2010

Do Applicants and Incumbents Respond to Personality Items Similarly? A Comparison using an Ideal Point Response Model

Erin L. O'Brien
Wright State University

DO APPLICANTS AND INCUMBENTS RESPOND TO PERSONALITY ITEMS SIMILARLY? A COMPARISON USING AN IDEAL POINT RESPONSE MODEL

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
By

ERIN LYNN O'BRIEN
B.S. Wright State University, 2007

2010
Wright State University

WRIGHT STATE UNIVERSITY

SCHOOL of GRADUATE STUDIES

May 17, 2010

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Erin O'Brien ENTITLED Do Applicants and Incumbents Respond to Personality Similarly? A Comparison Using an Ideal Point Model BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Committee of Final Examination

David LaHuis, Ph.D.
Thesis Director

David LaHuis, Ph.D.

Scott Watamanuik, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.

John Flach, Ph.D.
Chair, Department of Psychology

Herbert Colle, Ph.D.

Andrew T. Hsu, Ph.D.
Dean School of Graduate Studies

ABSTRACT

O'Brien, Erin Lynn. M.S., Department of Psychology, Wright State University, 2010. Do Applicants and Incumbents Respond to Personality Items Similarly? A Comparison Using an Ideal Point Model.

This study examined the extent to which applicants and incumbents use different response processes when responding to personality items. It was hypothesized that applicants' responses to personality items will be more similar to a dominance response model and that incumbents' responses will be more similar to an ideal point response model. I used item response theory to estimate sample data from applicants (N = 1509) and incumbents (N = 1568) who completed the Sixteen Personality Questionnaire Select. Differential item (DIF) and test functioning (DTF) analyses were conducted using the generalized graded unfolding model (GGUM), which is based on ideal point model assumptions. A number of items showed DIF; however, only about a quarter of those were in the hypothesized direction. DTF was significant for three of the twelve scales and two of those were in the hypothesized direction. Implications and limitations are provided.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Introduction

Research has indicated that concurrent and predictive validity designs produce similar validity estimates for cognitive ability assessments (e.g., Barrett, Phillips, & Alexander, 1981). However, there is less evidence for the comparability of validity designs for personality assessments. Hough (1998) reported that concurrent validity strategies provide higher validity coefficients for personality measures than predictive validity strategies. In addition, researchers have indicated that applicants' scale scores on personality assessments are higher than incumbents' scores (e.g., Barrick & Mount, 1996; Stokes, Hogan, & Snell, 1993; Viswesvaran & Ones, 1999). The primary explanation for these score differences between applicants and incumbents is that applicants intentionally distort their responses in an attempt to obtain the job. In contrast, incumbents are more likely to respond more honestly because they are less motivated to fake. Current research has focused on applicants engaging in purposeful impression management or answering in a more socially desirable way (e.g., Griffith, Chmielowski, & Yoshita, 2007).

These applicant-incumbent differences are potentially problematic because the development and validation of personality assessments often use incumbents. The extent to which results from incumbents generalize to applicants depends on how comparable applicants' and incumbents' responses are. Research has shown generally little evidence of applicant-incumbent differences affecting the psychometric properties of personality scales (e.g., Robie, Zickar & Schmit, 2001; Zickar, Gibby, & Robie, 2004). However, these studies have assumed that both groups use the same response process. That is, researchers estimated the same item response model for applicants and incumbents, and the focus was on if there were any differences in applicant and incumbent responses. For

example, Robie et al. compared applicants and incumbents using the graded response model (Samejima, 1969), and Zickar et al. used the partial credit model (Masters, 1982). The item response theory (IRT) models used to compare applicants' and incumbents' responses have assumed that both groups use a dominance response process where the probability of item endorsement relates monotonically to individuals' trait levels. This relationship is termed an item response function (IRF). However, these models may not be able to capture the differences between the applicants and incumbents.

Recently, Chernyshenko, Stark, Drasgow, and Roberts (2007) and Stark, Chernyshenko, Drasgow, and Williams (2006) suggested considering ideal point models for personality measures. These models suggest that individuals judge how well an item describes themselves in terms of the underlying trait and tend to endorse items that individuals feel match their level of the trait. The mismatch may be because individuals believe their trait level is less than that indicated by the item (disagreeing from below) or exceeds that indicated by the item (disagreeing from above). For example, consider the Conscientiousness item from the International Personality Item Inventory (IPIP; Goldberg, 1999), "I try to follow the rules." Individuals may not endorse the item because they hardly ever try to follow rules or because individuals always follow rules. This causes a bell-shaped IRF. Folding occurs when there is a decrease in the probability of endorsement associated with disagreeing from above.

In the present study, I suggest that applicants will be less likely to disagree from above than incumbents will. Thus, I expect the IRF's for applicants to exhibit less folding than IRF's for incumbents. Figure 1 shows an example of folding with two IRF's, where the dotted line exhibits less folding than the solid line. I am not arguing that applicants

and incumbents always use different response processes when responding to personality items. As I describe later, the ideal point response model can produce dominance-like IRF's when the location of the item and person matching occurs at very high levels of the trait. I do argue that situational differences lead to differences in the way that applicants and incumbents interpret items.

I tested for differential item functioning (DIF) across applicants' and incumbents' responses to dichotomous personality items using the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000). DIF refers to the ability to detect whether groups are responding differently across test items. If there is no DIF, then the IRF's would be equal across groups. My hypothesis is that the incumbent IRF will more closely resemble the solid line in Figure 1 and the applicant IRF will more closely resemble the dotted line. The GGUM assumes an ideal point response process and is applicable to personality data (Chernyshenko et al., 2007; Stark et al., 2006). Using the GGUM to model the data allows for the detection of the hypothesized differences in response processes. In the following sections, I briefly describe the dominance and ideal point response models. I then review the literature comparing applicants and incumbents responses to personality items and present the rationale for my hypothesis.

Thurstone and Likert Scaling Methods

The distinction between ideal point and dominance response models can be traced to differences between Thurstone (1928) and Likert (1932) scaling methods (Roberts, Laughlin, & Wedell, 1999; Stark et al., 2006). Thurstone scaling methods produce items that span across a range of trait levels, which results in IRF's that look like an ideal point

model. Likert scaling methods result in extreme items, which result in IRF's that look like a dominance model.

The Thurstone scaling methods involve two major steps (Thurstone, 1928). The first step is to create a large number of items that cover a wide range of opinions in terms of favorability or unfavorability towards the scale. For example, Conscientiousness items would be scaled in terms of how much Conscientiousness they reflect. There are several methods for scaling these items (e.g., pairwise comparisons, successive intervals). Participants judge how favorable or unfavorable an item is towards the attitude. The second step involves participants choosing which statements best reflect their attitude on that scale. Estimates of an individual's attitude are obtained by calculating the mean scale value. These scale values are then used to create empirical curves for each item. IRF's are created using these scale scores, which are used to determine relevant items. The final set of items under the Thurstone method are uniformly distributed across the attitude continuum. Individuals endorse a relevant item with construct levels consistent with the scaling for the item. For example, those with moderate levels of Conscientiousness should endorse neutral Conscientiousness items. Those with low levels of Conscientiousness should endorse items reflecting low levels of Conscientiousness. The result of this is a bell-shaped IRF similar to the solid line in Figure 1. The probability of endorsement reaches its zenith when the trait level equals the item location; the probability of endorsement decreases as the trait level becomes lesser or greater than the item location.

Coombs (1964) furthered the Thurstone method and proposed a response model based on ideal points and using an unfolding technique. The unfolded scale, often

referred to as the J scale, is the continuum of the construct. Each scale item is placed somewhere on the continuum based on how much or little it represents that construct. In addition, each examinee has an ideal point on the continuum that best represents their level of that construct. Examinees tend to endorse items that are proximal to their ideal point on the J scale and are less likely to endorse items farther from their ideal point. This produces data on an I scale where items are ordered in terms of the proximity to the ideal point. Figure 2 demonstrates the process for an individual with an ideal point of .4 and five items with difficulty levels of -3, -1, 0, 1, and 3. Items closer to the top of the I scale are more likely to be endorsed. Thus, in this example, the individual may endorse items with difficulty levels of 0, -1, and 1, but not the others. An unfolding technique is needed to map the response data onto the J scale to determine if the individual did not endorse an item from below or from above.

The Thurstone method of scale creation has been criticized as being too laborious. The use of a judgment group adds an additional step in the process that some researchers argue is unnecessary (Edwards & Kenney, 1946). The main benefit of this method, though, is the ability to distinguish different levels of the construct being measured. It provides information about a person's trait level at moderate levels of the construct versus just at extremely high or extremely low levels.

Likert scaling methods typically include collecting responses to a large number of items. Items are worded so that they express either positive or negative opinions and neutral items are avoided. Negatively worded items are reverse scored. There are several methods for analyzing the scores such as discrimination indexes, principal components, or item-total correlations. This process is an attempt to identify those items that are most

discriminating and highly reliable. This results in scales with high internal consistency and strong single factors. This process also assumes a dominance response process. An individual will endorse an item as long as that individual feels they have more of the underlying trait. For example, consider the Conscientiousness item from the IPIP (Goldberg, 1999), "I pay attention to details." An individual will endorse the item as long as they feel they always pay attention to details. Thus, total scores should be positively related to item endorsement.

The Likert method of scale development offers an alternative to the Thurstone method that is simpler to create. This method does not require an additional set of judges to review items. The drawback is that the information obtained is generally at extremely high and low levels of the construct, ignoring moderate levels. Another benefit to the Likert method is that it leads to higher test reliabilities. Edwards and Kenney (1946) argue that the Likert method leads to higher reliability with fewer items than the Thurstone method, although the reliabilities for the Thurstone method were acceptable at .79 (Edwards & Kenney, 1946). The benefits of the Likert method, higher reliabilities and simpler method, make it a more popular technique of scale development.

Dominance Response Models

The Likert method of scaling results in extreme scores, which leads to curves that look like the dominance model. Researchers have applied a number of dominance IRT models to personality item data. For dichotomous items, the two-parameter logistic model (2PLM) is often used. The 2PLM equation is

$$P_{ij}(Y=1|\theta_j) = \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1+\exp[\alpha_i(\theta_j - \beta_i)]} \quad (1)$$

In this equation, α and β are the respective item discrimination and difficulty parameters for item i. The discrimination parameter determines the steepness of the curve. In the personality context, item difficulty refers to the location of the item on the trait level continuum. Person j's trait level estimate is denoted θ. When trait levels are greater than item difficulty, there is greater than a 50% chance of endorsing the item. The dotted line in Figure 1 represents a typical 2PLM IRF in that the probability of endorsement relates monotonically with trait levels.

Ideal Point Response Models

A few ideal point IRT models exist; however, they are not applicable to personality data. In the present study, I focused on the GGUM (Roberts et al., 2000) because it appears to be the most applicable to personality data and because it is the most general (Stark et al., 2006). The GGUM equation for dichotomous data is:

$$P[Z_i = z \mid \theta_j] = \frac{\exp\left(\alpha_i\left[z(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M - z)(\theta_j - \delta_i) - \sum_{k=0}^{z}\tau_{ik}\right]\right)}{\sum_{w=0}^{C}\left[\exp\left(\alpha_i\left[w(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M - w)(\theta_j - \delta_i) - \sum_{k=0}^{w}\tau_{ik}\right]\right)\right]} \quad (2)$$

In this equation, Z is an observable response to item i, z equals zero for the strongest level of disagreement and C for the strongest level of agreement, α_i is the item discrimination parameter, θ_j is an individual's j trait level, δ_i is item location of item i on the trait level scale, τ_i is the subjective response category threshold on the trait scale, and M = 2*C+1. Both α_i and τ_i determine the shape of the curve. As α_i increases, the height of the curve increases (i.e., the probability of endorsement approaches 1) and the peak

becomes steeper. For example, Figure 3 plots two IRF's with different alpha values. The alpha for the solid line is three, and the alpha for the dotted line is one. The other parameters are identical. In contrast, the height of the curve increases but becomes less steep as τ_i increases. For example, Figure 4 plots two IRF's with different tau values. The tau for the solid line is -2, and the tau for the dotted line is -1. The other parameters are identical. The probability of endorsement is at maximum when θ_j is equal to δ_i. The probability decreases as the difference between the two increases. Studies by Roberts et al. (2000) and Stark et al. (2006) present more detailed descriptions of the GGUM parameters.

