Journal of Educational Research and Development
December 31, 2015 , Vol.11 No.4 pp. 1-22

Measurement Equivalence between Respondent Groups: A Non-Parametric Differential Item Functioning Analysis of Polytomous Personality Measures

Tzu-Ling Lai
Assistant Professor, Department of Counseling and I/O Psychology, Ming Chuan University

Abstract

The question of whether applicants respond to self-report personality measures differently when responding for selection purposes has been a crucial concern for decades. However, little research has focused on item-level measurement properties to identify the effect of testing situations on polytomous personality items. This study conducted a non-parametric poly-SIBTEST procedure to investigate both item-level and scale-level measurement equivalence on polytomous Likert-type personality scales between applicants and incumbents. The results indicated that several items exhibited differential item functioning (DIF); however, because DIF items did not systematically function with bias toward a particular group, substantial test functioning variations were not observed for all five scales. The items seemed to measure the same underlying constructs between applicants and incumbents.

Keywords: personality measures, measurement equivalence, differential item functioning (DIF), polytomous items

Tzu-Ling Lai's E-mail: sofia@mail.mcu.edu.tw

教育研究與發展期刊 DOI 10.3966/181665042015121104001
第十一卷第四期 2015年12月 頁1-22

多元計分人格測驗之測量恆等性：非參數方法之試題差異功能分析

賴姿伶 銘傳大學諮商與工商心理學系助理教授

摘　要

自陳式人格測驗經常以李克特式多元計分試題的方式呈現。然而，此類作答方式卻容易引起對於不同應試族群是否產生了不同的測量效果之疑慮，例如，當測量目的是為進行甄選時，受試者是否可能為了獲得錄取而刻意往高分的方向填答（亦即一般所稱的「作假」），而使得測量結果和其他情境下產生差異？過去已有大量研究探討應徵者在李克特式多分題的作答是否和一般學生或在職者不同，但卻多從整份測驗的層次著手，甚少針對試題層次的測量特性進行分析。本研究運用非參數的多分題同步試題偏差檢定法（poly-SIBTEST）來進行應徵者和在職者在試題層次以及量表層次的測量恆等性分析。研究結果發現：的確有若干試題對於不同的應試族群具有差異試題功能（DIF）；然而，由於差異試題功能並無系統性地偏利於某一族群，因此在所有的五個人格量表中皆未呈現差異測驗功能（DTF）。分析結果顯示多分題人格測驗應用於甄選情境時，所測量到的潛在特性和其他情境是相等的。

關鍵詞：人格測驗、測量恆等性、差異試題功能、多元計分試題

賴姿伶電子郵件：sofia@mail.mcu.edu.tw

Introduction

In recent decades, numerous studies have explored personality traits as predictors of job performance and other key predictors of job-related criteria (Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Salgado, 1997). In addition, personality tests have been suggested to add incremental validity to selection systems (Schmidt & Hunter, 1998), personality measures have increasingly been used in employee selection settings accordingly. However, the presence of Likert-type scales, which are frequently used in personality measures and the self-report process, raises concerns of respondents' conscious efforts to manipulate their responses. Applicants might be motivated to convey a contrived image that is positively biased and prototypic of an ideal employee when applying for a job, because responses can affect the probability of acquiring a job (Leary & Kowalski, 1990; Schmit & Ryan, 1993). Although research has constantly suggested that people are able to manipulate their scores on personality measures under experimental conditions (e.g., Frei, Griffith, McDaniel, Snell, & Douglas, 1997; Viswesvaran & Ones, 1999), evidences regarding actual response differences between applicants and incumbents in organizational settings are equivocal (Ellingson, Sackett, & Connelly, 2007).

Another concern derives from the development of personality measures. Because incumbents are often used for validating personality measures, the applicant-incumbent response variations are potentially problematic and pose a severe problem for constructing and interpreting personality measures. Thus, questions remain regarding the effect of applicant-incumbent response differences on the measurement properties of personality scales.

Over the years, various statistical and psychometric approaches have been devised to study applicant-incumbent response differences. Because applicants are considerably likely to manipulate their responses to acquire a job, scale score means comparison across applicants and incumbents has been used as a widespread method. Although meta-analysis has indicated that mean scale scores for applicants are generally higher than those for nonapplicants (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006), other studies have reported that applicants demonstrate a limited degree of scale score variation (O'Brien & LaHuis, 2011), even after repeated measures on the same applicants. For example, Hogan, Barrett, and Hogan (2007) observed that only 5.2% or fewer applicants improved their scores on the second occasion of completing the same personality measure when applying for the same job.

