DEVELOPING IDEAL INTERMEDIATE ITEMS FOR THE IDEAL POINT MODEL

BY

MENGYANG CAO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Advisor:

Professor Fritz Drasgow

ABSTRACT

The importance of intermediate items has been overlooked since the introduction of dominance-based Likert scales for measuring attitudes, personality, and vocational interests in 1932. Intermediate items have been discarded in dominance scale constructions beacause they have low item-total correlations and factor loadings (Chernyshenko, Stark, Drasgow, & Roberts, 2007). The current study aims to recognize the importance of intermediate items by showing that they can be successfully calibrated by an ideal point model. College students (*N* = 355) were selected to answer a series of personality and vocational interest measures including some newly written intermediate items. Results showed that personality and vocational interest scales demonstrated satisfactory model fits to the ideal point model, but not to dominance models. Intermediate items also provided more information than extreme items for respondents with extreme latent traits. Among the four domains (Frequency, Average, Condition, and Transition, “FACT”) of intermediate items, the Average domain was found to exhibit the best performance. The possibility of using the results of this study to develop guidelines for writing intermediate items, as well as constructing computerized adaptive tests based on the ideal point model, is also explored in the paper.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Back in the late 1920s, Louis Thurstone published a series of seminal works explaining his method for measuring people's attitudes (Thurstone, 1927, 1928, 1929). The core assumption of his scaling method, later known as *Thurstone scaling*, was that an individual would endorse an item only when his/her attitude was located close to the statement described in that item (Thurstone, 1929). Based on these assumptions, Thurstone developed a scale measuring people's attitude towards war (Thurstone, 1931), in which he included statements ranging from strongly opposing war (e.g. "*There is no conceivable justification for war*") to favoring war (e.g. "*War is glorious*"). Each statement was assigned a scale value in terms of its severity, and the score of an individual was determined by the average value of the statements that the respondent endorsed. In Thurstone scaling, intermediate items, referred to as statements with neutral values on the measured trait, were included as a crucial part of the scale, because those items were indispensable in measuring individuals with moderate attitudes towards war.

Though carefully designed, Thurstone scaling was not popular among researchers, as it was difficult to construct and score, especially compared to the *Likert scale*, which appeared at almost the same time (Likert, 1932). The Likert scale, as perhaps the most prevalent scaling method for measuring individual differences, usually consists of a number of statements, with several response categories ranging from "strongly disagree" to "strongly agree". The scores of the respondents are computed by averaging the endorsed response categories across the items. The psychometric model underlying a Likert scale is the *dominance model*, which proposes that the probability of endorsing an item increases as an individual’s latent trait moves towards the positive extreme (Coombs, 1964). The dominance assumption serves as the foundation of almost all statistical analyses in traditional *classical test theory* (CTT), including average scores, item-total correlations, and factor analysis (Drasgow, Chernyshenko, & Stark, 2010). For example, items with low item-total correlations are classified as bad items, because they indicate a weak relationship between item responses and the latent trait. Thus, intermediate items, which act as an essential part of Thurstone scaling, are excluded in dominance-based Likert scales, as those items consistently show low item-total correlations (Chernyshenko, Stark, Drasgow, & Roberts, 2007).

Recently, the appropriateness of using dominance models to construct and score personality and vocational interest tests has been challenged. Researchers argued that even though the dominance assumption holds for cognitive ability tests, where individual’s capacity or maximal performance is measured, it is not applicable to measuring attitudes or preferences, where respondents tend to use introspection and a matching strategy to choose the option that best describes themselves (Drasgow et al., 2010). That argument led researchers to reconsider and recognize the importance of the Thurstone scaling. Compared to the dominance assumption of the Likert scale, the psychometric model underlying Thurstone scaling is the *ideal point model*, which assumes that the probability of endorsing an item is inversely related to the distance between the item’s location and respondent’s latent trait (Coombs, 1964). Given that respondents are usually instructed to “choose the option that best describes themselves” when they respond to non-cognitive measures, the ideal point model seems to be more appropriate in describing the response process. In 2006, Stark and his colleagues empirically validated that the ideal point model exhibited satisfactory fits to personality items from the Sixteen Personality Factor (16PF) Scale (Stark, Chernyshenko, Drasgow, & Williams, 2006). Furthermore, Tay et al. showed that the ideal point model provided better descriptions of responses to vocational interest measures than the dominance model (Tay, Drasgow, Rounds, & Williams, 2009). Following the above two remarkable studies, researchers have consistently found that the ideal point model demonstrated better model fits for tests measuring attitudes or typical behaviors, including leadership-member exchange (Scherbaum, Finlinson, Barden, & Tamanini, 2006), job satisfaction (Carter & Dalal, 2010), and trait emotional intelligence (Zampetakis, 2011).