As noted by Stark et al. (2006), both the dominance and ideal point response models can produce similar IRF's. This is because the dominance model is a special case of the ideal point model where the ideal point approaches infinity. In this case, the curve reaches its peak at a trait level that is not typically observed in the population. For example, Figure 1 plots two IRF's based on the GGUM. The delta for the solid line is two, and the delta for the dotted line is four. The other parameters are identical. This graph shows that shifting the ideal point to a higher value causes the curve to resemble one based on a dominance model.

Modeling Applicants' and Incumbents' Item Responses

Research looking at differences between applicants and incumbents focuses on applicants' response distortion, or the idea that applicants do not answer personality items honestly. Several terms, such as response distortion, response bias, faking, and impression management, are used in the current body of research (e.g., Donovan, Dwight, & Hurtz, 2003; Griffith, Chmielowski, & Yoshita, 2007; Mueller-Hanson, Heggestad, &

Thornton, 2003; Ones, Viswesvaran,& Reiss, 1996). The impetus for this research is that applicants score higher on personality scales then do incumbents, and these differences lead to questions about the validity of these scales. Some research has shown that response distortion does not affect predictive validity (e.g., Hogan, Hogan, & Roberts, 1996; Ones, Viswesvaran, & Reiss, 1996), whereas some authors argue that faking may lead to weaker predictive validity estimates (e.g., Rosse, Stecher, Miller, & Levin, 1998). The result of these scale score differences can lead to organizations making hiring decisions that are not in line with organizational goals.

Differences between applicants' and incumbents' responses are viewed using two different paradigms (Zickar, 2000). The changing persons' paradigm of faking assumes that individuals fake by responding to items as if they had higher levels of the personality construct than they actually possess. For example, individuals may fake by endorsing Conscientiousness items that they would not endorse if they were responding honestly. In the changing items paradigm, trait levels are not affected by faking, but individuals interpret items differently and as a result there would be differences in item parameters. For example, research generally has supported the changing persons' paradigm. Several studies have indicated that applicants had higher levels of the trait compared to incumbents and that there were few differences in the item parameters (Robie et al., 2001; Stark, Chernyshenko, & Drasgow, 2004; Zickar & Robie, 1999). However, tests of the changing items paradigm have been limited to comparing item parameters of the same IRT model. To my knowledge, research has not explored the possibility that applicants and incumbents may interpret items differently and that these differences lead to different response models.

In the present study, I will use the changing items paradigm to suggest that the competitive nature of a selection process leads applicants to respond in a manner more consistent with a dominance response process whereas incumbents' responses will reflect an ideal point response process. As noted by Zickar (2000), changes in the way items are perceived may result from differences in frame of reference or the consequences of choosing particular options. Both of these differences would seem likely to be present between applicants and incumbents. For example, I suggest that when responding to some items, applicants evaluate whether they have at least the minimal level of the construct required for endorsing the item. In contrast, I believe that incumbents are more likely to respond to some items by gauging how well the item applies to them. They may not endorse the item if they feel that it reflects too little or too much of the construct. In terms of consequences of choosing particular options, applicants will likely be aware that not endorsing a positively worded item will lead to a lower score and less likelihood of being hired. In contrast, incumbents may perceive fewer negative consequences for not endorsing items. This leads to my hypothesis: Applicants' responses to personality items will exhibit less folding at high trait levels than incumbents' responses.

Method

Sample

Participants were applicants (N = 1509) and incumbents (N = 1568) who completed the 16PF Select, a shortened version of the 16PF Fifth Edition. The data were obtained from the Institute for Personality and Ability Testing. The 16PF Select contains 98 items distributed across 12 personality scales. These scales include Abstractedness (8 items, α = .75), Apprehension (8 items, α = .78), Dominance (8 items, α = .68), Emotional Stability (9 items, α = .77), Liveliness (8 items, α = .72), Openness to Change (9 items, α = .66), Perfectionism (8 items, α = .73), Rule-Consciousness (8 items, α = .75), Self-Reliance (8 items, α = .78), Social Boldness (8 items, α = .86), Vigilance (8 items, α = .73), and Warmth (8 items, α = .70) (Cattell, 2004). The 16PF Select uses a three-point scale (Agree, ?, Disagree). Consistent with Stark et al. (2006), the middle response option (?) was collapsed with the positive response option (Agree). This was necessary because the middle response option was not frequently endorsed.

For the applicant sample, 43% were female. Applicants were applying for housekeeping, hospitality, storekeeper, and wait staff positions. For the incumbent sample, 40% were female. These individuals were employed in positions including nurses and firefighters.

Analyses

Modified parallel analysis. Prior to estimating the IRT models, I examined the scales for unidimensionality using modified parallel analysis (MPA; Drasgow & Lissak, 1983). MPA was designed specifically to assess whether scales are unidimensional enough for IRT analyses. It consists of conducting a principal factor analysis on the

correlation matrix and comparing the resulting eigenvalues of the second factor against those obtained from factor analyzing data simulated to be unidimensional.

Model fit. Item fit was assessed using adjusted chi-square to degree of freedom ratios. I used GGUM 2004 (Roberts, 2001) to obtain the GGUM parameters for the applicant and incumbent samples. I also estimated dominance model fit using the 2PLM. To test this, I used Bilog (Zimowski, Muraki, Mislevy, & Bock, 2003) to obtain the 2PLM parameters. I expected that the dominance model would not fit the incumbent sample as well as the ideal point model. I expected the applicant sample to fit both models well. The fit of the models was tested using the software program Modfit, which is based on methods developed by Drasgow, Levine, Tsien, Williams, and Mead (1995). Modfit produces adjusted chi-square to degrees of freedom ratios that are based on comparing the model implied IRF's with empirical IRF's. According Chernyshenko, Stark, Chan, Drasgow, and Williams (2001), chi-square to degrees of freedom ratios below three indicate acceptable fit.

DFIT analyses. The Differential Functioning of Items and Tests (DFIT) framework (Raju, van der Linden, & Fleer, 1995) offers two indices of DIF and an index of DTF. All three indices compare true scores based on referent group item parameters with true scores based on linked focal group item parameters. At the item level, DIF can be assessed using a noncompensatory DIF index (NCDIF) or a compensatory DIF index (CDIF). The former assumes none of the other items on the test contains DIF. In contrast, CDIF does not make this assumption. The NCDIF index is the expectation over the focal group (E_F) of squared differences between the probability of endorsing an item using the focal item parameters, $P_{iF}(\theta)$, and using the referent group parameters, $P_{iR}(\theta)$. If d_i equals

the difference between the probabilities of item endorsement under the focal and referent group parameters then

$$NCDIF_i = E_F[P_{iF}(\theta) - P_{iR}(\theta)]^2 = E_F(d_i)^2 = \sigma_{di}^2 + \mu_{di}^2 \quad (3)$$

where σ and μ are standard deviations and means of d_i, respectively. Differences in true test scores can be calculated by summing the differences for the items

$$D = \sum_{i=1}^{n} d_i \quad (4)$$

DTF can be calculating by squaring these test level differences

$$DTF = E_F(D)^2 = \sigma_D^2 + \mu_D^2 \quad (5)$$

For example, to calculate DIF for an item, I used the applicant sample as the focal group and the incumbent sample as the referent group. I calculated the person parameters, or theta values, for the applicant sample and the item parameters for both the applicant and incumbent samples using GGUM. I then calculated the probabilities of endorsement using the applicant group person parameters and both sets of item parameters. Once I took the difference of the focal and referent group probability estimates, I then calculated the average squared difference plus the standard deviation of the difference squared to obtain the DIF value for each item. To obtain the DTF for each scale, I summed the probabilities, subtracted the referent sum from the focal sum, and then calculated the

average squared difference plus the standard deviation of the difference squared.

Linking. I conducted linking analyses for both models using the calibration samples for applicants and incumbents using a program called GGUMLINK (Roberts & Huang, 2003). Consistent with Raju et al. (1995) and Robie et al. (2001), I linked the incumbent parameters to the applicant parameter metric. Linking constants were obtained using an iterative process. The constants were computed using the ICC method (Haebara, 1980) and its extension to the GGUM (Koenig & Roberts, 2007) for all of the scale items. Items that exhibited DIF were removed and the ICC method was repeated until the same items were identified as having DIF on consecutive iterations.

DIF cutoffs. Although chi-square statistics have been proposed for NCDIF and DTF, these have shown to be overly sensitive (Fleer, 1993). To address this, several researchers have identified cutoff values. For NCDIF, the cutoff value for dichotomous items is .006 (Fleer, 1993). The DTF cutoff value is simply the number of items multiplied by the cutoff for NCDIF. However, these values have been shown to vary across conditions (Chamblee, 1998). In the present study, the 16PF Select items had three response categories that were collapsed resulting in dichotomous item responses. It was not clear what effect this would have on the DTF values. In addition, there have been no studies investigating cutoff values when the GGUM is estimated. Thus, I conducted several simulations to determine cutoff values for the present study. I used procedures similar to those outlined by Oshima, Raju, and Nanda (2006) to compute critical NCDIF values for each item and DTF critical values for each scale.

Oshima et al. (2006) developed the item parameter replication (IPR) method for calculating cutoffs for NCDIF statistics for dichotomous IRT models. The IPR method

uses the variance-covariance structures of estimated item parameters to produce separate cutoffs for each item. The variances and covariances are used to simulate item parameters for a large number of samples. NCDIF statistics are calculated for each sample, and then the distribution of the NCDIF values for each item is used to determine the cutoff. For example, the value that is the 95^{th} percentile would be the cutoff value for an α level of .05. I used a similar method to calculate cutoffs for the GGUM for the present study.

Results

Table 1 presents means, standard deviations, reliability (coefficient alpha) estimates, and intercorrelations of the scales for both the applicant and incumbent calibration samples. All but one, Warmth (.58), of the incumbent reliability estimates were at least .60. However, several of the scales for applicants displayed reliability estimates than were lower than population norms for the 16PF. Dominance (.48), Openness to Change (.43), Perfectionism (.56), Vigilance (.59), and Warmth (.40) had lower than expected reliability estimates for the applicant sample. One explanation for this may have been range restriction. A majority of the items (60) for applicants had endorsement rates exceeding 80%, whereas only 13 items for incumbents had similar endorsement rates.

Unidimensionality

An MPA was conducted for each scale separately for applicants and incumbents. Using the criteria outlined by Drasgow and Lissak (1983), I compared eigenvalues of simulated and real data. Table 2 displays the eigenvaules for the simulated and real data for each scale. Results for all of the scales indicated sufficient unidimensionality. The eigenvalues for the second factor for the 16PF Select scales were similar to those for the simulated data.

Model Fit

Table 3 presents the number of items for each measurement scale, the mean adjusted χ^2/df ratios for the applicant and incumbent samples for both the GGUM and the 2PLM. Initial calibrations of the GGUM model revealed unacceptably large standard errors for the item parameters. To address this, the tau parameters were constrained to be

equal across items within each scale. This yielded estimates with much smaller standard errors. For applicants, the mean adjusted χ^2/df ratios were 2.60 for the GGUM. For incumbents, the mean adjusted χ^2/df ratios were 2.13 for the GGUM. Only one scale, Apprehension, did not have acceptable model fit for the GGUM.

Table 3 presents the mean adjusted χ^2/df ratios for the applicant and incumbent samples for the 2PLM. For applicants, the mean adjusted χ^2/df ratios were 1.46. For incumbents, the mean adjusted χ^2/df ratios were 2.40. As expected, all of the scales had acceptable model fit for the applicant sample. However, only four of the scales (Emotional Stability, Liveliness, Rule-consciousness, and Warmth) did not have acceptable model fit for the incumbent sample.

