DETECTING INACCURATE RESPONSE PATTERNS IN KOREAN MILITARY PERSONALITY INVENTORY: AN APPLICATION OF ITEM RESPONSE THEORY

A Thesis

by

SEUNGHWA HONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Myeongsun Yoon
Co-Chair of Committee,	Robert Hall
Committee Member,	Victor Willson
Head of Department,	Victor Willson

August 2013

Major Subject: Educational Psychology

ABSTRACT

There are concerns regarding the risk of the inaccurate responses in the personality data. The inaccurate responses negatively affect in the individual selection contexts. Especially, in the military context, the personality score including inaccurate responses results in the selection of inappropriate personnel or allows enlistment dodgers to avoid their military duty. This study conducted IRT-based person-fit analysis with the dichotomous military dataset in the Korean Military Personality Inventory. In order for that, 2PL model was applied for the data and person-fit index l_z was used to detect aberrant respondents. Based on l_z values of each respondent, potentially inaccurate respondents was identified. In diagnosing possible sources of aberrant response patterns, PRCs was assessed. This study with the military empirical data shows that person-fit analysis using l_z is applicable and practical method for detecting inaccurate response patterns in the personnel selection contexts based on the personality measurement.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Personality measures have been increasingly used not only in psychological and educational contexts but also in organizational applications. A number of studies and meta-analytic research have incited more use of personality measures in personnel selection contexts. In fact, several personality traits such as conscientiousness, anxiety, emotional stability, nondelinquency, and extraversion have shown significant validity in predicting job-relevant variables such as success and effectiveness in performance, destructive behaviors (e.g. absenteeism and substance abuse), and conflict or violence at work (e.g., Barrick & Mount, 1991; Bernardin, 1977; Sparks, 1983; Tett, Jackson, & Rothstein, 1991).

However, there are concerns regarding the risk of inaccurate or aberrant responses in personality data since people might not respond honestly to personal or sensitive questions in even well-developed personality measures. It is also possible that examinees unknowingly answer questions in the fashion of poorly representing themselves. Much literature has argued that inaccurate responses in personality measures could occur by several reasons including social desirability, poor motivation, deliberate faking, or alignment error, etc (Hulin, Drasow, & Parsons, 1983; Reise, 1995; Schmitt, Cortina, & Whitney, 1993). For example, when the personality measures are used in an employment-selection, examinees might distort their answers on items in a socially desirable way that, they believe, allows them to have a better chance for selection. In

another case, some examinees may randomly or carelessly respond to items that they are not interested in when personality scores have little effect on their own interests.

The existence of those responses adversely influences not only individuals but also organizations. As for individuals, underestimation of their abilities because of inaccurate responses to some items might result in slim chances of getting a desirable job. For an organization, overestimation of examinees' true ability may have the organization spend more time and materials in order to train recruits who lack in their abilities or look for other satisfactory employees (Schmitt et al., 1999).

Inaccurate responses may raise more serious issues in the military personnel selection context. Military personality measures have been used to identify problematic personnel in the stage of recruitment (Choi et al., 2009). The failure in identifying people who have current or potential problems in personality might cause serious problems such as the demoralization and frequent conflicts in barrack life (Choi et al., 2009). The research using U.S. Army enlisted personnel provided evidence that the lack of traits in emotional stability and nondelinquency was highly related to the drop-out rate for their respective service term (White, Nord, Mael, & Young, 1993). Moreover, military personnel have a high chance of experiencing threatening conditions such as participation in war or deployment in conflict areas so that both mental strength and personality soundness are required for military personnel.

Although inaccurate responses in a faking-good manner is more frequent in a personnel selection context, deliberate faking-bad may be another problematic response behavior in the military (Carroll, 2003). For example, in some countries (e.g., Germany,

Turkey, and Israel) using the mandatory military service system (Pfaffenzeller, 2010), some draftees may try to avoid their military duty by malingering their mental illness or maladjustment on personality measures (Jones, Hyams, & Wessely, 2003; Lande & Williams, 2013). Thus, individuals' deliberate faking-bad behavior in military personality measures need to be given significant consideration, as well.

Traditionally, in order to identify inaccurate responses, several detection scales (e.g., Lie scale and Social Desirability scale) have been included in personality measures (Ferrando & Chico, 2001; Zickar & Drasgow, 1996). Some researchers (e.g., Drasgow, Levine, & McLaughlin, 1987; Meijer & Sitsma, 2001; Reise & Waller, 1993) have proposed that individuals' inaccurate responses also can be statistically identified by analyzing psychometric properties of parameters in the context of item response theory (IRT) (Meijer, Egberink, Emons, & Sijtsma, 2008). Psychological constructs are intrinsically latent so that they cannot be directly measured but be estimated through analyzing, in the given applied IRT models, the responses to a set of items designed to test the constructs (Meijer, 1997). Note that the estimation might be only valid when individuals accurately answer questions hence the responses properly represent their latent trait (Reige & Waller, 1993; Meijer, 1997). Usually, in IRT, the global fit of the applied model for data is assessed, then if the model has good fit, individual responses are meant to be also appropriately fitted to the parameters estimated by the given IRT model. This appropriateness analysis of individual responses is referred to as the person-fit analysis (Meijer & Sijtsma, 2001)

Person-fit statistics have been proposed by several IRT researchers as a useful

method to identify response patterns that were inappropriate for the expectation given the fitted IRT model (e.g., Drasgow, Levine, & McLaughline, 1987; Levine & Rubin, 1979; Meijer & Sijtsma, 1995, 2001; Reise & Flannery, 1996).

In this study, one of the widely used person-fit indices, l_z is applied to detect inaccurate response patterns in the Korean Military Personality Inventory (KMPI: Choi et al., 2009) data. As mentioned above regarding problems because of inappropriate personnel selection, studies to detect inaccurate response patterns in personality measures is continuously needed in military recruitment settings. Thus, the purpose of this study is to evaluate the applicability of the person-fit statistic in military personnel selection contexts using personality measures. More specifically, the research question is (a) how is the person-fit index l_z applied to the KMPI data? (b) are there inaccurate response patterns in the KMPI data? and (c) what are the possible reasons of inaccurate responses in the KMPI data?

This article is organized as follows. First, the overview of IRT and its models are explained. Second, fundamental background and practical equation of the person-fit index l_z are provided. Third, previous research regarding the use of the person-fit index l_z are reviewed in terms of its usefulness and applicability. Last, the practical use of person-fit index l_z is illustrated in the context of Korean Military personality data.

CHAPTER II

LITERATURE REVIEW

Response theory and measurement models

IRT is becoming popular approach to analyze tests in educational and psychological measurement area (Reise, Ainsworth, & Haviland, 2005; Chou & Wang, 2010). Because most of the IRT studies begin with decision of appropriate model, the basic knowledge regarding IRT models are needed in the application of IRT. However, before reviewing models, the assumptions that must be held in application of IRT models should be discussed in detail because the violation of the assumptions could lead to seriously biased parameter estimation (as cited at Chou & Wang, 2010).

Dimensionality and local independence

Dimensionality is referred to as the number of latent factors or attributes that influence examinees' responses to the items (Chou & Wang, 2010). This is a core assumption in IRT. Most IRT models assume unidimensionality, which implies that a set of items measures no more than one factor or attribute (Hambleton, Swaminathan, & Rogers, 1991; Chou & Wang, 2010). But, the unidimensionality assumption might not be strictly met in practice because several factors in a test taking affect examinees' performance (Hambleton, Swaminathan, & Rogers, 1991). These factors might include level of motivation, test anxiety, ability to work quickly, tendency to guess, and cognitive skills (Hambleton, Swaminathan, & Rogers, 1991). Thus, in order to meet the unidimensionality assumption, most IRT models evaluate the presence of a "*dominant*"

common factor that influences test performance (Hambleton, Swaminathan, & Rogers, 1991; Reise, Moore, & Maydeu-Olibares, 2011). Item response models in which a strong common factor is presumed sufficient to explain or account for examinees' performance are referred to as unidimensional models. The use of unidimensional models for an item set measuring more than one dominant factor might cause serious distortion in parameter estimations (Hambleton, Swaminathan, & Rogers, 1991; Reise, Moore, & Maydeu-Olibares, 2011). If there is more than one dominant factor which affects examinees' test performance, multidimensional IRT models should be fitted to data (as cited in Reise, Moore, & Maydeu-Olibares, 2011).

Local independence implies that when the abilities or traits influencing test performance are held constant at any given level, the probability that examinees endorse one item or a set of items is statistically unrelated to the probability that examinees endorse any other items (Hambleton, Swaminathan, & Rogers, 1991; Streiner, 2010). In other words, under local independence, the probability of endorsing items for examinees at a given ability or trait level is equal to the joint probability of endorsing the individual items (Hambleton, Swaminathan, & Rogers, 1991).

There is one point that should be noted regarding the relationship between these two assumptions. When the assumption of unidimensionality is true, local independence is necessarily obtained. When local independence assumption is true, however, the unidimenstionality cannot be guaranteed for the data set (Hambleton, Swaminathan, & Rogers, 1991).

IRT models for dichotomous item

The basic unit of IRT is the item response function (IRF; also known as the item characteristic curve [ICC]) (Reise, Ainsworth, Haviland, 2005). IRF is a mathematical function which indicates the probability that examinees with specific ability or trait (symbolized as θ) endorse a given item designed to measure the latent traits or abilities (Hambleton, Swaminathan, & Rogers, 1991; Reise, Ainsworth, Haviland, 2005). IRT modeling fundamentally focuses on determining an IRF for each item (Reise, Ainsworth, Haviland, 2005). IRFs are utilized to show how much information each item provide for a person's level of a particular trait in terms of three kinds of psychometric properties: difficulty, discrimination, and pseudo guessing (Reise, Ainsworth, Haviland, 2005). IRFs for dichotomously scored items (e.g., yes/no or agree/disagree) provide 'S' shape curves using these three parameters respectively labeled *a*, *b*, and *g* (DeMars, 2010). IRFs are mathematically expressed as follow equation.

$$\mathrm{P_i}(\theta) = \mathrm{g_i} + \frac{(1 - \mathrm{g_i})\exp[a_\mathrm{i}(\theta - b_\mathrm{i})]}{1 + \exp[a_\mathrm{i}(\theta - b_\mathrm{i})]}$$

i: item (i = 1, 2, …, k)

a_i: discrimination parameter

b_i: difficulty parameter

g_i: pseudo guessing parameter

The discrimination parameter, *a*, shows how steeply the probability of correct response changes at the steepest point on the curve (DeMars, 2010). High values of *a*-parameter result in IRFs that increase very steeply and low values of *a*-parameter lead to IRFs that increase gradually as a function of θ (Hambleton, Swaminathan, & Rogers,

1991). Items having a higher discrimination can better differentiate between an examinee with relatively high θ and relatively low θ (DeMars, 2010). Items that have IRFs with steeper slopes are more useful for separating examinees into different θ levels than are items that have IRFs with less steep slopes (Hambleton, Swaminathan, & Rogers, 1991).