However, arguments still exist among researchers on the necessity of using the ideal point model for non-cognitive ability measures. Most, if not all, of the existing non-cognitive ability Likert scales were developed based on a dominance scale construction approach, which excluded intermediate items with low item-total correlations, and leaving only items representing positive or negative latent trait values (Chernyshenko et al., 2007). Such extreme items make it less likely for respondents to disagree with the item from above the item’s location parameter, thus resulting in only trivial differences between the dominance and the ideal point models (Drasgow et al., 2010). Because of the dominance nature of the development of the scales used in the above studies, the dominance model also exhibited generally acceptable fits to the data, even though it did not provide theoretically correct descriptions of the underlying response process (e.g. Stark et al., 2006; Carter & Dalal, 2010). Thus, some researchers argued that it is not necessary to switch to the ideal point model, as the dominance model is already capable of describing existing Likert scales (e.g., Waples, Weyhrauch, Connell, & Culbertson, 2010; Reise, 2010).

The above controversy drives us to explore the properties of intermediate items, because the inclusion of such items in scales is necessary to demonstrate the distinction between the dominance model and the ideal point model. As stated in the ideal point model, when responding to intermediate items (e.g. "*I am about average in regard to details*"), individuals who are high on the latent trait tend to disagree, as they believe that their traits are higher than that described in the items. This is inconsistent with the dominance model, which stated that individuals with high latent traits would endorse intermediate items. Thus, intermediate items are crucial in differentiating dominance and ideal point models, as they can be fitted by the ideal point model, but not the dominance model. Therefore, to show that the ideal point model is advantageous in describing the responses of non-cognitive ability measures, we first need to develop items tapping intermediate trait levels.

So far, little is known about intermediate items, because they have been excluded from scales due to low item-total correlations. As intermediate items may be labeled as "bad double-barreled" items according to the dominance perspective, no study has been conducted to discover their psychometric properties, not to mention how to write them. Although intermediate items are theoretically important for the ideal point model, it still remains unknown whether their inclusion will lead to superior model fits for the ideal point model compared to the dominance model. Moreover, researchers generally believe that intermediate items are hard to write, (Brown & Maydeu-Olivares, 2010), but nobody has ever attempted to explore systematically on how to construct such items.

The current study serves as the first study to use *item response theory* (IRT) to investigate the psychometric properties of intermediate items, and to explore the best way of constructing intermediate items for personality and vocational interest measures. In the following introduction sections, we will first explain the difference between the dominance model and the ideal point model, followed by why we believe intermediate items are important in terms of improving the psychometric properties of personality and vocational interest tests. Finally, we will focus on the possible ways of constructing intermediate items that are ideal for the ideal point model.

Dominance Model or Ideal Point Model?

As discussed in the previous section, the disparity between the dominance model and the ideal point model can be traced to early last century when Likert and Thurstone developed their own scaling methods based on different assumptions. The dominance-based Likert scale demonstrated several advantages over the ideal-point-based Thurstone scale, such as higher item-total correlation, higher internal consistency reliability, and relative easiness in scale construction (Davison, 1977). Thus, even though dominance assumptions fail to describe the response process underlying non-cognitive measures, researchers are still inclined to use the Likert-scale format to measure personality and vocational interests, and adopt the dominance-based approaches, including average scores and item-total correlations to analyze responses. These traditional CTT methods are unable to detect whether the model used to analyze the data actually fits the data.