Another fundamental issue is the effect of applicant-incumbent difference on construct measurement and test validity. One approach is to test for measurement equivalence by using a combination of exploratory and confirmatory factor analyses (CFAs). Variations in the number of factors, factor-variable structural relations, or error variances between groups are interpreted as evidence of applicant distortion. Schmit and Ryan (1993) conducted a study using CFA to examine the fit of the five-factor model to test data from student and applicant samples. The five-factor structure fit the student data well, but an additional factor, the "ideal-employee factor," was obtained in the applicant sample. However, other studies have reported a relatively close measurement equivalence between applicants and incumbents (e.g., Ellingson, Smith, & Sackett, 2001; Smith & Ellingson, 2002).

The varying results suggested that applicants attempting to manipulate their responses on personality measures was a logical possibility, but not an empirical fact. The reason might be due to using mean variations and CFA, which are both at the scale or item-composite level, to identify applicant-incumbent difference. According to Schmit and Ryan (1993), different applicants might respond to particular items in distinct manners depending on their interpretation of the item regarding performance at work, resulting in complex loadings and factor intercorrelations. In addition, individuals used various strategies when responding to polytomous personality items (Zickar, Gibby, & Robie, 2004). The applicant-incumbent response difference is inherently an item-level phenomenon: People respond to individual items, not scales. Therefore, focusing on this question at the item level is crucial (Zickar & Robie, 1999).

In response to this, another approach used to examine item-level measurement property variations is the item response theory (IRT). The IRT-based methods for examining response variations between groups are known as the differential item and test functioning methods. Regarding polytomous items, differential item functioning (DIF) refers to the notion that a particular item may have different response functions for different groups so that two individuals from different groups may have different expected probabilities of choosing a particular option, even though they have identical thetas (θ, i.e. ability or trait level) (Camilli & Sheppard, 1994). The existence of DIF means that a particular item functions in different ways for distinct groups. Using DIF analysis, researchers can detect whether groups are responding test items differently. Differential test functioning (DTF) is the scale-level analog to DIF and refers to differences in expected scale scores by individuals with equal standings on the latent trait but sampled from varying subpopulations (Drasgow & Hulin, 1990). Because

item-level psychometric variations might or might not lead to scale-level psychometric variations (Robie, Zickar, & Schmit, 2001), DTF analysis is necessary to examine whether substantial DIF occurred to produce cumulative adverse effects at the scale level (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). If conscious item-level distortion accumulates to exhibit variation in the latent trait between groups, indicating that this scale measures distinct underlying traits, the measurement invariance is affected by response distortion.

Numerous procedures have been developed to detect DIF and classified as parametric and non-parametric approaches. Procedures for polytomous items, such as the graded response model (GRM; Samejima, 1969) and partial credit model (PCM; Masters, 1982), are examples of parametric approaches, whereas the Mantel-Haenszel method (Holland & Thayer, 1988; Mantel & Haenszel, 1959; Somes, 1986) and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) are non-parametric approaches.

Parametric methods compare item parameters estimated for a focal and reference group by using a particular IRT model after item parameters are placed on a common metric (i.e., linking). Variations in item parameters can provide insight into the nature of responding.

Previous IRT-based research on faking has generally adopted parametric procedures to identify response variations between groups (e.g., applicants or incumbents). Zickar and Robie (1999) used GRM to examine the effects of experimentally induced faking. They determined that approximately 22% of the items exhibited DIF, and the DTF existed across conditions. Follow-up research by Robie et al. (2001) also used GRM to examine measurement property variations between applicants and incumbents. The results indicated that moderately large mean variations existed in personality scale scores; however, only one of the six scales contained items that exhibited DIF, and no scale exhibited DTF.

These studies have estimated the same item response model for applicants and incumbents, and the focus was on whether any variations were present. However, they might not have been able to capture the differences between the applicants and incumbents. As indicated by a reviewer of Stark et al. (2001), “researchers might inherently assume that DIF, if caused by faking, acts uniformly against nonapplicants (i.e., DIF favors applicants), but that was not the case.” Zicker et al. (2004) used mixed-model item response theory to investigate within group response distinctions. They found that three classes were needed to model all response patterns across the applicant and incumbent data sets. In addition, there were a sizeable number of applicants who

appeared to be responding honestly and a sizeable number of incumbents who belonged to faking classes. Recently, O'Brien and LaHuis (2011) assumed that applicants and incumbents might interpret items differently and that their differences cause distinct item response functions (IRFs). In addition to finding that more than half of the items exhibited DIF, only 24% of the DIF items had IRFs in the hypothesized direction where the incumbent IRF demonstrated more folding than the applicant IRF (i.e., applicants faked and incumbents not), whereas 16% of the DIF items were opposite of the hypothesized direction. Nevertheless, DTF was exhibited for three of the 12 scales, but only two were in the hypothesized direction. Assumptions that applicants faked and incumbents responded honestly seem implausible based on these results.

