DIF in Personality Assessment: Does Cognitive Ability Influence Item Interpretation?

Amy DuVernet and Adam W. Meade
North Carolina State University

Chris Coughlin and Tracy Kantrowitz
PreVisor

The current study proposes and illustrates a method for examining item complexity during test construction using item response theory. We examined whether differential item functioning existed within 4 personality scales across cognitive ability groups. Results indicated that personality items do differentially function across highly disparate cognitive ability groups.

Organizations are increasingly relying on personality measures to make personnel decisions (Rothstein & Goffin, 2006). Various meta-analyses have established a link between conscientiousness and both overall performance (Dudley, Orvis, Lebiecki, & Cortina, 2006; Judge & Ilies, 2002; Ones, Viswesvaran, & Reiss, 1996b; Salgado, 1997; Tett, Jackson, & Rothstein, 1991; Vinchur, Schippmann, Switzer, & Roth, 1998) and specific facets of performance (Dudley et al., 2006; Ones, Viswesvaran, & Reiss, 1996a). Additionally, a pervasive finding in selection research is that cognitive ability (CA) test scores have trivially small correlations with measures of personality (e.g., Ackerman & Heggestad, 1997; Chamorro-Premuzic, Furnham, & Ackerman, 2006; Ones, Viswesvaran, & Schmidt, 1993; Schmidt & Hunter, 1998). This is a desirable trait of a selection procedure as uncorrelated predictors tend to account for more incremental variance in the criteria than do correlated predictors (Schmidt & Hunter, 1998).

Despite the lack of manifest correlation between personality and CA, a lack of correlation does not mean that CA cannot impact the measurement of personality or other survey measures. In this study, we examined the invariance of a series of personality measures for respondents with different levels of CA. Our purposes for doing so were twofold: (a) to examine if there is an impact of CA on personality measurement, and (b) to illustrate a technique that can be used to identify items that are overly complex. To accomplish these goals, we employ item response theory (IRT) methods of detecting differential functioning (DF) across low and high cognitive ability groups.

The Response Process

In order to understand how CA may impact personality and other survey-format measures, it is important to fully understand the survey response process. While there are several models of response available (Schwarz, 1999; Schwarz & Bless, 1992; Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000), there is a consensus that the process involves at least the following stages: literally understanding the item, understanding the pragmatic interpretation of the item, retrieving relevant information from memory, forming a judgment, and mapping the judgment to the response options. In the first stage, the respondent must interpret the semantic meaning of an item, which requires “the identification of words, the recall of lexical information from semantic memory, and the construction of a meaning of the utterance” (Schwarz, 1999, p 94). The respondent must also interpret the pragmatic meaning of a question, or what it is that the survey is asking. This requires the respondent to make assumptions about the intended meaning of the survey item which may go beyond the literal description (Schwarz, 1999). The third stage involves a search of short and long term memory seeking example behaviors that may be relevant to the content of the question (Schwarz & Bless, 1992; Tourangeau et al., 2000). During judgment formation, the retrieved information must be screened for relevance, integrated, and weighted in comparison to other retrieved information (Tourangeau & Rasinski, 1988; Tourangeau et al., 2000). Finally, when reporting a judgment, a respondent may attempt to appear consistent with responses to previous and related items (Schwarz, 1999). Clearly,
in each step of the response process, skills strongly related to CA are involved.