With the growing interest in using non-cognitive measures, especially personality inventories, as personnel selection tools, and the emerging needs for developing computerized adaptive tests, researchers started to apply item response theory (IRT) for analysis. The *two-parameter logistic* (2PL) model has been the most commonly used IRT model to calibrate dichotomous responses (e.g., Reise & Walker, 1990), whereas *Samejima's graded response model* (SGRM; Samejima, 1969) is usually adopted for personality measures with polytomous responses (e.g., Zickar & Robie, 1999). Both the 2PL model and the SGRM assume a dominance response process. As shown in Figure 1, the *item response function* (IRF) of an item based on the dominance assumption is an S-shaped curve, indicating a monotonic relationship between the latent trait and the probability of endorsing that item.

However, as fitting dominance-based IRT models to personality and vocational interest measures became more prevalent, researchers started to realize that some items failed to exhibit the monotonic IRF as theoretically expected (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). Stark et al. (2006) found that although constructed based on a dominance approach, some facets of personality scales were not fit well by the 2PL model, but instead showed adequate fits for the *generalized graded unfolding model* (GGUM; Roberts, Donoghue, & Laughlin, 2000), which is an ideal point model. They stated that, for personality inventories, people were usually asked to choose the options which "best describe" themselves. In that case, people could possibly disagree with an item from both above and below the item location (i.e., the "ideal point"). This response strategy would produce the non-monotonic, bell-shaped response function of the ideal point model, as shown in Figure 2.

Similar results have been found by Tay et al. (2009) in a study of vocational interest measures, in which they discovered that the ideal point model better fit vocational interest scales than the dominance model. More importantly, in that paper they introduced Cronbach’s (1949) classification of maximal behaviors and typical behaviors to describe the correspondence between the item content and the item response process. They summarized that the dominance model was more suitable for describing maximal behaviors, as they were “concerned about the limits of an individual’s capacity”. On the contrary, the ideal point model was more suitable for describing typical behaviors, as there was often “no notion of capacity limits”. That assertion provided a theoretical explanation for the ideal point model as the most appropriate model for describing the response process underlying non-cognitive measures, such as personality and vocational interest inventories.

Why Are Intermediate Items Important?

Although many researchers now agree that responding to personality items involves a process of introspection rather than challenging capacity limits, some are still reluctant to admit that responses to personality and vocational interest measures follow an ideal point process, as they argue that dominance-based IRT models (e.g., 2PL, SGRM) are also able to adequately fit current personality and vocational interest scales (e.g., Waples et al., 2010; Reise, 2010). In fact, this situation was caused by the dominance-based scale construction procedures, in which items with low item-total correlations were always identified as “poor” items and were excluded from the scale. As a result, only extreme items have been retained in the final versions (Chernyshenko et al., 2007). Since negatively worded items are usually reverse-scored when fitted to dominance-based IRT models, all the extreme items will exhibit high item location parameters, whereas very few respondents have latent traits that are higher than those item locations. In that case, the IRF based on the ideal point model will resemble a dominance IRF, as shown in Figure 1, leading to seemingly adequate fits for dominance-based IRT models.

Items deleted due to low item-total correlations are often intermediate items that describe behaviors with neutral extremity on the measured latent trait (Drasgow et al., 2010). Unlike extreme items, intermediate items have item location parameters around the middle range of the latent trait distribution. Hence, there will be considerable number of respondents with latent traits both below and above the item locations. Consider a cognitive ability test where respondents treat the items as hurdles that they attempt to overcome, individuals with high cognitive abilities will answer low and intermediate items correctly, as their abilities are high enough to overcome the problem posed in the question. That will lead to a monotonically increasing IRF as specified in the dominance model. However, in personality and vocational interest assessments, where respondents adopt an introspection process to answer questions (Tay et al., 2009), individuals with high traits tend to reject intermediate items, as the contents described in those items are far away from their latent traits. That will lead to a bell-shaped curve with unfolding, thus inconsistent with the monotonic curve speculated by the dominance model. In that case, the dominance model will fail to fit. Therefore, intermediate items are extremely important in that they can successfully distinguish whether the dominance model or the ideal point model should be used for a particular scale. In particular, we proposed the following two hypotheses.

Hypothesis 1: The ideal point model exhibits satisfactory model fit to responses to intermediate personality and vocational interest items.