These item-level results demonstrated that previous assumptions about the nature of applicant distortion on personality measures have been too restrictive. The assumption that individuals within a group respond in a similar way is questionable. Different styles and strategies exist for responding to polytomous personality items (Zickar et al., 2004).

Stark et al. (2001) investigated the effects of one parametric method, the Lord's (1980) chi-square method, and two non-parametric method, the Mantel-Haenszel method and SIBTEST method, to detect DIF between applicants and incumbents. Despite the disagreements regarding the DIF items identified by the three procedures, their work generally determined that each scale contained DIF, and DTF occurred for 13 of the 15 scales, the situational faking existed when responding to personality items. According to our review of relevant research, this is the only research that has adopted non-parametric procedures to examine item-level response variation between applicants and nonapplicants. However, the research investigated the situational effects on dichotomous items and adopted only the parametric procedure, the differential functioning of items and tests (DFIT; Raju, Van der Linden, & Fleer, 1995), to investigate DTF. It remains unclear how effective the non-parametric procedure is in examining DIF and DTF on polytomous items in organizational settings.

Parametric methods are intuitively appealing because item parameters, which are estimated and compared between groups, typically present simple psychological interpretations. Examining the variations in item parameters and the variations in the shapes of option response functions (ORFs) between groups might provide insights into the nature of responding. However, a problem of using parametric procedures is the potential influence of the extent of data-model fit (Stark et al., 2001). Nevertheless, regarding the polytomous personality items, the ORF comparisons between groups are complex. For example, an item that presents four options exhibits four ORFs that

correspond to each option. Such complex comparisons might limit application of parametric procedures to polytomous scales.

Non-parametric methods, by contrast, typically assume only monotonicity to an individual's trait levels (Stark et al., 2001), enabling the DIF results to be explained easily, particularly for polytomous items. SIBTEST for example, standardizes the two groups of interest to have a common distribution of the latent trait, and then estimates the expected difference in scores between the groups (Shealy & Stout, 1993). In addition, several researchers have suggested that non-parametric models fit Likert-type personality data better than do parametric models (e.g., Chernyshenko, Chan, Stark, Drasgow, & Williams, 2001; Maydeu-Olivares, 2005); thus, we believe that adopting non-parametric methods to investigate the response variations between applicants and incumbents might produce an enhanced understanding of applicant-incumbent difference on polytomous personality measures.

The SIBTEST Procedure

A widely used non-parametric DIF detection procedure is the SIBTEST, which was originally proposed by Shealy and Stout (1993) and used to detect DIF for dichotomous items. The SIBTEST was extended by Chang, Mazzeo, and Roussos (1996) to handle polytomous items and is utilized in the Poly-SIBTEST program. However, the SIBTEST procedure can also be conducted by widely applied statistical computer program like SAS or SPSS. The SIBTEST procedure can also be used to test for DTF (Doulas, Roussos, & Stout, 1996; Roussos & Stout, 1996). Although this approach has received little attention from applied psychologists, it is generally used among psychometricians who seek powerful alternatives to parametric methods. The primary advantage of this procedure is that it provides accurate DIF detection for samples of small to moderate sizes, even when the scales are multidimensional and the factors are highly correlated (Stark et al., 2001). Additionally, the SIBTEST is superior in controlling impact-induced type I errors, compared with the M-H and SMD procedures (Chang et al., 1996). Nevertheless, SIBTEST can be used for both estimates the amount of DIF in an item or set of items by using a regression correction technique, and has been extended for use with tests containing polytomous items as well as tests that are intentionally multidimensional (Bolt & Stout, 1996), which is beneficial to the analysis of multidimensional natured personality items. Because the SIBTEST method is non-parametric, DIF calculations can be performed directly using the scale scores without the linking procedure, thus, it

is easily applied and the estimated results is easily interpreted. The estimated value of DIF can be used to determine the favoring group of DIF item. Therefore, we adopted the SIBTEST procedure to analyze response variations on polytomous personality items between applicants and incumbents.