The difficulty parameter, *b*, explains how difficult the item is, which indicates the amount of the trait that is needed to be more likely to endorse the item (DeMars. 2010). Its value, in IRFs, equals the θ value where the slope of the function is steepest (DeMars, 2010). The greater value of the *b*-parameter indicates the greater ability that is required for an examinee to have a 50% chance of correctly answering on these items (Hambleton, Swaminathan, & Rogers, 1991).

The pseudo guessing parameter, *g*, provides the probability that an examinee with a very low level of θ will answer the item correctly (DeMars, 2010). Thus, in well-developed standardized tests, the *g*-parameter tends to be somewhat lower than chance because good test items have ability to keep low-ability examinees from correctly answering items by the guess (as cited in DeMars, 2010).

The most popular models used for dichotomous items are the one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL) (Hambleton, Swaminathan, & Rogers, 1991; DeMars, 2010). A primary distinction among these models is in the number of parameters used in the function that describe the relationship between θ and item responses (Hambleton, Swaminathan, & Rogers, 1991; DeMars, 2010).

Out of three logistic models, the 1PL model is one of the most widely used IRT models using only difficulty parameter to describe items (Hambleton, Swaminathan, &

Rogers, 1991). The 1PL model is often called the Rasch model ($a_i = 1$), in honor of its developer (Hambleton, Swaminathan, & Rogers, 1991). In this model, it is assumed that item difficulty is the only item characteristic that influences examinee performance (Hambleton, Swaminathan, & Rogers, 1991). Thus, IRFs differ only by their difficulty locations on the ability or trait and have equal slope and steepness (Hambleton, Swaminathan, & Rogers, 1991). Note also that the pseudo guessing parameter is zero: this specifies that examinees of very low ability have zero probability of correctly answering the item (Hambleton, Swaminathan, & Rogers, 1991). Equation and figure 1 below show the IRFs in the Rasch model.

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}$$

The 2PL model resembles the 1PL model except for the presence of an additional parameter used: the discrimination parameter, *a* (Hambleton, Swaminathan, & Rogers, 1991). The 2PL model is obviously a generalization of the 1PL model so that its IRFs have not only different difficulty location in the latent trait but also different slopes (Hambleton, Swaminathan, & Rogers, 1991). But note again that the pseudo guessing parameter is zero; hence, the 2PL model, like the 1PL model, does not allow the guessing behavior for examinees in the low ranges of latent trait (Hambleton, Swaminathan, & Rogers, 1991). IRFs of the 2PL model are described follow equation and figure 2.

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}$$

The 3PL model is the most generalized model among three logistic models for

dichotomous response items (Hulin, Drasgow, & Parsons, 1983). The 3PL model has an additional element called the pseudo guessing parameter that represents the probability of examinees with low ability answering the item correctly (Hambleton, Swaminathan, & Rogers, 1991). Hence, this model is particularly appropriate when individuals with low abilities or traits can occasionally respond correctly to difficult items (Hulin, Drasgow, & Parsons, 1983). IRFs of the 3PL model have different slopes as well as different lower asymptote like figure 3.

Person-fit measurement

Concerns that test scores in personality measures may not correctly reflect examinees' true traits being measured by items have driven test analyses using detective scales, such as Lie scale and Social Desirability scale (Ferrando & Chico, 2001; Zickar & Drasgow, 1996). However, the usefulness of those scales as a method of measuring inaccurate responses has been controversial. Some researchers argued that detective scales might have limitations because they were still in the self-report nature (MacNeil & Holden, 2006). According to their findings, if examinees had a very good skill in faking, they could keep away from being identified as a dishonest respondent by successfully faking on detection items as well. Thus, recently, statistical approaches using IRT to detect inaccurate test scores in personality measures are increasingly popular in various measurement settings. In IRT, the appropriateness of test scores can be measured by examining the consistency of individuals' response patterns with their trait (θ) estimated from assumed IRT model or the response patterns of majority examinees in given sample. The person-fit measurement simply uses the psychometric

information from the individual responses to detect inaccurate responses patterns. The person-fit statistics, then, numerically express how individuals' responses appropriately represent latent traits by quantifying the difference between an examinee's observed item response patterns and expected responses based on his or her latent trait (θ) as specified by IRT models.

For conducting the person-fit measurement, several person-fit indices have been developed. Among several indices, l_z index has been one of the most popular person-fit indices since the l_z index has showed relatively consistent detection power and has been easily interpreted in several person-fit studies (e.g., Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, & Williams, 1985; Ferrando, 2012; Zickar & Drasgow,1996).

Researcher should follow several steps in order to use l_z index. First, the fitness of the assumed model is obtained for the data. Second, each examinee's latent trait level (θ) is estimated given fitted model. Third, the log likelihood statistic l_0 is computed by compounding probabilities of individuals' endorsing and non-endorsing items given θ level. Mathematical equation of the l_0 is as follow.

$$l_0 = \sum_{i=1}^{k} \{X_i \ln P_i(\theta) + (1 - X_i) \ln [1 - P_i(\theta)]\}$$

k: the number of items in the measure

X_i: the response of the individual to the gth item (e.g., 1=Yes, 0=No in dichotomous items)

$P_i(\theta)$: the probability of the response to item i given the estimated person trait level

Then, in order to standardize l_0 statistics like z-score, the following formula is used:

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{\frac{1}{2}}}$$

For computing l_z, E(l_0) (the expected value of l_0) and Var(l_0) (the variance of l_0) use the following equations (Drasgow, Levine, & Williams, 1985).

$$E(l_0) = \sum_{i=1}^{k} \{P_i(\theta) \ln[P_i(\theta)] + [1 - P_i(\theta)] \ln[1 - P_i(\theta)]\}$$

$$Var(l_0) = \sum_{i=1}^{k} P_i(\theta)\,[1 - P_i(\theta)] \left[\ln \frac{P_i(\theta)}{1 - P_i(\theta)}\right]^2$$

Ideally, the null distribution of the l_z is the standard normal so that it has 0 of the expected mean and 1 of the variance like z-distribution (Drasgow, Levine, & Williams, 1985). Once determining the distribution of the l_z, researchers should set the cut-point which is used for classifying individuals whose l_z values are below the cut-point as the inaccurate respondents. Reise's simulation study indicated that different null distribution of the l_z might call for different cut-point (Reise, 1995). Depending on researchers' decision of the false positive value which refers to the rate that honest respondents become classified as faking respondents, the cut-point can be set.

Person-fit research using l_z

Previous research regarding the usefulness of the l_z statistic for identifying invalid responses has provided mixed results. Meijer (1997), in the simulation study, examined the influence of the invalid responses on test validities using the differently simulated

data set in the respect of test length, the size of the correlation between the predictor and the criterion tests, the proportion of invalid stimulees, and the type of invalid responding (cheating and guessing). Results indicated that l_z had little detection power by showing that the test validity was moderately increased by removing aberrant stimulees identified using l_z. More specifically, Meijer found that approximately 40% of total invalid stimulees remained in the sample regardless of the applied proportion of invalid stimulees. However, he suggested that l_z might be used to identify the group of respondents whose responses did not fit well to the IRT model so that their test validities were relatively lower than the validities of the groups of respondents with all valid responses.

Using empirical data, Ferrando and Chico (2001) examined whether l_z index could detect deliberate dissimulation in three kinds of personality scales (Extraversion, Neuroticism, and Psychoticism) of the Eysenck Personality Questionnaire Revised. By analyzing the normal group and the instructed faking good group, they also compared the usefulness of the index to the scale-based approach using the Lie and the Social Desirability scales. According to the results, the index was less powerful to detect dissimulating respondents than the Lie and Social Desirability scales.

Conversely, in the empirical study using U.S. Army data set, Zickar and Drasgow (1996) evaluated the effectiveness of the l_z index for identifying faking good respondents in compared with the scale-base approach using the Social Desirability scale. The data set included that three groups of respondents: the first group was asked to honestly answer all items, the second group was asked to answer in the manner of making

themselves more attractive, the last group was instructed how to answer questions in order to present themselves more attractive. By analyzing these data sets, they found that the l_z index performed better than the Social Desirability scale at the lower false positive rate. At the higher false positive rate, however, the Social Desirability scale was more effective than the index in detecting faking respondents.

Taken together the all finding from above research, the perspectives regarding effectiveness and usefulness of the l_z index are somewhat controversial. However, in those studies, l_z showed that it had still some ability to detect inaccurate responses even though its detection rate was low. The l_z index may play a role as a additional tool to screen possible problematic responses by supplementing limitation of self-report detection scales.

Some researchers pointed out the insufficiency of the person-fit index as a method to diagnose the reason of inaccurate responses (e.g. Reise & Waller, 1993). For supplementing that, Sijtma and Meijer (2001) suggested that the person response curve (PRC) could be used for diagnosing purpose in person-fit analyses. In practice, Ferrando (2012) used PRC analysis as a graphical procedure to suggesting possible sources of non-fitting response in personality measures (Neuroticism and Extraversion). He analyzed the discrepancies between expected and observed PRCs of 39 non-fitting response patterns which were identified using the l_z index. From carrying out PRC analyses, he classified the sources of aberrant responses into an idiosyncratic interpretation of certain items, low person reliability, and deliberate distortion. For example, 25 respondents endorsed certain items in the higher location compared to their

estimated latent trait level or did not endorse certain items in the lower location compared to their estimated latent trait level. As for them, the possible source of aberrant responding was classified as an idiosyncratic interpretation. In another case, 12 response patterns showed flat PRCs, which implied that those respondents were insensitive with item locations and had very low person reliability so that they almost randomly responded to items. Lastly, 2 participants were identified as a dishonest respondent by responding in opposite way from majority of participants doing. They endorsed items that were less frequently agreed and did not endorse items that were most frequently disagreed.