Hypothesis 2: The dominance model does not fit responses to intermediate personality and vocational interest items.

Besides a considerable contribution to distinguishing dominance and ideal point models, intermediate items are also believed to improve measurement precision (Drasgow et al., 2010). In IRT models, measurement precision is usually described by item and test information functions. Test information is inversely related to the standard error of trait estimates (Lord, 1980). Thus, the higher the information, the more precise the measurement at a specific latent trait level. According to the ideal point model, scales with extreme items typically provide high information for individuals standing in the middle range of the trait distribution. In other words, extreme personality items are not as precise in measuring people at the extremes as in measuring people in the middle. In personnel selection practice, however, the selection rates usually vary a lot across different organizations. Thus, a selection tool needs to show approximately equal measurement precisions across the whole trait continuum to guarantee that no matter what cutoff points are used, measurement is accurate. Yet, this goal cannot be achieved by including only extreme items in the tests.

By adopting an ideal-point-based scale construction approach that does not discard intermediate items, Chernyshenko et al. (2007) obtained a set of items that were different from those generated from dominance assumptions. Moreover, they found that personality scales constructed based on the ideal point model provided more information than dominance-based scales. However, they did not directly test whether the increased measurement precision was due to the inclusion of intermediate items. Thus, we intended to validate that assumption in this study.

> *Hypothesis 3*: Personality tests with intermediate items provide more accurate estimates for individuals at the extremes of trait continuum than personality tests without intermediate items.

How to Write Intermediate Items?

Despite the crucial role that intermediate items play in the ideal point model, little is known about what types of such items perform best. Thus, how to write good intermediate items turns out to be a completely new and challenging topic. Traditional item development guidelines such as avoiding “double-barreled” items are based on the dominance assumptions (Hinkin, 1998), which do not seem appropriate for generating ideal-point-based intermediate items. By looking at intermediate personality items developed by Chernyshenko et al. (2007) for measuring “order”, as well as existing intermediate personality items from other sources, we argue that intermediate personality items can be constructed in the following four different ways, each representing a unique domain of item wording.

a. *Specifying frequencies (“Frequency”)*. Words or phrases indicating moderately low frequencies are included in the item, so that respondents who always perform the designated behaviors tend to reject the item. For example, "*I seldom make detailed 'to do' lists*".

b. *Comparing to average ("Average")*. Respondents are asked to compare themselves to people on average to indicate a matching with moderate item locations. For example, "*My room neatness is about average*".

c. *Specifying conditions ("Condition")*. A particular condition is specified in the item so that the respondent's endorsement becomes conditional on certain occasions. For example, "*When busy, I spend little time cleaning and organizing things*".

d. *Using transitions ("Transition")*. Items are structured in a transition format such that extreme descriptions are avoided. For example, "*Although I have a daily organizer, I have a hard time keeping it up to date*".

There has been no research studying the above four domains (Frequency, Average, Condition, Transition, or simply "FACT"), and consequently we do not know which ones will perform well as intermediate items. Even though items created based on the FACT domains seem to capture neutral intensities at the trait continuum, some problems may arise when those items are used in practice (Brown & Maydeu-Olivares, 2010). For example, items in the Condition and Transition domains may lead to a "double-barreled" presentation, which may be confusing for people to answer; for items in the Average domain, the reference group to which the respondents are asked to compare is ambiguous, which may lead to different standards across respondents. There is also evidence suggesting that items from some of the FACT domains may have influence on individuals' responses. For example, Nye, Newman, and Joseph (2010) manipulated the Frequency domain by artificially including the words "always" and "typically" in the items, and showed that different wordings did affect people's response to emotional intelligence tests. Thus, it is a research question as to which of the FACT domains will generate items that demonstrate unfolding and perform as intermediate items.

> *Research Question 1*: Among four different domains (i.e., comparing to average, specifying conditions, specifying frequencies, and using transitions) to create intermediate items, which method(s) perform well in practice?