The procedure is based on the assumption that DIF occurs when

$$E_R\left[Y|\theta\right] \neq E_F\left[Y|\theta\right] \quad (1)$$

where Y denotes the score on the studied item, θ denotes the true score on the matching subtest, and the subscripts refer to either the reference group or the focal group. To conduct the SIBTEST to detect DIF, the two groups must be matched regarding the total scale score, where the ability level (θ) is the observed sum of all item responses within a scale. Because we tested the personality trait variations across groups in this study, the ability level θ also represent the trait level.

Let $P_{k.R}(\theta)$ denote the item category response function (ICRF), the probability of getting score k for a randomly sampled examinee with proficiency θ from group g (g = R for the reference group or F for the focal group). The regression of item score on ability can be defined as a weighted sum of ICRFs:

$$E_R\left[Y|\theta\right] = \sum_{k=1}^{m} kP_{k.R}(\theta) \quad (2)$$

Where Y is the studied item score which has $m+1$ ordered categories. $X_1, X_2, ..., X_n$ are the item scores for the n matching items. $m_1, m_2, ..., m_n$ are the maximum possible scores for $X_1, X_2, ..., X_n$, respectively. $X = \sum_1^n X_i$ is the matching score, $X = 0,1,...,n_H$, where $n_H \equiv \sum_{j=1}^{n} m_j$ is the maximum possible matching score for the n matching items. DIF could be estimated locally by the values of

$$d_k^* = \overline{Y}_{Rk}^* - \overline{Y}_{Fk}^* \quad , \quad k = 0, 1, 2, ..., \ n_H \quad (3)$$

$\overline{Y}_{gk}^*$ is the average score on the studied item for all group g (g = F or R) examinees for which $X = k$.

The expected amount of DIF at θ is measured by β:

$$\hat{\beta}^* = \sum_{k=0}^{n_H} p_k d_k^* \quad (4)$$

The positive β estimate represents which item is in favor of reference group.

In the present study, if an individual exhibiting a particular θ (i.e., a particular scale score) in the applicant sample tends to endorse a higher scored option compared with an individual exhibiting an identical θ in the incumbent sample, this item would evidence DIF regarding group membership. In this example, the item functions in favor of applicants (or, against incumbents).

To test for DTF, consider a subtest of items, *S*, which are obtained from some real-valued scoring function $h(\vec{U})$ that is applied to the original item scores, $(U_1,U_2,...,U_N)$. If the studied subtest does not exhibit DTF, $E_R[h_S(\vec{U})|\theta]-E_F[h_S(\vec{U})|\theta]\approx 0$.

The Present Investigation

Variations between the applicants and incumbents' responses at the item level were observed using two paradigms: the changing items paradigm and the changing persons paradigm (Zickar & Robie, 1999). The changing persons paradigm assumes that individuals fake by responding to items as if they had higher levels of trait than they actually possess. In other words, the respondent's true θ is temporarily and consciously changed to improve the personality test scores. In the changing items paradigm, perceptions of the same item might differ between groups. The changes in perceptions might relate to changing expectations of the consequences of choosing particular options or different frames of reference. In other words, trait levels are not affected by faking, but individuals interpret items differently.

We believe that trait levels are relatively stable and are unrelated to the testing situation. However, we argue that situational differences cause variations in the manner that respondents interpret items. In our study, we evaluated the extent to which individuals responded consistently to a personality measure across real selection and development contexts; thus, we adopted the changing items paradigm to explain response variations across testing situations.

This study examined the measurement equivalence at both the item-level and scale-level of polytomous personality measures between incumbent and applicant groups by using the non-parametric SIBTEST procedure. In particular, the current study presented two primary goals. First, we investigated the DIF prevalence and what group the items were functioning in favor of. Second, we determined whether the scale-level measurement properties were influenced across testing situations.

Method

Sample

Data were obtained from the database operated by a large human resource management consultancy company in Taiwan. The samples included applicants and incumbents who completed the Employee Personality Inventory either when they were applying for a job (applicants) or when they were asked by their company for research, counseling, or employee development purposes (incumbents).

Because of the limited amount of samples used in SIBTEST software and the data use policy restrictions of the company, we randomly selected 7,000 incumbents and 7,000 applicants for this study. Regarding the incumbent sample, 61% were men and 39% were women. Regarding the applicant sample, 48% were men and 52% were women. The applicants and incumbents applied for or worked jobs in various fields, including administration, finance, sales, engineering, art, and customer service jobs. The job category proportions for each sample were nearly identical.

