UNIVERSITY OF CAMBRIDGE

Faculty of Politics, Psychology, Sociology, and International Studies

Michal Kosinski

Application of the dominance and ideal point IRT models to the Extraversion scale from the IPIP Big Five Personality Questionnaire

Dissertation for the degree of MPhil in Social and Developmental Psychology

Supervised by: Professor John Rust

(word count: 12,700)

Cambridge 2009

mk583@cam.ac.uk

TABLE OF CONTENTS:

Application of the dominance and ideal point IRT models to the Extraversion scale from the IPIP Big Five Personality Questionnaire

Abstract

In the present study, the ideal point and dominance graded response Item Response Theory (IRT) models (respectively: General Graded Unfolding Model, GGUM, Roberts, Donoghue, & Laughlin, 2000; and Samejima's Graded Response Model, SGR, Samejima, 1969) were applied the Extraversion scale from the Goldberg's 100 item Big Five personality questionnaire (Goldberg, et al., 2006). Data from 20,000 individuals were used to calibrate the models, and another 20,000 cases were used to evaluate model-data fit and measurement accuracy. Additionally, optimized scales were developed independently for ideal point GGUM and dominance SGR models by removing items that showed poor fit and low discrimination from the original 20-item Extraversion scale. Results revealed that application of the IRT models did not improve measurement accuracy of the original Extraversion scale or any of the optimized Extraversion scales developed in this study. The comparison of the model-data fit showed that ideal point GGUM demonstrated worse fit to the original and optimized Extraversion scales than dominance SGR model. Moreover, whereas scale optimization significantly improved data-model fit of the dominance SGR model, attempts to improve the data-model fit of the ideal point GGUM were unsuccessful. Several implications for the IRT models' application to the existing personality inventories were discussed.

Introduction

The increasing use of personality constructs in career planning, training, and selection have drawn attention to the accuracy and quality of personality tests. Their results have serious consequences for individuals and organizations. Undeniably, there has been a considerable progress in the quality and availability of the personality measures (Costa & McCrae, 2005). Moreover, researchers and practitioners can choose from an abundance of well-established instruments that allow for assessment and comparison of individual personalities under various theoretical approaches. For instance, Goldberg's International Personality Item Pool (IPIP) initiative provides open and free access to thousands of personality items catalogued in nearly 300 scales (Goldberg, et al., 2006). However, surprisingly little has changed in the area of personality test scoring and development methods (Chernyshenko, Stark, Drasgow, & Roberts, 2007). The great majority of personality instruments in use today are still based on the approach proposed by Likert (1932). They are composed of many long scales developed using Classical Test Theory (CTT) and are scored by simply summing endorsed response options across the items. It is assumed that item responses follow a *dominance process,* that the higher the level of the respondent's individual characteristic (i.e. *latent trait*), the higher the probability that he or she will endorse highly scored response options.

Likert's approach has been shown to have many disadvantages that can be addressed by Item Response Theory (IRT), often referred to as Modern Test Theory (Reise & Henson, 2003). The IRT uses mathematical modelling to describe the relationship between latent trait, item characteristics (e.g. difficulty), and individual response patterns (Drasgow & Hulin, 1990). IRT methodology was developed in the area of achievement and aptitude tests that were composed of the items that could be scored in the “right/wrong” fashion (e.g. Scholastic Aptitude Test; Lord, 1968), but it was also successfully applied to personality, attitude, and other inventories employing dichotomous items (e.g. Drasgow & Hulin, 1990). Later on, the IRT methodology was applied to instruments employing polytomous response scales, such as nominal scales (e.g. Bock, 1972; Samejima, 1979; Thissen & Steinberg, 1985) or ordered-category scales (e.g. Likert-type; Drasgow, Levine, Tsien, & Williams, 1995). As a result, the IRT approach is currently applicable to virtually any kind of standardized psychometric measurement and is able to deal with items that are scored in either dichotomous or polytomous fashion. Consequently, the IRT approach has been gradually replacing the CTT in many areas of assessment, especially in educational testing (Reise & Henson, 2003).

Although IRT is still virtually absent from applied personality assessment, personality researchers have expressed a growing interest in IRT methodology. Several attempts were made to fit various IRT models to the personality scales from the established questionnaires (e.g.: Rauch, Schweizer, & Moosbrugger, 2008; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001) and to develop new personality scales using an IRT approach (e.g. Waller, Tellegen, McDonald and Lykken, 1996; Chernyshenko, Stark, Drasgow, & Roberts, 2007). Moreover, IRT has been used in studies regarding the relation between testing situation and item responding (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001), context effects (Steinberg, 2001), and computer adaptive testing (Hol, Harrie, & Mellenbergh, 2008). Others have applied IRT to detect dishonest respondents (e.g. Zickar & Robie, 1998) and to address the issue of item bias (Collins, Raju, & Edwards, 2000). More recently, several IRT models allowing for the non-monotonic relation between responses and the latent trait were developed (Roberts, Donoghue, & Laughlin, 2000) and applied to personality data (e.g. Weekers & Meijer, 2008; Stark, Chernyshenko, Drasgow, & Williams, 2006; Chernyshenko, Stark, Drasgow, & Roberts, 2007).

Although recent studies have answered many questions related to the application of IRT in the context of personality assessment, a number of fundamental methodological issues are still not adequately addressed. These issues confine the application of IRT methods in practical personality assessment. Consequently, few, if any, popular personality inventories have been constructed or scored with IRT methods. The first and most conspicuous problem with the application of IRT in the personality assessment is the lack of a universally accepted or, at least, well studied IRT model (or set of models) that could be applied to personality data. Although the majority of the personality scales rely on the graded response scales (e.g. Likert-type scales), in most of the previous studies binary IRT models were examined, often by simply dichotomizing original responses on a graded response scale (e.g. Stark, Chernyshenko, Drasgow, & Williams, 2006; Weekers & Meijer, 2008; Chernyshenko, Stark, Drasgow, & Roberts, 2007). As a result, little is known about the applicability of the graded response IRT models to personality data. Moreover, although it was suggested that recently proposed non-monotonic IRT models (Roberts, Donoghue, & Laughlin, 2000) should outperform dominance models in the assessment of non-cognitive scales, ensuing studies showed only small differences between the ideal point and the dominance approaches to test development and scoring. Finally, despite of the widespread opinion that *"Advances in measurement models have the potential to increase accuracy and efficiency in practical measurement applications."* (Baker, Rounds, & Zevon, 2000, p. 253), the author has not found any results that would demonstrate the potential of the IRT methods to increase the measurement accuracy in the area of personality assessment.

This study was an attempt to address these three methodological problems. Two graded response IRT models, ideal point and dominance, were selected from among those that showed the best performance in the previous studies. They were applied to the Extraversion scale from the IPIP Big Five questionnaire (Goldberg, et al., 2006) to compare their suitability for personality data. Moreover, the measurement accuracy offered by both of the above models was compared with the measurement accuracy of the CTT approach. Due to the limited length of the current paper, only the Extraversion scale was analysed here, but there is little indication that the results of this study cannot be generalised to the remaining four Big Five scales.

Item Response Theory as the solution to the classical test theory problems

Since CTT offers a simple methodology and is easily understandable, it is still the prevailing approach to psychological measurement (Murphy & Davidshofer, 2001). However, there are several problems inherent in CTT, most of them related to the overly simplistic concept of reliability (Weiss, 1995). First, the single reliability coefficient estimated for each CTT-based scale overlooks the fact that the test's measurement precision varies with the level of the measured trait (Feldt & Brennan, 1989). For instance, a neuroticism scale that is highly accurate in the population of moderately neurotic individuals will fail to differentiate well in a group composed of respondents with extreme levels of neuroticism. Second, the variance in the measurement precision across the levels of the latent trait implies that the instrument's reliability will vary with the average level of the latent trait in a sample used in the estimation of reliability. Third, the standard error of measurement (SEM) of the individual scores is simply derived from the scale's reliability, which does not account for the fact that individual scores differ in accuracy (Rouse, Finger, & Butcher, 1999). Fourth, since in the CTT the respondent's level of the measured trait is estimated by comparing his or her score with the distribution of the scores in the sample, the interpretation of the individual score in the CTT is sample dependent. Fifth, as the item selection procedure in CTT favours items characterized by average facility and high discrimination (Rust & Golombok, 2009), accuracy of the CTT instruments is high around the mean levels of latent trait and decreases significantly at the extremes of the trait. Finally, CTT does not make any assumptions about the relation between response to single item and the measured trait, but simply treats the sum of the item scores as an approximation of the true score. Consequently, items that vary in their ability to measure are equally weighted against the total score (Rouse, Finger, & Butcher, 1999).

The IRT approach addresses each of these CTT-related problems. First, the reliability of the scale (and each item) is estimated at every level of the latent trait (Baker, 2001). This not only reflects the fact that measures are not equally effective across the latent trait continuum, but also allows tailoring of a scale that is efficient at a desirable level of the latent trait. Second, in the IRT approach the SEM of the score is estimated individually for each case with regard to the response options endorsed by the respondent and their potential to measure at respondent's level of the latent trait (Baker, 2001). Third, the psychometric properties of the IRT measures are not sample dependent because item and test properties are directly related to the latent trait (Lord, 1980). Fourth, direct relation between individual item properties and the latent trait implies that items contribute in the estimation of the results according to their ability to measure on a given level of the latent trait. Finally, direct relation between individual item properties and the latent trait together with the individual estimation of the score's SEM imply that the individual score is not sample dependent and it is constant regardless of the items on which the assessment is based (Baker, 2001).

Dominance and ideal point approaches to measurement

The majority, if not all, of the personality scales that are widely used today were constructed under the assumption that item responses follow a dominance process whereby the higher the level of the respondent's latent trait, the higher the probability that he or she will endorse highly scored response options. However, it was suggested that the relation between response to a non-cognitive item and the latent trait might be more complex. Thurstone (1928) proposed that a respondent answers in the keyed direction only if he or she is located near the item on the latent trait continuum. For instance, the item "I like meeting with friends in quiet cafés" would not be endorsed by some introverts, as they feel uncomfortable in public places, but it would neither be endorsed by some extreme extraverts who find meeting in quiet cafes extremely boring. Consequently, the likelihood function of endorsing this item at different levels of the Extraversion is bell-shaped; it is low on the low and high levels of the latent trait and peaks somewhere in between. The hypothetical shape of this function is illustrated in *Figure 1.* A non-monotonic relation between item response and latent trait is called an *ideal point response process*, because it assumes that the probability of endorsing an item increases as the distance between a respondent's location on the latent continuum (called ideal point) and the item's location on the latent continuum decreases. There is some evidence of the ideal point approach's supremacy over dominance methods in the context of personality assessment. For instance, Chernyshenko and colleagues (2007) developed a Conscientiousness scale under both the ideal point and dominance approaches, which showed that the ideal point approach offered more precise measurement at the extremes of the latent trait continuum.

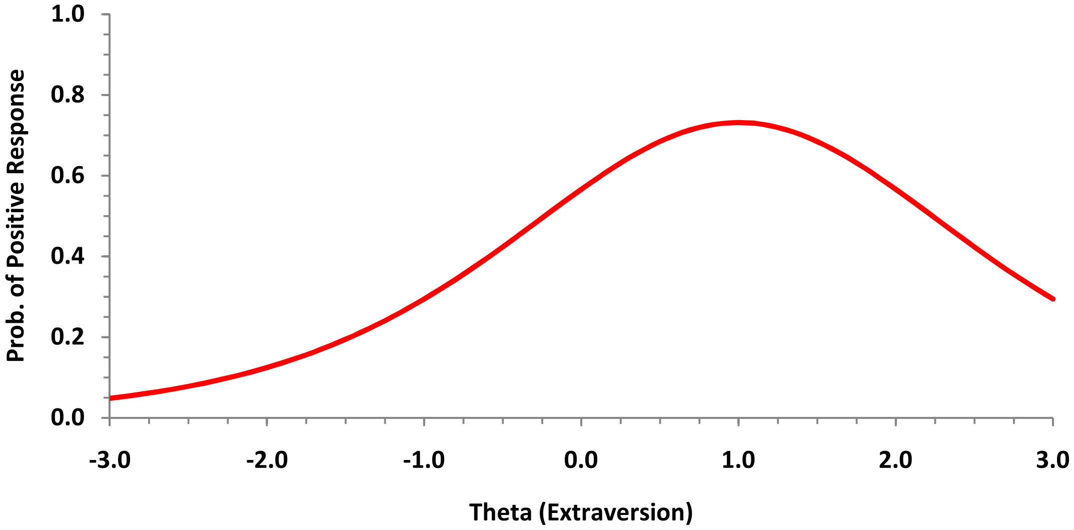

Figure 1. Hypothetical likelihood function of endorsing of the response “I like meeting with friends in quiet cafés” on the Extraversion (item parameters: $\propto = 1.0, \delta = 1.0, \tau = -1.0$)

CTT and the majority of the IRT models assume a strict monotonic relation between item responses and the latent trait. Scale development processes under the dominance approach ascertain that scales contain only items monotonically related to the latent trait. The hypothetical non-monotonic item presented in *Figure 1* would likely be discarded from the dominance scale as it is characterized by a low or negative item-total correlation, and a weak factor loading. Nevertheless, recent psychometric studies involving the MMPI (Meijer & Baneke, 2004) and the Sixteen Personality Factor Scale (Stark, Chernyshenko, Drasgow, & Williams, 2006) revealed that those instruments contained some items that followed ideal point response process. Since ideal point models can fit both monotonic and non-monotonic items (Chernyshenko, Stark, Drasgow, & Roberts, 2007), it seems that ideal point models could provide a better fit even in the scales constructed under the dominance assumption.

Brief introduction to the main concepts of Item Response Theory

Option Response Function. While CTT focuses on the properties of the whole test, such as test score or overall reliability coefficient, IRT is much more focused on the properties of the individual item, or more precisely, individual response option. The core concept of IRT is the relationship between the response to a test item and the latent trait. This relationship is expressed by the Option Response Function (ORF), which is a nonlinear regression of the probability of choosing a particular response category on the latent trait (Baker, 2001). There are several families of ORFs that can be used to model unidimensional or multidimensional data using either dichotomous or polytomous response formats.