CHAPTER III

METHODOLOGY

Instrument

The Korea Institute of Defense Analyses (KIDA) constructed the revised version of Korean Military Personality Inventory (KMPI) in order to supplement several personality measures which were previously used in the Korea military personnel selection system, such as the shorter version of MMPI, the group Rorschach test, and MPI (Military Personality Inventory) (Choi et al., 2009). The KMPI originally has 183 items which contain dichotomously scaled response options. The KMPI has been administered to all new draftees within the first week after they join the military to screen their psychiatric disorders and military adaptability problems in the Korea Military since 2009.

The KMPI consists of five main content scales including Pathology scale, Accident Prediction scale, Accident scale, Emotion & Military Adaptability scale, and Response Distortion scale. The Pathology scale, developed with an experiential approach, measures for six factors: draftees' Anxiety, Depression, Somatization, Personality Disorder, Schizophrenia, and Paranoia. The Accident Prediction scale measures three factors to predict the degree of risk regarding Suicide, Desertion, and Mental Disorder. The Accident scale was designed to detect individuals who have possible problems related to the military accidents in terms of four different factors: Desertion, Adaptation Problem, Behavior Delay, and Acting Out. The Emotion & Military Adaptability scale

assesses five main factors such as Emotional Stability, Physical Discomforts, Personal Relationship, Attitude toward Military Life, and Action Control (Choi et al., 2009).

Except the Pathology scale, three substantive scales (Accident Prediction, Accident, and Emotion & Military adaptability) were developed with a factor-analytic approach. The KMPI includes a Response Distortion scale (34 items) which was designed to detect draftees who responded inaccurately to questions.

Among five main substantive scales, Accident scale and Emotion & Military Adaptability scale were taken for this research because these scales not only have significantly related items to the performance in the military but also have relatively obvious factor structures. Based on the output of Confirmatory Factor Analysis (CFA) using MPLUS, the Accident scale has fair fit for the four factor model (chi-square=5105.973*, df=1,704; RMSEA=.02; CFI=.908; WRMR=1.855). Also, the Emotion and Military adaptability scale shows fair fit for the five factor model (chi-square=2145.504*, df=979; RMSEA=.016; CFI=.968; WRMR=1.380).

Out of the 9 factors in Accident and Emotion & Military Adaptability, five were chosen for the person-fit analysis based on item length and relevance to the military personnel selection: Desertion (11 items), Adaption Problem (19 items), Behavior Delay (13 items), Acting Out (17 items), and Emotional Stability (17 items). Most items in these five factors are negatively worded so that the endorsement on those items indicates respondents' negative attitude or opinions. However, 2 items in the Desertion scale and 8 items in the Adaptation Problem scale are positively worded.

Sample

Data were obtained from the Korea Air Force Reserve Wing which is in charge of the recruitment, specialty classification, and reserve forces training. Individuals' scores on the KMPI are used as a reference in final selection and their future military lives. Draftees are provided written instruction and are asked to complete the pencil-and-paper KMPI without time limit. Their responses to each item are saved in the data process program using the KIDA server (Choi et al., 2009). The real response patterns of draftees who were recruited in July and August 2012 were collected from the server. The data don't have any personal information. The sample data consisted of 4,825 males in their early 20's.

Procedure and data analyses

For conducting the person-fit analysis, the model-data fit was investigated with raw data of the draftees' responses to each item in five subscales. The 2PL model was used in this research because it is not only less restrictive than the 1PL model but also without estimation problem regarding the pseudo guessing parameter in the 3PL model. The 2PL model has also provided relatively consistent fit with personality data in several previous studies (e.g., Reise & Waller, 1990; Zickar & Drasgow, 1996). Before applying 2PL model, in order to evaluate that each scale are indeed unidimensional, factor eigenvalues were obtained by EFA in MPLUS. Once large percentage of variance for the first factor eigenvalues (about three times as large as their respective second factor eigenvalues), it is concluded that the personality test is reasonably unidimensional (Schmitt, Cortina, & Whitney, 1993). As for another method to check the

unidimesionality assumption, CFA for each factor was also conducted in MPLUS. Based on values of the multiple fit indices, the fit of the single factor model was evaluated.

Once the 2PL model fitted well in the global level, the person-fit statistic l_z was computed for each respondent. By examining the standardized l_z values of the draftee' response patterns, individuals who might inaccurately answer questions were identified. If a draftee has the value of l_z below the cut-point, he is detected as an inaccurate respondent. In setting the cut-point, the normality of l_z distributions should be examined. In this study, if the distribution of l_z can be reasonably presumed as a standard normal distribution, the cut-point was set by usual cut-off value -2.0 on the left tail of l_z distribution (Ferrando, 2012; Schmitt et al., 1999). This value also corresponds with the point that approximately 98% of respondents who honestly answer questions should have the value of l_z above -2.0 (as cited in Brown & Villarreal, 2007). In other words, around 2% of response patterns from perfectly honest respondents might be classified as an aberrant response, which means the false positive rate (α) is .02. If the normality of l_z distribution cannot be presumed, the cut-off point was computed based on the critical value at a certain false positive rate. However, there haven't been a conventional use of false positive rates in the person-fit analysis and several studies argued that l_z statistics performed better than the social desirability scale at the lower false positive rate ($\alpha < .04$) (Zickar & Drasgow, 1996). Thus, in this study, the same false positive rate that was used for normal distribution ($\alpha = .02$) was applied to non-normal distributions of l_z, then the cut-point was set by the value corresponding with the l_z score of the respondents who were in the second percentile on l_z statistics.

After detecting inaccurate response patterns based on the cut-point of l_z, Person Response Curves (PRCs) were plotted for each problematic respondent in order to diagnosing possible sources of each aberrant responding (Emons, Sijtsma, & Meijer, 2005). In order to plot PRCs, first, entire items in the scale are divided several sets which includes 4 to 5 items ordered in item difficulty levels. Then, expected probabilities are calculated by computing the average of probabilities, which was estimated based on item and person parameters, of endorsing each item in every 4 to 5 set. Observed probabilities are calculated based on actual responses on items within each set. After that, PRC is plotted by connecting dots representing probabilities that respondents endorse each item in the expected and observed level (Trabin & Weiss, 1983; Ferrando, 2012). Generally, in the expected PRCs, the probability of endorsing the items decreases as the item difficulty level increases. In the same manner, the observed probability should decrease according to the increase of item difficulty level. However, PRCs of aberrant respondents are not accordance with this general concept.

In this study, by analyzing PRCs' patterns, the possible sources of each aberrant response pattern were suggested. For example, the flat line of the observed PRC might indicate a respondent's random responding because of relatively low motivation in taking a test. Also, the big discrepancy between two curves in the high or low end of difficulty location, which means endorsement of most frequently rejected items or rejection of most frequently endorsed items, might indicate other sources. Some respondents can have the big discrepancy due to idiosyncratic understanding on particular items. Some possible reasons of the idiosyncratic interpretation on personality

items have been suggested in previous studies. According to those studies, in self-report measures, respondents could idiosyncratically understand items because of different interpretation of words, different standards used, extreme responses, or reading error (McCrae, Stone, Fagan, & Costa, 1998). In another case, some respondents can have the big discrepancy by deliberately distorting their answers on particular items in faking-good or faking-bad manners. For example, a respondent can deliberately reject easy items to provide unfavorable impression to others or conversely he can endorse every item in the favorable way. Both idiosyncratic understanding and deliberate distorting on particular items can be sources for the big discrepancy of PRCs.

As another method to identify possible sources of inaccurate response patterns, the examination of actual responses on some particular items was used in this study. Adaptation Problem scale, for instance, item 105 and 141 are related to respondents' perception of the group life: item 105 is 'I can concede my point for the group where I belong to.' and item 141 is 'In a group life, collaboration is more important than competition.' If someone do not endorse these items and do endorse other items in similar difficulty locations with these two items, they might intend to behave themselves like an inappropriate person for the military. In other words, deliberate faking bad might be a possible source for that kind of response patterns.

CHAPTER IV

RESULTS

Descriptive statistics of KMPI scales

Most items of KMPI are negatively worded items so that 'Yes (1)' on those items indicates that the respondent might have possible problem in personality traits. However, out of five scales, Desertion and Adaptation Problems scales include positively worded items (items 12 and 130 in Desertion scale, items 59, 65, 70, 105, 120, 133, 134, 141 in Adaptation Problems scale). As for the positively worded items, 'No (0)' refers that the respondent possibly have problem in personality traits. Thus, in this study, to treat every items into the same direction those positive items were re-coded reversely ($1 \rightarrow 0$, $0 \rightarrow 1$) for allowing them to be easily interpreted.

Table 1 provides information of characteristics in each scale in terms of the number of items, included items, and the KR20 internal consistency estimate. The KR-20 was in ranging from .543 to .83. Except the Desertion scale ($\alpha = .543$), four scales showed reasonable reliability over .6 (Raykov & Marcoulides, 2011). Means and standard deviations (SDs) of summed (the sum of all 0 and 1for each individual) scores on each scale are given in Table 2.

In examining the unidimensionality, the results of factor analyses (EFA and CFA) supported that each scale was indeed unidimensional. Table 3 supports the existence of one strong dominant factor in each scale and the reasonable fit of the single factor model for data by providing several fit indices. All the first factor eigenvalues of each scale are

at least three times as large as the second factor eigenvalues, which supports that each scale has one strong dominant factor measured by items. Besides, the values of fit indices indicate that all scale had reasonable fit of one single factor model for data.

Results of 2PL model

In several previous studies, 2PL model had provided consistent results and fitted well for dichotomously scored personality data. However, to ensure the use of 2PL model, the fit evaluation was conducted using the IRT software. Based on the output of running IRTPRO for the 2PL model, the 2PL model was fitted well to the data in each scale. Table 4 shows fit statistics of 2PL model for each dataset of scales.