Overview of DFIT Analyses

DTF was observed for three of the 12 scales (Dominance, Liveliness, and Warmth). Although, significant DTF was not found for many of the scales, I examined differences in applicant and incumbent test characteristic curves (TCC's) for each of the scales (except for the Abstractedness and Apprehension scales) to gauge how the expected test score related to trait levels (see Figures 9-18). I did not include TCC's for the Abstractedness scale because no items had DIF and I did not include TCC's for the Apprehension scale because it did not have acceptable model fit. The DTF index is averaged across all levels of the trait. In my study, I primarily expected differences for higher values of the trait and such differences may be present even when the DTF index does not exceed the critical values. The TCC's for four of the scales resembled the hypothesized differences. Inspection of the curves also demonstrates why significant

DTF was not detected: There are little differences between applicants and incumbents for low to moderate levels of the trait resulting in a low average difference between groups.

The results of the DFIT analyses are presented in Table 4. A total of 51 items exhibited DIF according to the NCDIF index. The form of DIF fell into four distinct patterns: in the hypothesized direction (Figure 5), opposite of the hypothesized direction (Figure 6), both curves exhibiting folding (Figure 7), and neither curve exhibiting folding (Figure 8). Table 5 presents the frequencies of the items with DIF. Twelve of these (24%) had IRF's in the hypothesized direction, where the incumbent IRF demonstrated more folding than the applicant IRF. Eight (16%) of the items with DIF were opposite of the hypothesized direction, where the applicant IRF demonstrated more folding than the incumbent IRF. Six of the curves that were in the opposite direction came from one scale, Dominance. Five (10%) of the items with DIF had IRF's where both applicants and incumbents demonstrated folding. Twenty-six (51%) of the items with DIF had IRF's where neither applicants nor incumbents demonstrated folding. These IRF's are more representative of a dominance model than an ideal point model.

DFIT of Dimensions

Abstractedness. None of the items for this scale had DIF and therefore, the DTF for this dimension was not significant.

Apprehension. The Apprehension scale did not have acceptable model fit for either the applicants (14.55) or incumbents (4.71). I did not look at DIF or DTF since this scale did not fit the model.

Dominance. Seven of the items for this scale had DIF. Six of the items were opposite of the expected direction and the other item showed no folding for either group.

Figure 9 shows the TCC's for this scale. The DTF for this dimension was significant; however, the curves are opposite of the expected direction.

Emotional Stability. Three of the items for this scale had DIF. One item was in the hypothesized direction and two of the items were similar to Figure 8, where neither applicants nor incumbents demonstrated folding. Figure 10 shows the TCC's for this scale. DTF was not significant.

Liveliness. Six of the items for this scale had DIF. Three items were in the hypothesized direction, two items were similar to Figure 8, where both IRF's looked more like a dominance model, and one item had IRF's where both groups demonstrated folding. Figure 11 shows the TCC's for this scale. DTF was significant, therefore, the Liveliness scale showed support for my hypothesis.

Openness to Change. Seven of the items for this scale had DIF. Six of the items had IRF's that did not demonstrate folding and one item had IRF's similar to Figure 7, where both applicants and incumbents demonstrated folding. Figure 12 shows the TCC's for this scale. DTF was not significant.

Perfectionism. Five of the items for this scale had DIF. Two of the items were opposite of the hypothesized direction, two of the items have IRF's where both groups reflected a dominance model, and one item showed both groups folding, similar to Figure 7. Figure 13 shows the TCC's for this scale. DTF was not significant.

Rule-consciousness. Four of the items for this scale had DIF. One item was in the hypothesized direction, as in Figure 5, and the remaining three items showed curves where both applicants and incumbents exhibited folding. Figure 14 shows the TCC's for this scale. DTF was not significant.

Self-reliance. Five of the items for this scale had DIF. One item was in the hypothesized direction, three items were showed neither group demonstrated folding, and one item showed both groups folding. Figure 15 shows the TCC's for this scale. DTF was not significant.

Social Boldness. Four of the items for this scale had DIF. Two items were in the hypothesized direction, and the other two items were more like Figure 8, where neither group showed signs of folding. Figure 16 shows the TCC's for this scale. DTF was not significant.

Vigilance. Five of the items for this scale had DIF. Four of the items showed no folding for either group, and one item was more similar to Figure 7, where both groups showed folding. Figure 17 shows the TCC's for this scale. DTF was not significant.

Warmth. Five of the items for this scale had DIF. Four of the items were in the hypothesized direction, similar to Figure 5, and the other item showed neither group demonstrated folding. The DTF for this scale was significant (Figure 18) and the TCC's show greater folding for incumbents. Overall, results for the Warmth scale were in the hypothesized direction.

Discussion

In the current study, I examined the differences between applicant and incumbent response patterns to a personality measure. The results showed that out of the 12 dimensions of the 16PF Select, only Dominance, Liveliness, and Warmth were significantly different between the two groups. And only Liveliness and Warmth were in the hypothesized direction. This provides some support for my hypothesis that incumbents will show more folding at higher levels of the trait than applicants. I tested this assumption using an ideal point response model. This model is flexible enough to resemble either an ideal point or a dominance model. These results do not suggest that applicants never use an ideal point response process. As evidenced by the Dominance scale, applicants may also express folding at higher trait levels. However, further support for my hypothesis is that four of the twelve scales did not fit the simpler dominance model but did fit the ideal point model. And, two of those had significant DTF using the GGUM. This supports the idea that incumbents are more likely to respond using an ideal point model than a dominance model, and that applicants are more likely to respond using a dominance model.

One reason for the lack of support for my hypothesis may be the use of the 16PF Select, which was based on dominance model assumptions. Many of the personality measures used in selection are based on these assumptions (i.e., Hogan Personality Inventory, NEO-PI). However, the problem is that folding is less likely when scales are developed based on a dominance response process. Of all the items with DIF, 51% did not show folding for either group. A scale developed using ideal point model ideally addresses both high and low levels of the trait, like the dominance model, but it also

includes more neutral items. This offers a more complete collection of items that spans a broader range of trait values than the dominance model.

Another problem that stems from this is that folding should only occur for more neutral items. Both the ideal point and the dominance models do well identifying items at extremely high or extremely low levels of the trait. In those cases, we would not expect to see any folding occurring. However, only the ideal point model has the ability to distinguish moderate levels of theta. At these levels of theta, we would expect to see instances of folding. In my study, it is interesting that folding occurred given the overall dominance of this data set.

It is interesting to speculate about dimensions of the 16PF Select individually. The dimension Dominance, for example, measures a continuum with deferential and dominant on opposite poles. When examining the scales with DTF, I argue that it makes sense that Dominance is in the opposite direction of what I hypothesized. Applicants should want to appear like they listen and can take direction well, instead of always wanting to take charge of a situation. However, applicants would likely rather be seen as more warm than reserved (Warmth) and more lively than serious (Liveliness) when applying for a job. Based on this more thorough review of the scales, it seems that I should have looked at each scale individually to determine a hypothesized direction rather than a blanket hypothesis about twelve different dimensions. That is, the differences between applicants and incumbents may be scale-specific.

Whereas the remaining scales did not have significant DTF, several of the IRF's (Emotional Stability, Self-reliance, and Social Boldness) showed slight incumbent folding at higher trait levels. Again, I would argue that applicants would like to appear

more emotionally stable than reactive, more self-reliant than group-oriented, and more socially bold than reserved. These qualities would likely change depending on the position the applicants were trying to obtain. For example, if someone were applying for a job to work on a specific group project, they would be less likely to endorse the Self-reliance dimension.

Although only three scales had significant DTF, I still found that 51 out of the 90 items did have DIF. This is contradictory to what some researchers have found. Previous research did not find such differences on personality items between applicants and incumbents (Robie et al., 2001; Stark et al., 2004; Zickar & Robie, 1999). One reason this could be is that I used an empirically-derived cutoff value to determine DIF. Mead, Lautenschlager, and Johnson (2007) found that using the initial cutoff may be too conservative and lead to under identification of items with DIF. Several researchers have suggested the use of empirically-derived cutoffs like those used in this study (Chamblee, 1998; Fleer, 19993).

Incumbents are often used in the development of personality assessments. Previous research has shown that differences between applicants and incumbents do not affect the psychometric properties of personality scales (e.g., Robie, Zickar & Schmit, 2001; Zickar, Gibby, & Robie, 2004). However, it is important to note that the extent to which incumbents use an ideal point response process would lead to underestimates of the criterion-related validity. Furthermore, this research suggests that scales based on dominance model assumptions and using incumbent samples would have similar psychometric properties when given to applicants.

Limitations and Future Research

A more thorough examination of my hypothesis would be to conduct the same analysis using a personality measure developed based on an ideal point response process. Chernyshenko et al. (2007) developed a personality test based on an ideal point response process; however, this is the only one of its kind, to my knowledge. It will take a long time to collect enough data to conduct an IRT-based analysis on the new scale because this type of analysis requires a very large sample size. Also, this measure has not been used for selection, so it is impossible to compare applicants and incumbents.

Another limitation of this research was the use of a between subjects design. This limits the ability to compare the applicant and incumbent samples. Applicants were applying for jobs such as housekeeping, hospitality, and wait staff whereas incumbents held positions such as nurses and firefighters. The incumbent positions arguably require more training than the positions applicants were applying for; however, I cannot see any reason that would lead the applicants to use a dominance response process or lead incumbents to use an ideal point response process. Regardless, future research should use a matched sample, where applicants are applying for positions held by the incumbents.

A final limitation of this study is that DIF cannot be detected at trait levels that are not observed in the data set. This poses a problem for finding DIF at higher levels of the trait. For example, an item from the Liveliness scale showed a large separation of the applicant and incumbent lines at higher values of the trait, but this item did not have DIF. It would be useful to have a way to calculate DIF for targeted trait levels. This would expand our ability to detect DIF at all levels, or more specific levels, of the trait instead of the limited range we have now.

References

Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology, 66,* 1-6.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81,* 261-272.

Chamblee, M. C. (1998). A Monte Carlo investigation of conditions that impact Type I error rates of DFIT. Unpublished doctoral dissertation, Georgia State University.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research¸ 36,* 523-562.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88-106.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81-106.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*. *68,* 363–373.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting

polytomous items response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Edwards, A. L., & Kenney, K. C. (1946). A comparison of the Thurstone and Likert techniques of attitude scale construction. *Journal of Applied Psychology, 30*(1), 72-83.

Fleer, P. F. (1993). A Monte Carlo investigation of the conditions that impact Type I error rates of DFIT (Doctoral dissertation, Illinois Institute of Technology, 1993). *Dissertation Abstracts International*, 54-04, 2266B.

Goldberg, L. R. (1999, March 17). *IPIP*. Retrieved May 4, 2003 from http://ipip.ori.org/lew.

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

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist, 51*(5), 469-477.

Hough, L. M. (1998), Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*, 109-244.

Koenig, J. A., & Roberts, J. S. (2007). Linking parameters estimated with the generalized graded unfolding model: A comparison of characteristic curve methods. *Applied Psychological Measurement, 31*¸504-524.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5-53.

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

Mead, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430-455.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348-355.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660-679.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

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.

Roberts, J. S. (2001). GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 25*, 38.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory

model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Roberts, J. S., & Huang, C. (2003). GGUMLINK: A computer program to link parameter estimates of the generalized graded unfolding model from item response theory. *Behavior Research Methods, Instruments & Computers, 35*(4), 525-536.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59*, 211-233.

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.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*(4), 634-644.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrica Monographs*, *34*.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89,* 497-508.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied*

Psychology, *91*, 25-39.

Stokes, G. S., Hogan, J. B., & Snell, A. F. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: the influence of social desirability. *Personnel Psychology*, *46*, 739-62.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529-554.

Viswesvaran, C., & Ones, D. S. (1999). Meta-Analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *54*, 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*, 168–190.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [Computer Software]. Chicago, IL: Scientific Software International.