The response process is cognitively demanding; therefore, if a respondent is unable to complete any of the steps in the response process, he or she may become frustrated or lack the ability or motivation to accurately complete the questionnaire (Stone, Stone, & Gueutal, 1990). For this reason, most item writing guidelines (e.g., Furr & Bacharach, 2008; Hinkin, 1998) suggest that items be written in a familiar language and be as short and simple as possible. While there has been a paucity of research that specifically investigates the way that cognitive ability relates to respondents’ ability to understand items and their response options (Stone et al., 1990), a few studies have indicated that CA could play a role in survey item response. For example, Meisenberg and Williams (2008) found that education, which they used as a proxy for CA, was negatively related to question acquiescence (generally agreeing with items without cognitively processing their content) and extreme responding (favoring one end of the response scale) on a personality inventory. Similarly, Stone et al. (1990) investigated the possibility that CA influences respondents’ ability to provide accurate questionnaire responses. Stone et al. found that reliability estimates were significantly larger when they were derived from high cognitive ability respondents. The authors attributed this finding to differential item interpretation across levels of CA. They concluded by calling for more research investigating this issue using organizationally relevant measures; however, an examination of the extant literature reveals that little subsequent research has responded to their call.

The current study expands upon Stone et al.’s (1990) investigation by examining the extent to which personality items function differently across different CA levels. In addition, this study demonstrates a new method of using IRT-based DF analyses in test construction. Generally, test developers attempt to eliminate poorly worded items during the item generation and revision phases. In addition, they typically rely on factor analyses, reliability estimates, and item discrimination indices to identify and refine poorly performing items (Hinkin, 1998). The DF analyses illustrated in this study add to the test constructor’s tool box. Investigating DF in non-cognitive scales across high and low CA groups provides a means for identifying items that need revision. If item properties do not vary across high and low CA samples, then items are likely written at an appropriate level.

**Differential Functioning**

IRT models the relationship between the probability of an observed response and an examinee’s level of some latent trait or attitude. Perhaps the most commonly used IRT model in organizational research has been Samejima’s (1969) graded response model (GRM). In the GRM, a series of boundary response functions (BRFs) are graphed that relate the probability of responding at or above a given response category to the respondent’s underlying level of the latent trait or attitude (such as conscientiousness). A set of parameters is estimated for each item in the GRM, which is then used to plot the BRFs. These parameters can also be used to plot category response curves, which represent the probability of responding with a given response category across the range of the latent trait. We refer the reader to Embretson and Reise (2000) for a much more thorough discussion of the GRM.

When multiple samples (e.g., groups, time periods, response formats, etc.) respond to the same set of items, item parameters can be estimated separately for those samples. The extent to which the item functions differently for the groups is commonly referred to as differential item functioning, though we prefer the more generic term DF which applies to both items and scales. DF analyses account for any distributional differences in the latent variable between samples and assess the likelihood of response given the level of the latent trait or attitude for each group. In other words, regardless of mean differences in the observed or latent variable, DF analyses assess whether persons in different groups with equal levels of the latent trait have different probabilities of response. The current study will examine DF in four personality scales across groups of different CA.

**Method**

**Sample**

Participants were 7,020 job applicants who completed pre-employment assessments for entry-level call center positions within a large healthcare organization. Applicants voluntarily responded to demographic items. Of those who responded, 481 (21%) were male and 1,865 (79%) were female. In terms of self-reported race/ethnicity, 1,162 (50%) were African American, 1,011 (43%) were Caucasian, 135 (6%) were Hispanic, 25 (1%) were American Indian/Alaskan Native, and 12 (<1%) were Asian/Pacific Islander. This initial sample was reduced to a sample of 6,861 respondents due to missing data on one or more measures.

**Procedure**

Applicants completed the assessments from April 2005 to May 2006. Assessments were completed on-site and under proctored testing conditions at various locations of the host
organization across the United States. Participants' identities were verified before taking the assessments. All identifying applicant information was removed prior to data analysis to protect applicants' confidentiality and anonymity. The position for which applicants were applying was competitive as evidenced by a very low selection ratio. As such, applicants’ motivation to demonstrate maximal effort was likely to be very high.