Compared to personality scales in which most items describe a type of attitude or behavior, vocational interest measures usually contain items describing some types of activities within a specific interest domain, and the respondents are asked to choose one of the options ranging from "strongly dislike" to "strongly like" that best describes their preference (e.g. Deng, Armstrong, & Rounds, 2007). Therefore, in order to use the same domains as in personality measures to construct intermediate items for vocational interest scales, we first need to revise the vocational interest items so that they can fit to the same response format ("strongly disagree" to "strongly agree"). This is another research question: do the same intermediate domains also perform well for vocational interest measures?

Research Question 2: Will the intermediate domains (i.e. FACT) for the personality measures work well for vocational interest measures in which respondents' preferences are measured?

CHAPTER 2

METHOD

Participants

Participants were undergraduate students recruited from the subject pool of the Psychology department at a large mid-western university in the United States. Since our study required a precise understanding of item wordings, only students who identified themselves as native English speakers were eligible for this study. After signing up for the study, participants received a link that directed them to the personality and vocational interest scales residing on a professional online survey website. Overall, 375 students successfully finished our survey and were granted course credit.

To rule out the effects of careless responses, we randomly embedded 5 quality control items, which asked the participants to choose a certain response option (e.g., *"Please click 'Agree' for this item."*). Participants who incorrectly answered more than one quality control items were identified as careless respondents and therefore excluded from the sample. Of the 375 students, 355 passed the quality control checks, yielding a valid response rate of 94.7%.

Among those remaining participants, 71% were female. The age of participants ranged from 18 to 28, with a mean of 19.35. Most participants identified themselves as White (70.1%). Other self-reported race/ethnicity included Asian (15.2%), Hispanic or Latino (5.9%), Black or African American (5.6%), and others (3.1%). 90.4% of the respondents reported that they had concurrent or previous work experience.

Measures

Personality items were adapted from the Comprehensive Personality Scale (CPS) which is being developed through years of work. The CPS was constructed based on the Big-Five Model, with 22 underlying facets. Items for each facet were developed via an ideal point scale construction approach (Chernyshenko et al., 2007). In the current study, 3 facets, including the “Order” facet of Conscientiousness, the “Dominance” facet of Extraversion, and the “Curiosity” facet of Openness, were selected, as they have shown satisfactory model fits when analyzed with an ideal point model (i.e., the GGUM2004 software developed by Roberts et al. (2004)).

In this study, 20 items were included for each facet, including 4 positive items and 4 negative items that were directly selected from the original CPS. These items were chosen based on high item-total correlations and high factor loadings in a previous validation sample. Intermediate items in the original scales were also selected. In addition, additional intermediate items were created for the four intermediate domains (Frequency, Average, Condition, and Transition; “FACT”) mentioned in the previous section, so that there were 3 items from each intermediate domain for each personality facet. The contents, locations (positive, negative, or intermediate), and domains (for intermediate items only) of all personality items are listed in Table 1 through Table 3.

Vocational interest scales were adapted from the Interest Item Pool (IIP) developed by Liao, Armstrong, Rounds, and Su (2007). In particular, 4 vocational interest facets of RIASEC, including Realistic, Investigative, Artistic, and Social, were selected for the current study. Each of the Realistic and Social scales contained 13 items, including 5 positive items, 5 intermediate items, and 3 negative items. The Artistic and Investigative scales were constructed based on 5 item triads. Each item triad referred to one specific activity described in the original measures, with modifiers included to indicate different extents of preference for that activity. For example, based on the item "draw pictures" in the original Artistic scale, a positive item was phrased as "*I love drawing pictures*", an intermediate item was phrased as "*I like to draw pictures only when I have time*", and a negative item was phrased as "*I have no interest in drawing*". Overall, there were 15 items in each of the Artistic and Investigative scales, with equal numbers of positive, intermediate, and negative items. All intermediate items were written in terms of the FACT intermediate domains described in the previous section. The contents, locations (positive, negative, or intermediate), and domains (for intermediate items only) of all vocational interest items are presented in Table 4 through Table 7.

In the final online survey, all personality items were presented in a random order, and then the vocational interest items were presented in a random order. Participants were asked to rate each item on a 4-point Likert-scale, where 1 = "*Strongly Disagree*", 2 = "*Disagree*", 3 = "*Agree*", and 4 = "*Strongly Agree*".