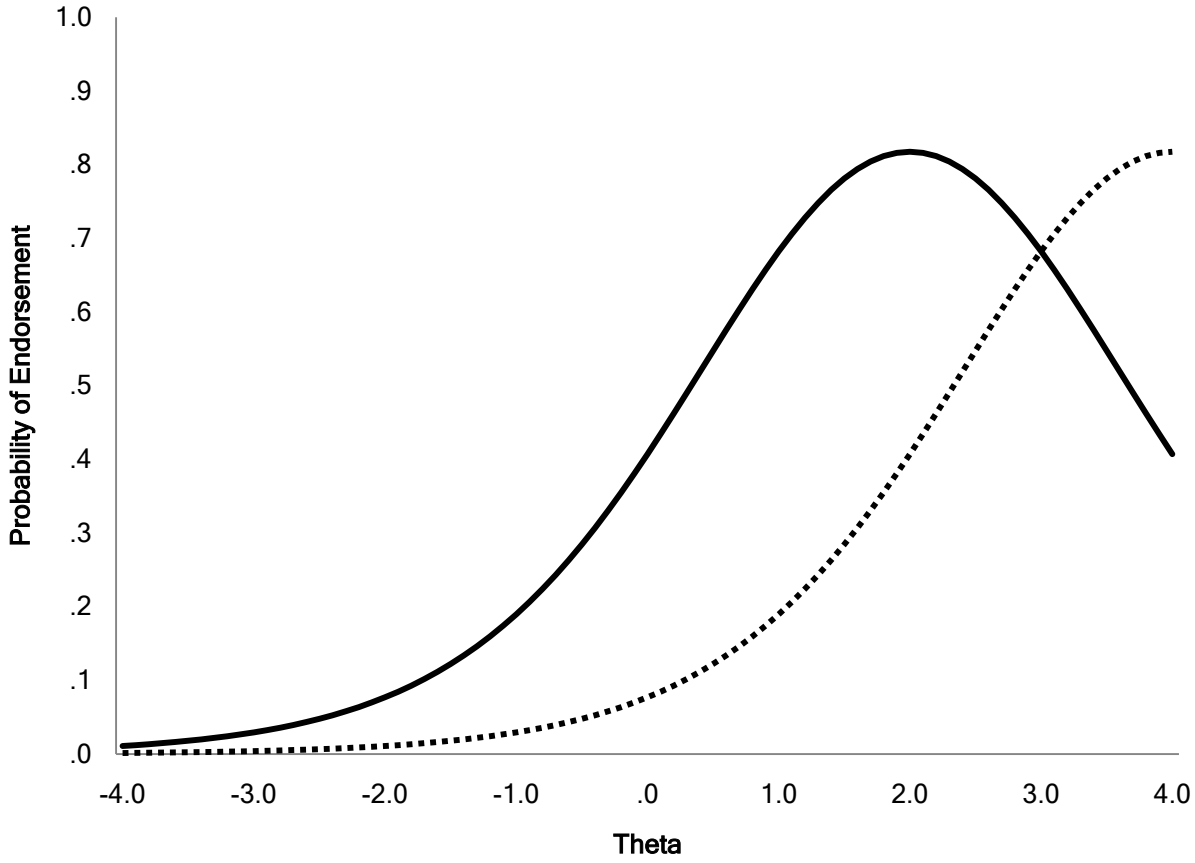

Figure 1. Item response functions based on two different ideal points. For both curves, $\alpha = 1$, $\tau = -1$. The δ for the solid line equals 2; the δ for the dotted line equals 4.

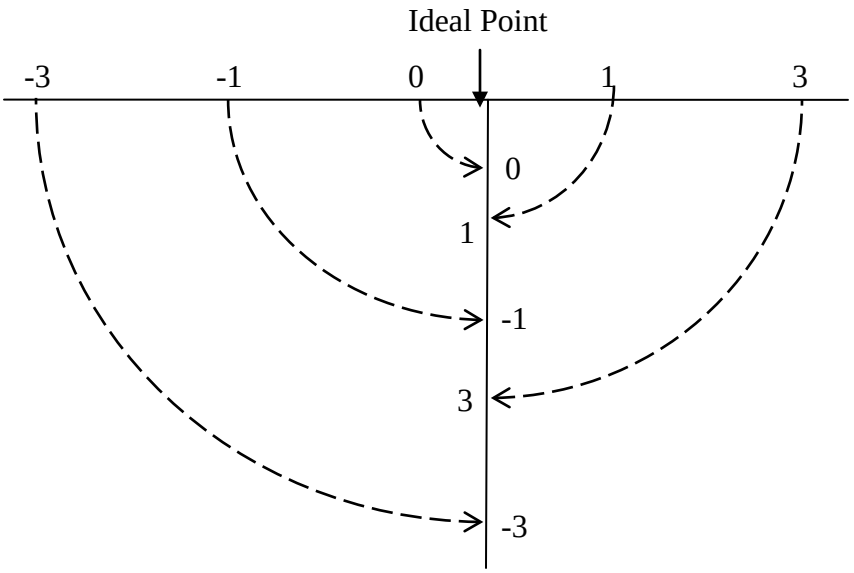

Figure 2. Unfolding technique for ideal point responses.

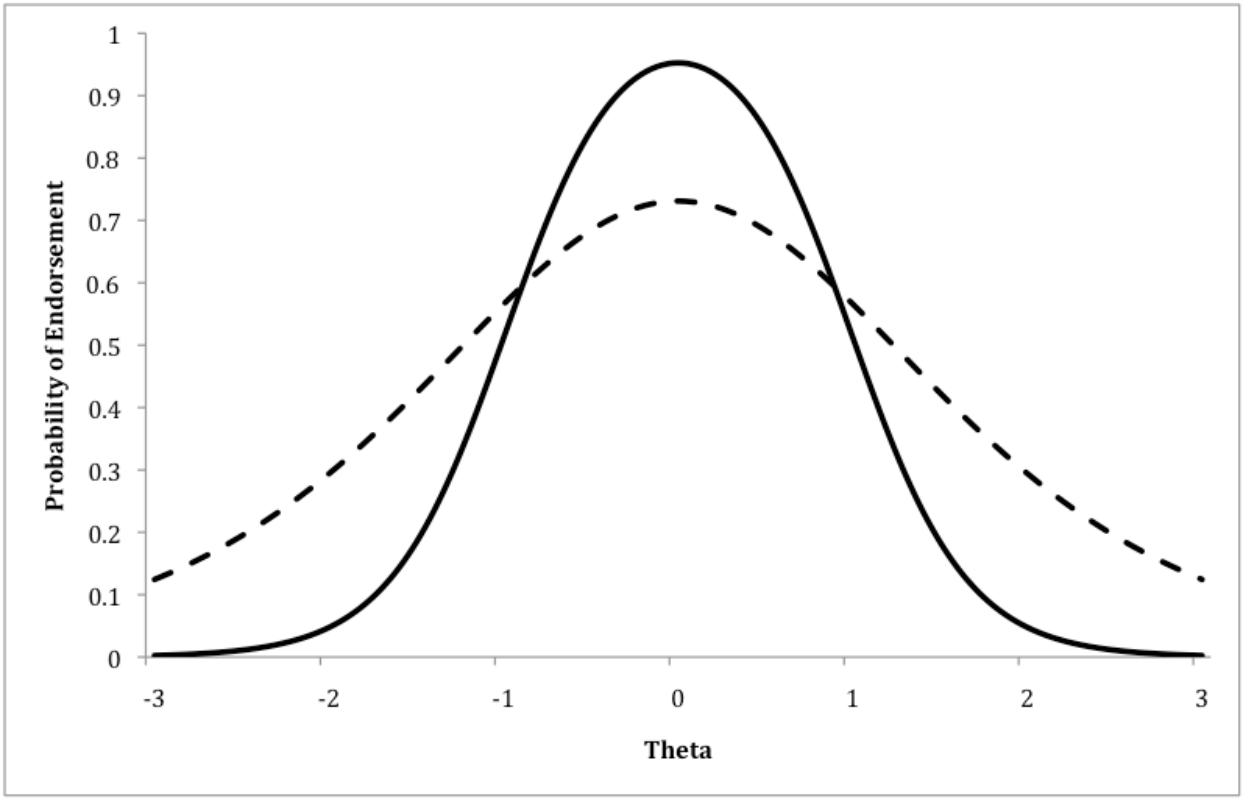

Figure 3. Item response functions based on two different alpha parameters. For both curves, $\delta = 0$, $\tau = -1$. The α for the solid line equals 3; the α for the dotted line equals 1.

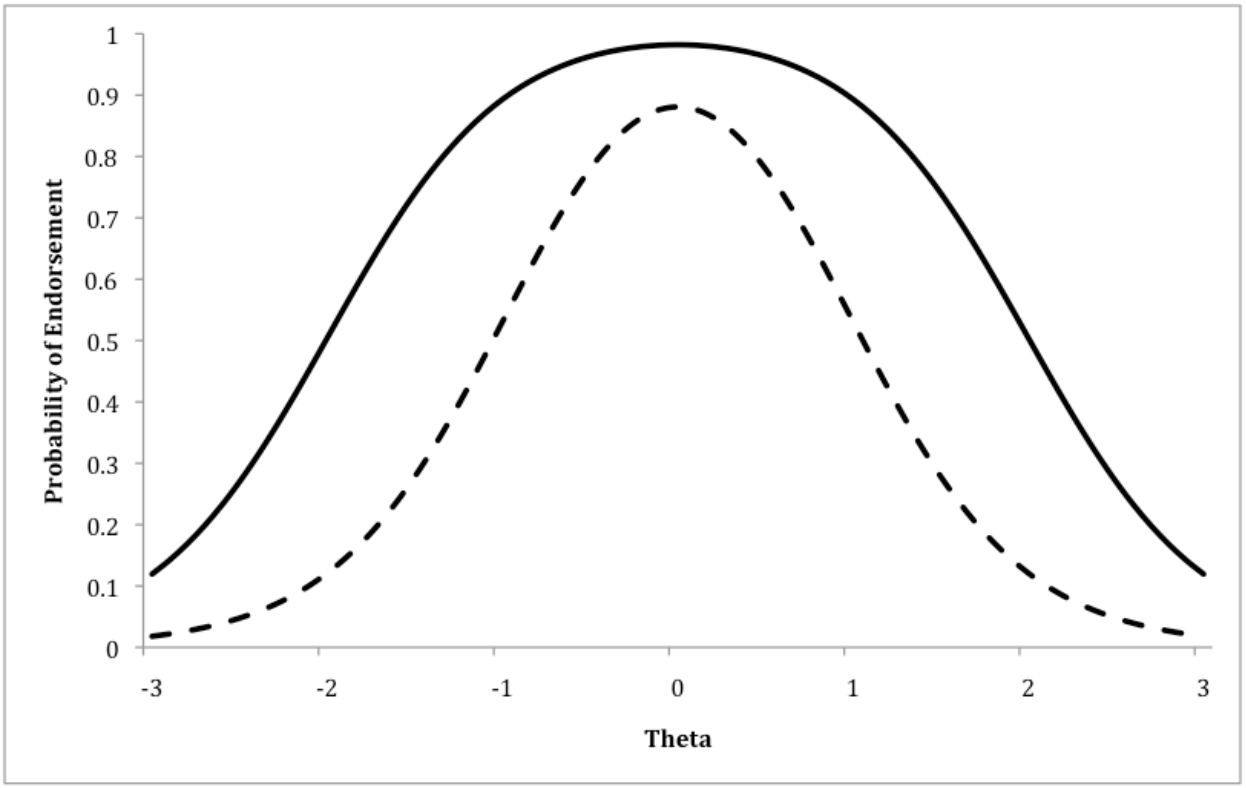

Figure 4. Item response functions based on two different tau parameters. For both curves, $\alpha = 2$, $\delta = 0$. The τ for the solid line equals -2; the τ for the dotted line equals -1.