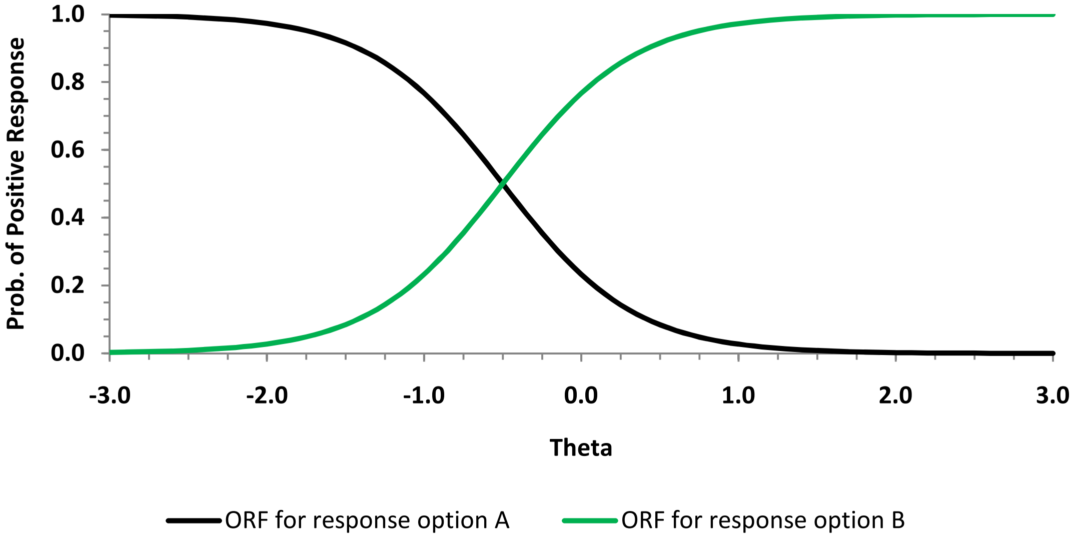

Figure 2. Option Response Functions for responses A and B to the dichotomous item (item parameters: a = 1.4, b = - 0.5)

Although various kinds of ORFs can take different shapes, they usually resemble ORFs such as the examples presented in *Figure 2*, and *Figure 3*. The horizontal axis displays the standardized (*SD* = 1) level of the latent trait, and is referred to as theta (θ).

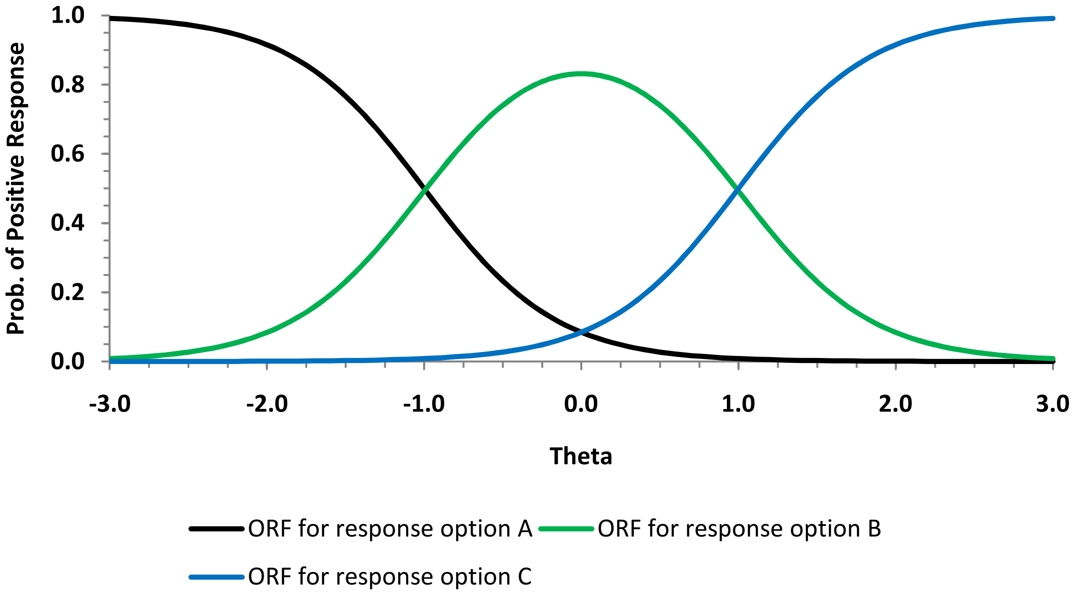

Figure 3. Option Response Functions for responses A, B and C to the polytomous item (item parameters: a = 1.4, b1 = -1.0, b2 = 1.0)

The probability that a respondent will choose a particular response at any level of the latent trait, ranging from 0 to 1, is plotted on the vertical axis. The item presented on *Figure 2* has two response options (e.g. "Agree" and "Disagree"), whereas item presented on *Figure 3* has three response options.

Interpretation of the ORF function is relatively straightforward. For example, the relation between response to an item and the latent trait presented in *Figure 2*, indicates that a respondent at the level of a trait equal to 0.5 (θ = 0.5) will chose the response option B with probability of 95% and response option A with probability of 5%. Accordingly, in the item presented in *Figure 3*, the same respondent will choose response A with around 5% of probability, response B with probability of 75%, and response C with probability of 20%. Note that since the sum of the probabilities of endorsing all of the response options on a single item equals 1 (100%), one of the ORFs is always determined. For instance, in the case of the dichotomous items, the probability of choosing the second response option can be expressed as one subtracted by the probability of choosing the first response option.

Estimation of the individual scores. IRT scoring methods, such as maximum likelihood or Bayesian estimation, are beyond the scope of the current paper (for a detailed description, see Embretson & Reise, 2000). The procedure looks for the value that maximizes the agreement between estimated latent trait and individual's responses. It could be graphically represented as combining the ORFs of the responses chosen by participants. Assume that a hypothetical respondent endorsed response options B in item X and item Y. Appropriate ORFs are plotted in *Figure 4*. By multiplying these two ORFs at each level of the latent trait, a curve is obtained that describes the likelihood of a respondent having each of the latent trait levels. As can be seen in *Figure 4*, the likelihood curve for this particular respondent peaks at approximately -0.5, indicating

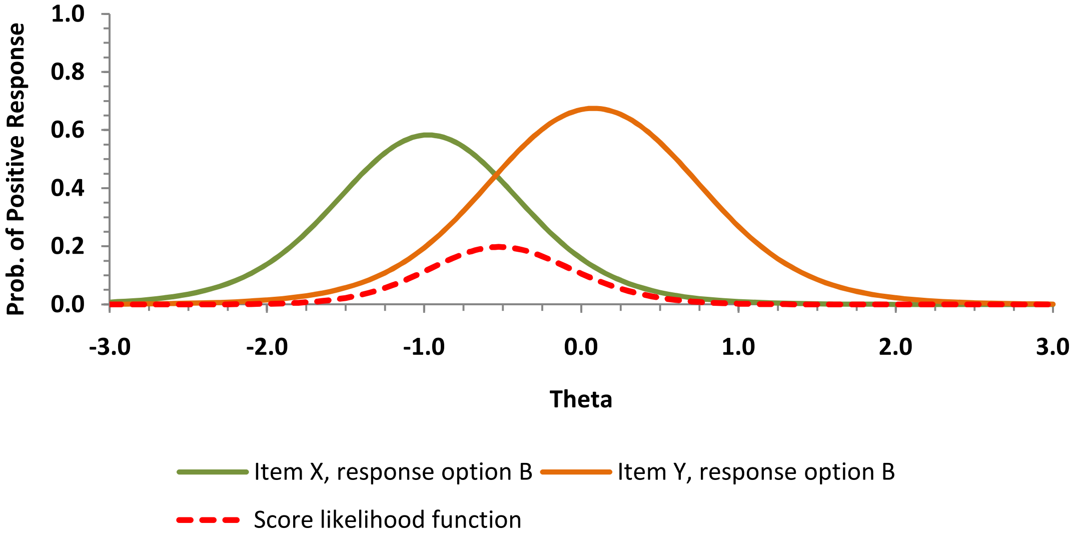

Figure 4. Option response curves endorsed by hypothetical respondent with the relevant score likelihood function

the most likely level of the respondent's latent trait[1]. The steepness of the score likelihood function indicates the accuracy of the measurement. Note that combining highly overlapping ORFs produce a score likelihood function that is noticeably peaked, whereas combining inconsistent ORFs would produce a flat score likelihood function. This accounts for the fact that the score based on the responses unanimously indicating a certain level of the latent trait is more accurate than the score produced by responses that indicate various levels of the latent trait. The score likelihood curve presented in *Figure 4* is flat, which suggests that the score estimation based on the two responses was imprecise. However, if the following responses are consistent, the estimation would be more precise, and the likelihood curve would become more peaked.

Another important property of IRT scoring procedure is that response items are not equally weighted in latent trait level estimation. As shown above, a score likelihood curve is estimated by combining ORFs that differ in shapes. ORFs characterized by steep slopes located close to individual latent trait level will affect the shape of score likelihood curve, whereas flat-shaped ORFs and/or ORFs with slopes far from the individual latent trait level decrease estimation accuracy or do not contribute heavily to the estimation of the score. The direct relation between response and the latent trait in IRT allows for a comparison of results between instruments and samples.

Information Function. As is discussed above, the ability to differentiate between respondents varies across response options and levels of the latent trait. Consider the ORF presented in *Figure 5*. A small change in the level of the latent trait around any of the steep slopes of this ORF causes a dramatic change in the probability of endorsing this response option, whereas the same change in the latent trait in the areas where the ORF is flat does not affect this probability much. Consequently, the response option

[1] Score calculated with the maximum likelihood procedure: $\theta = -0.43$, conditional SEM $= 0.41$

presented in *Figure 5* provides some information regarding the respondent's location on the latent trait continuum around the steep slopes of its ORF, but yields little information about respondent located in the areas where the ORF is flat. The amount of the information provided by a given response option on every level of the latent trait is expressed by the Option Information Function (OIF).

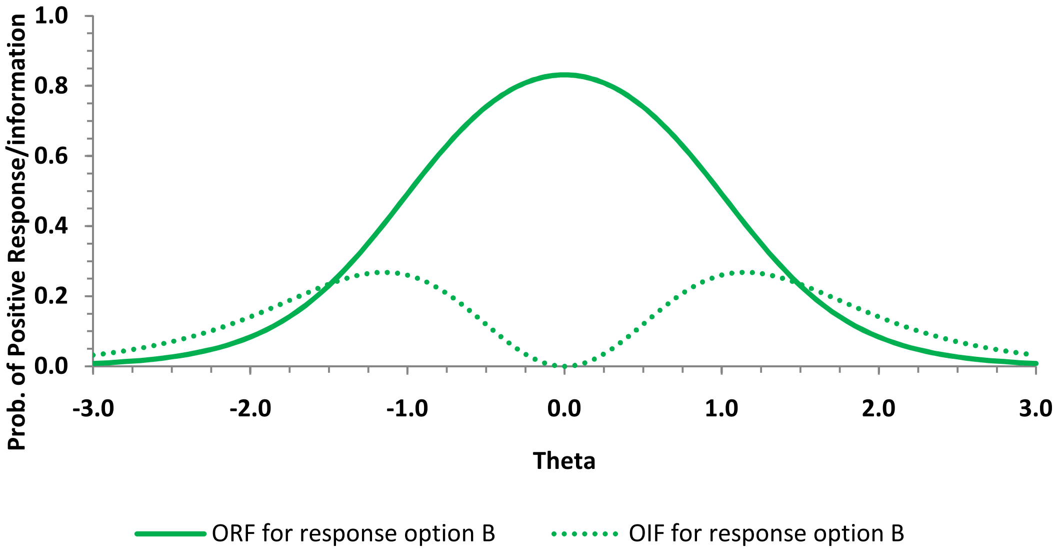

Figure 5. Option Response Function with relevant Option Information Function

The sum of the Option Information Functions is called the Item Information Function (IIF), and the sum of the IIF of the items belonging to one scale is called the Test Information Function (TIF). Option Information Functions and the Item Information Function for the item presented in *Figure 3* are presented in *Figure 6*.

The information function is a powerful characteristic of the IRT approach. Test information function is very useful in selecting the appropriate test for a particular task. Moreover, the precise information about individual item's accuracy at every level of the latent trait, and the fact that accuracy of the whole test is the sum of the individual item's accuracies allows constructing scales for a specific purpose, employing pools of pre-calibrated items (Waller, Tellegen, McDonald, & Lykken, 1996). Similarly, tests that are administered using computers can be modified during administration using computer adaptive testing (CAT; Waller & Reise, 1989). In the CAT approach, the computer estimates the respondent's score after each response and uses item pool to select the next item that provides the maximum information near the expected level of

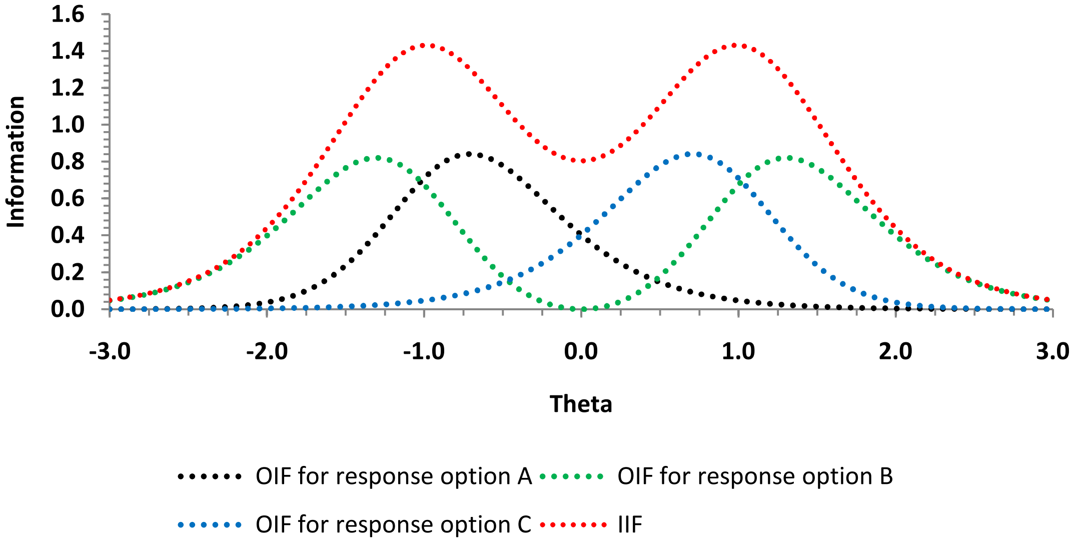

Figure 6. Option Information Functions and Item Information Function for the item presented in *Figure 3*

respondent's latent trait. This process continues until the predefined length or predefined level of the test's accuracy is reached. As respondents are presented with most discriminating items at their level of latent trait, CAT makes it possible to increase a test's accuracy (or decrease the number of items, while keeping the same accuracy; Weiss, 1985).

