items, male-directed DIF items, and female-directed DIF items. Scores from each cluster can be obtained and given to individuals. If scores are computed including DIF items, means and variances of scores would vary across groups. So different norms(for norm-referenced tests) or criterion scores(for domain-or criterion- referenced tests) should be prepared for different groups. The disadvantage of this option is the difficulty in scoring and interpretation.

In terms of MPQ, since there are not many DIF items, it will be desirable to discard DIF items. At this point, however, information on which items should be discarded and which items should be retained is not enough from the results of the present study. To do this job, we need to have thorough investigation with a bigger sample size. Systematic DIF analysis with various cultural groups will be needed as well.

In summary, the contribution of the present study lies in systematic and holistic analyses of item-level DIF in MPQ, including whole 11 scales. Even if there have been studies on the DIF analysis in the personality area, the majority

was limited to analysis of just one subscale or a few scales of the test. There are, however, limitations of the present study, which might have been made up for; (1) small sample size, (2) inability to detect non-uniform DIF, (3) use of only one method, Mantel-Haenszel statistic, for the analysis. Further research in this field may include non-uniform DIF analysis with different analytic methods, factor analyzing the 11 scales of MPQ and examine if there is another dimension or factor that influences DIF. By using the same MPQ, application of DIF analysis to different cultural populations such as Koreans might have different results, which may be another area of further investigation.

Rerences

Ackerman, T. A., Gierl, M. J., & Walker, C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22,* 37-53.

Becker, G. (2006). NEO-FFI scores in college men and women: A view from McDonald's unified treatment of test theory. *Journal of Research in Personality, 40,* 911-941.

Bouchard, T. J. Jr. (1994). Genes, environment, and personality, *Science*, 264(5166), 1700-1.

Buss, D. (1991). Evolutionary personality psychology. *Annual Review of Psychology, 42,* 459-491.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 25-29). Hillsdale NJ: Lawrence Erlbaum Associates.

Francis, L. J. (1993). The dual nature of the Eysenckian neuroticism scales: a question of sex differences? *Personality and Individual Differences, 15(1),* 43-59.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Holland, P. W. & Thayer, D. T. (1986). *Differential item performance and Mantel- Haenszel procedure.* A paper presented at the Annual Meeting of the American Educational Research Association(67th, San Francisco, CA. April 16-20, 1986).

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

Horney, K. (1967). The distrust between the sexes. In H. Kelman (Ed.) *Feminine psychology* (pp.107-118). New York: Norton. (Original work published 1930).

Houtman, I. (1990). Personal coping resources and sex differences. *Personality and Individual Differences, 11,* 53-63.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: differential item functioning in the NEO personality inventory. *Journal of Cross-cultural Psychology, 28(2),* 192-218.

Kwak, N. H. (1997). *MHITER version 1.0.* University of Minnesota. Minneapolis, MN.

Lykken, D. T., Bouchard, T. J. Jr., McGue, M., & Tellegen, A. (2000). The Minnesota Twin Family Registry: some initial findings. *Acta Geneticae Medicae et Gemellologicae, 39*, 35-70.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Marascuilo, L. A. & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on χ^2 statistics. *Journal of Educational measurement, 18*, 229-248.

Madden, M. E., Allee, M. & Smith, K. (1989). *Jealousy, gender, sex roles, and dependency.* Paper presented at the Eastern Psychological Association, Boston, March, 1989.

McBroom, W. H. (1986). Changes in role orientations of women: A study of sex role traditionalism over a five-year period. *Journal of Family Issues,* 7, 149-159.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics,* 7, 105-118.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17(2),* 105-116.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education, 2(3),* 255-275.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16(3),* 143-152.

Schnohr, C. W., Kreiner, S., Due, E. P., Currie, C., Boyee, W., & Diderichsen, F. (2008). Differential Item Functioning of a family affluence scale: validation study on data from HBSCC 2001/02. *Social Indicators Research, 89,* 79-95.

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin, 28,* 754-763.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: an IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75(5),* 1350-1362.

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota.

Tellegen, A. (1985). Structure of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma and J. D. Maser(eds.) *Anxiety and the Anxiety Disorders*. Hillsdale, NJ: LEA, pp.681-706.

Tellegen, A., Lykken, D. T., Bouchard, T. J. Jr., Wilcox, K., Segal, N. L., & Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology, 54(6)*, 1031-1039.

Tellegen, A. & Waller, N. G. (2008). Exploring personality through test construction: development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G.

Matthews, & D. H. Saklofske(eds.) *The SAGE Handbook of Personality Theory and Assessment, Vol 2.* (pp.261-292). Sage Publications Ltd.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun(Eds.), *Test validity*(pp.147-169). Hillsdale NJ: Erlbaum.

Wilson, D., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121-140.

1차원고접수 : 2009. 4. 20.
수정원고접수 : 2009. 5. 21.
최종게재결정 : 2009. 6. 7.

한국심리학회지 : 일반
Korean Journal of Psychology: General
2009, Vol. 28, No. 1, 263-281

다면적 성격검사 척도 반응의 성차 분석: 맨틀-핸젤(Mantel-Haenszel) 통계방법을 통해 본 차별적 문항 기능 (DIF)

이 정
충남대학교 심리학과

이 순 묵
성균관대학교 심리학과

본 연구에서는 만텔-헨젤(Mantel-Haenszel; MH) 방식을 사용하여 Tellegen의 다면적 성격검사(Multidimensional Personality Questionnaire; MPQ) 척도들에 대하여 성별에 따라 다르게 기능하는 문항들이 있는지를 조사하였다. 본 연구의 목적을 위하여, 300 문항의 MPQ 검사 반응 표본에 MHITER 프로그램을 적용하였다. 본 연구의 참여자는 300명의 성인들로(남: 137, 여: 163), 평균연령은 39.72 세였다. 분석결과 다음과 같은 결과가 산출되었다. 우선, 11개의 MPQ 척도 중, 사회적 영향력, 사회적 친밀, 스트레스 반응, 공격성, 위해 회피, 몰두 척도들은 점수평균에 있어 유의미한 남녀차를 나타냈다. 차별적 문항기능(DIF) 분석에 있어서는, 11개 척도들 중 6개의 척도가 세 개 이상의 DIF 문항을 보유하고 있음이 드러났다. 이 중, 가장 많은 DIF 문항을 보여준 척도는 전통주의였고, 그 뒤를 이어 스트레스 반응, 공격성, 사회적 영향력, 성취, 그리고 위해 회피 척도들에서 DIF 문항들이 나타났다. 본 연구의 결과는 MPQ 성격검사에 있어 간과될 수 없는 숫자의 DIF 문항들이 존재함을 제시하며, 또한 다른 성격 검사들에도 많은 DIF 문항들이 있을 수 있음을 암시한다. 남성과 여성이 어떠한 문항들에 다르게 반응한다는 사실은 여러 가지로 해석될 수 있으나, 사회/문화적 관점에서 보는 시각이 주로 언급되었다.

주요어 : 차별적 문항기능, 성격 검사, 성차, 만텔-헨젤(Mantel-Haenszel), 텔레겐 성격검사