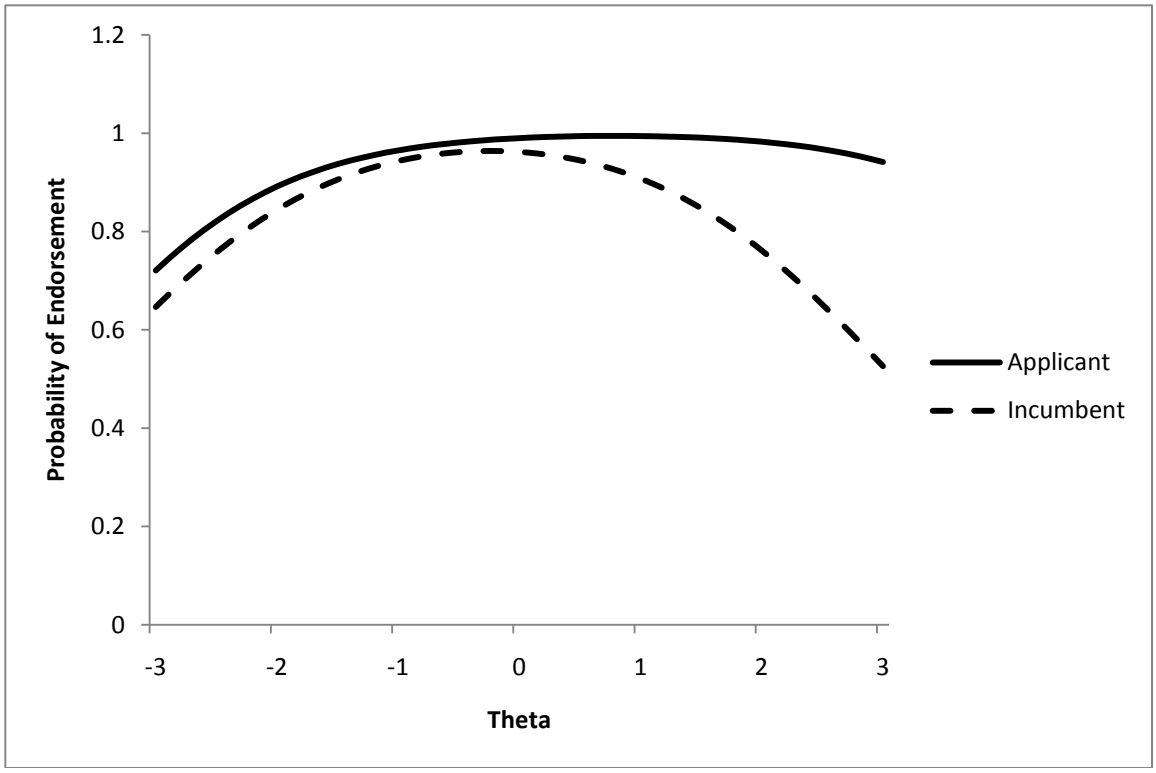

Figure 5. IRF's for an item from the Warmth scale.