Mathematically, information function for the response option is defined as:

$$I_{i,k}(\theta) = -\frac{\partial^2}{\partial\theta^2} logORF_{i,k}(\theta)$$

where $I_{i,k}(\theta)$ is the amount of information provided by the response option k on item i at a particular level of theta (θ), $ORF_{i,k}(\theta)$ is the Option Response Function of option k on item i, and $\frac{\partial^2}{\partial\theta^2}$ indicates the second partial derivative of the function with respect to θ (Samejima, 1969).

Measurement accuracy. In the IRT approach, measurement accuracy is expressed by the *conditional SEM* function that is inversely related to amount of information regarding the respondent's level of the latent trait. The conditional SEM is a model-based estimate of the observed theta values' variability around the respondent's unknown true latent trait level. In contrast to the CTT-based SEM, conditional SEM varies across items, response options, and levels of the latent trait (Baker, 2001). Conditional SEM is expressed as:

$$Conditional\, SEM(\theta) = \frac{1}{\sqrt{I(\theta)}},$$

where $I(\theta)$, is the amount of information provided by a particular response option, item or the whole scale at the particular level of latent trait. *Figure 7* presents conditional SEM of the polytomous item whose Information Function was presented in *Figure 6.* The conditional SEM can be also calculated for an individual score estimation. In this case $I(\theta)$ is the amount of information provided by the response options endorsed by the individual at the latent trait level estimated for him or her. As conditional SEM of the score estimation is directly related to individual responses, it is possible to compare conditional SEM of the score values across individuals and instruments. Moreover, in IRT two identical latent trait scores can be characterized by the very different conditional SEM values. This is possible because various sets of response options can produce the same latent trait score with different accuracy. For example, latent trait score estimated for the individual protocol composed of random responses is likely to be average (θ value around 0). However, as random responses are informative on various levels of latent trait, the conditional error of this score will be higher than the conditional error of the similar score based on the honest and consistent responses.

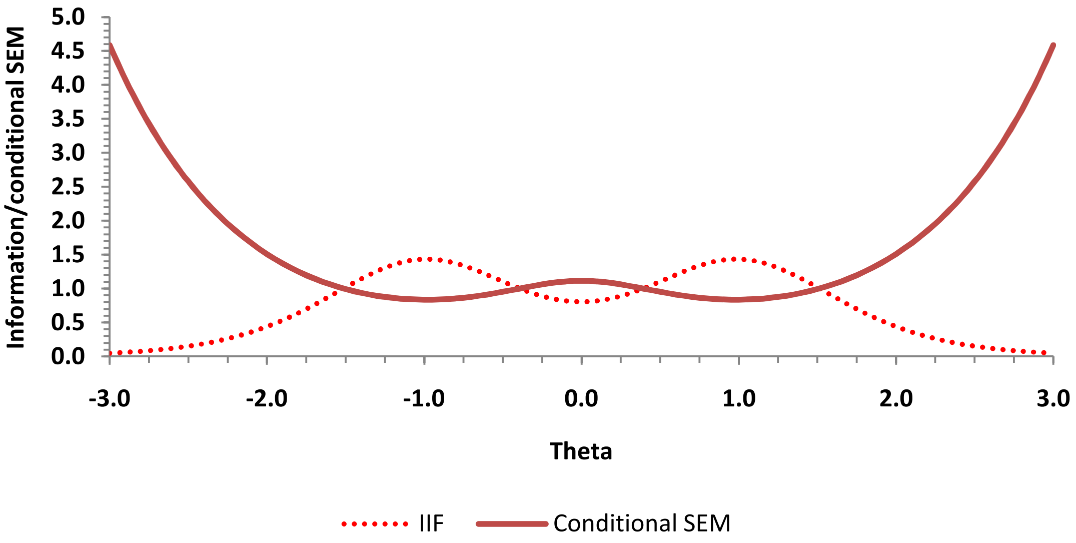

Figure 7. Conditional SEM and item information function for the item presented in *Figure 3*

In some of the previous studies on the IRT models application to the personality scales the conditional SEM values were directly compared with SEM based on the CTT reliability estimates (e.g. Rauch, Schweizer, Moosbrugger, 2008). Nevertheless, it should be emphasized that IRT-based conditional SEM and CTT-based SEM cannot be simply compared, as conditional SEM does not account for the fact that the model is itself estimated with some amount of error and does not fit the real data perfectly. There are methods available to estimate conditional SEM that is comparable with reliability-based SEM (Qualls-Payne, 1992), however IRT-based conditional SEM is not one of them.

Unidimensionality and local independence. Most of the IRT models make two inter-related assumptions required for estimating item parameters: *unidimensionality*[2] and *local independence*. Unidimensionality requires that the responses to items belonging to one scale must be based on a single latent trait (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). This also implies that on any fixed level of the underlying factor, the responses to the different items are unrelated, which is referred to as local independence (Reise & Henson, 2003).

In personality assessment, the theoretical assumption of absolute scale's unidimensionality (and hence local independence) is hard, if not impossible, to achieve (Reise & Henson, 2003). Various factors tend to contaminate responses, such as acquiescence, social desirability, self-enhancement, and many others. Fortunately, previous research showed that unidimensional IRT models are relatively immune to the moderate violations of the unidimensionality assumption (e.g., Drasgow & Parsons, 1983). Therefore, it was suggested that it is sufficient to demonstrate that there is a single dominant factor underlying participants' responses and that the single-factor model fits well the data (Embretson & Reise, 2000).

A wide variety of tests of unidimensionality is available (Hattie, 1985). Popular methods of unidimensionality testing include modified parallel analysis (Drasgow & Lissak, 1983), dimensionality testing (Stout, 1987), and confirmatory factor analysis (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). Nevertheless, the above methods could not be used in the current study, as they assume monotonic relation between item response process and levels of the latent trait, which violates the assumptions of the ideal point IRT model used here.

[2] Several multidimensional IRT models were also proposed (Embretson, 1991; Knol & Berger, 1991; McKinley & Reckase, 1983), however their application is limited as the theory behind these models is still in an early stage of development.

An alternative approach to testing unidimensionality and local independence, appropriate for both dominance and ideal point approaches, was proposed by Chernyshenko et al. (2007). They suggested that if a unidimensional IRT model can fit patterns of observed responses, than a single latent trait is underlying item responses. Therefore, if it can be shown that model fits the observed data, there is a little reason to suspect that the scale is multidimensional. The issue of model-data fit is discussed in more detail in the following section.

Model-data fit.

The linear relationship between latent trait and individual responses expressed by the ORF is a simplification. First, individual responses are always affected by a certain degree of uncontrolled factors that lead to discrepancies between the unidimensional model and empirical data. Second, the actual relationship between responses and the latent trait is rarely a perfect match with the ORF shape, as the latter is limited by its mathematical foundations. Consequently, some differences between model predictions and real data are unavoidable. Nevertheless, significant misfit would indicate that the model does not properly describe the relation between predictors and the latent trait, and its results are inaccurate. Moreover, as it was mentioned before, appropriate model-data fit is a good indicator of the scale's unidimensionality. As a result, proper assessment of the model-data fit is crucial in the application of the IRT methods.

There are two widely accepted and complementary methods of assessing the model-data fit in the IRT. The first is based on the Pearson chi-squared goodness-of-fit test that compares observed and expected frequencies across the response patterns, and the second is the visual inspection of the observed vs. expected item response curves.

The application of the chi-squared goodness-of-fit test in IRT is described in detail in Bartholomew et al. (2008, pp. 218-222). The main problem with the application of this method, is that even in the simple IRT models many response patterns will have expected frequencies that are very small or zero (depending on sample size), whereas it is recommended that all expected frequencies should be at least five for the chi-squared goodness-of-fit test to be valid. Therefore, instead of examining the whole set of response patterns, only the lower order margins for single items (singlets), pairs of items (doublets), and three items (triplets) are examined. To compute the lower order margins the observed frequencies of endorsing a particular response are calculated for all single items, all item pairs, and all item triplets. The comparison between observed and expected frequencies is made using *chi-squared residuals:*

$$\chi_i^2 = \sum_{\mathrm{k=1}}^{\mathrm{s}} \frac{[\mathrm{O_{i,k}} - \mathrm{E_{i,k}}]^2}{\mathrm{E_{i,k}}} ,$$

where $\mathrm{O_{i,k}}$ is the observed frequency of the response *k* to the item *i*, $\mathrm{E_{i,k}}$ is the expected frequency of response *k* to item *i*, and s is the number of response categories (Bartholomew, Steele, Moustaki, & Galbraith, 2008, pp. 218-222). In order to allow for comparisons between various models, samples and studies, chi-squared residuals are adjusted to a constant sample size of 3,000 and divided by their degrees of freedom. Previous studies found that adjusted residuals lower than 3 are the indicators of good model-data fit (Drasgow, Levine, Tsien, & Williams, 1995; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001).

Chernyshenko et al. (2001) demonstrated that chi-squared analysis is not sensitive to some kinds of model-data misfit. Therefore chi-squared analysis should be supported by the visual inspection of the observed vs. expected item response curves (*Fit-plots*; Drasgow, Levine, Tsien, & Williams, 1995). Fit-plot consist both the ORF and *empirical curves* plotted as described in Drasgow et al. (1995). The model-data misfit is indicated by the discrepancy between empirical curve and ORF that exceeds the empirical curve's 95% confidence interval. The example of the fit-plot can be found in *Figure 8*. The item presented in this figure shows considerable misfit at the high levels of the latent trait.

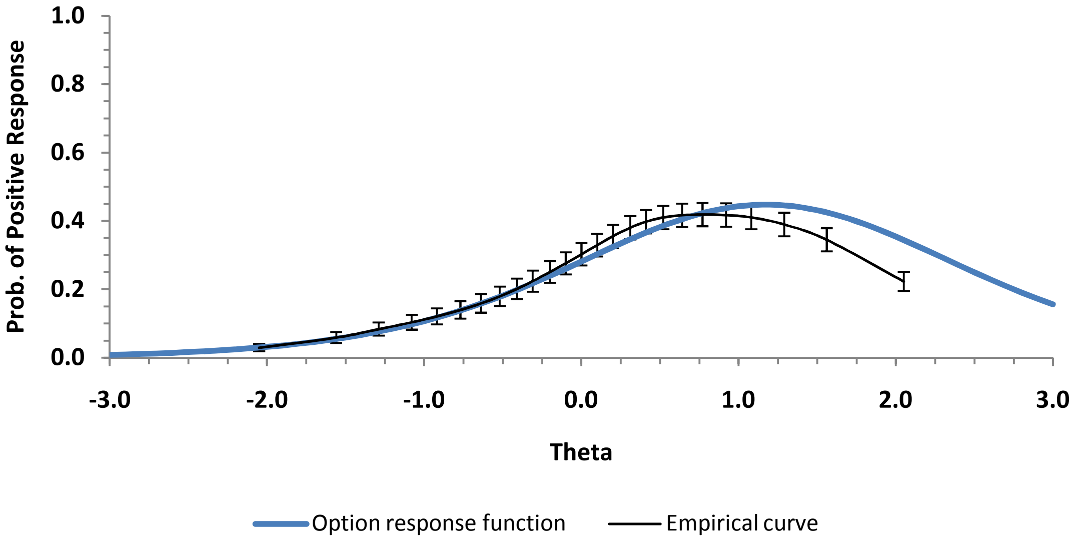

Figure 8. Example of the fit-plot plotted as described in Drasgow et al (1995)

IRT models used in this study

Dominance model for polytomous data: Samejima's Graded Response Model. Samejima's Graded Response Model (SGR; 1969), an enhancement of the two-parameter logistic IRT model for dichotomous data (see Hambleton, Swaminathan, & Rogers, 1991 for a discussion of logistic models), was one of the most popular IRT polytomous models in personality research. In SGR, m_i*-1* ORFs are estimated for each item, where m equals number of response options for item *i* (as it was mentioned earlier one of the ORFs is always determined). Each of the ORFs is characterized by two parameters. Difficulty parameter, *b*, is defined as a point on the latent trait at which a respondent has 50% probability of selecting a given response option or any of the subsequent higher ordered options. In other words, difficulty determines the vertical distribution of the ORFs, and consequently it determines the range on the latent trait continuum where the item is most informative. Discrimination parameter, *a*, describes the steepness of the ORF slope and it is assumed to be the same for each response option within a particular item. As previously stated, the steepness of the ORF slope determines the ability of an item to discriminate among the respondents at different levels of latent trait.

To fit a binary model to polytomous data, SGR transforms a polytomous response set into a series of binary response sets by gradually merging response options. Likelihood function of selecting given response option or any higher response is estimated using a two-parameter logistic model. Resulting “merged response options function” is called *boundary response function* (BRF). For an item measured via four response categories (e.g. 1 to 4), SGR will estimate three BRFs. First BRF describes the probability of selecting response categories: 2, 3, and 4; second BRF: 3, and 4, and third BRF: 4. BRFs estimated by two-parameter logistic model are used to calculate single Option Response Functions. Note that the probability of endorsing response option k equals the probability of endorsing response option k or any higher response option decreased by the probability of endorsing response option k+1 or any higher response option. In other words, ORF of option k equals BRF of option k decreased by BRF of option k+1. For instance, ORF for response option 3 equals BRF for options 3, 4, and 5, decreased by BRF for options 4 and 5. Mathematically, the probability of selecting option *k* on item *i* is expressed as:

$$P_{i,k}(\theta) = BRF_{i,k}(\theta) - \ BRF_{i,k+1}(\theta) =$$

$$= \frac{1}{1 + \exp\,[-1.702 a_i(\theta - b_{i,k})} - \frac{1}{1 + \exp\,[-1.702 a_i(\theta - b_{i,k+1})}$$

where $P_{i,k}(\theta)$ is the probability of a respondent at a particular level of theta (θ) responding to option k on item i, $BRF_{i,k}(\theta)$ is the boundary response function for option k on item i, a_i is the item discrimination parameter, $b_{i,k}$ is the difficulty parameter of option k on item i; 1.702 is a scaling constant, and exp represents an exponential function. Note that ORF for option 1 equals one subtracted by BRF for options 2 to 5, and ORF for option 5 equals BRF for option 5. A detailed description of fitting the SGR model to Likert-type data can be found in Muraki (1990).