In fitted the 2PL model, the descriptive statistics of the person (θ) and item (discrimination *a* and difficulty *b*) parameters estimated are shown in Table 5. According to the descriptive statistics of parameters, the Emotional Stability scale, among five scales, has the highest ability (a = 2.12) in discriminating draftees with regard to their level of traits. On the other hand, the Desertion scale is the least discriminating (a = 1.27) in these scales. Five scales have relatively similar difficulty locations from 1.84 to 2.87. The Adaptation Problem scale includes the most difficult items for draftees to endorse and the Emotional Stability scale includes the easiest items comparing to other scales. In terms of the difficulty range of items, the Acting Out scale has items covering the widest range of difficulty location from 1.28 to 7.31.

Detection of inaccurate response patterns

The l_z values for each draftee was obtained with EXCEL by using parameters given the fitted 2PL model. Table 6 shows the descriptive statistics (means, standard

deviation, skewness, maximum, and minimum) of l_z distribution for each scale. The distribution of the l_z values in five datasets has a negative skewness. In common, the observed l_z distribution is negatively skewed (Reise & Flannery, 1996).

In setting the cut-off criteria for determining normality of l_z distributions, skewness and kurtosis should be within the +2 to -2 and the +3 to -3 when the values are normally distributed, respectively (Garson, 2012). Thus, four datasets except the Adaptation Problem scale were presumed to be normally distributed based on their skewness and kurtosis.

Under the presumed normality of l_z distribution in four scales' datasets, the cut-point was set by -2.0 (Ferrando, 2012; Schmitt et al., 1999), which indicates that draftees whose l_z values are lower than -2.0 should be identified as aberrant respondents. In the case of the Adaptation Problem scale, as I mentioned in the procedure part of the methodology, the cut-point was computed as the critical value of -1.5 at the false positive rate $\alpha = .02$.

Applying the each cut-point, draftees who gave aberrant responding to each scale were detected. The number and observed percentage of draftees whose l_z values were below the cut-point are shown in Table 7. The observed percentages were far below the nominal level 2.5%, except the percentage in Adaptation Problem scale. Only in the dataset of the Adaptation Problem scale, the percentage of aberrant response patterns (5%) was higher than the nominal level.

Diagnosis of possible sources

In diagnosing possible sources of each aberrant response pattern, PRCs and actual

response patterns on particular items which may be highly related to specific perspectives of respondents were analyzed. The results regarding possible sources of aberrant responding were arranged in each scale.

Desertion scale

In the Desertion scale, 40 draftees were detected as respondents who have inaccurate response patterns.

40 draftees' PRC were plotted based on the observed and expected probabilities of endorsing four sets of items (1st set: 130, 18, 29; 2nd set: 124, 43, 2; 3rd set: 12, 4, 33; 4th set: 30, 15) ordered in the difficulty level, then was assessed to identify the possible sources of the aberrancy. From analyzing PRCs (Emons, Sijtsma, & Meijer, 2005; Ferrando, 2012), two different aberrant patterns were classified: a type with the big discrepancy and the other type with the flat observed line. Figure 4 shows the illustrative examples of two draftees with two types of PRCs.

The former type of PRCs have big discrepancies between the observed and expected probabilities curves of endorsement in very low or high difficulty level of items sets. 7 draftees' response patterns were classified by this type of PRCs.

In Figure 4, the left graph corresponds to the draftee number 3019 who had -2.85 of the l_z value. Looking at the two lines, the expected curve (red line) is properly decreasing, which means that the probability of endorsing items decreases as the difficulty of items increases. However, the observed curve (blue line) provides the different pattern. Draftee 3019 yielded the extreme probability (1) of endorsement in the last two item sets with high difficulty level. In other words, this draftee endorsed the

last five items with high difficulty levels from 2.58 to 4.14. In considering the estimated θ level of this draftee, -.125, this extreme probability refers his response patterns did not accurately represent his trait level. This PRC with big discrepancy in the high end of difficulties might be because of the draftee's idiosyncratic understanding. In the last two set, several items which are possibly related to draftees' unique experiences are included. For example, as for item 15, 'I have vomited blood', item 33, 'I have caused as legal problem', and item 30, 'My family is displeased with the job that I had or want to have', draftees who had experienced the situation related to those items might have high probabilities of endorsement by uniquely understanding those items. Also, as for the item 4, 'Because I have a habit of wandering around, I become happy when I go around and travel', draftees could endorse or reject that item according to the standard they used. The use of different standards and the existence of unique experience might result in draftees' idiosyncratic interpretation on those four items.

On the other hand, 3 draftees were detected as an inaccurate respondent with the flat observed curves. In Figure 4, the right side of graph illustrates the example of the flat PRCs. Draftees number 141 whose l_z value was -2.32 has properly decreasing expected line but have relatively flat observed line as a function of the item difficulty. Those flat lines might be led by random responding because of the low test-taking motivation. (Emons, Sijtsma, & Meijer, 2005; Ferrando, 2012).

As another method to identify inaccurate response patterns, in this scale, three items (item 2, 4, and 29) were figured out, which are related to the possible problems regarding the relationship with surrounding people or the adjustment. Statements of

three items are as follow.

2. Unconsciously, I often argue against the other.

4. Because I have a habit of wandering around, I become happy when I go around and travel.

29. I once ran away from home without a parent's permission or noticing parents.

By examining actual responses on these items, one aberrant draftee was identified: draftee number 6 (l_z = -2.76). This draftee endorsed all three items, but did not endorse other items which have similar difficulty level as these items. Table 8 provides this draftee's actual responses in the Desertion scale. As looking at his actual responses, item 124 and 43 have the identical and lower difficulty location comparing to item 29 and 2, respectively. This response pattern may indicate that he tried to show himself as an inappropriate individual who might have possible problem in making relationship with other people, that is, he might answer these questions in faking bad manner. His score on the Detection scale, especially faking bad detection items, provided evidence for this by showing relatively high score (69, mean = 45.1, SD = 6.5). He also was detected as a aberrant respondent in the dataset of the Adaptation Problem scale.

Adaptation Problem scale

For this Adaptation Problem scale, the l_z distribution was non-normal so that the cut-point was set by -1.5 after the calculation based on the false positive rate (α = .02). Using this cut-point, the person-fit analysis detected 234 draftees as respondents who have inaccurate response patterns.

After analyzing PRCs which connected the probabilities of endorsing five sets of

items (1st set: 65, 133, 14, 59; 2nd set: 52, 8, 72, 120; 3rd set: 69, 47, 134, 132; 4th set: 61, 57, 51, 105; 5th set: 70, 141, 99) based on their difficulty orders, two kinds of aberrant PRCs were identified.

Firstly, 13 draftees had big discrepancy between observed and expected PRCs on the low or high difficulty levels of items. The discrepancy may result from the draftee's idiosyncratic understanding on those items. Some of them endorsed most items in the last two groups (relatively difficult items) and some others rejected most items in the first group (relatively easy items).

Another possible reason of the discrepancy might be draftees' faking-bad or faking-good responding on the high or low end of difficult items. In the last two groups, there are three items (items 105, 70, and 141) which highly indicate individuals' perspectives against a group life and the military. These three items are stated as follow.

105. I can concede my point for the group where I belong to.

70. I decided to follow the rule while serving in the military.

141. In a group life, collaboration is more important than competition.

Note that these three items are positively worded so that these items were re-coded reversely. Hence, '1' indicates that the draftee did not endorse those items in reality. Conversely, '0' means that the draftee endorsed the items actually.

Most of draftees who yielded big discrepancy in the difficult item set provided '1' on these three items (actually non-endorsement on three items), which refers that these individuals might try to give others unfavorable impression that they have negative attitude about the group or the military lives. That is, these individuals might answer

questions in the faking bad manner that, they believed, allows them to avoid their military duty. Figure 5 shows two illustrative examples of those individuals. Looking at their PRCs, the expected lines are decreasing as the difficulty levels are increasing in both graphs. However, the observed line of the draftee number 3289 (l_z = -2.93) is picked in the last set of items and the observed line of the draftee 1161 (l_z = -4.02) indicates his all endorsement in the 4th set of items. Also, draftee 3289 yielded '1', which means 'No', on items 105, 70, and 141 but yielded '0' on the easier item 51. Moreover, draftee number 1161 provided '1' on 5 items out of 7 but gave '0' on item 132 with the lower difficulty level than others. These response patterns might show draftees' intention of the deliberate faking bad in answering items. In looking at their observed scores on the faking bad detection items in the Detection scale, draftees 3289 and 1161 provided relatively high scores, 58 and 63 respectively (mean = 45.1; SD = 6.5), comparing to others' scores.

On the other hand, three items (items 65, 133, and 59) in the first set are highly related to the adaptability or the leadership. The endorsement of those items may allow respondents to be seen as a favorable individual. Again, these items are positively worded as follow so that they were re-coded reversely. Thus, '0' on these items indicates the agreement from draftees.

65. A lot of people follow me in the meeting that I belong to.

133. I'm quick at perceiving others in a group life.

59. I can lead my subordinates well when I become a veteran.

Aberrant draftees who had PRCs with the big discrepancy in the easy item set

provided ‘0’ (actually endorsement on these items) on these three items but yielded ‘1’ on several items which are more difficult than these three. Figure 6 are two examples of those draftees. Looking at both observed PRCs, the probabilities on the first set are zero, that is, draftees 3025 and 816 (their l_z values were -3.72 and -3.31) endorsed all three items, but the next probabilities are higher than expected ones. They answered those three items in the desirable way, which means that they might yield faking-good responses.

Secondly, 28 draftees had relatively flat PRCs. Again, flat PRCs refer possible random responding because of low test taking motivation. Out of 28, two examples are shown in Figure 7. According to two graphs, the expected curves of both draftees are decreasing in proper way, but their observed curves are almost flat. These kinds of flat observed PRCs might indicate that the draftees randomly answered questions with somewhat repeated patterns (Emons, Sijtsma, & Meijer, 2005; Ferrando, 2012). Looking at the draftees’ actual responses before they were re-coded, the random response patterns were more clearly appeared. Draftee number 1258 whose l_z value was -2.32 in the left graph had the same observed probabilities (.25) from the 1st to 4th set of items. His actual scores pattern on each item was 1010010010100110100. Even though the repeated patterns are not perfect, it seems to be obvious that this response pattern should be get warning which the draftee’s score might not accurately represent the trait (Emons, Sijtsma, & Meijer, 2005).