**Measures**

**Personality.** The Global Personality Inventory (GPI) is a 300-item cross-cultural assessment of personality used for selecting, coaching, feedback, training, and succession planning (Schmit, Kihm, & Robie, 2000). It measures 9 performance factors (Thinking, Planning & Execution, Facilitating Leadership, Debilitating Leadership, Interpersonal, Motivation, Individual Work Orientation, Collective Work Orientation, Self-Management) and 37 facets of personality. Four facets, comprising 30 items, were selected for inclusion in the pre-employment test battery on the basis of job analysis findings. The facets included Empathy, Optimism, Emotional Resilience, and Responsibility. Applicants indicated their level of agreement with each statement on a 5-point Likert-type (1 = strongly disagree, 5 = strongly agree) scale.

**Cognitive Ability.** The 30-item Working with Information (WINFO) assessment (PreVisor, 2005) was used as the measure of CA. WINFO is a 15-minute timed, face valid measure of customer service information processing ability and has been shown to predict performance in call center positions. For each of the 30 items, a stimulus is presented to the test taker in which he or she has to locate the most salient information from which to determine the correct answer to a simulated customer’s inquiry in a timely manner. Items use a multiple choice format with five response options. The score is derived by summing the total number of correct responses across the items.

**Analyses**

**Groups Compared.** In order to determine the effect of CA on the measurement properties of the personality scales, groups were formed based on multiple values of CA in two sets of formations. In the first set of group formations, groups at the extremes of CA score distribution were compared for different sizes of samples. There were three conditions of such groupings: 200 persons per group pulled from the extremes of the CA distribution, 400 persons per group at the extremes of the CA distribution, and persons more than one standard deviation above and below the mean on CA (N approximately 1000 per group). These analyses address whether or not CA can impact the psychometric properties of the scales.

In the second set of group formations, groups were formed based on adjacent CA scores. In most hiring scenarios, it is most critical to make differentiations between persons with similar scores. Decision making is easy for persons with the highest and lowest scores; however, precision is more important when judging those with similar scores or those near a cutoff point in the selection system. As such, we examined the DF of items for persons with adjacent levels of CA at different points on the distribution of CA scores. Specifically, CA cutoffs for different selection ratios were calculated and the 200 respondents with scores just above and below the cutoff score were compared with respect to the DF of the personality measures.

**IRT Analyses.** DF of the four personality scales was examined using the Likelihood Ratio Test (LRT) (Thissen, Steinberg, & Wainer, 1993) in which nested models have item parameters constrained across groups. The IRTLRDIF program (Thissen, 2001) was used for these analyses. The decrement in model fit associated with these constraints is distributed as chi-square, allowing parametric tests of DF. As with all chi-square based statistics, the LRT is sensitive to sample size and has very high power when samples are large (Rivas, Stark, & Chernyshenko, 2009), potentially detecting even trivial DF. As such, the adjacent group analyses and one extreme group comparison used samples of 200 respondents per group.

We also computed DF effect sizes to indicate the difference in observed scale scores that is due to DF alone. The DTFR (Stark, Chernyshenko, & Drasgow, 2004) was computed for all analyses and represents the difference in observed scale scores that can be attributed to DF alone. DTFR is a scale level index that ranges from zero to a value near the observed scale score range (where scale scores are computed as a sum of the item responses). As such, larger DTFR indices are expected based on larger numbers of items. Also, as a scale-level index, DF present in individual items can cancel out across different scale items. We expect that more extreme groups will have a larger effect size, but the lower power associated with smaller more extreme groups may not result in any more DF items than larger and less extreme groups.

**Results**

Means, standard deviations, coefficient α’s, and the inter-correlations amongst all study variables are presented in Table 1. As indicated in the table, the
correlations between the personality scales and CA were very small (<.05 in all cases).

**Extreme Groups**

Results of the comparisons for those with the highest and lowest CA scores in the sample revealed substantial and significant DF for several items (see Table 2). Each of the four personality scales showed at least one DF item in all comparisons, and as many as 7 of 9 items showed DF in one comparison.