SGR was selected to be used in this study as an example of the dominance IRT model, as it was used in the majority of previous studies on polytomous personality data (e.g. Maydeu-Olivares, 2005; Baker, Rounds, & Zevon, 2000) and generally showed good model-data fit (one significant exception can be found in Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). Consequently, results achieved in the present study could be compared with results achieved to date.

Ideal point model for polytomous data: General Graded Unfolding Model. Although an ideal point IRT approach to personality assessment is promising, appropriate models were only developed recently. Among those few previous studies on the ideal point approach to personality assessment, most employed the model that was proposed by Roberts and colleagues (2000). General Graded Unfolding Model (GGUM) is based on four premises about the response process (Roberts, Donoghue, & Laughlin, 2000). The first premise is that respondents endorse the item that is closest to their location on the unidimensional latent trait continuum. A second premise states that observable responses can result from two different subjective responses. For example, observable response “disagree” to an Extraversion item “I like meeting with friends in quiet cafés” can be interpreted as subjective response “disagree from below” (respondent’s level of Extraversion is too low to agree), or “disagree from above” (respondent’s level of Extraversion is too high to agree). Similarly, respondents closer to the item on the latent trait continuum can “agree” or “strongly agree” from “below” or from “above”. The third premise of the GGUM is that subjective responses (as opposed to observable) to attitude-type items are monotonically related to the latent trait. The fourth premise of the GGUM is that ORFs are symmetric about the item’s location on the latent trait. In other words, a respondent is just as likely to agree with an item located either h units below or h units above his or her ideal point on the latent trait continuum.

In GGUM subjective response functions are estimated using Muraki’s (1992) Generalized Partial Credit Model. A set of Subjective Response Functions (SRF) to the hypothetical item with three response options (“Disagree”, “Neither agree nor disagree”, “Agree”) is presented in *Figure 9*. Note that, consistent with the third and fourth GGUM premises, SRFs are monotonic and symmetrical about an item’s location. The probability of selecting a given observable response is calculated by summing the probabilities of the two subjective responses associated with that observable response options. This can be graphically illustrated as summing appropriate SRFs. The observable option response functions for hypothetical item presented in *Figure 9* are presented in *Figure 10*. Mathematically, the GGUM model is defined as:

$$P(Z_i = z|\theta) = P(Y_i = z|\theta) + P[Y_i = (M - z)|\theta] =$$

$$= \frac{exp\{\alpha_i\ [z(\theta - \delta_i) - \sum_{k=0}^{z} \tau_{ik}]\} +\ exp\{\alpha_i\ [(M - z)(\theta - \delta_i) - \sum_{k=0}^{z} \tau_{ik}]\}}{\sum_{w=0}^{C} \left\{exp\{\alpha_i\ [w(\theta - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\} +\ exp\{\alpha_i\ [(M - w)(\theta - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\}\right\}}$$

Where Z_i is an observable response to item i, z can take values from 0 to C (0 is the strongest level of disagreement, and C is the strongest level of agreement; C equals the number of observable responses minus 1), M is the number of subjective response categories minus 1 (M = 2*C + 1), α_i is the discrimination of the item i, τ_{ik} is the location of the *k*th subjective response function threshold on the theta continuum relative to the location of the item i, and δ_i is the location of the item i on the latent trait continuum.

Three parameters are used to describe observable Option Response Function in GGUM. Discrimination parameter, ∝, determines the steepness of the ORF slope and it is assumed to be the same for each response option within a particular item. As in the

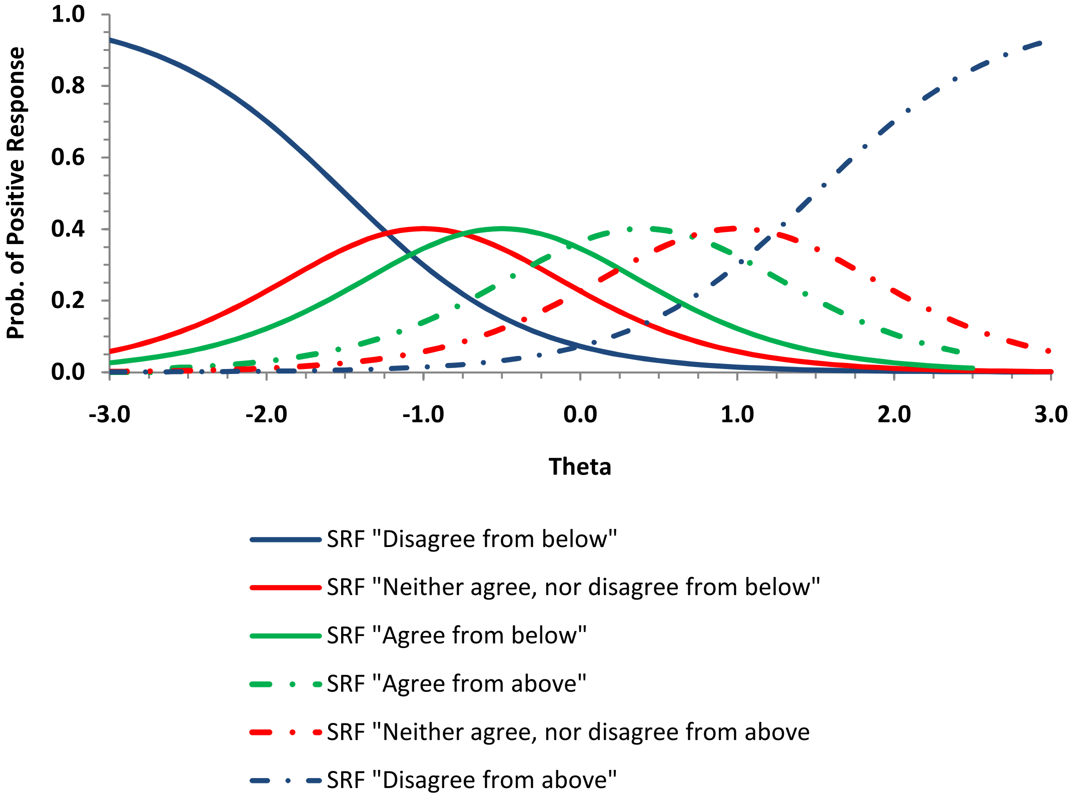

Figure 9. Subjective response functions for a hypothetical item with 3 response options (“Disagree”, “Neither agree nor disagree”, “Agree”)

SGR model, the higher discrimination of the item, the better it differentiates among the respondents at different levels of latent trait. The ORF presented in *Figure 10* is characterized by discrimination parameter of $\propto = 1$. The item location parameter, δ, identifies the location of the item on the theta continuum. It can be easily detected as a common symmetry point of the item's ORFs plot. The item presented in *Figure 10* is described by location parameter $\delta = 0.0$. The subjective response threshold parameter, τ, is defined as the location (relative to the item location parameter) where successive subjective response functions intersect. According to the fourth GGUM premise, subjective response functions are symmetric about item location; therefore, one threshold describes both subjective response functions associated with one observable response. The item that is presented in *Figure 10* is characterized by the threshold parameters: τ_2 = -1.25, τ_1 = -0.7, τ_0 = 0.0.

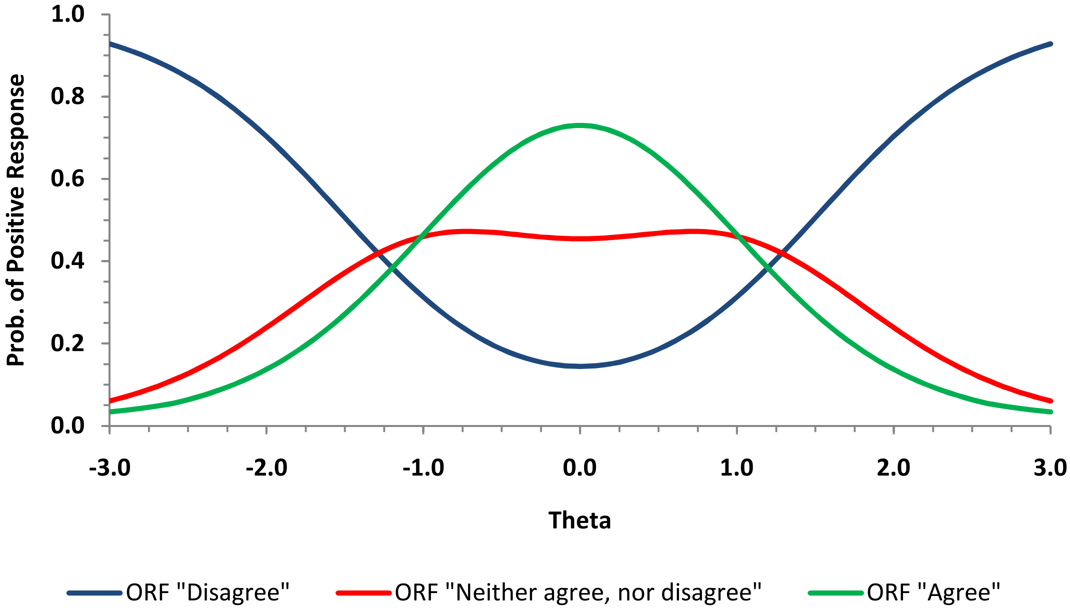

Figure 10. Observable Option Response Functions based on subjective response functions presented in *Figure 9*

Hypotheses

The goal of this research was to examine the applicability of the IRT methods in the context of personality assessment. Considering the advantages of the IRT approach discussed before, such as the advanced score estimation procedures, and comprehensive modelling of the relationship between individual response and latent trait level, it was hypothesized that:

H1: The IRT approach offers greater accuracy than the CTT approach in estimating individual scores for IPIP Extraversion scale.

As it was discussed before, some studies suggest that ideal point IRT models offer better fit to the personality data than traditional dominance IRT models. Therefore, the secondary goal of this study was to compare the model-data fit between the ideal point and the dominance IRT models. Thus, the second hypothesis was:

H2: The ideal point IRT model (GGUM) fits the IPIP Extraversion scale better than dominance IRT model (SGR).

The summary of the present research plan

The interpretability of the IRT scores is limited by the degree to which a given IRT model fits the data. Since the CTT-based Extraversion scale used in this study was not originally designed to be scored with the IRT models, the attempt was made to maximize the model-data fit of the models used in this study. The Extraversion scale was optimized independently for each of the 3 methods used in this study, by removing the items characterized by the poor performance, misfit, or local independence violation. To allow for easy comparisons, lengths of the optimised scales were fixed to 10 and 4 items. Consequently, the performance of the three methods tested in this study was compared on the scales of the three lengths (20, 10, and 4 items), which were not necessarily composed of the same items. Next, the primary hypothesis was tested by comparing the accuracy of the CTT and IRT models on the original 20-item scale and optimized scales. Finally, the secondary hypothesis was tested by comparing the model-data fit statistics and fit-plots between GGUM and SGR models on the original and optimized scales.

Methods

Personality measure

The personality measure used in this study was the IPIP 100-item Big Five personality questionnaire (Goldberg, et al., 2006). IPIP Big Five scales represent Costa and McCrae's Five Factor Model employed in the NEO-PI-R (Costa & McCrae, 2005), and they correlate highly with the corresponding NEO-PI-R domain scores with correlation coefficients ranging from 0.88 to 0.93 (IPIP.org, 2009). The IPIP Big Five measure is widely used in both traditional and online personality assessment and proved to be useful and reliable in the research (e.g. Buchanan, Johnson, & Goldberg, 2005; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). Moreover, IPIP domain scales have been shown to outperform the matching NEO-PI-R constructs as predictors of a number of self-reported behavioural indices (Goldberg, et al., 2006).

The IPIP instrument consisted of five 20-items scales measuring Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Respondents were asked to decide how much a particular item (e.g. "Am the life of the party") describes themselves and gave their answer using five-point Likert scale ("*Very Inaccurate*", "*Moderately Inaccurate*", "*Neither Accurate nor Inaccurate*", "*Moderately Accurate*", or "*Very Accurate*"). The design of the IPIP measure and the underlying personality model are typical for numerous other instruments; therefore, conclusions reached in this study can be to some extant relevant to number of other instruments. Moreover, the IPIP is available in the public domain, which implies that it is not only free but could also be modified and its items could be listed in this paper without the publisher's consent (Goldberg, et al., 2006). Consequently, the results of this study might be applied in practice by creating new Big Five scales or item banks optimized for IRT models. As mentioned before, due to the space limitations, only one of the scales, Extraversion, was analysed in this paper. The list of items belonging to Extraversion scale can be found in *Table 1*.

Item	Item content
e1	Do not mind being the centre of attention
e2	Make friends easily
e3	Keep in the background
e4	Avoid contact with others
e5	Cheer people up
e6	Don’t talk a lot
e7	Warm up quickly to others
e8	Have little to say
e9	Talk to a lot of different people at parties
e10	Keep others at a distance
e11	Know how to captivate people
e12	Don’t like to draw attention to myself
e13	Am the life of the party
e14	Find it difficult to approach others
e15	Am skilled in handling social situations
e16	Retreat from others
e17	Feel comfortable around people
e18	Would describe my experiences as somewhat dull
e19	Start conversations
e20	Am hard to get to know

Table 1. List of Extraversion scale items

Participants

The participants (N = 182,922) completed the personality test between 1 January 2008 and 9 January 2009 on Facebook.com, a popular social networking website. The test was administered via the MyPersonality.org Facebook application (Stillwell, 2009) that was offering its users free on-line personality assessment together with extra features such as comparisons with friend’s personality profiles. More details regarding the MyPersonality application and testing procedures can be found in Appendix 1.

Protocol validity

Protocol validity refers to whether an individual record can be scored with the standard scoring schema (Kurtz & Parrish, 2001). The accuracy of psychometric properties estimation of any instrument can be severely reduced by invalidity of individual protocols (Buchanan, Johnson, & Goldberg, 2005). Even a well-established and validated personality measure can produce invalid results in individual cases due to technical problems, linguistic incompetence, inattentiveness, and deliberate misrepresentation. It is especially relevant to samples collected on the internet (Johnson, 2005). There is some evidence that web instruments may be affected by an increased number of unreliable and unmotivated participants, who would respond in a careless, dishonest, or mischievous way (Buchanan & Smith, 1999; Reips, 2000). However, Gosling et al. (2004) compared the results from online and paper-and-pencil measures and concluded that Internet-based findings are consistent with those based on traditional methods.