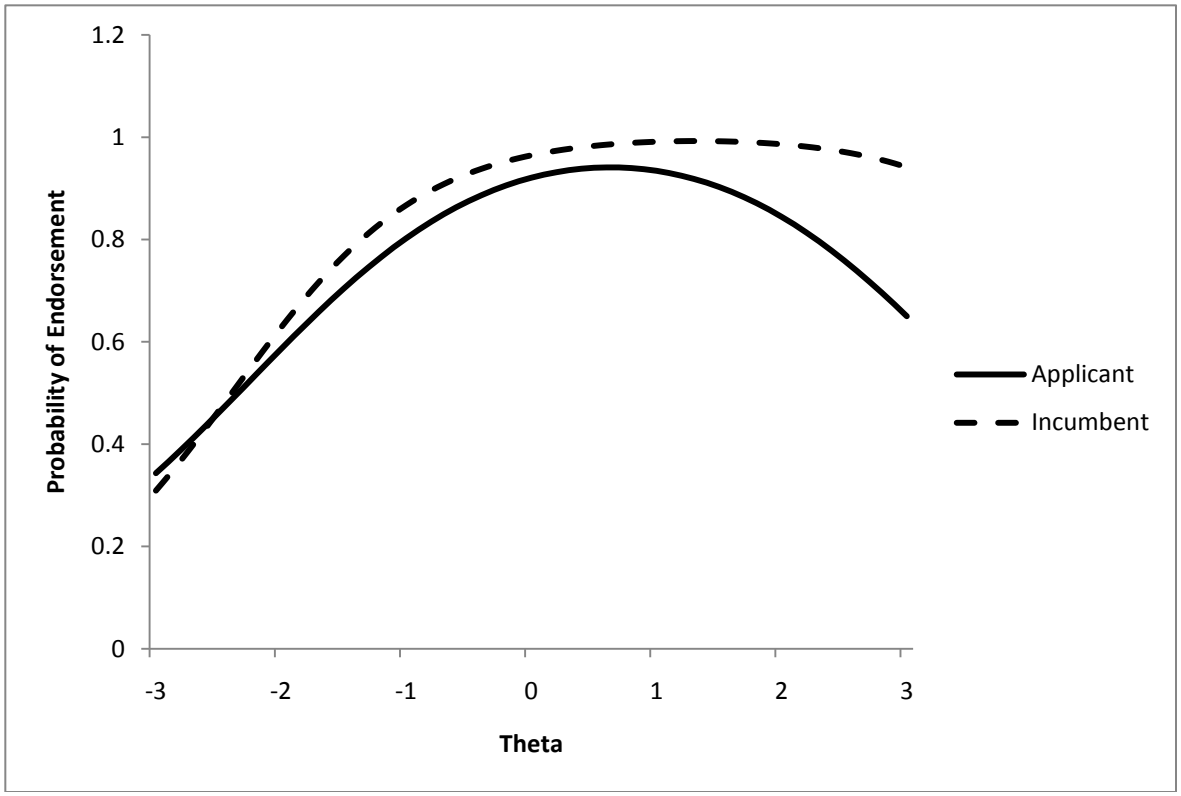

Figure 6. IRF's for an item from the Dominance scale

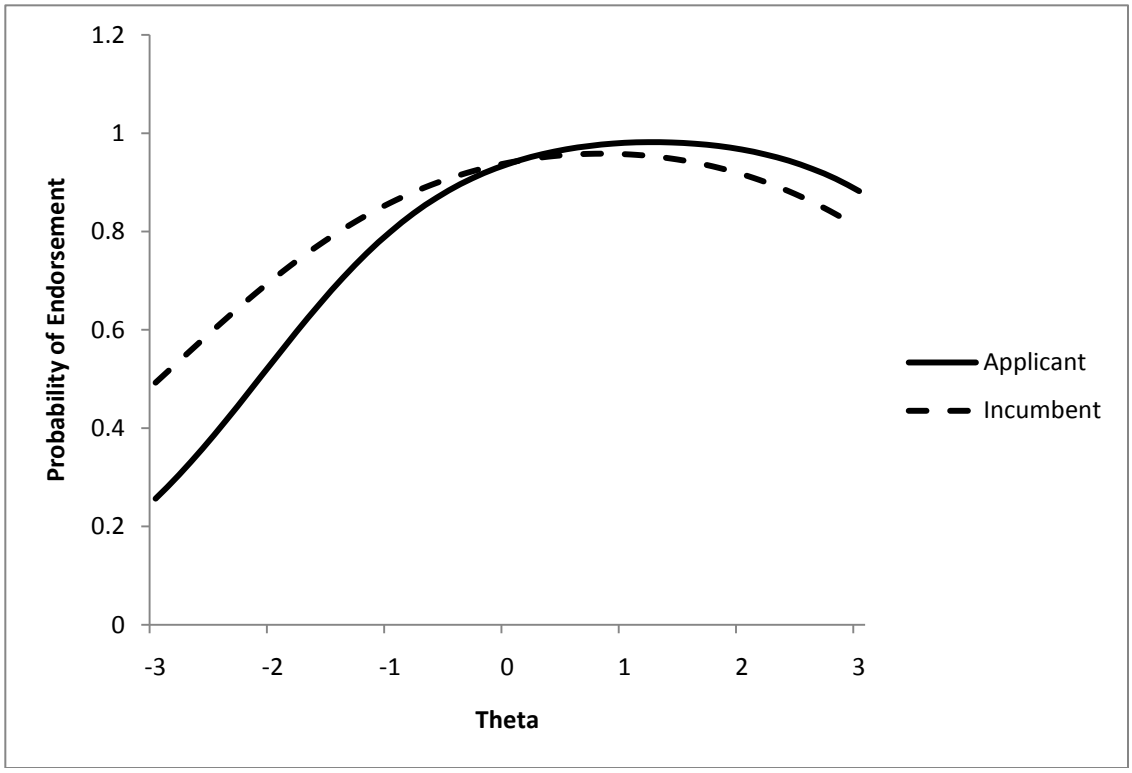

Figure 7. IRF's for an item from the Perfectionism scale

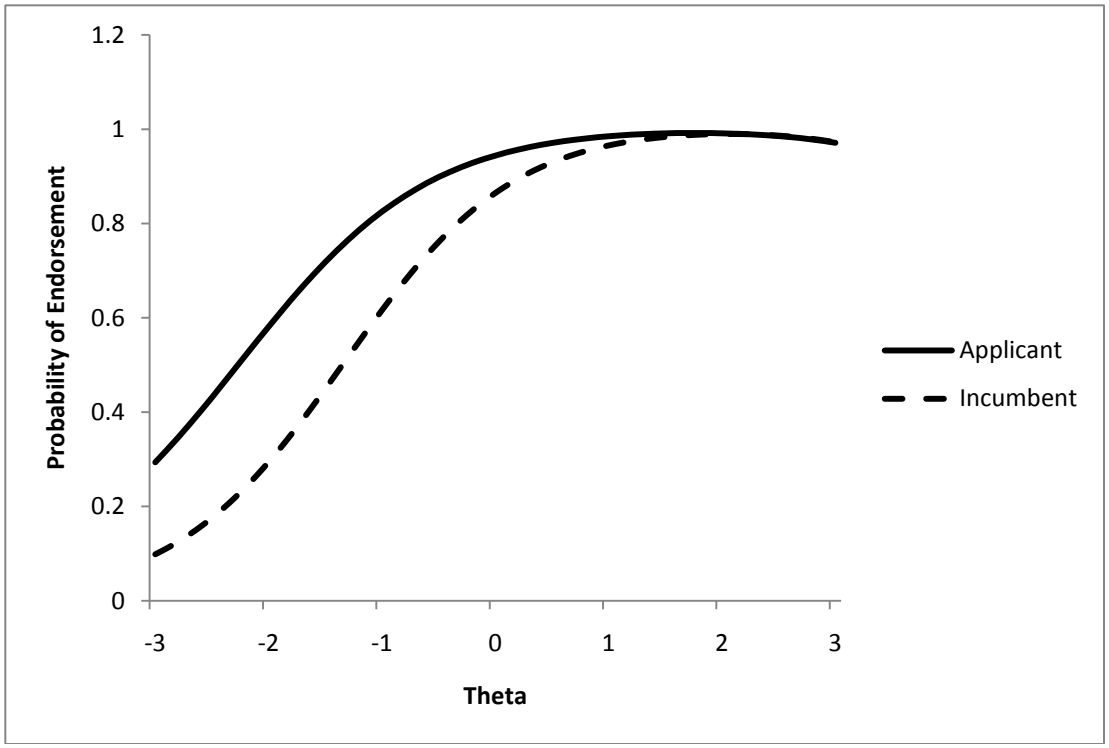

Figure 8. IRF's for an item from the Liveliness scale

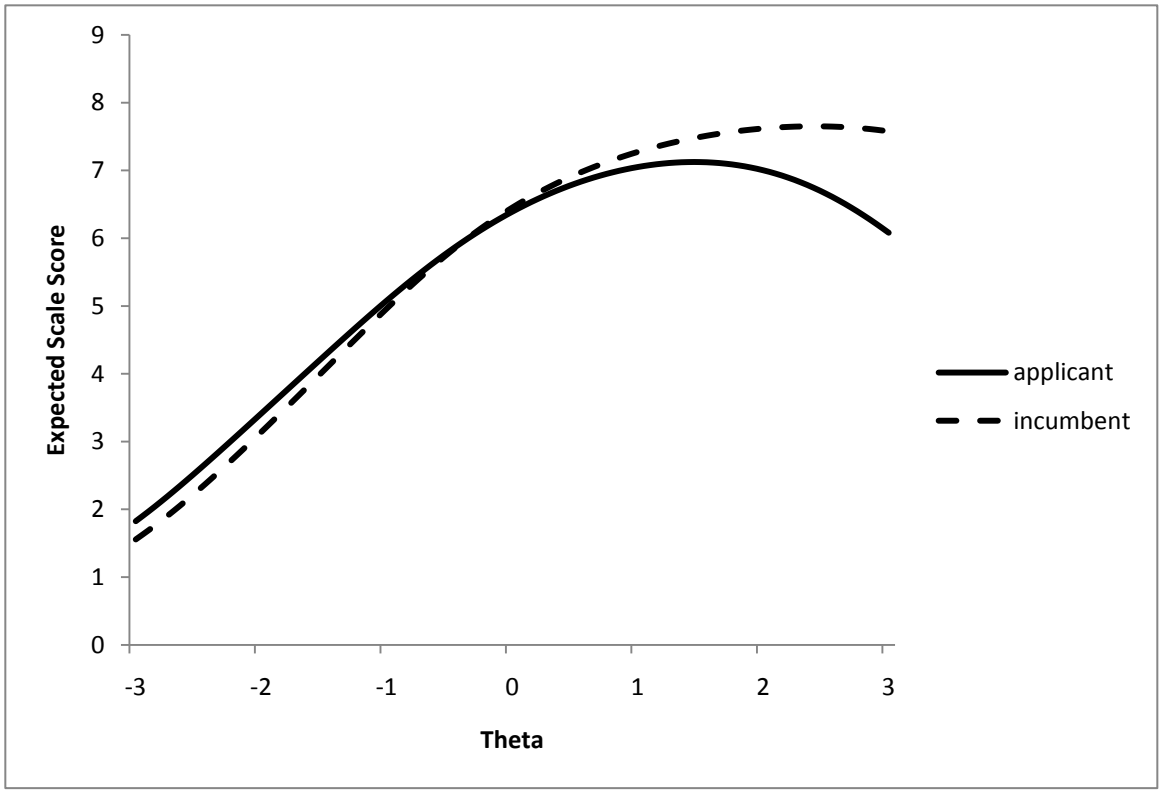

Figure 9. TCC's for the Dominance scale.

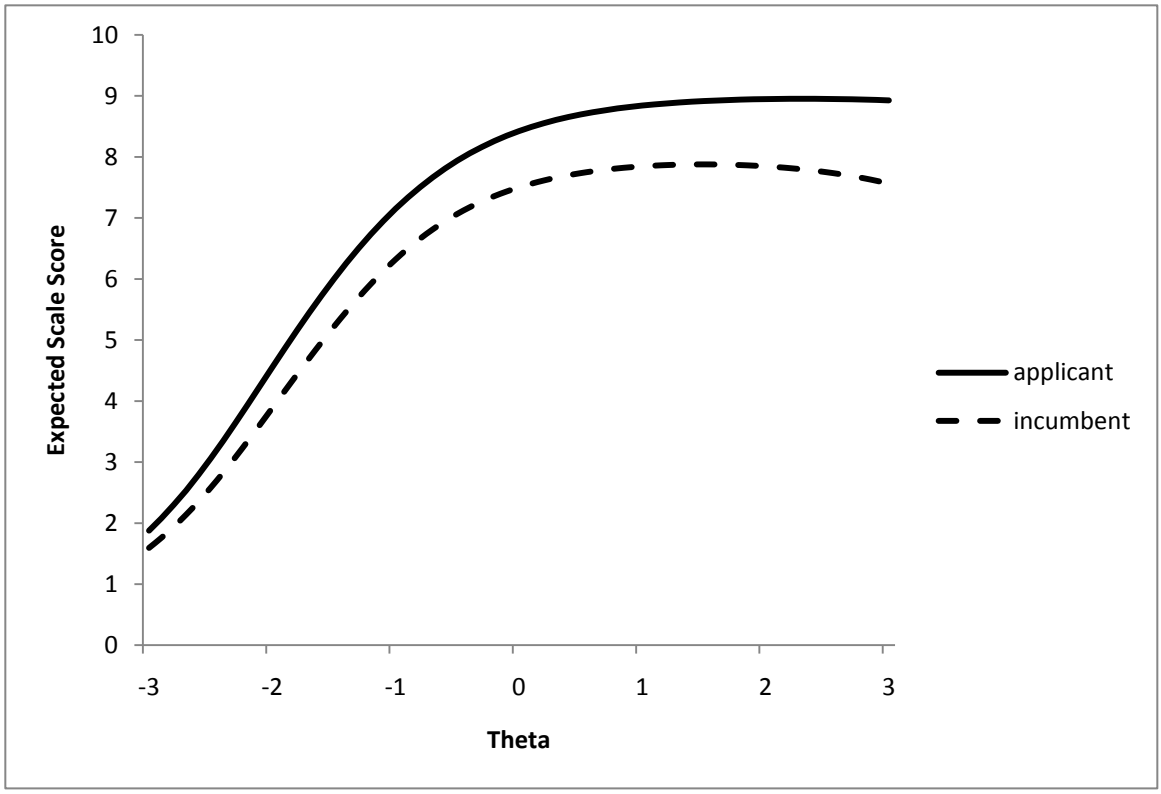

Figure 10. TCC's for the Emotional Stability scale.

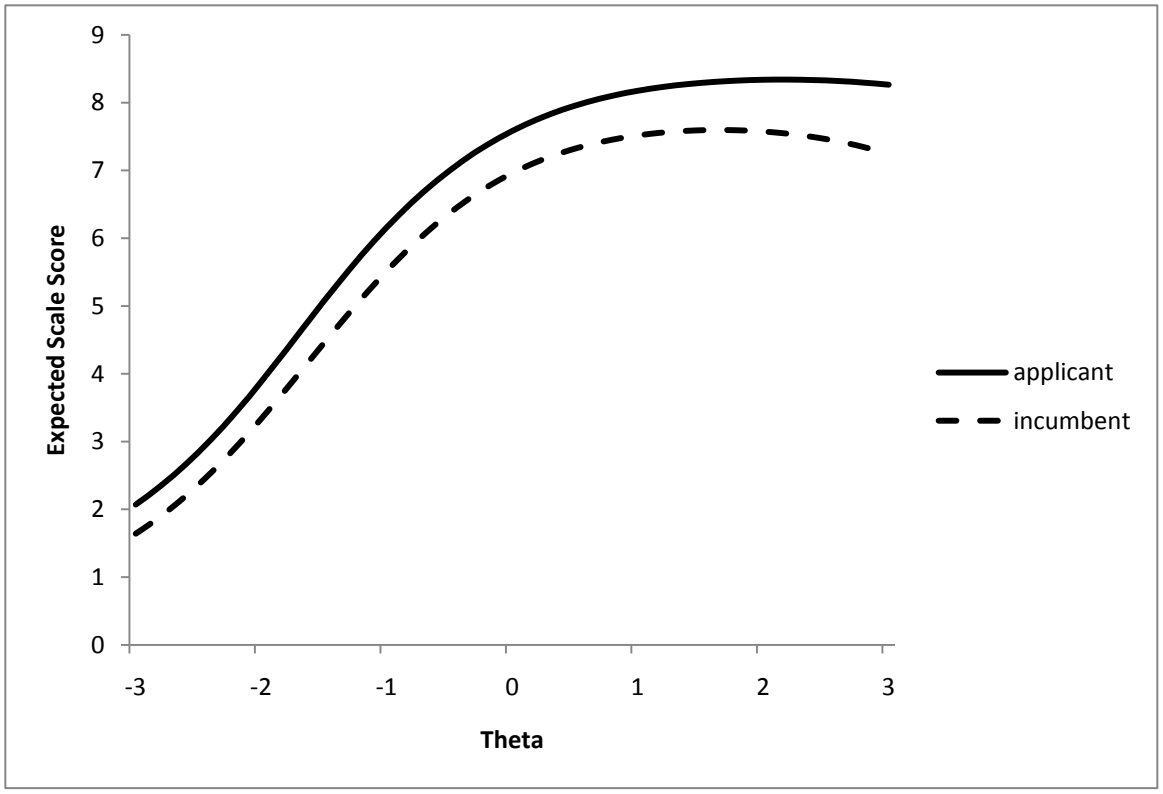

Figure 11. TCC's for the Liveliness scale.

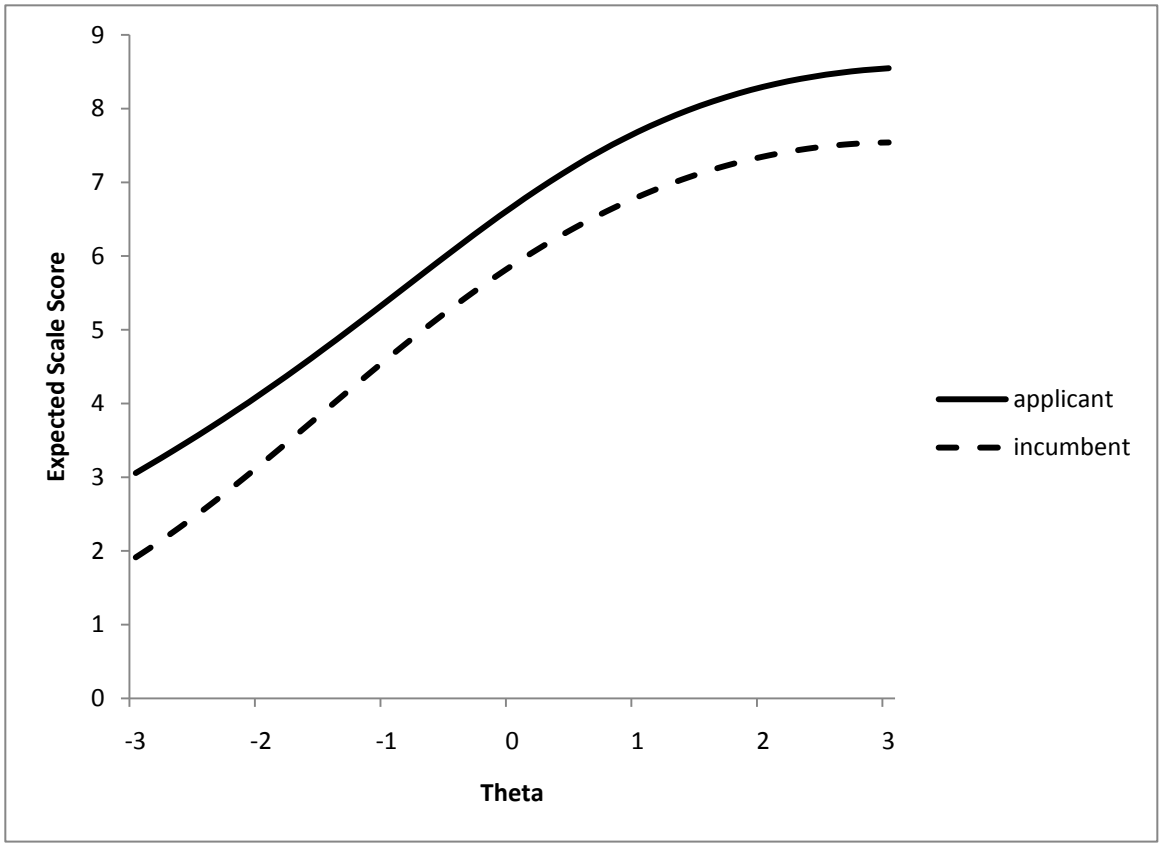

Figure 12. TCC's for the Openness to Change scale.