Behavior Delay scale

The person-fit analysis for personality data in the Behavior Delay scale detected

only 7 individuals with large negative l_z values, which means those draftees' responses might be aberrant.

To plot PRCs, 13 items ordered in the difficulty level were divided 4 groups (1st set: 6, 20, 118, 122; 2nd set: 107, 94, 116; 3rd set: 139, 22, 126; 4th set: 71, 45, 5), then probabilities to endorse each set of items in both observed and expected level were calculated. After analyzing PRCs connecting those probabilities, two draftees yielded PRCs that showed large discrepancy between observed and expected probabilities in the most difficult item set out of 4. Two individuals' PRCs are shown in Figure 8. In the left graph, draftee number 425 whose l_z value was -2.82 endorsed all items in the last group which included most difficult items but did not endorse all items in the 3rd group which consisted of easier items than the last group of items. Also, in the right graph, draftee number 1653 with -2.16 of l_z value provided similar pattern of PRCs with draftee number 425. More extremely, 1653 did not endorse any items in the 2nd and 3rd group. In both graphs, the expected curves are decreasing as item difficulties are increasing, which is accordance with concepts of PRCs in previous literature (Sijtsma & Meijer, 2001). However, the observed curves were picked at the high end of difficult item set and led to big discrepancy between two lines. These PRCs could indicate two possible sources of inaccurate responses. First, these draftees might idiosyncratically understand items in the high end of difficulty level. Second, they might try to deliberately endorse all items that are related negative perspectives against the trait. Actually, the last group of items, 71, 45, and 5 are stated as 'It is annoying that people seem to watch me', 'No one seems to understand me', and 'I hardly have a close relationship with others',

respectively. These three items are relatively difficult compared other items in this scale, which means that most draftees did not endorse these items. However, above two draftees endorsed all three items, which indicates that these two individuals might distort their answers those questions in the faking bad manner to give an unfavorable impression to others. Actually, draftee 425 had relatively high score on faking bad detective items in the Detection scale (58; mean = 45.1; SD = 6.5) and was detected as an inaccurate respondent in the Action Out scale dataset.

Acting Out scale

In analyzing l_z values of draftees' responses in the Acting Out scale, 21 recruits were identified as aberrant individuals who gave inaccurate response patterns.

17 items were ordered according to their difficulty level then divided 4 sets (1st set: 128, 127, 77, 74; 2nd set: 135, 113, 23, 82; 3rd set: 96, 73, 137, 88; 4th set: 27, 116, 115, 76, 106). The observed probabilities based on actual responses and the expected probabilities based on estimated parameters were connected as PRCs.

After analyzing each PRC, two aberrant patterns were identified, which are picked observed curve at the last item group and relatively flat observed curve. Among 21, one respondent (draftee number 4444) had the former PRC and two respondents (draftee numbers 4227 and 938) had the latter one. Figure 9 shows the PRCs of draftees 4444 and 938. Two draftees in the Figure 9 had l_z values of -3.86 and -2.13. Looking at both graphs, their expected probabilities are decreasing appropriately. However, the observed probability in the left graph was highly different from the expected probability at the last item set. Draftee 4444 endorsed 4 items out of the last 5 items, hence his

observed PRC was picked at the last end. Again, this aberrant PRC pattern might be due to a respondent's idiosyncratic understanding on particular items. For example, he might extremely respond on item 76, 'I probably commit something terrible to myself with guiltiness about what I have done in the past years', or answer item 27, 'It seems that a lot of people don't like me and behave unkindly to me' with perceiving contrast of self with others or using different standards. Also, another possible reason of this aberrant PRC is faking-bad responding. Actually, he agreed 12 items among 17. However, his response pattern is the least likelihood in the entire data, which might indicate that his responses inaccurately represent his latent trait. In looking at his observed score on faking bad detective items in the Detection scale, the score was 58 which was relatively high considering mean (45.1) and standard deviation (6.5).

In the right graph, the observed lines are almost flat because of almost equable probabilities of endorsing items in each set. The flat observed line possibly results from draftees' random responding regardless of item difficulty levels. The main reason of the random responding is the draftee's low test taking motivation (Emons, Sijtsma, & Meijer, 2005; Ferrando, 2012).

Emotional Stability scale

Person-fit analysis in Emotional Stability scale detected 27 inaccurate response patterns.

Like the analysis of PRCs in the Acting Out scale, 17 items were grouped in 4 sets (1st set: 75, 183, 160, 42; 2nd set: 124, 151, 66, 135; 3rd set: 180, 107, 153, 10; 4th set: 92, 98, 9, 58, 45) according to the order of their difficulty levels. Then, both

observed and expected PRCs were plotted by connecting probabilities of endorsing each set were calculated.

In scrutinizing PRCs of each aberrant respondent, 8 draftees yielded aberrant PRCs which has the big discrepancies between observed and expected probabilities in the high or low end of difficulty levels. Two illustrative examples of PRCs with big discrepancies are shown in Figure 10. Draftee number 390 (l_z = -2.43) in the left graph has large difference of probabilities in the first set of items. That is, he rejected all items that most draftees frequently endorsed. Conversely, draftee number 3172 (l_z = -3.21) in the right graph has high probability of endorsing items of the last set, which means he endorsed most items that other draftees frequently rejected. Observed PRCs of both draftees are not properly decreasing as item difficulties increase. Again, for those PRCs, two possible sources of inaccurate responses can be identified. The first source is the idiosyncratic interpretation on particular items and the second one is the response distortion. Looking at the actual response patterns, draftee number 390 rejected the easiest item 75, 'I have more concerns than others', but agreed the most difficult item 45, 'No one seems to understand me.' This response pattern might be due to draftee 390's idiosyncratic understanding on those items by using different standards compared to others. Also, draftee number 3172 endorsed the last 4 items in the most difficult item set. Those items include 'Sometimes, I felt that I seem to be shattered into pieces' and 'I do not want to remind of most of my memories in past.' These items might result in a respondent' extreme responding or various responding according to the use of different time frame adopted. Besides, as for the second source, the endorsement on most

difficult items might be because of the draftee's deliberate distortion. Respondents can more easily realize that particular items are related to the negative perspectives as the items are more extreme and difficult. Thus, they might answer those items in the unfavorable way to avoid specific situations that they do not want to be involved. In the examples above, draftee number 3172 endorsed 4 items out of 5 in the most difficult item set, which might indicate his faking-bad responding on those items.

Another aberrant PRCs identified is the flat ones. 3 draftees had relatively flat PRCs. In those PRCs, observed lines were almost flat regardless of decreasing expected lines. Again, those flat lines might come from draftees' random responding due to the low test taking motivation (Emons, Sijtsma, & Meijer, 2005; Ferrando, 2012).

CHAPTER V

SUMMARY

The main goal of this study was to evaluate the applicability of the person-fit measurement for detecting inaccurate response patterns of KMPI data in the Korean military personnel selection context. The person-fit analysis was conducted with real response data of new draftees' in five scales using l_z, one of the widely used person-fit statistics. This study is meaningful in the aspect that the person-fit measurement was practically applied to significantly large military sample (4,825). Besides, this study is finding that inaccurate response patterns exist in the military personality data.

In applying l_z index, inaccurate response patterns were identified for each scale. Possible reasons (e.g. idiosyncratic interpretation on items, faking, and low test-taking motivation) of the aberrant responding were classified for each individual through the analysis of PRCs and the examination of actual responses on particular items. In conducting person-fit analyses for datasets in five scales, 307 draftees were identified as an individual who provided inaccurate response patterns in total. Among them, 22 draftees were flagged as an aberrant respondent in two scales. 8 gave problematic answers on items of both Adaptation Problem and Emotional Stability scales and 4 did inaccurate responses in both Desertion and Adaptation Problem scales. Since they were draftees who gave aberrant responses over more than one scale, it is definite they should have more attention in interpreting their total scores on this military personality inventory. Nevertheless, only 7 draftees were detected as an aberrant respondent based

on their observed scores in the Detection scale according to the test result data from the Korea Air Force. The number of inaccurate respondents detected by both person-fit analysis and the Distortion scale was summarized in Table 9. In looking at the numbers, it seems to be deficient if only one method is used in detecting an inaccurate respondent. Several previous studies also provided evidence that the person-fit analysis based on IRT supplemented some limitations of scale-based methods (Social desirability and Lie scales) in detecting aberrant respondents (MacNeil & Holden, 2006; Zickar and Drasgow, 1996). That is, it appears that person-fit analysis based on IRT should be additionally conducted in order to detect problematic respondents in the military personnel selection contexts.

However, there are some limitations of this study. First, it is difficult to determine the detection power associated with the effectiveness and usefulness of the l_z. Since this study simply quantified the existence of aberrant response patterns, it is hard to examine whether the l_z detected entire inaccurate respondents. In looking at the observed scores in the Detection scale for those who gave aberrant responses, only 40 draftees were also included in those who were identified as an aberrant respondent by the detection scale. Nevertheless, this study can be a point of departure for the future validity study of l_z. Because most draftees used in this sample are currently serving in the Korea Air Force, other sources of information such as face-to-fact interview with officers or observation of their barrack lives might be used for investigating the validity of l_z statistics. Furthermore, during their two years military term, tracking of job performances such as the drop-out rate or disciplinary punishment records also might

be useful for the future validity study.

Second, in the Desertion and Adaptation Problem scales, there is the method effect because they are including some reversely worded items. It is possible that some draftees who are lack of concentration or are highly anxious of test unwillingly provide inaccurate responses without being aware of reversely worded items. If so, the aberrant responses might be due to wording effect rather than individuals' intention. Also, note that the Desertion and Adaptation Problem scales had relatively low internal consistency compared to others. In one previous study, researchers argued that the questionnaire which was worded in only one direction (negative or positive) had significantly higher internal consistency (Eys, Carron, Bray, & Brawley, 2007). The Desertion and Adaptation Problem scales included conversely worded items, which might affect their internal consistency. Besides, the Desertion scale consisted of only 11 items. In considering that longer tests generally have higher internal consistency, only 11 items of the Desertion scale led to its low internal consistency rate. Low internal consistency brings about one significant problem in applying a unidimensionality IRT model because the internal consistency is a measure of inter-correlatedness of test items (Schmitt, 1996). The internal consistency is not sufficient for unidimensionality (a set of items can measure one latent construct) but necessary for that (Schmitt, 1996). Thus, the low internal consistency of those scales cannot guarantee whether both scales held the unidimensionality assumption. If those scales violated the unidimensionality assumption, the parameter estimations based on applying unidimensionalty model (2PL) might be distorted and l_z values also based on the parameters estimated given that

model might not be assured.