To minimize the negative influence of invalid protocols on the results of the current study, the dataset was screened for potentially invalid records using methodology proposed by Johnson (2005). SPSS scripts were written to detect and discard protocols: (1) containing long strings of consecutive identical responses longer than 6, 6, 8, 8, and 7 items (for response categories "*Very Inaccurate*", "*Moderately Inaccurate*", "*Neither Accurate nor Inaccurate*", "*Moderately Accurate*", or "*Very Accurate*" respectively); (2) containing long strings comprised of the recurring sets of consecutive responses longer than 10, 11, 12, 13, 14, and 15 items (for strings based on 2, 3, 4, 5, 6, and 7 items long sets respectively); (3) characterized by very low internal consistency, as indicated by Goldberg's Psychometric Antonym coefficient lower than 0.2 and Jackson's Individual Reliability coefficient lower than 0.4; (5) containing one or more missing answers. This procedure decreased the sample size by 36%, leaving 115,766 records.

Calibration and cross-validation samples

Calibration and cross-validation samples, both containing 20,000 records, were randomly selected from the dataset. The size of above samples was limited to 20,000 records, because that was the maximum number of cases that could be analyzed in the software used to examine model-data fit. Note that since reverse scoring of the negatively worded items is not needed under GGUM, all GGUM analyses were calculated using raw item-level data. However, negatively worded items were reverse scored for CTT and SGR analyses.

SGR model calibration and individual scores estimation

SGR item parameters were estimated on the calibration sample using the LTM package for latent variable modelling and Item Response Theory analyses (version: 0.9-0; Rizopoulos, 2006), running on R statistical package (version 2.8.1; Hornik, 2009). The same software was used to estimate individual scores of 2,000 randomly chosen individuals from the cross-validation sample that were used in the validity analysis. Both item parameters and individual scores were estimated using Maximum Likelihood Estimation. No constraints were used in parameter estimation.

GGUM calibration and individual scores estimation

GGUM item parameters were estimated on the calibration sample using the GGUM2004 computer program (v1.1; Roberts, Donoghue, & Laughlin, 2000; Roberts, Fang, Cui, & Wang, 2006). Although GGUM2004 limits the size of the calibration sample to 2,000 records, Roberts (personal communication, 2009) proposed a method to overcome this limitation. The calibration sample was randomly split in 10 subsamples of 2,000 cases and item parameters were estimated separately for each group. To ascertain that the theta scale did not vary among 10 calibrations, GGUMLINK (Roberts, 2001) computer program was used to equate parameter estimates derived from separate calibrations. The final GGUM item parameters were calculated by averaging item parameters estimated in 10 calibrations. GGUM2004 was also used to calculate individual scores that were used in the validity analysis. Both item parameters and individual scores were estimated using Maximum Likelihood Estimation. No constraints were used in parameter estimation.

Model-data fit

Model-data fit of both IRT models was examined using fit-plots and chi-squared residuals. MODFIT v2.0 software (Stark, 2001) was used to plot fit-plots and calculate chi-squared residuals for single items (singlets), pairs of items (doublets), and three items (triplets) in the cross-validation sample. To allow for comparisons with other studies chi-squared residuals were adjusted to a constant sample size of 3,000 and divided by their degrees of freedom. Adjusted residuals higher than 3 were treated as a proof of a poor fit (Drasgow, Levine, Tsien, & Williams, 1995). To summarise the large number of adjusted chi-squared/df ratios, they were sorted into six intervals: very small (<1), small (≥1 and <2), medium (≥2 and <3), moderately large (≥3 and <4), large (≥4 and <5), and very large (≥5). Means and standard deviations of adjusted chi-squared/df ratios were also computed for each of the scales.

Fit-plots of each response category of each item were analyzed. A discrepancy between the empirical curve and ORF exceeding the empirical curve's 95% confidence interval was treated as a misfit (Drasgow, Levine, Tsien, & Williams, 1995).

Development of the optimized scales

Traditional CTT-based scale optimisation. To select candidates for the 10-item and 4-item scales, the standard approach to personality scale construction as described for instance by Rust and Golombok (2009) was followed. First, an item's facility and item's discrimination were calculated. Facility in the Likert-style items equals the average response option for a particular item. The discrimination coefficient was the item-scale correlation (not including the item in the scale prior to computing the correlation; Murphy & Davidshofer, 2001). Second, items were ranked based on their discrimination parameter. Equal numbers of negatively scored and positively scored items characterized by the highest discrimination were chosen for the 10-item and 4-item versions of the scale. Finally, the distribution of the total scores was analyzed and some items were swapped to normalise the distribution of total scores. The item difficulty parameter was used as a cue to replace items that lead to total score distribution skewness. The final version of the scale was subjected to the principal axis factoring to ensure that it was unidimensional.

Dominance and ideal point IRT scales optimization. To select items for IRT optimized scales, item parameters were estimated using calibration sample, and model-data fit was verified on the cross-validation sample. Optimized scales were constructed separately for dominance and ideal point IRT models, by removing the items that were violating the assumption of unidimensionality while trying to maintain items contributing highly to the test's information. The following procedure was employed. First, negatively and positively worded items were ranked separately based on their discrimination parameter. Second, fit-plots and chi-squared goodness-of-fit statistics for singlets were analyzed and items showing poor fit were removed. Third, adjusted chi-squared/df ratios tables for item pairs and triplets were analyzed in order to identify possible violations of local independence. In case of violation, the item of lowest discrimination rank was removed from the given item pair or triplet. Fourth, the five most discriminating, positively worded items, and five most discriminating, negatively worded items were selected from among the remaining items as candidates to a final scale. Item difficulty parameters were analyzed and some swapping was done in an effort to produce a scale that was informative across a wide range of latent trait. Finally, models were recalibrated on the 10-item and 4-item optimized scales and model-data fit was examined once again to ascertain that IRT assumptions were met.

Comparison of the CTT and IRT scales accuracy

As discussed before, the standard error of measurement estimates in IRT and CTT cannot be simply compared. Therefore, it was decided to apply CTT based split-half reliability analysis to both IRT and CTT scales. To calculate the split-half reliability, the scales investigated in this study were split into two half-scales (consisting odd and even numbered items). Next, individual scores were estimated for each of the halves using cross-validation sample. Note that IRT models were not recalibrated on the half-scales, instead the item parameters estimated for the whole scale were used. Subsequently, half-scale scores were correlated using The Pearson product-moment correlation. Resulting correlation coefficient was equivalent to the scale's split-half reliability (for the proof see: Lord & Novick, 1968). Finally, scale's standard error of measurement was calculated using the standard equation:

$$SEM = \sqrt{1 - Reliability}.$$

Standard error of measurement estimated in the way described above (*split-half SEM*) was compared between CTT and IRT scales. Moreover, split-half SEM was compared with conditional SEM function and average conditional SEM of the scores.

Validity

Conventionally, validity is described as the extent to which a psychometric instrument measures what it is supposed to measure (Rust & Golombok, 2009). Psychometric instrument cannot be valid and inaccurate (because inaccurate instrument does not measure anything well), but can be very accurate and completely invalid (as it could accurately measure something unrelated to the concept under examination). It implies that increase of the measurement accuracy is advantageous only when it is accompanied by the increase of the instrument's validity.

Although the current study was not focused on the issue of validity, validity must have been controlled in order to ascertain that comparisons between accuracy of the various methods make sense. To ascertain that the application of the IRT does not decrease the amount of valid information in the scores, the validity of all the scales developed in this study (including original 20-item scored using IRT methods) was examined by correlating individual scores with scores achieved in the 20-item CTT scale (concurrent validity; Murphy & Davidshofer, 2001).

Results

20-item Extraversion scale

CTT, GGUM, and SGR item parameters that were estimated on the original 20-item calibration sample are presented in *Table 2.* Negatively worded items were in general much less discriminating than positively worded items. Note that GGUM location parameter was located far from the average levels of the latent trait (the minimum absolute value of the location parameter is 2.39, for item e17). Extreme values of the GGUM location parameter showed that IPIP items were highly monotonic.

	CTT		SGR					GGUM					
Item	Discrimination	Difficulty (1-5)	Discrimination (a)	Difficulty parameters (b)				Discrimination (∝)	Location (δ)	Response thresholds (τ)			
				b1	b2	b3	b4			τ_1	τ_2	τ_3	τ_4
e1	.55	3.41	0.75	-2.51	-1.12	-0.15	1.37	0.70	3.56	-5.63	-4.28	-4.16	-2.24
e2	.73	3.75	1.31	-2.16	-1.16	-0.50	0.62	1.58	2.43	-4.30	-3.30	-3.13	-1.95
e3	**.69**	**3.15**	**1.09**	**-1.88**	**-0.60**	**0.28**	**1.43**	**1.19**	**-3.17**	**-4.40**	**-3.29**	**-2.71**	**-1.44**
e4	**.66**	**3.89**	**1.08**	**-2.77**	**-1.45**	**-0.66**	**0.43**	**1.19**	**-3.81**	**-3.93**	**-3.03**	**-2.78**	**-1.17**
e5	.51	4.19	0.71	-4.44	-3.00	-1.69	0.45	0.87	2.80	-6.02	-5.07	-4.77	-2.44
e6	**.68**	**3.57**	**1.08**	**-2.00**	**-0.99**	**-0.28**	**0.66**	**1.09**	**-3.47**	**-3.73**	**-3.04**	**-2.82**	**-1.74**
e7	.58	3.61	0.84	-2.80	-1.38	-0.45	1.20	0.87	2.94	-5.43	-3.87	-3.83	-1.74
e8	**.58**	**3.83**	**0.83**	**-2.77**	**-1.52**	**-0.71**	**0.57**	**0.83**	**-4.09**	**-4.27**	**-3.02**	**-3.18**	**-1.60**
e9	.72	3.25	1.27	-1.45	-0.52	0.04	0.98	1.33	2.61	-3.79	-2.84	-2.93	-1.80
e10	**.60**	**3.42**	**0.82**	**-2.52**	**-0.92**	**-0.15**	**1.11**	**0.80**	**-3.72**	**-4.47**	**-3.16**	**-3.50**	**-1.13**
e11	.54	3.56	0.75	-3.08	-1.55	-0.28	1.44	0.79	3.30	-5.79	-4.66	-3.86	-1.89
e12	**.57**	**3.00**	**0.78**	**-1.85**	**-0.48**	**0.45**	**1.90**	**0.74**	**-3.49**	**-5.21**	**-3.48**	**-3.33**	**-1.85**
e13	.70	2.80	1.16	-1.15	-0.23	0.65	1.68	1.27	2.78	-3.56	-3.08	-2.26	-1.27
e14	**.75**	**3.32**	**1.42**	**-1.59**	**-0.56**	**-0.05**	**0.88**	**1.59**	**-2.90**	**-3.60**	**-2.57**	**-2.63**	**-1.49**
e15	.71	3.60	1.21	-2.13	-1.11	-0.36	0.93	1.42	2.72	-4.49	-3.58	-3.32	-1.90
e16	**.66**	**3.61**	**1.04**	**-2.55**	**-1.13**	**-0.31**	**0.83**	**1.10**	**-3.55**	**-4.12**	**-3.05**	**-2.79**	**-1.06**
e17	.76	3.80	1.56	-2.30	-1.27	-0.57	0.64	1.99	2.39	-4.45	-3.41	-3.09	-1.85
e18	**.50**	**3.82**	**0.63**	**-3.23**	**-1.65**	**-0.80**	**0.57**	**0.58**	**-4.32**	**-4.11**	**-2.90**	**-3.75**	**-1.48**
e19	.75	3.75	1.47	-2.24	-1.19	-0.54	0.72	1.90	2.51	-4.51	-3.42	-3.23	-1.89
e20	**.53**	**3.12**	**0.71**	**-1.72**	**-0.51**	**0.14**	**1.34**	**0.57**	**-4.11**	**-4.71**	**-3.39**	**-4.63**	**-2.70**

Note: parameters of the negatively worded items are written in bold.

Table 2. CTT, SGR, and GGUM item parameters of the 20-item Extraversion scale

Consequently, in the moderate areas of the latent trait, ideal point GGUM ORFs resembled monotonic SGR ORFs as can be seen in *Figure 11* and *Figure 12* representing SGR and GGUM ORFs for the same item. Option Response Functions and fit-plots for SGR and GGUM models can be found in Appendices 2 and 3.