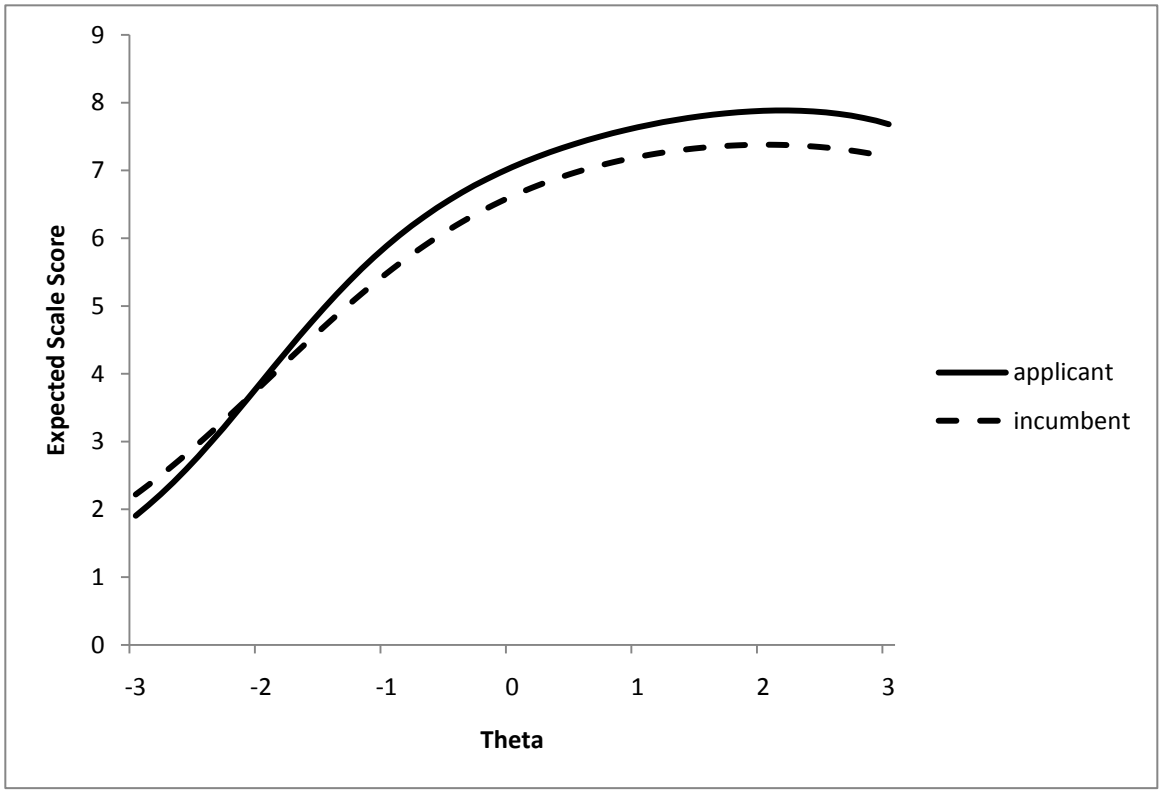

Figure 13. TCC's for the Perfectionism scale.

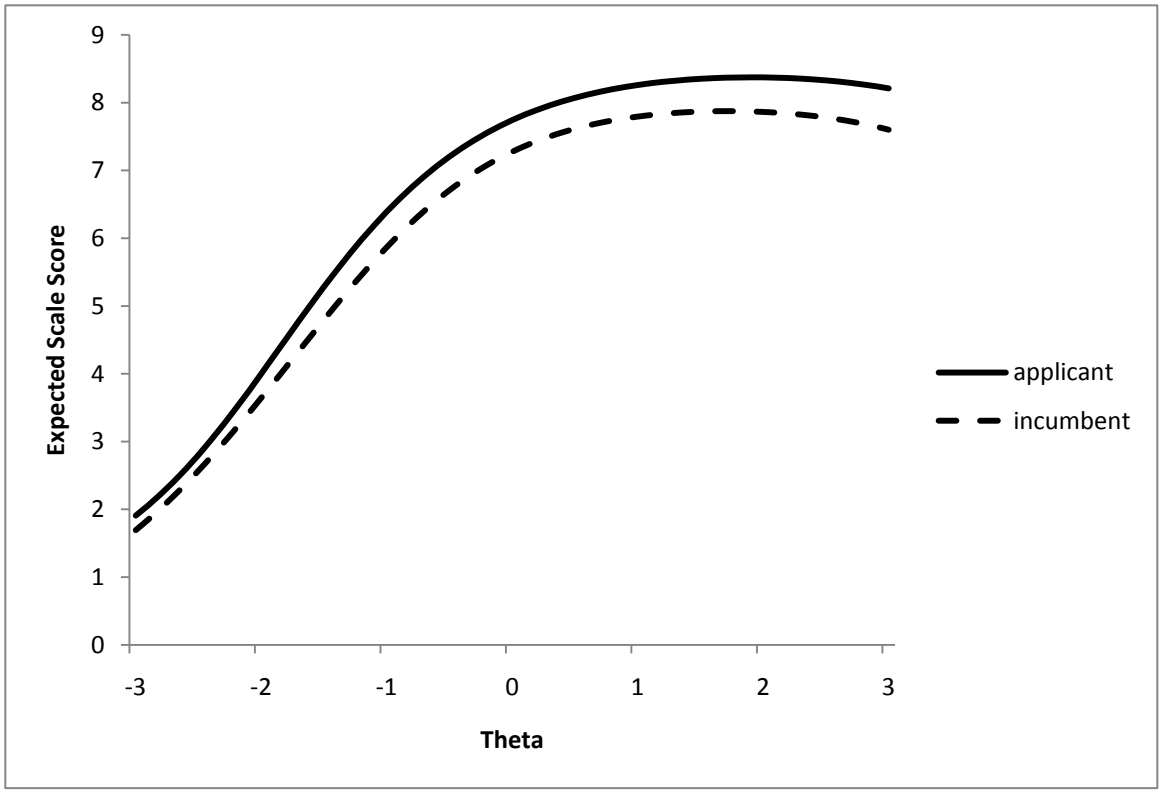

Figure 14. TCC's for the Rule-consciousness scale.

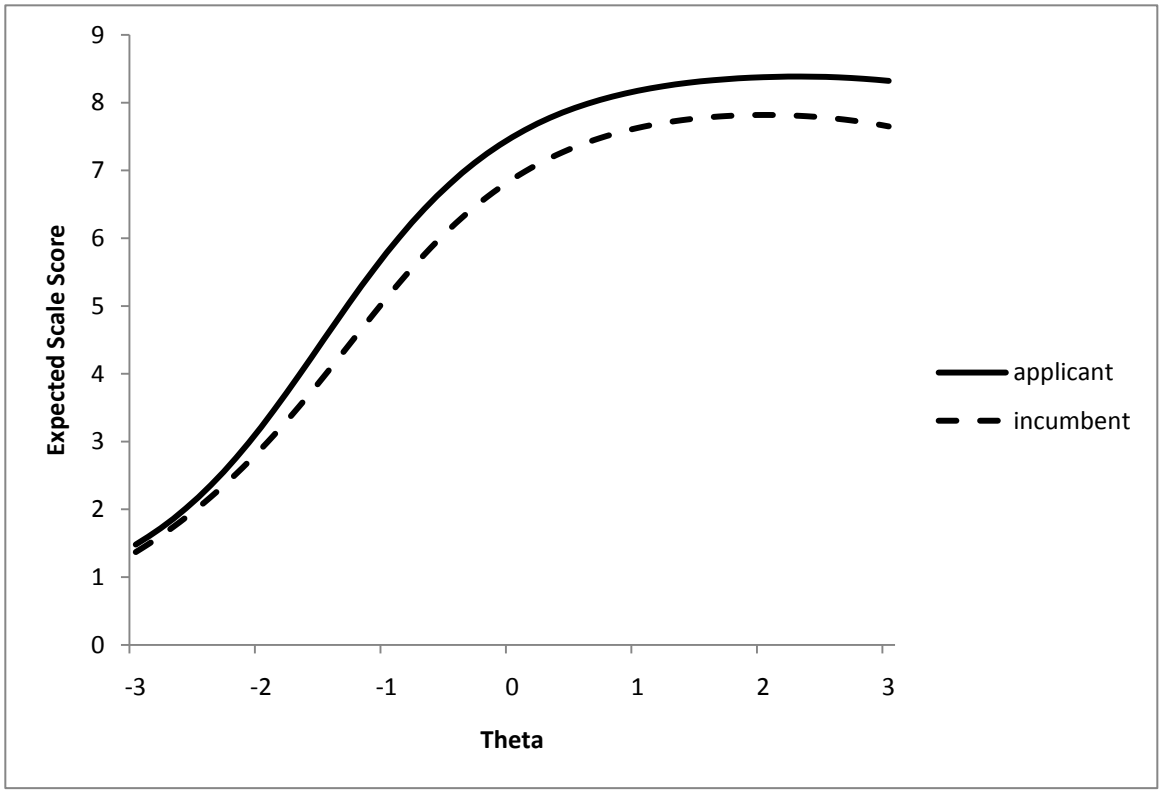

Figure 15. TCC's for the Self-reliance scale.

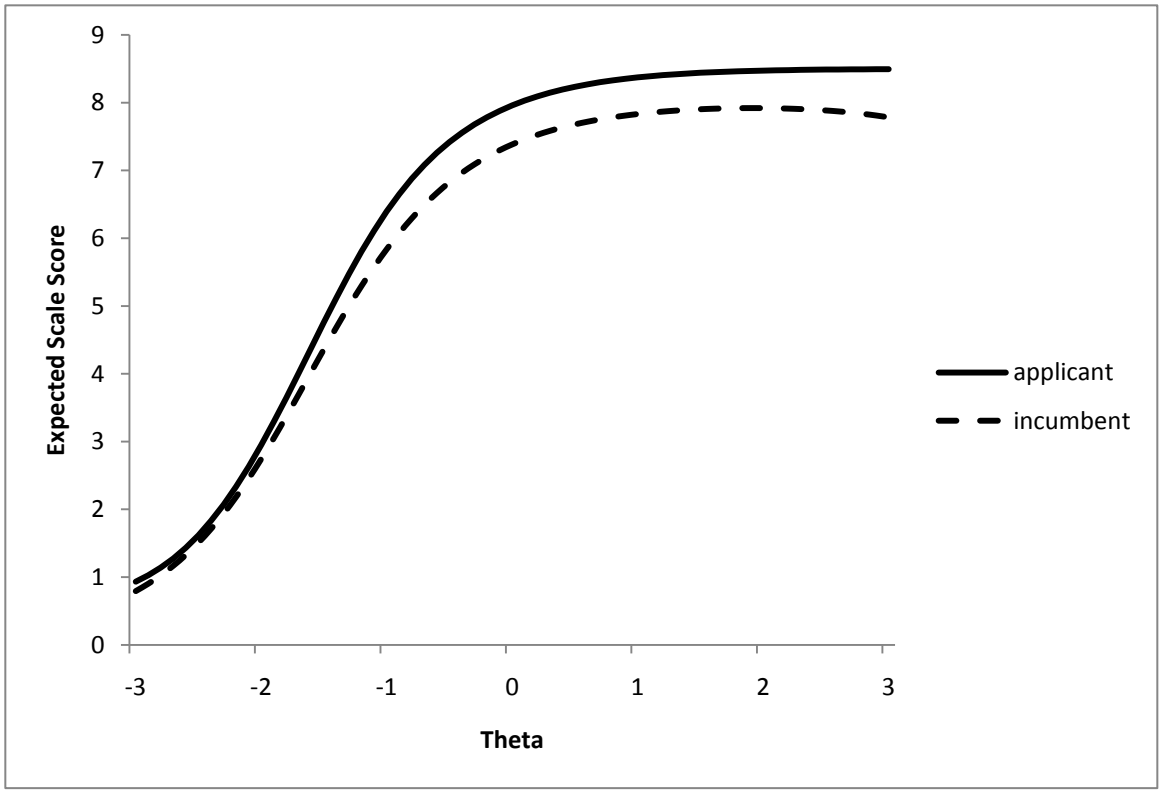

Figure 16. TCC's for the Social Boldness scale.