Notwithstanding these limitations, this study provides important implications for the organization. This research demonstrated that the person-fit analysis should be conducted as an additional method to flag inaccurate response patterns before test scores are interpreted and used for the final decision. The person-fit analysis actually detected larger number of aberrant respondents than the detection scale did and removed the concerns regarding the self-report nature of the detection scale. However, simply detecting inaccurate responses should not be final goal for organizations. Organizations have to continuously deal with problems regarding what to do for individuals who were detected by both person-fit analysis and detection scale. For example, organizations might administer in-depth interview by professionals or more precise measures with them.

In sum, this study provided evidence that the person-fit measurement is applicable and feasible method for identifying inaccurate response patterns in the military personality data.

REFERENCES

Barrick, M. R., & Mount, M. K. (1991). The big-five personality dimensions job performance: A meta-analysis. *Personnel Psychology*, 44, 1-26.

Bernardin, H. J. (1977). The relationship of personality variables to organizational withdrawal. *Personnel Psychology*, 30, 17-27.

Brown, R. S. & Villarreal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing*, 7, 1-25.

Carroll, Matthew, F. (2003). Malingering in the military. *Psychiatric Annals*, 33(11), 732-736.

Choi, K., Jung, S., Choi, K., Moon, C., Kim, J., Park, B., Shin, E., Yuk, S., & Bae, J. (2009). New military personality inventory development report (KIDA Report 7-2699). Seoul; Korea Institute for Defense Analyses.

Chou, Y., & Wang, W. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, 70(5), 717 – 731.

DeMars, C. (2010). Item response theory. New York, NY; Oxford University Press.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59-79.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British*

Journal of Mathematical and Statistical Psychology, 38, 67–86.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person-response functions. *Psychological Methods*, 10, 101–119.

Eys, M. A., Carron, A. V., Bray, S. R., & Brawley, L. R. (2007). Item wording and internal consistency of a measure of cohesion: The group environment questionnaire. *Journal of Sport & Exercise Psychology*, 29, 395–402.

Ferrando, P. J. & Chico, E. (2001). Detecting dissulation in personality test score: A comparison between fit indices and detection scales, *Educational and Psychological Measurement*, 61(6), 997 – 1012.

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, 52. 6: 718 – 722.

Garson, G. D. (2012). Testing statistical assumptions. Retrieved from http://www.statisticalassociates.com/assumptions.pdf

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

Hough, L. M., Eaton, N. K., Dunnctte, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, 75, 581-595.

Hulin, C. L., Drasgow, F, & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin

Jones, E., Hyams, K. C., & Wessely, S. (2003). Screening for vulnerability to psychological disorders in the military: An historical study. *Journal of Medical Screening*, 10(1), 40-46.

Lande, R. G., Williams, L. B. (2013). Prevalence and characteristics of military malingering. *Military Medicine*, 178(1), 50 – 54.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269–290.

McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T., Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality*, 66, 285-313.

MacNeil, B.M., & Holden, R.R. (2006). Psychopathy and the detection of faking on self-report inventories of personality. *Personality and Individual Differences*, 41, 641-651.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement*, 21, 99–113.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, 90, 227–238. doi:10.1080/00223890701884921

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education*, 8, 261–272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107–135.

Pfaffenzeller, S. (2010). Conscription and Democracy: The Mythology of civil-military relations. *Armed Forces & Society*, 36, 481–506.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. Taylor & Francis. New York: Routledge.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, 19(3), 213-229.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, 14, 95 – 101.

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, 9, 9–26.

Reise, S. P., Moore, T., & Maydeu-Olibares, A. (2011). Target rotations and assessing the impact of model violations on the parameters of unidimensional item response theory models. *Educational and Psychological Measurement*, 71, 684 – 711.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, 14, 45–58.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, 65, 143– 151.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353.

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, 17(2), 143-150

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, 23(1), 41-53

Sijtsma, K. & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, 66, 191-208.

Sparks, C. P. (1983). Paper and pencil measures of potential. In G. P. Dreher & P. R. Sackett (Eds.), Perspectives on employee staffing and selection (pp. 349-368). Homewood, IL: Dow-Jones Irwin.

Streiner, D. L. (2010). Measure for measure: new developments in measurement and item response theory. *Canadian Journal of Psychiatry*, 55, 180–186.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44, 703-742.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item characteristic curve models. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptative testing (pp. 83-108). New York: Academic Press.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (1993). The assessment of background and life experiences (ABLE). In T. Trent & J. H. Laurence (Eds), *Adaptability screening for the armed forces* (pp. 101-162). Washington, DC: Office of the Assistant Secretary of Defense.

Zickar, M. J. & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, 20(1), 71-87.

APPENDIX

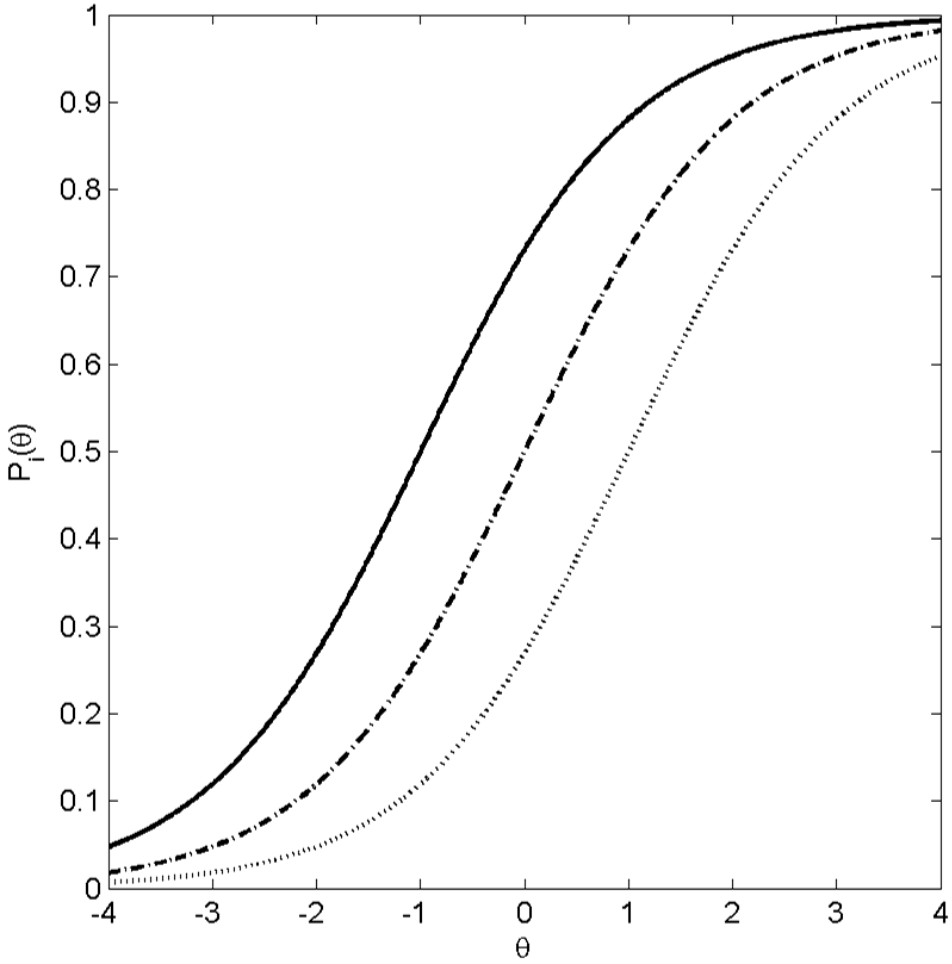

FIGURE 1. 1PL, Three item response functions with differences in difficulty (b = -1, 0, 1; a = 1)

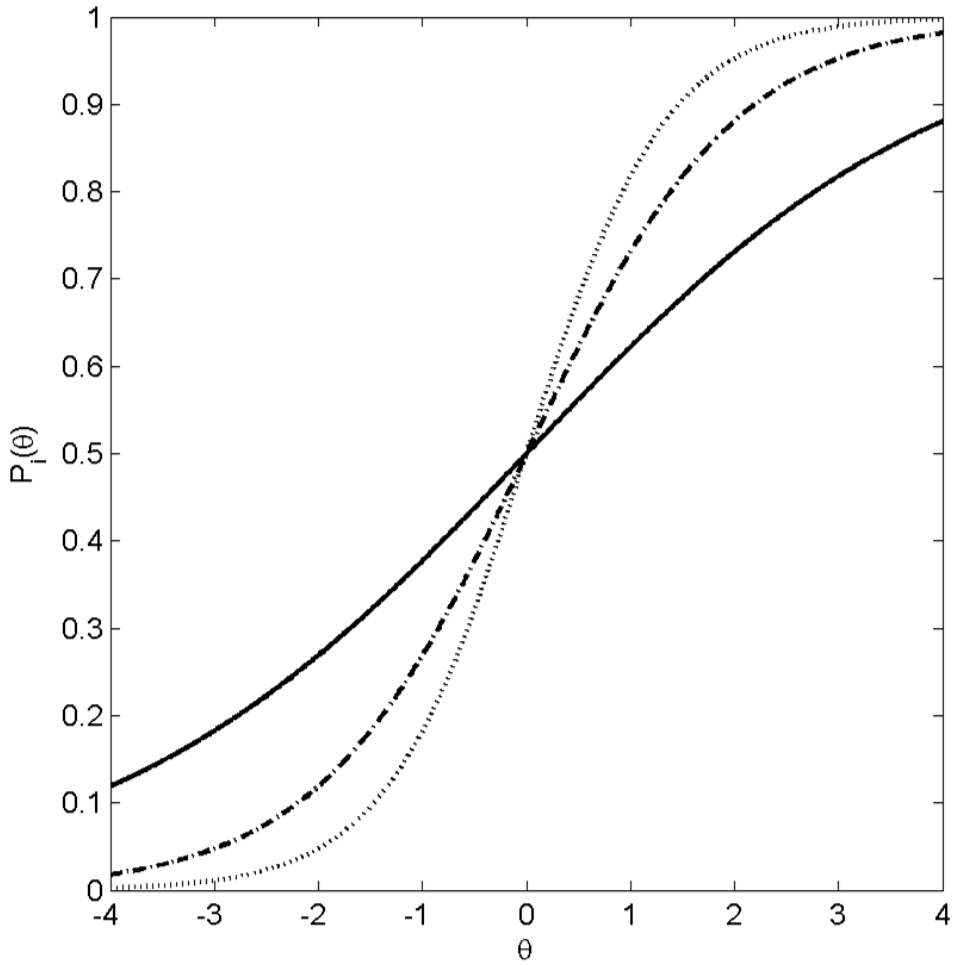

FIGURE 2. 2PL, Three item response functions with differences in discrimination (a = .5, 1, 1.5; b = 0)