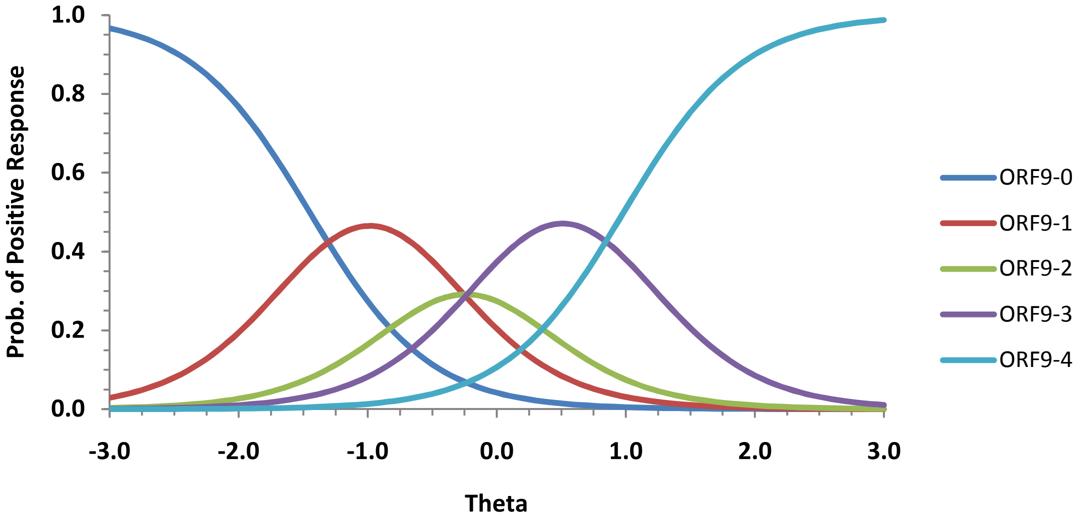

Figure 11. SGR Option Response Functions for item e9

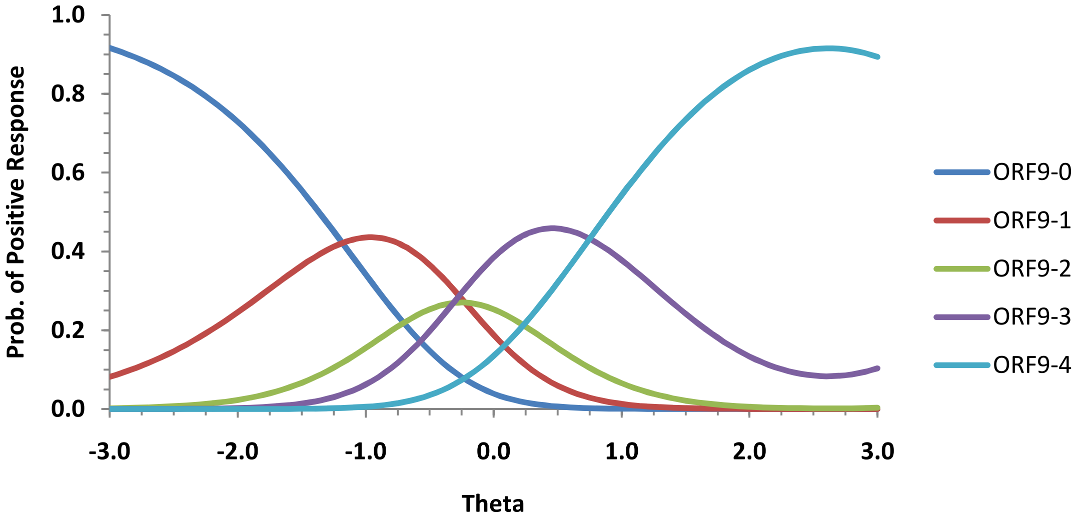

Figure 12. GGUM Option Response Functions for item e9

Optimized 10-item Extraversion scales

Item properties, model-data fit statistics, and fit-plots (fit-plots and model-data fit statistics are discussed in the next section) were used to select items for optimized 10-item scales. Please note that tables consisting adjusted chi-squared/df ratios for singlets, doublets, and triplets that were used in assessing model-data fit are very large (190 pairs and 1140 triplets for single 20-item scale) and are not presented in this paper. The summary of adjusted chi-squared/df ratios can be found in *Table 5*. Since GGUM model did not show significant improvement in its data-model fit, a number of alternative scale versions were tested and one that offered the best fit was selected. The resulting optimized 10-item scales were similar, and they all shared seven items: e2, e3, e4, e6, e9, e14, and e17. CTT and IRT item parameters estimated for the optimized 10-item scales are presented in *Table 3*. Option Response Functions and fit-plots for 10-item scales optimized for SGR and GGUM models can be found in Appendixes 4 and 5.

	CTT		SGR					GGUM					
Item	Discrimination	Difficulty (1-5)	Discrimination (a)	Difficulty parameters (b)				Discrimination (∝)	Location (δ)	Response thresholds (τ)			
				b1	b2	b3	b4			τ_1	τ_2	τ_3	τ_4
e2	3.75	.72	1,39	-2,08	-1,13	-0,49	0,60	1,66	2,32	-4,20	-3,20	-3,00	-1,83
e3	**3.15**	**.67**	**1,05**	**-1,86**	**-0,62**	**0,29**	**1,44**	**1,11**	**-3,32**	**-4,57**	**-3,44**	**-2,86**	**-1,54**
e4	**3.89**	**.66**	**1,07**	**-2,76**	**-1,45**	**-0,66**	**0,43**	**1,17**	**-3,84**	**-3,97**	**-3,05**	**-2,80**	**-1,16**
e6	**3.57**	**.66**	**1,06**	**-1,97**	**-0,96**	**-0,26**	**0,67**	**0,98**	**-3,59**	**-3,83**	**-3,14**	**-2,95**	**-1,79**
e7	-	-	-	-	-	-	-	0,86	2,74	-5,26	-3,68	-3,63	-1,51
e9	3.25	.73	1,33	-1,40	-0,50	0,03	0,95	1,41	2,78	-3,97	-3,02	-3,07	-1,98
e13	2.80	.68	-	-	-	-	-	1,24	3,09	-3,88	-3,39	-2,56	-1,57
e14	**3.32**	**.75**	**1,60**	**-1,48**	**-0,54**	**-0,05**	**0,84**	**1,57**	**-2,91**	**-3,62**	**-2,58**	**-2,64**	**-1,47**
e15	-	-	1,30	-2,02	-1,06	-0,33	0,88	-	-	-	-	-	-
e16	**3.61**	**.65**	-	-	-	-	-	-	-	-	-	-	-
e17	3.80	.75	1,57	-2,25	-1,25	-0,56	0,62	1,95	2,22	-4,32	-3,25	-2,93	-1,66
e18	-	-	**0,66**	**-3,06**	**-1,58**	**-0,78**	**0,55**	**0,55**	**-4,42**	**-4,18**	**-2,93**	**-3,84**	**-1,47**
e19	3.75	.75	1,63	-2,13	-1,14	-0,51	0,68	-	-	-	-	-	-

Note: Parameters of the negatively worded items are written in bold.

Table 3. CTT, SGR, and GGUM item parameters of the 10-item optimized Extraversion scales

Optimized 4-item Extraversion scales

Similarly for the 10-item scale, the procedures discussed in the methods section were used to select items for the 4-item scales. Again, GGUM model did not show improvement in model-data fit, and many alternative scales were tested in order to select the best fitting one. Resulting optimized scales were identical for both IRT models and contained items: e2, e3, e14, and e19. In the scale optimized for CTT, item e2 was replaced by item e17. CTT and IRT item parameters that were estimated on the optimized 4-item scales are presented in *Table 4*. Option Response Functions and fit-plots for 4-item scales can be found in Appendixes 6 and 7.

	CTT		SGR					GGUM					
Item	Discrimination	Difficulty (1-5)	Discrimination (a)	Difficulty parameters (b)				Discrimination (∝)	Location (δ)	Response thresholds (τ)			
				b1	b2	b3	b4			τ_1	τ_2	τ_3	τ_4
e2	-	-	1.36	-2.1	-1.1	-0.5	0.59	1.44	-2.42	-4.4	-3.3	-3.2	-1.9
e3	**3.15**	**.61**	**1.00**	**-1.9**	**-0.6**	**0.29**	**1.46**	**0.96**	**3.40**	**-4.7**	**-3.5**	**-2.9**	**-1.5**
e14	**3.32**	**.72**	**1.75**	**-1.4**	**-0.5**	**-0.1**	**0.81**	**2.04**	**3.10**	**-3.9**	**-2.9**	**-2.8**	**-1.7**
e17	3.80	.69	-	-	-	-	-	-	-	-	-	-	-
e19	3.75	.71	1.64	-2.1	-1.1	-0.5	0.67	1.91	-2.45	-4.5	-3.4	-3.2	-1.8

Note: Parameters of the negatively worded items are written in bold.

Table 4. CTT, SGR, and GGUM item parameters of the 4-item optimized Extraversion scales

Model-data fit

The summary of the adjusted chi-squared/df ratios is presented in *Table 5*. Model-data fit in the current study was comparable or better than model-data fit obtained in the previous applications of the IRT models to the IPIP Big Five personality scales. For instance, Chernyshenko et al. (2001) reported that mean adjusted chi-squared/df ratios for SGR model and 10-item IPIP Big Five scales ranged from 3.18 to 8.28. Considerably better mean fit statistics achieved in this study (2.24 to 3.71) can be attributed to the fact that the calibration sample was over 10 times bigger than that used by Chernyshenko and colleagues (2001).

Attempts to improve model-data fit of the GGUM undertaken in this study proved ineffective. GGUM recalibrated on the 10 and 4 item long optimized scales created by discarding the least accurate and poorly fitting items, did not fit the data better than the model calibrated on the original 20-item scale. Mean adjusted chi-squared/df ratios for optimized 4-item GGUM scale (1.53, 4.34, and 3.12, for singlets, doublets, and triplets respectively) were only slightly better than those achieved on the 20-item original scale (1.51, 4.78, and 2.93, for singlets, doublets, and triplets respectively). However, the

		Frequency distribution of the adjusted chi-squared/df ratios								
Model		<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
					20-item original scale					
SGR	Singlets	1	15	3	1				1.78	0.62
	Doublets		24	27	69	53	9	8	3.71	1.45
	Triplets		272	475	356	36	1		2.53	0.66
GGUM	Singlets		16	4					1.51	0.48
	Doublets			31	32	39	72	16	4.78	1.53
	Triplets		283	451	362	41	3		2.93	0.57
					10-item optimized scales					
SGR	Singlets	3	6	1					1.41	0.45
	Doublets		4	21	14	4	2		3.09	0.99
	Triplets		33	80	7				2.24	0.36
GGUM	Singlets		9	1					1.46	0.37
	Doublets			6	11	12	14	2	4.52	1.40
	Triplets		1	54	61	4			3.06	0.53
					4-item optimized scales					
SGR	Singlets		4						1.30	0.30
	Doublets			4	2				2.93	0.53
	Triplets		1	3					2.25	0.25
GGUM	Singlets		3	1					1.53	0.41
	Doublets				3	1	2		4.34	0.80
	Triplets			2	2				3.12	0.27

Note: all chi-squared values were adjusted for sample size N=3000 and divided by degrees of freedom.

Table 5. Frequency, Means and SD of adjusted chi-squared/df ratios for Extraversion scales of different length optimization procedure proved fruitful in the case of the SGR model. Fit statistics of the 4-item SGR scale (1.30, 2.93, and 2.25, for singlets, doublets, and triplets respectively), were better than those of the 10-item SGR scale (respectively: 1.41, 3.09, and 2.24). Accordingly, the original 20-item SGR scale showed poorest fit (respectively: 1.78, 3.71, and 2.53). In general, examination of the adjusted mean adjusted chi-squared/df ratios showed that dominance SGR model fit the data better than ideal point GGUM.

Inspection of the fit-plots that can be found in Appendices 2 to 7 showed that most of ORFs fitted data well, even though due to the huge sample size, empirical curve's confidence intervals were very narrow. It could be seen that discrepancies between ORF and empirical curves were present mostly on the extreme levels of the latent trait and that GGUM showed slightly worse fit than SGR model. *Figure 13* and *Figure 14* present the fit-plots for both models and the same response option. Note that SGR ORF was closer to the empirical curve than GGUM ORF was.

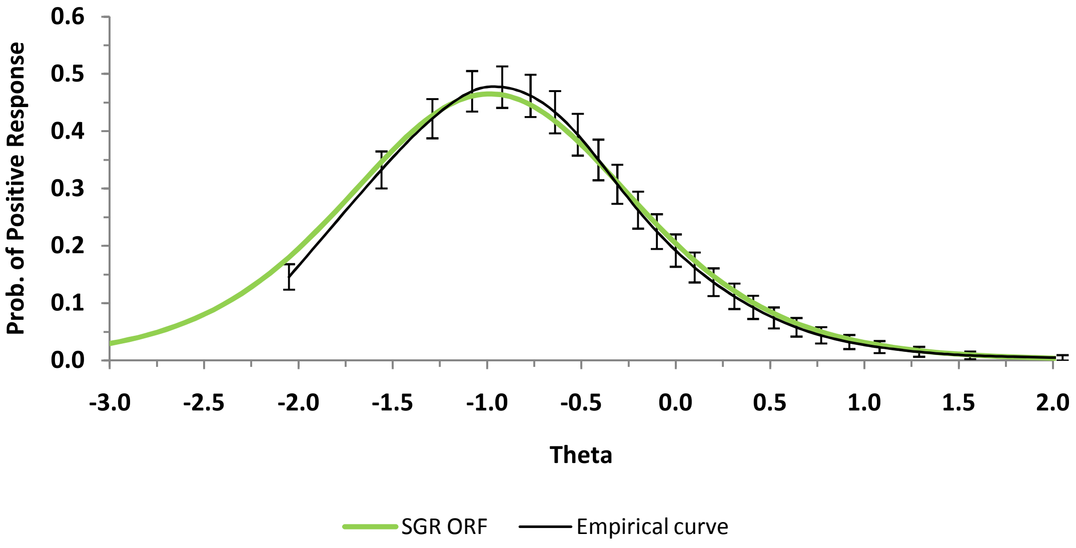

Figure 13. Fit-plot for the 20-item scale, representing SGR's ORF of the response option 1 ("Inaccurate") on item 9, together with an empirical curve

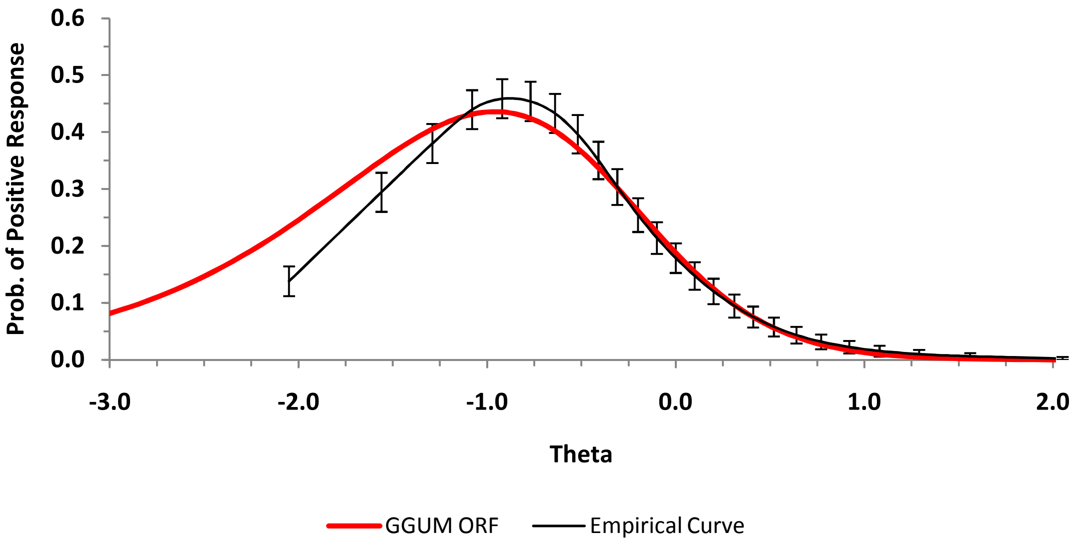

Figure 14. Fit-plot for the 20-item scale, representing GGUM's ORF of the response option 1 ("Inaccurate") on item 9, together with an empirical curve

Comparison of the CTT and IRT accuracy

Split-half SEM estimates, and the average conditional SEM of IRT scores are presented in *Table 6*. Comparison of the split-half SEM values between measurement methods revealed that IRT methods did not offer higher accuracy than the CTT approach in estimating individual scores for IPIP Extraversion scale. Application of the SGR model decreased the split-half SEM on the both optimized scales by 0.01 compared with CTT, whereas scores estimated with GGUM had between 0.02 and 0.03 more SEM than same scores estimated with CTT. It was apparent that GGUM offers less measurement accuracy than both CTT and SGR on every length of the scale.