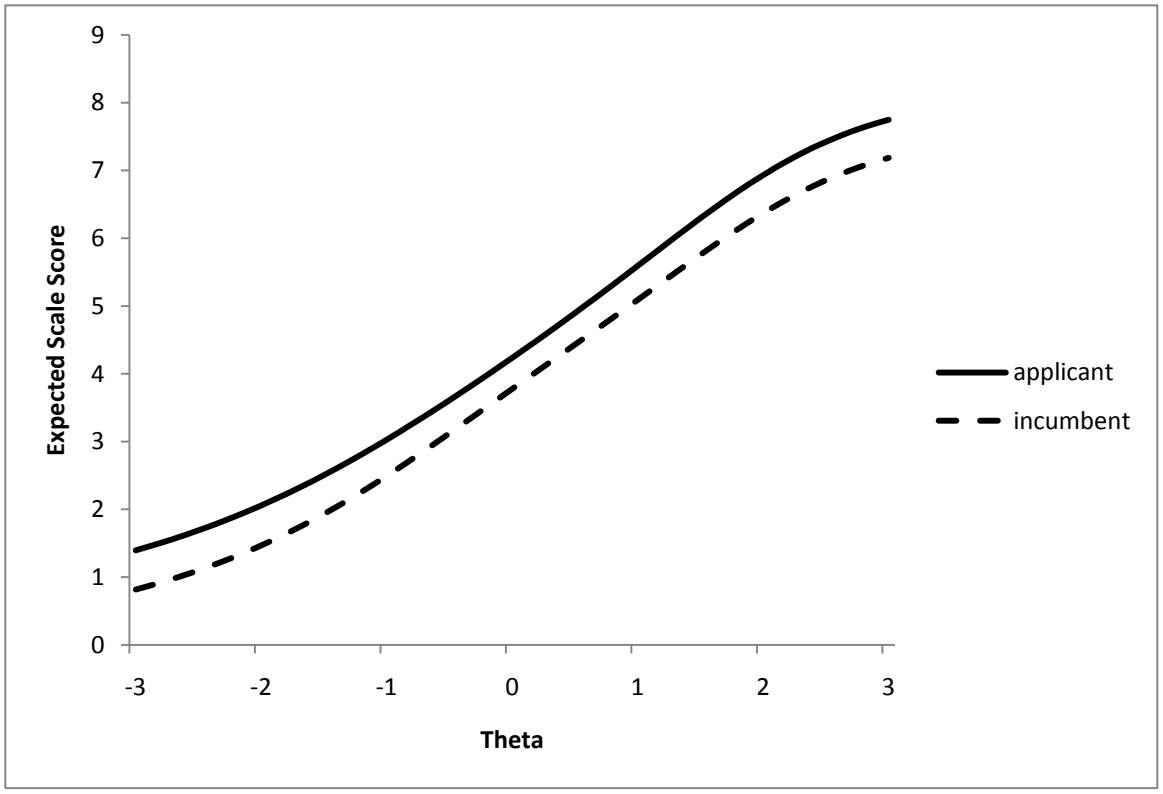

Figure 17. TCC's for the Vigilance scale.

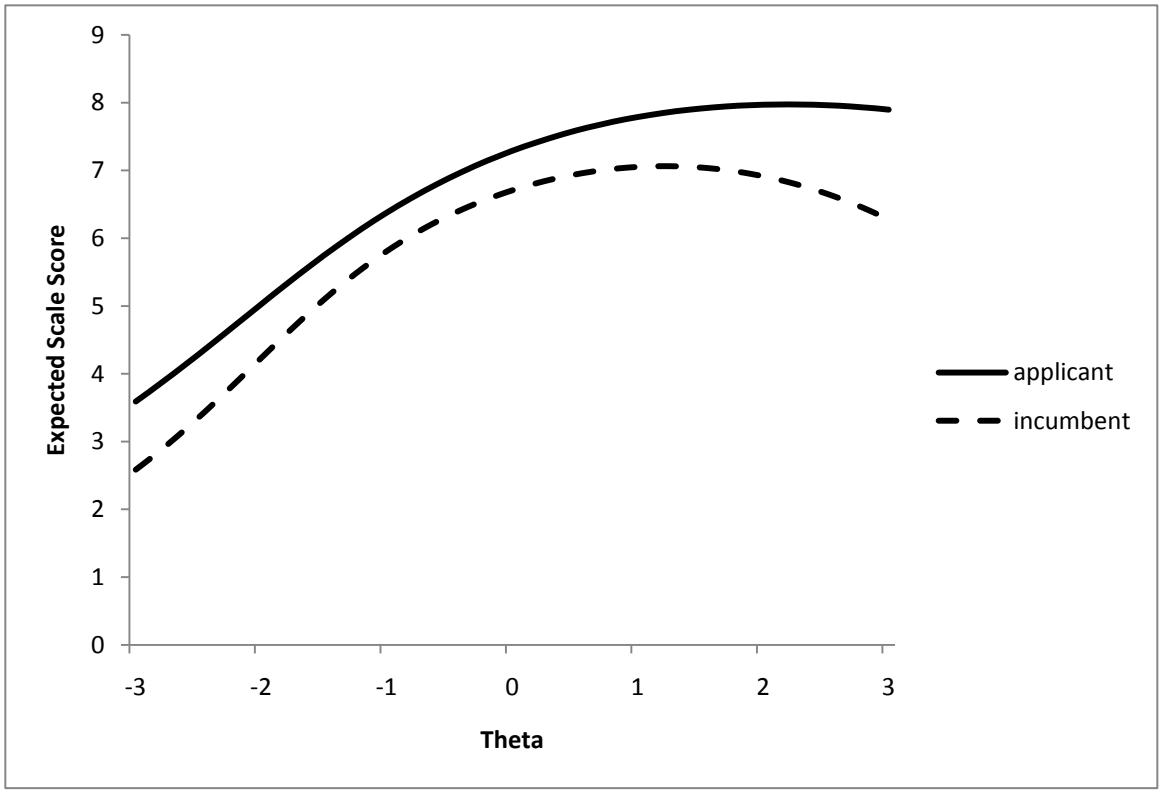

Figure 18. TCC's for the Warmth scale.

Table 1

Means, Standard Deviations, Reliability Estimates, and Intercorrelations for the 16PF Select Scales

Scale	Mean	SD	Reliability	1	2	3	4	5	6	7	8	9	10	11	12
1. Abstractedness	.80 (.76)	.22 (.24)	.68 (.70)	-	0.25	0.14	0.35	0.00	-0.14	0.43	0.41	-0.29	0.17	-0.25	0.12
2. Apprehension	.33 (.46)	.25 (.28)	.66 (.70)	-0.27	-	0.28	0.41	0.09	0.09	0.13	0.16	-0.23	0.35	-0.28	0.05
3. Dominance	.76 (.74)	.18 (.22)	.48 (.64)	0.07	-0.07	-	0.19	0.23	0.31	0.25	0.21	-0.28	0.38	-0.04	0.26
4. Emotional stability	.89 (.63)	.16 (.24)	.62 (.67)	0.37	-0.38	0.02	-	0.18	0.18	0.24	0.33	-0.41	0.33	-0.34	0.30
5. Liveliness	.84 (.58)	.18 (.26)	.60 (.67)	-0.05	-0.14	0.11	0.12	-	0.19	0.02	0.03	-0.37	0.41	-0.08	0.40
6. Openness to change	.73 (.57)	.18 (.23)	.43 (.60)	-0.27	0.09	0.13	0.00	0.13	-	0.06	-0.04	-0.14	0.20	-0.10	0.30
7. Perfectionism	.78 (.65)	.18 (.25)	.56 (.66)	0.36	-0.12	0.15	0.29	-0.07	-0.07	-	0.43	-0.16	0.14	-0.09	0.10
8. Rule-consciousness	.86 (.71)	.18 (.25)	.62 (.70)	0.38	-0.17	0.11	0.41	-0.01	-0.12	0.45	-	-0.28	0.15	-0.14	0.19
9. Self-reliance	.81 (.32)	.21 (.27)	.69 (.74)	0.26	-0.18	0.03	0.28	0.28	-0.09	0.02	0.14	-	-0.36	0.28	-0.42
10. Social boldness	.86 (.60)	.21 (.34)	.77 (.85)	0.24	-0.34	0.19	0.42	0.43	0.05	0.15	0.25	0.36	-	-0.21	0.39
11. Vigilance	.46 (.64)	.22 (.23)	.59 (.61)	-0.16	0.34	0.07	-0.32	-0.12	0.03	-0.12	-0.20	-0.21	-0.26	-	-0.21
12. Warmth	.82 (.66)	.15 (.22)	.40 (.58)	0.09	-0.11	0.02	0.25	0.31	0.13	0.07	0.18	0.29	0.36	-0.21	-

Note. Values in parentheses and above the diagonal are for the incumbent sample. Based on calibration samples of 1000 applicants and 1000 and 1000 incumbents.

Table 2

MPA Eigenvalues

Scale	Applicant MPA		Incumbent MPA	
	Simulated Data	Sample Data	Simulated Data	Sample Data
Abstractedness	3.55 (0.91)	3.74 (1.13)	3.50 (0.97)	3.72 (1.07)
Apprehension	3.49 (0.87)	3.32 (1.03)	3.30 (0.88)	3.53 (0.95)
Dominance	2.28 (1.34)	2.59 (1.41)	3.03 (1.05)	3.46 (0.95)
Emotional stability	4.23 (0.93)	4.37 (1.07)	3.38 (1.02)	3.49 (1.36)
Liveliness	3.43 (1.06)	3.46 (1.05)	3.31 (0.89)	3.45 (1.27)
Openness to change	2.09 (1.19)	2.17 (1.21)	2.59 (1.03)	2.79 (1.09)
Perfectionism	3.20 (1.15)	3.46 (1.03	3.14 (1.02)	3.41 (1.04)
Rule-consciousness	3.22 (1.02)	3.62 (1.34)	2.95 (1.13)	3.35 (1.19)
Self-reliance	4.23 (0.84)	4.14 (1.14)	4.23 (0.76)	4.32 (0.80)
Social boldness	5.31 (0.64)	5.40 (0.69)	5.26 (0.62)	5.44 (0.74)
Vigilance	3.13 (0.92)	3.19 (1.06)	2.94 (0.94)	3.22 (0.93)
Warmth	2.18 (1.97)	2.32 (2.14)	2.99 (0.94)	3.15 (1.22)

Note. Values in the parentheses are eigenvalues for the second factor.

Table 3

Model fit for the GGUM and the 2PL for the Applicant and Incumbent Samples.

		GGUM		2PL	
Scale	No. of items	Applicants χ^2/df	Incumbents χ^2/df	Applicants χ^2/df	Incumbents χ^2/df
Abstractedness	8	2.16	2.43	2.15	2.41
Apprehension	8	14.55	4.71	2.25	1.17
Dominance	8	1.55	0.30	1.85	1.56
Emotional stability	9	0.69	2.30	0.70	4.02
Liveliness	8	1.21	3.20	1.21	4.61
Openness to change	9	0.88	1.02	0.85	2.00
Perfectionism	8	0.86	0.91	0.75	1.85
Rule-consciousness	8	2.85	2.12	2.88	4.00
Self-reliance	8	1.99	1.85	1.94	1.78
Social boldness	8	0.29	2.47	0.27	1.18
Vigilance	8	2.02	1.25	0.54	0.57
Warmth	8	2.17	2.96	2.15	3.62

Average	2.60	2.13	1.46	2.40
Total				

Note. The number of items containing DIF is based on the NCDIF index. * indicates that the DTF index exceeded the critical value

Table 4

DFIT Results and Number of Ideal Point Items for the Applicant and Incumbent Samples.

	DFIT results	
Scale	No. of items with DIF	DTF
Abstractedness	0	.003
Apprehension	NA	NA
Dominance	7	.019*
Emotional stability	3	.002
Liveliness	6	.018*
Openness to change	7	.004
Perfectionism	5	.012
Rule-consciousness	4	.002
Self-reliance	5	.013
Social boldness	4	.007
Vigilance	5	.001
Warmth	5	.015*

Total	51

Note. The number of items containing DIF is based on the NCDIF index. * indicates that the DTF index exceeded the critical value.

Table 5

Frequency Distribution of Items with DIF.

DIF Patterns	# of items with DIF	% of items with DIF
Hypothesized Direction	12	23
Opposite of Hypothesized Direction	8	16
Both Folding	5	10
Neither Folding	26	51
Total	51	100