Instrument

The Employee Personality Inventory, a test designed to measure personality traits useful for predicting performance in various jobs, was used in the current study. The inventory contained 124 items that were collapsed into scales that corresponded to the Big Five personality traits: openness (32 items), conscientiousness (23 items), extraversion (27 items), agreeableness (24 items), and emotional stability (18 items). A 6-point rating scale was provided, with scores ranging from 1 (strongly disagree) to 6 (strongly agree). The coefficient α of each scale was greater than .75, indicating adequate internal consistency reliabilities for the instrument. Results of the CFA suggested that a five-factor model adequately represented the data. Standard model fit measures were within acceptable ranges, the root-mean-square error of approximation (RMSEA) = .057, comparative fit index (CFI) = .95, and incremental fit index (IFI)=.94, implying sufficient construct validity (see also Lai & Yu, 2009, for more details). Because the Big-Five personality factors are frequently applied in organizational settings and most previous studies analyzed response variations across applicants and incumbents by the Big-Five structure, it is beneficial to adopt the response data of Employee Personality Inventory for comparing and discussing our findings to other research.

Procedure

This test was administrated online for both the applicants and incumbents. The applicants were administered a battery of online measures, including the aforementioned personality measure. The applicants were aware that the scores on the personality measure would be considered in the hiring process. A written warning also accompanied the instructions for the applicants that stated that distorted self-descriptions would invalidate the respondents' test results. The incumbents were allotted time during work to complete the personality measure. The incumbents were assured that their responses would not be used for any administrative purposes.

Analyses

We conducted several analyses to assess the differential functioning of the items and scales across testing situations. First, we calculated mean and standard deviations for each item and scale. A t test analysis was then performed to compare mean variations between the groups at the item and scale levels.

Subsequently, we used a procedure similar to that recommended by Byrne, Shavelson, and Muthén (1989) and Raju, Laffitte, and Byrne (2002) to examine whether data were sufficiently unidimensional to apply the DIF and DTF analysis by using CFA. This procedure was conducted on each of the five personality scales by using the LISREL software program. Once the unidimensionality was verified, the Poly-SIBTEST procedure was used to assess DIF and DTF.

To detect DIF, we conducted the one-item-at-a-time analysis of each item (i.e., the suspect item) and matched samples regarding total scale score (i.e., the two samples were matched by the same θ). As suggested by Shealy and Stout (1993), we adopted a critical p value of .001 divided by the number of items in each scale to identify DIF. For example, the critical p value for items on the openness scale was 0.001/32. Recall that if , the studied item exhibits DIF; in particular, a positive value indicates DIF against the focal group. The subsequent DTF analyses were conducted for each of the five scales to examine whether DIF changed scale measurement properties (Raju et al. 1995).

Consistent with previous studies, we referred to the applicants as the focal group and the incumbents as the reference group

Results

Descriptive Statistics and Reliabilities

Table 1 presents the resulting scale means, standard deviations, and reliabilities (coefficient α). Every reliability estimate of the scales, except for the agreeableness scale, was greater than .80. Reliabilities for the agreeableness scale were slightly lower (.76 for the incumbents and .77 for the applicants), but remained within an acceptable range. In general, the reliabilities were consistent for the incumbent and applicant samples. The standard deviations of the five scales for the incumbents (ranging from .42 to .60) were all slightly higher than those for the applicants (ranging from .41 to .58), suggesting that responses were slightly diverse in the incumbent samples. This result is consistent with that of other studies (O'Brien & LaHuis, 2011; Stark et al., 2001), indicating that scores when applying for a job tend to exhibit few variances.

Table 1

Means, Standard Deviations, and Reliabilities for The Employee Personality Inventory

Scale	No. of Items	Incumbents			Applicants		
		Mean	*SD*	*α*	*Mean*	*SD*	*α*
Openness	32	4.28	0.55	.92	4.41	0.53	.92
Conscientiousness	23	4.37	0.48	.85	4.49	0.47	.85
Extraversion	27	4.03	0.58	.91	4.06	0.54	.90
Agreeableness	24	3.63	0.42	.76	3.82	0.41	.77
Emotional Stability	18	3.77	0.60	.85	3.93	0.58	.86

The scale score means for the applicants were higher for four of the five scales. Regarding the extraversion scale, the scale means did not significantly differ between the incumbents and applicants. A further item-level mean variation investigation indicated that the applicants did not consistently score higher than the incumbents.

Table 2 presents the number of items that exhibited significant mean variation between applicants and incumbents. The incumbents reported no item-level higher means on three of the five scales (conscientiousness, agreeableness, and emotional stability). However, for the openness and extraversion scales, the incumbents scored higher for two items (6%) and five items (19%) respectively. Although the number of items for which incumbents scored higher was less than that of the applicants, the results of this item-level analysis indicated that the applicants scored higher on some items, whereas incumbents scored higher on other items.