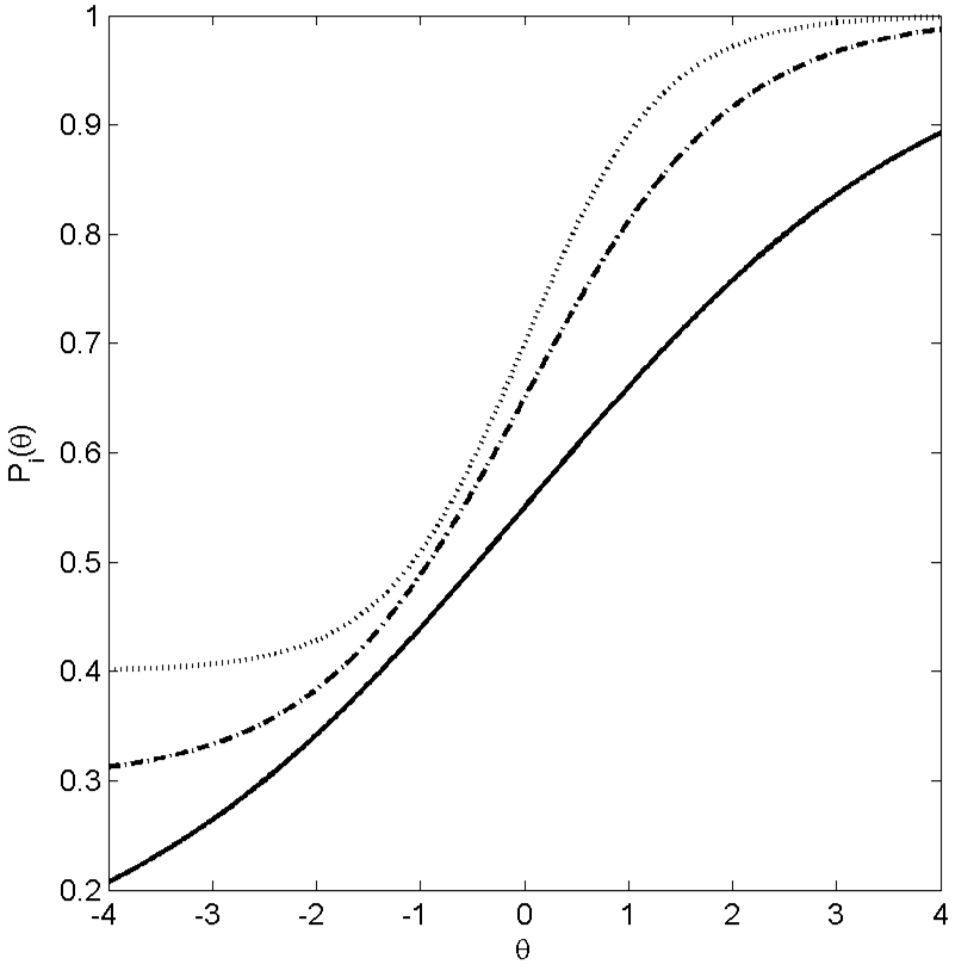

FIGURE 3. 3PL, Three item response functions with differences in pseudo guessing and discrimination (g = .1, .3, .4; a = .5, 1, 1.5; b = 0)

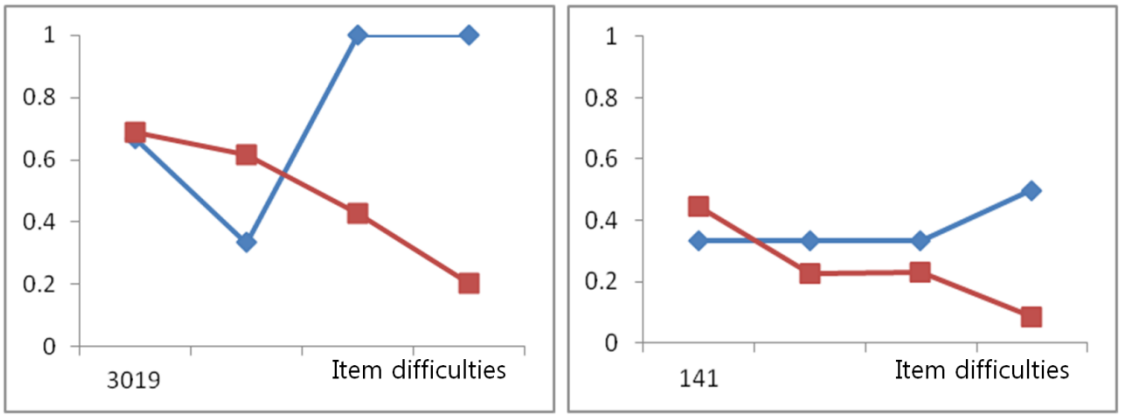

FIGURE 4. Two examples of PRCs with the big discrepancy on the difficult item sets and the flat observed probabilities in the Desertion scale (the blue line is the observed PRC and the red line is the expected PRC)

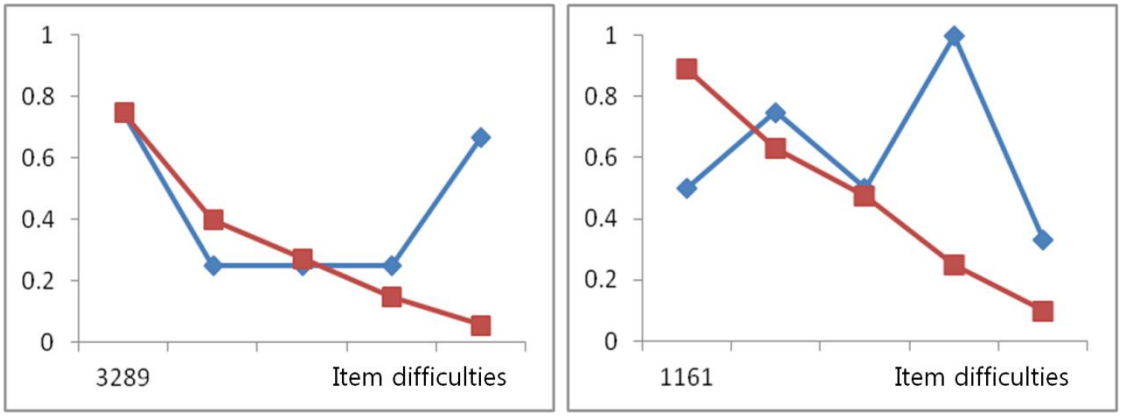

FIGURE 5. Two examples of PRCs with the big discrepancy on the difficult item sets in the Adaptation Problem scale (the blue line is the observed PRC and the red line is the expected PRC)

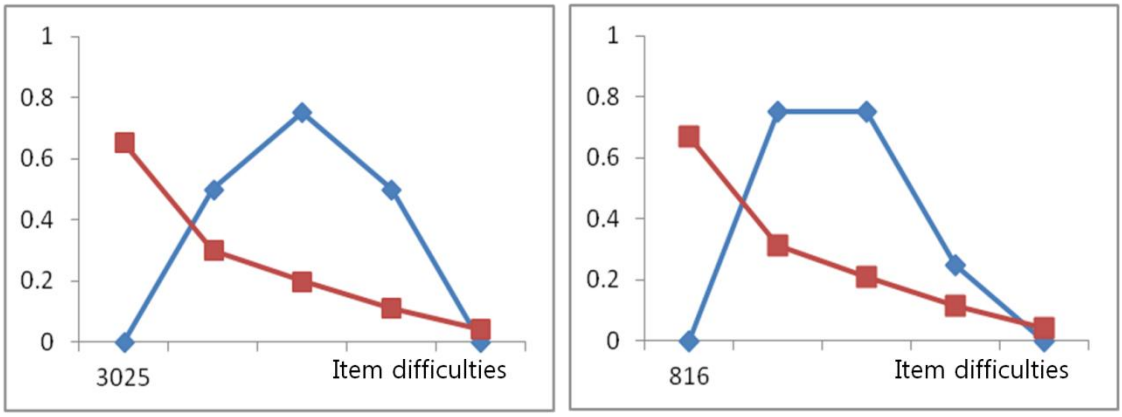

FIGURE 6. Two examples of PRCs with the big discrepancy on the easy items set in the Adaptation Problem scale (the blue line is the observed PRC and the red line is the expected PRC)

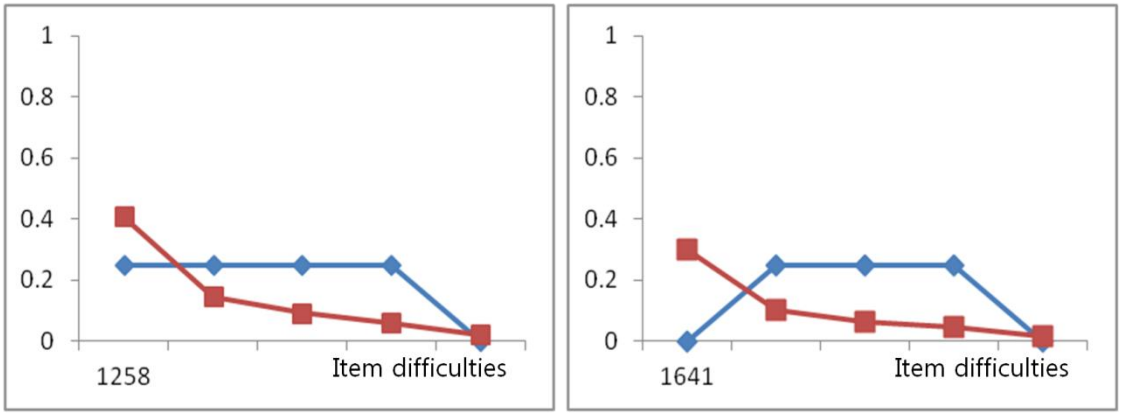

FIGURE 7. Two examples of PRCs with the flat observed probabilities in the Adaptation Problem scale (the blue line is the observed PRC and the red line is the expected PRC)