Perhaps more telling are the DTFR effect size estimates, which varied widely (see Table 2). DTFR indices indicate, the difference in observed scores that can be expected due to DF alone. DTFR indices were relatively low for Empathy and Responsibility across all three conditions, indicating that very little DF at the scale level is expected despite some DF at the item level. DTFR was modest for most conditions of Emotional Control and was largest for the Optimism scale. For comparisons of the 200 persons per group that were most extreme on CA, an expected observed scale score difference due to DF alone was .637. The scale contained 9 items, thus the range of possible scale scores was from 9 to 45. Even though the DTFR indicated that observed high and low CA group means should differ by less than one point, there was a half-point difference for two of the three conditions. Based on these results, it is clear that CA can and does have an impact on the measurement of personality at the extremes of CA.

**Adjacent Groups**

The adjacent group comparison saw much lower levels of DF than did the extreme group comparisons (see Table 3). This is not surprising as persons in these samples tended to have CA scores that differed by only a point or two (out of 30). Still, there was some DF, though it tended to be relatively small in effect. The DF that was present tended to be at the extremes of the distribution (i.e., those with nearly perfect CA scores and those just below and also for those with the minimum CA score and those just above). Inspection of the DTFR effect size estimates indicates that even at the extreme selection ratio cutoff scores, DF effect sizes were quite small.

**Discussion**

There were two important findings from this study that have implications for both theory and practice in personality testing and survey development. First, it is clear that when comparing groups with very different levels of CA, many personality items have different psychometric properties. These findings parallel those of Stone et al. (1990) who found that high and low CA can impact the psychometric properties of survey items. Whereas Stone et al. investigated coefficient alpha based on observed scores, we investigated DF which accounts for observed and latent mean score differences. Like Stone et al., we believe that for persons of low ability, the cognitively laden response process of item interpretation, memory retrieval, and judgment may either fundamentally differ or may be sufficiently discouraging that respondents get frustrated. These results were interesting in light of relatively pervasive findings that most personality measures do not correlate with CA... however, clearly there is a relationship between CA and personality measurement. Thus while the constructs themselves may not relate, the personality measurement process is associated with, and likely impacted by, CA.

The second main finding was that although CA can impact personality measurement in the extremes, the impact for comparing persons with relatively similar CA is minimal. Even though there were some DF items found, rates were low and Type I error alone should account for some DF items with so many comparisons. Moreover, effect sizes for adjacent groups were very small, never more than .15 observed score points for measures with a potential observed score range from 7 to 35.

Perhaps the greatest benefit of this study is that it illustrates a methodology for use in scale development. Items suspected of being confusing due to advanced language can be screened by investigating DF across groups with high and low CA. Should DF be found, the item may be written in a manner that is excessively difficult for those low in CA. It would also be interesting to compare those with the highest CA and those at the mean of the CA distribution. DF in this case may indicate that wording is excessively verbose or utilizes words that are at a high reading level. Such measures would likely not work well for much of the population. Similarly, comparing those with moderate CA and those lowest in CA may be useful for determining whether the item difficulty is appropriate for those lowest in CA. Depending on the use of the measure, this information may be important when working with low CA samples (e.g., instances of screening at risk persons in medical surveys).

**Limitations**

As with all studies, there were limitations to this research. First, we were limited to only four relatively short homogeneous scales, potentially limiting the generalizations that can be made from this study. Although both the CA and personality measures investigated are actively used in employment contexts and our sample included actual job applicants, this study represents but a single instance of many such potential measures.
Additionally, because this sample consisted of applicants to a healthcare organization, future research is needed to determine whether our results generalize to other types of samples or organizations. Also, LRT DF analyses can be sensitive to very small DF effects with large sample sizes. To counteract this, we reported DTFR (Stark et al., 2004) effect sizes. However, DF effect size indices are only recently developed and expectations about what constitutes small, medium, and large effects still do not exist. Hopefully future research will more fully clarify what constitutes practically important DF effects.