Table 2

Number of items evidenced mean differences for The Employee Personality Inventory

Scale	No. of Items	No. (%) of items exhibited higher means for incumbents	No. (%) of items exhibited higher means for applicants
Openness	32	2 (6%)	29 (91%)
Conscientiousness	23	0 (0%)	16 (70%)
Extraversion	27	5 (19%)	12 (44%)
Agreeableness	24	0 (0%)	22 (92%)
Emotional Stability	18	0 (0%)	16 (89%)
Total	124	7 (6%)	95 (77%)

Unidimensionality

To examine unidimensionality, we used the RMSEA, CFI, and IFI as the indices. RMSEA is an index of overall model fit (Steiger & Lind, 1980). Browne and Cudeck (1993) suggested a good model fit if the RMSEA ≤ .08 and an adequate fit if the RMSEA is between 0.05 and 0.08. Regarding the CFI and IFI, greater than .90 should be the accepted threshold for a good model fit (Bollen, 1989).

The results indicated that RMSEAs ranged from 0.054 to 0.079 for each scale separately for applicants and incumbents, indicating an adequate fit. In addition, the CFI and IFI for each scale were all higher than .90. Thus, the results suggested that each scale exhibited sufficient unidimensionality.

Differential Item/Test Functioning Analyses

Table 3 presents a summary of DIF and DTF analyses for the comparisons between applicants and incumbents for the five scales. Overall, significant DIF occurred for 10, 12, 15, 7, and 6 items on the openness, conscientiousness, extraversion, agreeableness, and emotional stability scales, respectively. However, not all DIF items favored the applicants. Of the 10 DIF items on the openness scale, 7 (22%) items favored the applicants, whereas the remaining 3 (9%) items favored the incumbents. Regarding the conscientiousness scale, 6 (26%) items favored the applicants, whereas the remaining 6 (26%) items favored the incumbents. Regarding the extraversion scale, 7 (26%) items favored the applicants, whereas the remaining 8 (30%) items favored the incumbents. Regarding the agreeableness scale, 3 (13%) items favored the applicants, whereas the remaining 4 (17%) items favored the incumbents. Regarding the emotional stability scale, 4 (22%) items favored the applicants, whereas the remaining 2 (11%) items favored the incumbents. For example, the differential functioning of the first item on the

openness scale favored the applicants (but was against the incumbents), indicating that the applicants were more likely to choose the higher scored options than the incumbents having the same level of ability. By contrast, the sixth item on the openness scale favored the incumbents, indicating that the incumbents were more likely to choose the higher scored options on this item than were the applicants exhibiting the same ability level.

On average, DIF items that favored the applicants or incumbents existed for all five scales, suggesting that both the applicants and incumbents were likely to endorse higher scored options on personality items. However, the varying DIF prevalence that favored the applicants or incumbents across scales might indicate that each item meant different to applicants or incumbents. These might be influenced by the various perception or interpretation of items for the respondents under different testing situations. Because certain items favored the incumbents and others favored the applicants on the same scale, we further examined whether DIF systematically influenced the measurement properties of the scale-level. The results suggested that although several items were determined to exhibit DIF, substantial test functioning variations were not observed for every scale; the items seemed to measure the same underlying constructs between the groups, and the scale measurement properties for the incumbents and applicants were equivalent.

Table 3

Results of DIF / DTF Analysis across Incumbents (I) and Applicants (A)

Scale/item	*DIF/DTF β Estimates*	In favor of applicant (A) or incumbent (I)
Openness	--	--
Item 1	-0.114	A
Item 2	-0.094	A
Item 6	0.236	I
Item 9	0.105	I
Item 17	-0.049	A
Item 24	-0.122	A
Item 25	-0.081	A
Item 27	-0.050	A
Item 28	-0.103	A
Item 32	0.198	I
Conscientiousness	--	--
Item 1	0.089	I
Item 5	-0.078	A
Item 6	-0.092	A
Item 7	0.118	I

Note: Only items that exhibit significant DIF across groups are listed.