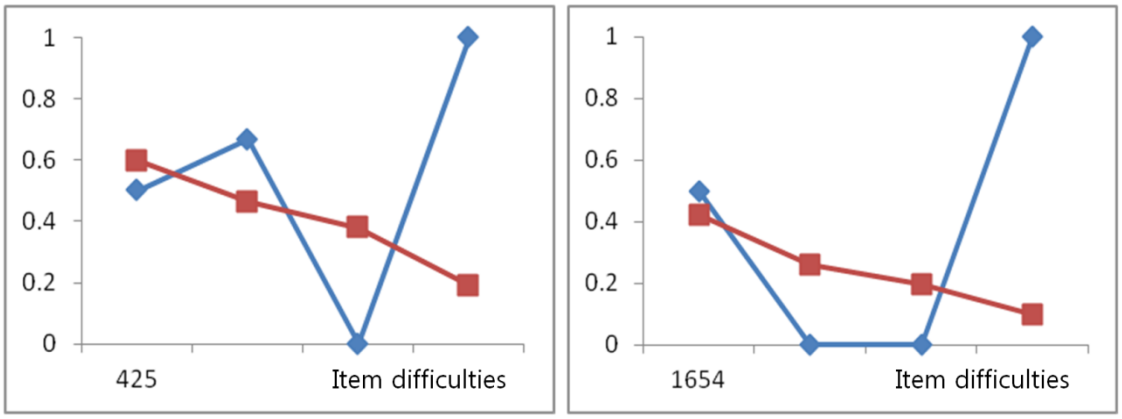

FIGURE 8. Two examples of PRCs with the big discrepancy on the difficult item set in the Behavior Delay scale (the blue line is the observed PRC and the red line is the expected PRC)

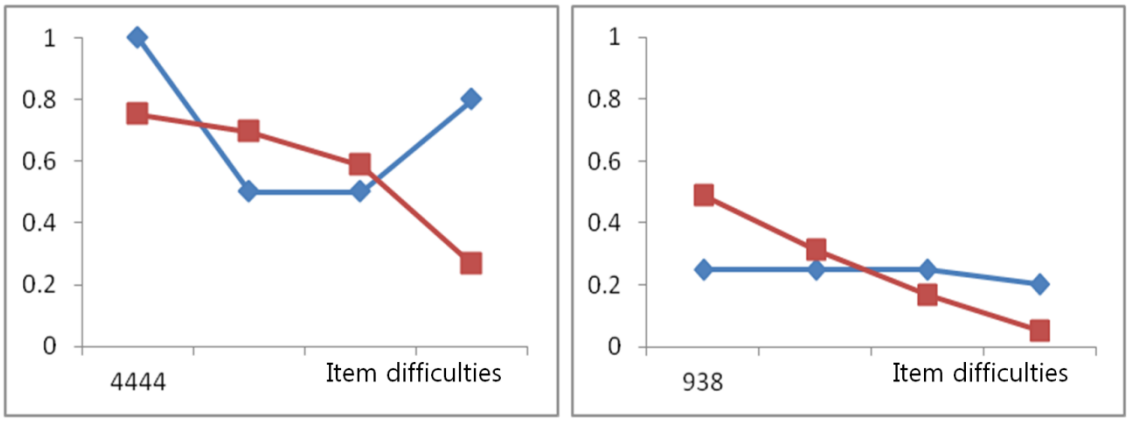

FIGURE 9. Two examples of PRCs with the big discrepancy on the difficult item set and the flat observed probabilities in the Acting Out scale (the blue line is the observed PRC and the red line is the expected PRC)

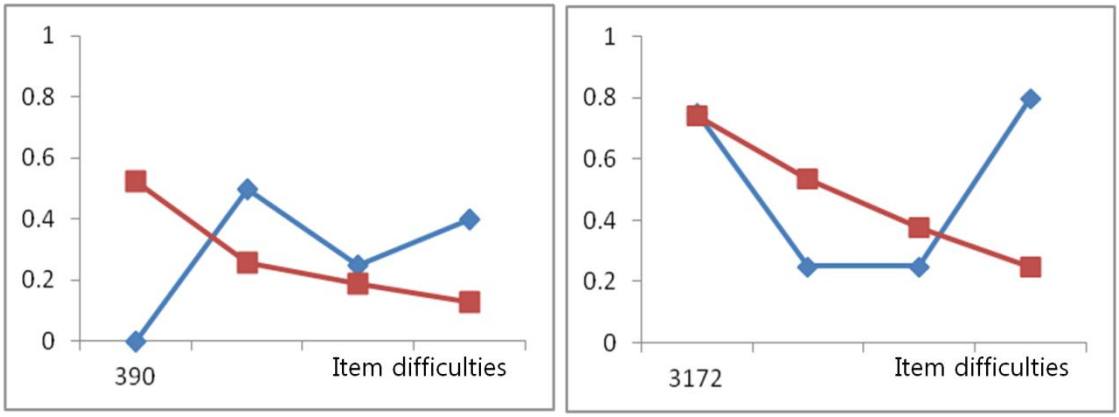

FIGURE 10. Two examples of PRCs with the big discrepancy on the difficult and easy item sets in the Emotional Stability scale (the blue line is the observed PRC and the red line is the expected PRC)

TABLE 1. Number of items, KMPI item numbers, and KR-20 for the scales

Scale	Number of Items	KMPI # (negative / positive)	KR-20
Desertion	11	2 4 15 18 29 30 33 43 124 / 12 130	.543
Adaptation Problem	19	8 14 47 51 52 57 61 69 72 99 132 / 59 65 70 105 120 133 134 141	.682
Behavior Delay	13	5 6 20 22 45 71 94 107 116 118 122 126 139	.705
Acting Out	17	23 27 73 74 76 77 82 88 96 106 113 115 119 127 128 135 137	.732
Emotional Stability	17	9 10 42 45 58 66 75 92 98 107 124 135 151 153 160 180 183	.830

TABLE 2. Means and standard deviation of summed scores on each scale

Scale	Mean	SD	Maximum	Minimum
Desertion	1.15	1.38	10	0
Adaptation problem	1.52	1.90	16	0
Behavior Delay	0.97	1.58	12	0
Acting Out	1.12	1.81	15	0
Emotional Stability	1.64	2.57	17	0

TABLE 3. Results of factor analyses (EFA & CFA)

Scale	EFA (eigenvalues)				CFA			
	1st factor	2nd factor	3rd factor	4th factor	x^2	RMSEA	CFI	WRMR
Desertion	3.924	1.072	1.028		147.705* (df=44)	.022	.959	1.342
Adaptation Problem	7.085	2.010	1.296	1.032	1058.157* (df=152)	.035	.864	2.221
Behavior Delay	6.249	1.279			507.460* (df=65)	.038	.931	2.035
Acting Out	8.557	1.379	1.051		304.089* (df=119)	.018	.979	1.309
Emotional Stability	9.342				375.055* (df=119)	.021	.986	1.280

TABLE 4. Fit statistics in IRTPRO

Scale	M2	df	p	RMSEA
Desertion	143.71	44	.0001	.02
Adaptation problem	1041.17	152	.0001	.04
Behavior Delay	440.15	65	.0001	.04
Acting Out	332.47	119	.0001	.02
Emotional Stability	392.88	119	.0001	.02

TABLE 5. Descriptive statistics for item and person parameters

Scale		Mean	SD	Maximum	Minimum
Desertion					
	a	1.27	.40	1.82	.73
	b	2.46	.76	4.14	1.35
	θ	-.00018	.68	3.24	-.60
Adaptation Problem					
	a	1.41	.44	2.23	.77
	b	2.87	1.22	5.06	.86
	θ	.00	.77	3.58	-.73
Behavior Delay					
	a	1.75	.32	2.27	1.15
	b	2.23	.49	2.98	1.31
	θ	-.00027	.72	3.17	-.53
Acting Out					
	a	2.03	.76	3.92	.79
	b	2.50	1.31	7.31	1.28
	θ	.00	.74	3.37	-.56
Emotional Stability					
	a	2.12	.51	3.66	1.61
	b	1.84	.50	2.78	.69
	θ	.00	.81	3.23	-.68

TABLE 6. Descriptive statistics of l_z distribution

Scale	Mean	SD	Skewness	Kurtosis	Maximum	Minimum
Desertion	.04	.68	-.91	.51	1.74	-3.16
Adaptation Problem	.06	.78	-1.77	3.75	1.32	-4.66
Behavior Delay	.03	.57	-1.01	.39	2.14	-2.82
Acting Out	.03	.59	-1.29	2.11	1.77	-3.86
Emotional Stability	.06	.59	-1.31	2.15	1.95	-3.25

TABLE 7. The number of inaccurate respondents

Scale	#	Observed %
Desertion	40	.86
Adaptation Problem	234	5
Behavior Delay	7	.15
Acting Out	21	.44
Emotional Stability	27	.57
Total	329	

*22 draftees were detected as an inaccurate respondent in two scales.

TABLE 8. Draftee number 6's actual responses

ID	v130	v18	v29	v124	v43	v2	v12	v4	v33	v30	v15	l_z
	1.35	1.5	2.09	2.09	2.13	2.35	2.58	2.61	2.9	3.37	4.14	
6	1	0	1	0	0	1	1	1	1	0	1	-2.76

*Second row indicates the difficulty parameters of each item

TABLE 9. The number of inaccurate respondents detected by both distortion scale and person-fit analysis

		Distortion scale	
		Accurate	Inaccurate
Person-fit analysis	Accurate	4511	7
	Inaccurate	307	0