**Conclusions**

Though the constructs of CA and personality may be relatively independent (Chamorro-Premuzic et al., 2006; Schmidt & Hunter, 1998) it is clear that the measurement of these constructs is not. This study provides further support for Stone et al.’s (1990) theory that persons of different CA may approach the response process differently. Such persons may either get frustrated or simply have more difficulty successfully completing the cognitively loaded response process of item interpretation, memory retrieval, judgment, and response. Perhaps more importantly, this study illustrates a way in which scale developers can screen items for excessively high reading level.

**References**


Author Contact Info:

**Amy DuVernet**  
Department of Psychology  
North Carolina State University  
Campus Box 7650  
Raleigh, NC 27695-7650  
Phone: 919-515-2251  
Fax: 919-515-1716  
E-mail: amyduv@gmail.com

**Adam W. Meade**  
Department of Psychology  
North Carolina State University  
Campus Box 7650  
Raleigh, NC 27695-7650  
Phone: 919-513-4857  
Fax: 919-515-1716  
E-mail: awmeade@ncsu.edu

**Chris Coughlin**  
1805 Old Alabama Road, Suite 150  
Roswell, GA 30076  
E-mail: ccoughlin@previsor.com

**Tracy Kantrowitz**  
1805 Old Alabama Road, Suite 150  
Roswell, GA 30076  
E-mail: tkantrowitz@previsor.com
Table 1

Means, Standard Deviations, Coefficient α’s, and Intercorrelations for Study Measures

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional Resilience</td>
<td>28.85</td>
<td>3.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.79)</td>
</tr>
<tr>
<td>Empathy</td>
<td>29.58</td>
<td>3.11</td>
<td>0.49**</td>
<td></td>
<td></td>
<td></td>
<td>(.77)</td>
</tr>
<tr>
<td>Responsibility</td>
<td>31.59</td>
<td>2.81</td>
<td>0.56**</td>
<td>0.57**</td>
<td></td>
<td></td>
<td>(.80)</td>
</tr>
<tr>
<td>Optimism</td>
<td>37.42</td>
<td>4.13</td>
<td>0.61**</td>
<td>0.57**</td>
<td>0.60**</td>
<td></td>
<td>(.81)</td>
</tr>
<tr>
<td>Cognitive Ability</td>
<td>18.75</td>
<td>6.15</td>
<td>0.01</td>
<td>0.05**</td>
<td>0.04**</td>
<td>0.03**</td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 6861. * p < .05. ** p < .01. Scale reliabilities are presented in parentheses.
### Table 2

*Number of DF Items and Effect Size Estimates by Condition*

<table>
<thead>
<tr>
<th>Scale</th>
<th># items</th>
<th>Condition</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional Control</td>
<td>7</td>
<td># DF Items</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
<td>-0.165</td>
<td>-0.089</td>
<td>-0.088</td>
</tr>
<tr>
<td>Empathy</td>
<td>7</td>
<td># DF Items</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
<td>0.066</td>
<td>0.088</td>
<td>0.102</td>
</tr>
<tr>
<td>Optimism</td>
<td>9</td>
<td># DF Items</td>
<td>6</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
<td>0.637</td>
<td>0.519</td>
<td>0.444</td>
</tr>
<tr>
<td>Responsibility</td>
<td>7</td>
<td># DF Items</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
<td>0.028</td>
<td>-0.032</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Conditions: A = approximately 200 per group drawn from extremes. B = approximately 400 per group, drawn from extremes. C = groups are >1 SD above mean and < 1 SD below mean (N approx. 1000 per group).
Table 3

Number of DF Items per Scale for Different Potential Selection Ratios

<table>
<thead>
<tr>
<th>Scale</th>
<th># items</th>
<th>Selection Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>Emotional Control</td>
<td>7</td>
<td># of DF items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
</tr>
<tr>
<td>Empathy</td>
<td>7</td>
<td># of DF items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
</tr>
<tr>
<td>Optimism</td>
<td>9</td>
<td># of DF items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
</tr>
<tr>
<td>Responsibility</td>
<td>7</td>
<td># of DF items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTFR</td>
</tr>
</tbody>
</table>

Note. N=200 per group for all analyses.