(continued on next page)

Table 3 (continued)

Results of DIF / DTF Analysis across Incumbents (I) and Applicants (A)

Scale/item	*DIF/DTF β Estimates*	In favor of applicant (A) or incumbent (I)
Conscientiousness	--	--
Item 8	-0.114	A
Item 10	-0.179	A
Item 12	-0.067	A
Item 13	0.092	I
Item 14	0.098	I
Item 15	0.090	I
Item 17	-0.076	A
Item 23	0.069	I
Extraversion	--	--
Item 2	-0.138	A
Item 3	-0.125	A
Item 5	-0.108	A
Item 7	0.110	I
Item 8	0.113	I
Item 11	0.077	I
Item 13	0.151	I
Item 14	0.218	I
Item 17	-0.110	A
Item 21	0.179	I
Item 22	0.096	I
Item 23	-0.086	A
Item 24	0.078	I
Item 26	-0.240	A
Item 27	-0.126	A
Agreeableness	--	--
Item 2	-0.144	A
Item 3	-0.068	A
Item 7	0.110	I
Item 9	0.104	I
Item 14	-0.101	A
Item 17	0.189	I
Item 21	0.113	I
Emotional Stability	--	--
Item 8	0.094	I
Item 9	0.162	I
Item 13	-0.093	A
Item 16	-0.080	A
Item 17	-0.164	A
Item 18	-0.052	A

Note: Only items that exhibit significant DIF across groups are listed.

Discussion

This study applied a non-parametric SIBTEST procedure to examine the measurement equivalence between applicants and incumbents regarding a polytomous personality measure. The SIBTEST procedure has been used by only a few applications related to organizational research. Our results provided a unique insight into responding to polytomous personality items through the lens of a non-parametric procedure.

The scale-level mean variations indicated that the applicants reported higher means on four of the five scales. This was consistent with previous meta-analysis findings (Birkeland et al. 2006). However, item-level analyses indicated that not only the applicants scored higher on personality items, but the incumbents also scored significantly highly on certain other items, although the proportion was relatively low. There were also several items presenting no mean differences. The mean variations between the groups were often interpreted as evidence of applicant faking by previous studies (Burns & Christiansen, 2011; Griffith et al., 2007); however, the job applicants did not simply inflate their responses for all of the items on the test according to our findings.

DIF analysis indicated that several items exhibited DIF across groups. However, in contrast to previous studies that have assumed that DIF items act against nonapplicants, our findings suggest that DIF items did not systematically function with bias toward a particular group. Each scale contained a group of items that favored the applicants and another group that favored the incumbents. The varying DIF prevalence among the scales further suggested that the concept of intentional distortion was not an either/or dichotomy. One alternative might be that certain respondents responded honestly on certain items and dishonestly on other items. The reason might be that items differed on feathers related to job performance or social desirability in varying situations (Zickar & Ury, 2002), or that various frames of reference alter how items are perceived (Zickar, 2000). For example, the item 24 of openness scale, “I like to travel around and experience different things” favored applicants, whereas the item 14 of extraversion scale, “I like to talk about my achievements” favored incumbents. Such varying responses might reflect that incumbents intensively exhibited their value to the job, but applicants may avoid presenting too aggressive trait so that they strengthen their open-minded image alternatively. All of these variations were present between the applicants and incumbents such that, for individuals having the same ability but in distinct testing situations, the applicants tend to endorse the higher scored options on certain positive dimensions

(e.g., the item 1 of openness scale), whereas the incumbents tend to endorse the higher scored options on other positive dimensions (e.g., the item 7 of conscientiousness scale). Respondents under a certain situation do not simply extend their response to all of the items. The basic assumption of previous studies that applicants fake and incumbents do not appears to be untenable, based on the results of this study.

Although a few items function differently, the DTF results suggested that the measurement properties on the scale level were not changed across testing situations; the items measured the same underlying construct for the applicants and incumbents. Nevertheless, the lack of DTF under this condition suggested that the DIF was not systematically in favor or against one group in particular.

Combined, our analyses indicated that although the nature of responding to personality items clearly differed in the applicant and incumbent samples, the scale-level measurement properties were not substantial. This finding might also support the research of Ellingson et al. (2001) and Smith and Ellingson (2002), whose examinations of factor structure suggested that responding distortion might not affect the construct validity of personality measures. All of these suggested that applicant-incumbent response variations on personality measures are more likely an item-level issue.

This study used a nonparametric procedure, the SIBTEST, to detect response variations on polytomous items between applicants and incumbents and determined that a relatively moderate proportion of items exhibited DIF, and no DTF existed. Regarding item-level investigations using real applicant and incumbent samples, our findings indicate little agreement with those of O'Brien and LaHuis (2011), Robie et al. (2001), and Stark et al. (2001). In Robie et al. (2001), relatively few DIF items and no DTF were identified, whereas in O'Brien and LaHuis (2011) and Stark et al. (2001), numerous items were identified as exhibiting DIF and DTF. Previous studies generally suggested a large proportion of items seemed to exhibit DIF and DTF for dichotomous items, regardless of whether a parametric or non-parametric approach was conducted; however, DIF and DTF were rarely observed when a parametric approach was adopted for analyzing polytomous items.