The comparison of the average conditional SEM of the IRT scores and split-half SEM showed that SEM values stemming from the information function are significantly lower than those based on the true score theory. This is also illustrated in *Figure 15*, *Figure 16*, and *Figure 17*, which compare conditional SEM curves together with split-half SEM values. For example, conditional SEM in 20-item SGR scale is below 0.25 for the broad range of theta values (-2, to 1), which is consistent with average conditional SEM in the scores level (0.24), whereas the split-half SEM was more than 30% higher and equalled 0.33.

Interestingly, inspection of the conditional SEM curves presented on *Figure 15*, *Figure 16*, and *Figure 17* shows that SGR scales were accurate on the broader intervals of theta continuum than GGUM scales.

Method	Split-half SEM[1]	Average conditional SEM[2]
20-item original scale		
CTT	0.33	-
SGR	0.33	0.24
GGUM	0.36	0.24
10-item optimized scales		
CTT	0.41	-
SGR	0.40	0.28
GGUM	0.43	0.30
4-item optimized scales		
CTT	0.52	-
SGR	0.51	0.37
GGUM	0.55	0.40

[1] SEM calculated using split-half reliability coefficient
[2] Average conditional SEM in the IRT score estimates
Average conditional SEM and split-half SEM values are different with the significance level of *0.001*

Table 6. Estimates of the standard error of measurement

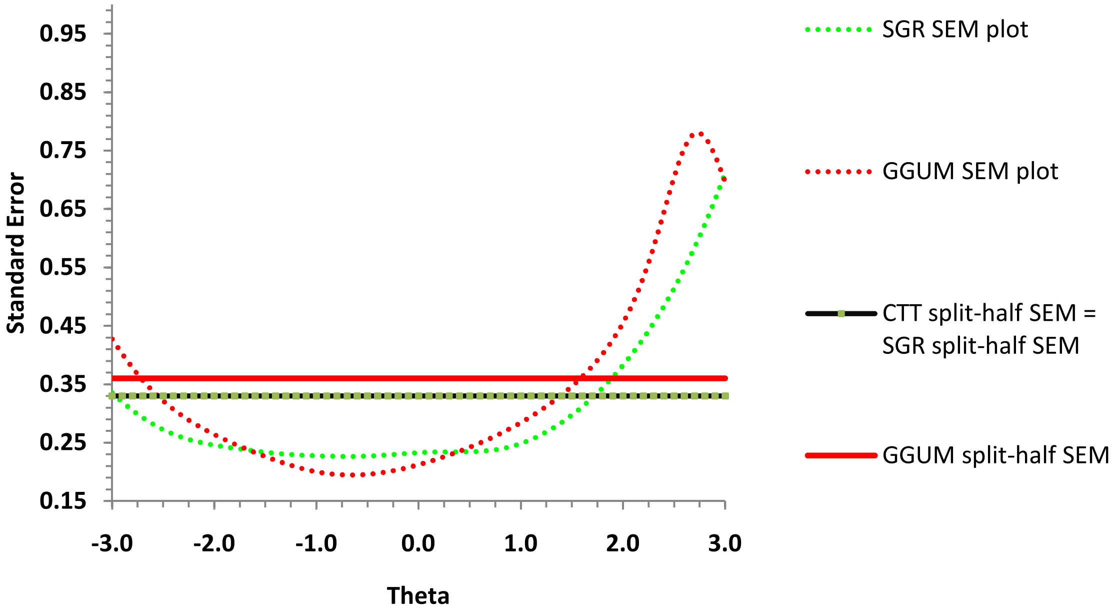

Figure 15. Split-half SEM and conditional SEM plots for 20-item scales

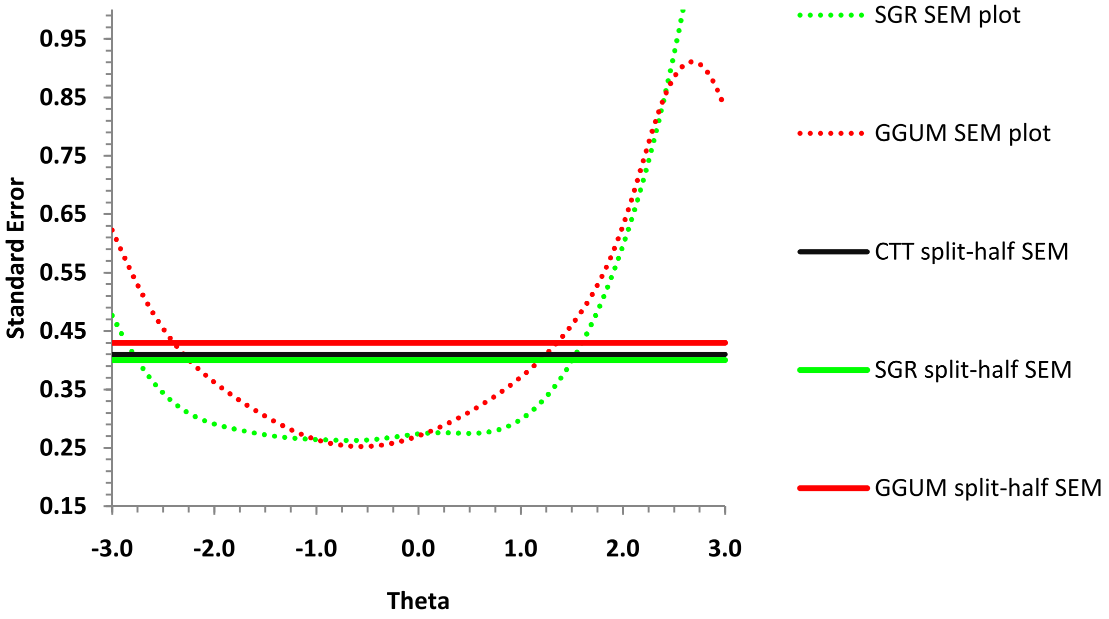

Figure 16. Split-half SEM and conditional SEM plots for 10-item scales

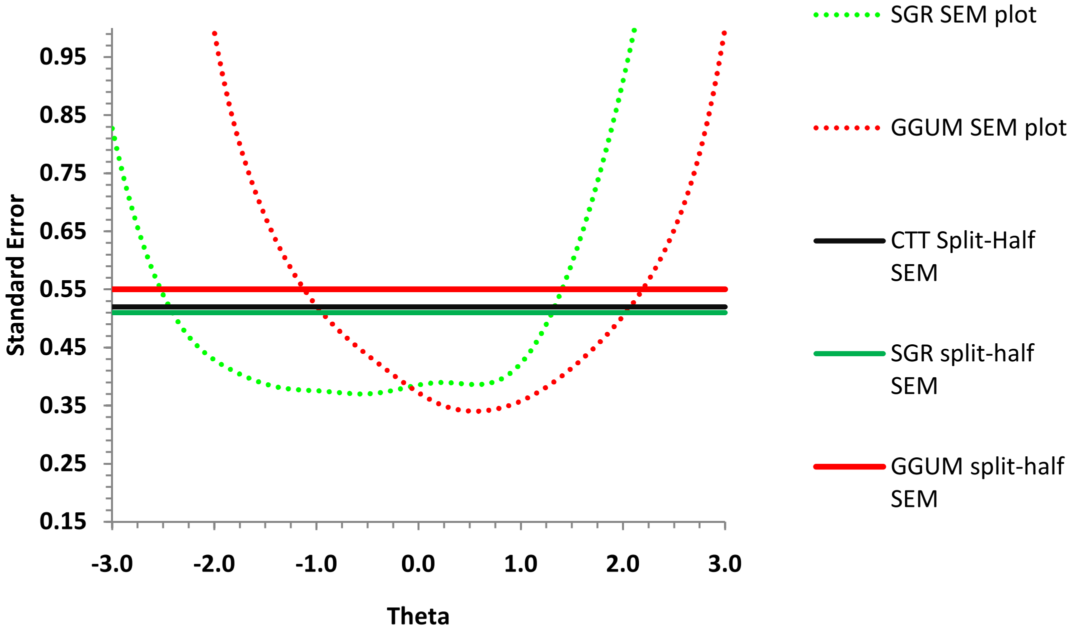

Figure 17. Split-half SEM and conditional SEM plots for 4-item scales

Validity

The inter-correlations between scores estimated on different scales are presented in *Table 7*. It can be seen that estimated scores were very similar, regardless of the method and scale length. The highest correlation was found between 20-item scales (.99 to .98) and the lowest between 4-item and 20-item scales (.91 to .94).

High levels of concurrent validity with well-established IPIP Extraversion scale showed that IRT scales were valid. However, high concurrent validity and similar or lower accuracy in comparison with CTT implied that IRT methods failed to deliver new quality of measurement in the scale under examination.

Examination of the scatter plot of the SGR and standardized CTT scores on the 20-item scale presented in *Figure* 18 revealed that the distribution of the scatter-points was S shaped. The similar distribution was found for every other IRT scores paired with CTT scores. It shows that the IRT methods were characterized by the broader range of results on the extremes of the latent trait continuum.

Method		SGR			GGUM			CTT		
	Scale	20	10	4	20	10	4	20	10	4
SGR	20	-	.98	.94	.99	.97	.93	.99	.97	.93
	10		-	.96	.97	.97	.96	.96	.97	.95
	4			-	.93	.93	.93	.92	.94	.96
GGUM	20				-	.97	.93	.98	.96	.92
	10					-	.93	.95	.97	.92
	4						-	.91	.94	.96
CTT	20							-	.97	.93
	10								-	.95
	4									-

Table 7. Correlations between stores estimated on different scales

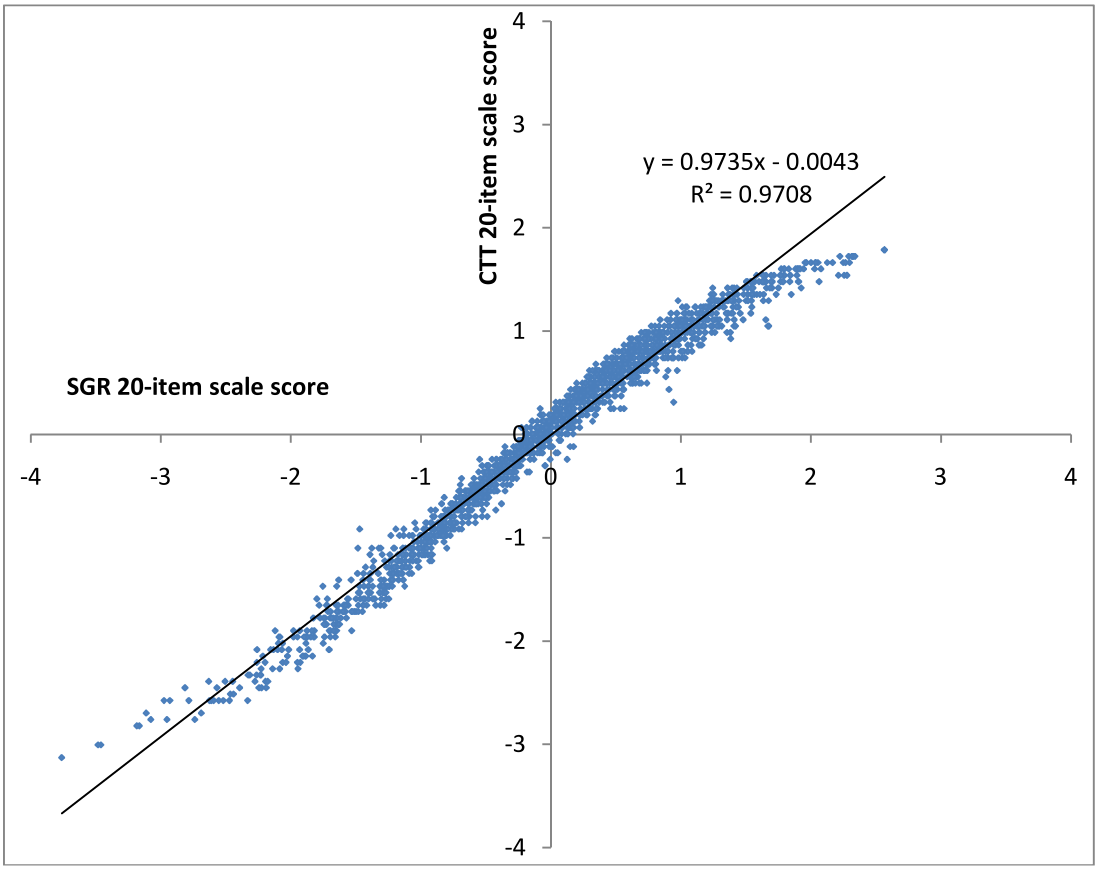

Figure 18. Scatter plot of the 20-item SGR and CTT scale scores with the trend-line

Discussion

In author's opinion, the most interesting finding of this study is that, in spite of the fact that IRT apply sophisticated models to describe the relationship between individual responses and the latent trait, it fails to increase the measurement accuracy of the 20-item Extraversion scale. Moreover, the individual results estimated with IRT methods were practically identical to those calculated by an unweighted linear sum of item responses. This suggests that currently available IRT methods do not increase the measurement accuracy and validity of the CTT-based scales.

In addition, numerous inconveniences of the IRT methodology application become apparent during the course of this research. First, since IRT is far more complex than CTT, its application requires much more effort and time than the application of CTT methods. Second, to perform the model calibration and verify the model-data fit of the 2 IRT models used in this study, the minimum number of 4 different computer programs[3] must be applied. Furthermore, IRT software is at times undependable and poorly documented. Third, IRT analyses tend to be very laborious. For example, assessment of the data-model fit of the single model on the 20-item scale includes examination of around 100 item parameters, 100 fit-plots, and the tables consisting 1350 adjusted chi-squared/df ratios. After selecting poorly fitting items to be removed, the model must be recalibrated and the whole procedure repeated on the remaining items. In order to develop a well-performing scale the above process has to be repeated several times. Finally, huge datasets are required in order to calibrate the IRT model. The importance of the calibration sample size can be easily recognized by analyzing

[3] Programs that were necessary to apply GGUM and SGR models: GGUM2004, GGUMLINK (DOS-based), MODFIT v2.0 running on Microsoft Excel, ltm package running on the R programme.