One reason could be that certain personality measures did not fit the IRT model well, such that parametric procedures on item-level measurement invariance studies appeared to underestimate the prevalence of DIF items in organizational settings (Stark et al., 2001). A second potential source of disagreement concerns the scale types that were dichotomous or polytomous. However, comparisons among the DIF procedures or scale types were far beyond the scope of this study. Perhaps further research could examine the interaction between DIF procedures and scale types.

The finding of this study suggests a moderate magnitude of DIF, indicating that using a non-parametric procedure to examine measurement invariance on polytomous measures applied in organizational settings might be advantageous. Further research is suggested to apply similar techniques to the development and validation of other tests (for example, vocational inventories or attitude testing) that are often used in organizational settings.

Limitations and Future Research

This study is one of the first to focus on response processes and provides a unique look at response variations on polytomous items in organizational settings through the lens of a non-parametric technique. Several response variations were observed between applicants and incumbents. We argued that the situational variations were caused by differences in the frames of reference that applicants and incumbents use or varying perceptions related to job performance; however, we were unable to attribute those variations to respondents' or items' characteristics. Because of the exploratory property of this analysis, further research is required to understand why these variations occur. In particular, future research is suggested to compare responses among applicants who are applying for various jobs to clarify job-related response variations.

Another limitation of this study is that only one personality inventory was investigated. It is possible that these items might be more susceptible or less susceptible to applicants' perception than those of other inventories. Further research is required to investigate the effect of the SIBTEST procedure when used on other inventories. We also suggest that researchers use multiple procedures to compare the accuracy of DIF/DTF detection methods. We hope these efforts might reduce problems for constructing and interpreting personality measures in organizational settings.

References

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

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14* (4), 317-335.

Bolt, D. & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23* (1), 67-95.

Bollen, K.A. (1989). A new incremental fit index for general structureal equation models. *Sociological Methods and Research, 17*, 303-316.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

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

Chang, H. H., Mazzeo, J., Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333-353.

Chernyshenko, O. S., Chan, K. Y., Stark, S., Drasgow, F., & Williams, B. (1999, April). *Fitting item response theory models to personality data.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Doulas, J. E., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial & organizational psychology* (pp. 577-636). Palo Alto, CA: Consulting Psychologists.

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*(2), 386-395.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*(1), 122-133.

Frei, R. L., Griffith, R. L., McDaniel, M. A., Snell, A. F., & Douglas, E. F. (1997). Faking non-cognitive measures: Factor invariance using multiple groups LISREL. In G. Alliger (Chair), *Faking matters.* Symposium conducted at the annual meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Griffith, R. L., Chmielowski, T., Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36,* 341–355.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*(5), 1270-1285.

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

Hough, L. M., Eaton, N. K., Dunnette, 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.

Hough, L. M., & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 31-88). San Francisco, CA: Jossey-Bass.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40*(2), 261–279

Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. In G. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 13, pp. 153-200). Greenwich, CT: JAI.

O'Brien, E., & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment, 19*(2), 109-118.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

Roussos, L. A., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Salgado, J. F.（1997）. The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82*(1), 30-43.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 34, (Suppl.17).*

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and non-applicant populations. *Journal of applied psychology, 78,* 966-974.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item Bias/DIF. *Psychometrika, 58,* 159-194.

Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of applied psychology, 87*(2), 211-219.

Somes, G. W. (1986). The generalized Mantel- Haenszel statistic. *The American Statistician, 40,* 106-108.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86* (5), 943-953.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Zickar, M. J. (2000). Modeling faking on personality tests. In D. R. Ilgen, & C. L. Hulin (Eds.), *Computational modeling of behavior in organizations: The third scientific discipline* (pp. 95–113). Washington, DC: American Psychological Association.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational research methods, 7*(2), 168-190.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*(4), 551-563.

Zickar, M. J., & Ury, K. L. (2002). Developing an interpretation of item parameters for personality items: Content correlates of parameter estimates. *Educational and Psychological Measurement, 62*, 19-31.

賴姿伶、余民寧、徐崇文（2009）。員工甄選人格量表的編製及其信效度考驗之初步報告。**教育研究與發展期刊，5**（4），269-304。