Table 8, which is discussed in more detail below. Note that the adjusted chi-squared/df ratios for singlets decreased by half when the size of the calibration sample was increased from 2,000 to 20,000.

The inconvenience of using IRT methods combined with the near-identical accuracy of IRT and CTT approaches are strong arguments against abandoning well-established CTT methods in the area of the personality assessment (see also Hambleton & Jones, 1993; or Reise & Henson, 2003 for the discussion of choosing between IRT and classical methods). However, IRT methodology offers numerous advantages over CTT, so its application could be still beneficial in some cases. It requires a large sample, but it can produce less biased estimates from unrepresentative samples (Embretson & Reise, 2000), allows controlling measurement accuracy at different levels of the latent trait (Baker, 2001), offers convenient methods of item and test bias detection (Rust & Golombok, 2009), and can be used in CAT applications of the personality measures (Waller & Reise, 1989). Moreover, the results of this study showed that a well-fitting IRT model was as valid and as accurate when applied to CTT-based scale as well-established CTT methods. Consequently, it is possible that scales developed completely using an IRT approach and scored with IRT methods will be more accurate, more valid, and less biased than present CTT-based measures.

Although it was hypothesized that the ideal point GGUM would fit the IPIP Extraversion scale better than dominance SGR, the results of this study show that GGUM offered a worse fit than SGR model when applied to IPIP Extraversion scale. Despite the fact that the adjusted chi-squared/df ratios for singlets were similar between the GGUM and SGR models, inspection of the fit-plots revealed that GGUM did not fit data equally well as SGR. Moreover, GGUM showed significantly worse fit than SGR when adjusted chi-squared/df ratios for doublets and triplets were examined. There are two possible explanations of the GGUM relative misfit. First, the source of GGUM's relative misfit could stem from the GGUM calibration methodology used in this study. Note that due to the limitations of the software used in GGUM model calibration, item parameters were estimated by averaging the results of 10 independent sub-calibrations. This was not an issue in the preceding studies, as the calibration samples used previously did not exceed the GGUM2004 software limits of 2,000 cases. Above hypothesis can be easily verified by comparing the model-data fit of the SGR and GGUM models calibrated on the sample small enough to be processed by the GGUM calibration software in one cycle. *Table 8* contains the comparison of data-model fit between GGUM and SGR models calibrated on the calibration sample containing 20,000 and 2,000 cases. It can be seen that regardless of the calibration sample size the relationship between GGUM and SGR models fit was constant. GGUM had lower mean adjusted chi-squared/df ratios for singlets, much higher mean adjusted chi-squared/df ratios for doublets and higher adjusted chi-squared/df ratios for triplets. The fact that model-data fit of the SGR and GGUM models had increased proportionally with the calibration sample size indicated that the relatively poor fit of the GGUM was not related to calibration methodology used in this study.

A second possible explanation of GGUM's relatively poor fit might be related to the fact that the Extraversion scale studied here was composed of highly monotonic items. Although previous studies showed that GGUM could fit both monotonic and nonmonotonic items equally well as dominance models (Chernyshenko, Stark, Drasgow, & Roberts, 2007), a visual inspection of the fit-plots and adjusted chi-squared/df ratios for doublets and triplets suggests that this was not the case in this study. Consequently, the fit of the GGUM model to the highly monotonic items should be explored in more detail in future studies.

Model		Frequency distribution of the adjusted chi-squared/df ratios								
		<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
					20-item original scale					
SGR 20,000	Singlets	1	15	3	1				1.78	0.62
	Doublets		24	27	69	53	9	8	3.71	1.45
	Triplets		272	475	356	36	1		2.53	0.66
GGUM 20,000	Singlets		16	4					1.51	0.48
	Doublets			31	32	39	72	16	4.78	1.53
	Triplets		283	451	362	41	3		2.93	0.57
SGR 2,000	Singlets		7	2	5	3	2	1	3.31	1.81
	Doublets			4	55	47	4	8	4.32	1.78
	Triplets		19	475	475				2.73	0.78
GGUM 2,000	Singlets		5	6	5	3	1		2.93	1.32
	Doublets			24	16	32	87	31	5.38	1.62
	Triplets			38	76				3.17	0.57

Note: all chi-squared values were adjusted for sample size N=3000 and divided by degrees of freedom

Table 8. Frequency, Means and SD of adjusted chi-squared/df ratios for IRT models calibrated on the samples of different size

Development of optimized scales under CTT and IRT was another important part of this study. The author supposed that improvement in model-data fit achieved by removing items violating IRT model assumptions, would lead to an increase in the measurement accuracy of the IRT scale in relation to the CTT scale of the same length. However, the results of this study showed that significant improvement in data-model fit (e.g. between SGR 20-item scale and SGR 4-item scale) did not change the ratio between CTT scale’s SEM and IRT scale’s SEMs. Moreover, while the procedures aimed at improving the data-model fit turned out to be successful in case of the SGR model, the same procedures failed in the case of GGUM model. It shows that GGUM poor model-data fit cannot be attributed to the several particularly poor-fitting items, but is a property of the whole Extraversion scale.

This study shows that significant work must be done before IRT methods could be widely applied in the area of personality assessment. First, since the model-data fit of the popular IRT models used in this study was at most average, it seems that more IRT models should be applied to different kinds of personality scales in order to identify those providing the best fit. Second, results presented here suggest that suitability of the ideal point GGUM model to the monotonic items could be worse than previously suggested (e.g. Stark, Chernyshenko, Drasgow, & Williams, 2006). Third, the application of the IRT models would become significantly easier if the software and procedures aimed at improving the model-data fit were improved. Finally, the IRT methods used in personality assessment should be enhanced with the score's reliability measure reflecting the actual score reliability on different levels of the latent trait.

This study was original in several ways. First, the measurement accuracy was compared between the IRT and CTT approaches, whereas in previous studies on the application of IRT in personality assessment, the supremacy of the IRT approach over CTT methods has been taken for granted (e.g. Rauch, Schweizer, & Moosbrugger, 2008). Second, in this study, polytomous IRT models were applied, whereas formerly the binary models were typically used. Finally, the current study was based on a dataset several times larger than those used in preceding studies on the subject (e.g. Baker, Rounds, & Zevon, 2000), which increases the power of significance tests and decreases the likelihood of sampling error.

There were numerous limitations of this study, and the two most significant in the author's opinion are listed below. First, items examined here were developed using dominance CTT methods. It is likely that analysis based on the items designed to be used with dominance or ideal point IRT methods would yield very different results. Consequently, the conclusions reached here are limited to the IRT applications to the CTT-based scales and items of instruments similar to IPIP Big Five questionnaire. Second, limited space allowed examining only one of the five personality scales from the Big Five questionnaire. Although, provisional analyses indicate that present findings can be generalized to the other scales, it would be desirable to supplement the results presented here with the similar results from the other scales.

In conclusion, the IRT methods did not show significant improvement over the well-established CTT approach. On the contrary, IRT application proved complex while results were roughly equivalent to the CTT approach of summing item responses. Although in some areas, such as computer adaptive testing, advantages of the IRT justify its application, this research indicates that CTT, despite having some notable weaknesses, is still appropriate for evaluating personality assessment data.

References

Baker, F. B. (2001). *The Basics of Item Response Theory (2nd edition).* ERIC Clearinghouse on Assessment and Evaluation.

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A Comparison of Graded Response and Rasch Partial Credit Models with Subjective Well-Being. *Journal of Educational and Behavioral Statistics , 25* (3), pp. 253-270.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data.* Boca Raton, FL: Taylor & Francis Group, LLC.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika , 37*, pp. 29-51.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology , 90*, pp. 125-145.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a Five-Factor Personality Inventory Personality Inventory. *European Journal of Psychological Assessment , 21*, pp. 115-127.

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

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing Personality Scales Under the Assumptions of an Ideal Point Response Process: Toward Increasing the Flexibility of Personality Measures. *Psychological Assessment , 19* (1), pp. 88-106.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology , 85*, pp. 451-461.

Costa, P. T., & McCrae, R. R. (2005). The revised NEO Personality Inventory (NEO-PI-R). In S. Briggs, J. Cheek, & E. Donahue, *Handbook of Adult Personality Inventories.*

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2 ed., pp. 577-636). Palo Alto, CA: Consulting Psychologists' Press.

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*, pp. 363-373.

Drasgow, F., & Parsons, C. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement , 7*, pp. 189–199.

Drasgow, F., Levine, M. V., Tsien, S., & Williams, B. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement , 19*, pp. 143-165.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Feldt, L., & Brennan, R. (1989). Reliability. In *Educational Measurement 3rd Ed.* (pp. 105-146). New York: Macmillan.

Goldberg, L. R., Johnson, L. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality , 40*, pp. 84-96.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist , 59*, pp. 93-104.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice , 12* (38).

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

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement , 9*, pp. 139-164.

Hol, A. M., Harrie, C. M., & Mellenbergh, G. J. (2008). Computerized Adaptive Testing of Personality Traits. *Journal of Psychology , 216*, pp. 12-21.

Hornik, K. (2009). *The R FAQ*. Retrieved 06 08, 2009, from http://www.r-project.org/: http://CRAN.R-project.org/doc/FAQ/R-FAQ.html

IPIP.org. (2009, 1 1). *A Comparison between the 5 Broad Domains in Costa and McCrae's NEO Personality Inventory (NEO-PI-R) and the Corresponding Preliminary IPIP Scales Measuring Similar Constructs*. Retrieved 1 14, 2009, from International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences: http://ipip.ori.org/newNEO_DomainsTable.htm

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality , 39*, pp. 103–129.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic Response Consistency and Protocol Validity in Structured Personality Assessment: The Case of the NEO–PI–R. *Journal of Personality Assessment , 16*, pp. 315-332.

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

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement , 28*, pp. 989-1020.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* MA: Addison-Wesley.

Maydeu-Olivares, A. (2005). Further Empirical Results on Parametric Versus Non-Parametric IRT Modeling of Likert-Type Personality Data. *Multivariate Behavioural Research , 40(2)*, pp. 261–279.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods , 9*, pp. 354–368.

Muraki, E. (1992). Ageneralized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement , 16*, pp. 159–176.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement , 14*, pp. 59-71.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and Applications. (5th ed.).* Upper Saddle River, NJ: PrenticeHall.

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement , 29* (3), pp. 213–225.

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2008). An IRT Analysis of the Personal Optimism Scale. *European Journal of Psychological Assessment , 24* (1), pp. 49-56.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. *Psychological Experiments on the Internet* , pp. 89-117.

Reise, S. P., & Henson, J. M. (2003). A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales. *Journal of Personality Assessment , 81* (2), pp. 93–103.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software , 17* (5), pp. 1–25.

Roberts, J. S. (2001). Equating parameters of the generalized graded unfolding model. *Paper presented at the 2001 annual meeting of the American Educational Research Association.* Seattle, WA.

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*, pp. 3-32.

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters of the generalized graded unfolding model. *Applied Psychological Measurement , 34*, pp. 64-65.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in Clinical Personality Measurement: An Item Response Theory Analysis of the MMPI–2 PSY–5 Scales. *Journal of Personality Assessment , 72(2)*, pp. 282-307.

Rust, J., & Golombok, S. (2009). *Modern Psychometrics (3rd edition).* Hove: Routledge.

Samejima, F. (1979). *A new family of models for the multiple-choice item.* Research report. 79-4 prepared under Office of Naval Research contract N00014-77-C-360, NR 150-402. Department of Psychology, University of Tennessee.

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

Stark, S. (2001). *MODFIT: A computer program for model-data fit. Unpublished manuscript.* University of Illinois at Urbana-Champaign.

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

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (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* (1), pp. 25–39.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology* , *81*, pp. 332–342.

Stillwell, D. (2009, 1 1). *MyPersonality Research.* Retrieved 1 1, 2009, from Mypersonality.org: http://mypersonality.org/research/interested-in-collaborating/

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* , *52*, pp. 589-617.

Thissen, D., & Steinberg, L. (1985). A response model for multiple choice items. *Psychometrika , 49*, pp. 501-519.

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

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring Nonlinear Models In Personality Assessment: Development and Preliminary Validation of a Negative Emotionality Scale. *Journal of Personality , 64* (3), pp. 545-576.

Waller, N., & Reise, S. (1989). Computerized adaptive personality assessment: An illustration with the Absorption Scale. *Journal of Personality and Social Psychology , 57*, pp. 1051–1058.

Weekers, A. M., & Meijer, R. R. (2008). Scaling Response Processes on Personality Items Using Unfolding and Dominance Models: An Illustration with a Dutch Dominance and Unfolding Personality Inventory. *European Journal of Psychological Assessment , 24(1)*, pp. 65-77.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology , 53*, pp. 774–789.

Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. Lubinski, & R. V. (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 49–79). Palo Alto, CA: Davis-Black.

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

List of Tables

List of Figures

List of Appendices

(Lord, 1968) (Bock, 1972) (Samejima, 1979) (Thissen & Steinberg, 1985) (Drasgow, Levine, Tsien, & Williams, 1995) (Rauch, Schweizer, & Moosbrugger, 2008), (Chernyshenko O. S., Stark, Chan, Drasgow, & Williams, 2001) (Waller, Tellegen, McDonald, & Lykken, 1996) (Chernyshenko O. S., Stark, Drasgow, & Roberts, 2007) (Zickar & Robie, 1999) (Weiss, 1985) (Weekers & Meijer, 2008) (Stark, Chernyshenko, Drasgow, & Williams, 2006)(Chernyshenko O. S., Stark, Drasgow, & Roberts, 2007)(Samejima, 1969)(Embretson & Reise, 2000)(Waller & Reise, 1989)(Drasgow & Parsons, 1983)(Bartholomew, Steele, Moustaki, & Galbraith, 2008)(Hambleton, Swaminathan, & Rogers, 1991) (Baker, Rounds, & Zevon, 2000)(Buchanan & Smith, 1999); (Reips, 2000)(Rizopoulos, 2006)(Roberts, Fang, Cui, & Wang, 2006)(Hornik, 2009)(Roberts J. S., 2001) (Lord & Novick, 1968)(Hambleton & Jones, 1993) *(Roberts, Donoghue, & Laughlin, 2000)* (Maydeu-Olivares, 2005), (Baker, Rounds, & Zevon, 2000) (Chernyshenko O. S., Stark, Chan, Drasgow, & Williams, 2001))