Analysis

CTT statistics, including item means, standard deviation, internal consistency (Cronbach's α), and item-total correlations, were computed with SPSS 20.0. Negative items were reverse-coded before any statistics were computed. Because responses to intermediate items were not expected to be monotonically related to true scores or total scores computed from the positive and negative items, Cronbach's α indices were computed with all intermediate items excluded from the scales. Similarly, the total scores in item-total correlations were computed based on only positive and negative items. In particular, for positive and negative items, the item-total correlations were corrected by excluding the target item in computing the total score in order to avoid spurious linear trends (Allen & Yen, 1979). For intermediate items, the item-total correlations were computed by correlating the responses to the target items with the total scores computed based on the positive and negative items.

Following the practice of previous studies (e.g., Stark et al., 2006; Chernyshenko et al., 2007; Tay et al., 2009), we dichotomized the responses to facilitate the IRT analyses, such that "*Strongly Disagree*" and "*Disagree*" were coded as "0", and "*Agree*" and "*Strongly Agree*" were coded as "1". The Generalized Graded Unfolding Model (GGUM; Roberts et al., 2000) was used for the ideal point model. The GGUM for dichotomous data can be expressed in the following equation:

$$\mathrm{P}\,[U_i = 1 \mid \theta_j\,] = \frac{\exp\{\alpha_i[(\theta_j - \delta_i) - \tau_i]\} + \exp\{\alpha_i[2(\theta_j - \delta_i) - \tau_i]\}}{1 + \exp\{\alpha_i[3(\theta_j - \delta_i)]\} + \exp\{\alpha_i[(\theta_j - \delta_i) - \tau_i]\} + \exp\{\alpha_i[2(\theta_j - \delta_i) - \tau_i]\}},$$

where θ_j denotes the latent trait of respondent *j*, and α_i, δ_i, τ_i, respectively refer to the discrimination parameter, the location parameter, and the subjective response category parameter of item *i*. The GGUM parameters were estimated by the GGUM2004 program (Roberts, Fang, Cui, & Wang, 2006), which uses the marginal maximum likelihood (MML) method in estimating item parameters. Note that the estimation of GGUM parameters does not require the negative items to be reverse-coded, as the location parameters in GGUM can provide a direct reference to the location of the item content.

The responses were also calibrated by the Two-Parameter Logistic (2PL) model, which has been widely used as the dominance-based model for personality and vocational interest assessments (e.g., Chernyshenko et al., 2001; Stark et al., 2006). The 2PL is stated as:

$$\mathrm{P}\,[U_i = 1 \mid \theta_j\,] = \frac{1}{1+\exp[-Da_i(\theta_j - b_i)]}\,,$$

where θ_j denotes the latent trait of respondent j, a_i, and b_i refer to the discrimination and difficulty parameters, and D is a scaling constant set equal to 1.702 for historical reasons. The 2PL item parameters were estimated by the BILOG program (Mislevy & Bock, 1991) with default settings.

To assess model fit, we calculated chi-square fit indices using Stark's (2001) MODFIT 3.0 computer program. The item singles chi-square statistics examined the difference between the observed responses and the expected responses of a single item. Item pairs and triplets were also considered, as they are sensitive to violations of local independence and multidimensionality (Chernyshenko et al., 2007). All values were adjusted to a sample size of 3,000 for comparisons across different sample sizes, with small chi-square values indicating good model fit. It is suggested that the mean adjusted χ^2/df ratios across all items should be less than 3.0 in order to conclude that model fit is satisfactory (Drasgow, Levine, Tsien, Williams, & Mead, 1995). A previous simulation has shown that using adjusted χ^2/df ratios to compare relative fits can differentiate ideal-point-based responses from dominance-based responses (Tay, Ali, Drasgow, & Williams, 2011).

CHAPTER 3

RESULTS

Traditional CTT analysis

As we mentioned in the Method section, due to the non-monotonic relations between responses to intermediate items and true scores, traditional CTT methods for assessing internal consistency (e.g., Cronbach's α) and item discrimination (e.g. corrected item-total correlation, ITC_c) cannot be directly applied to scales with intermediate items. Thus these CTT statistics were computed by only including positive and reversed-coded negative items, or by removing intermediate items when computing total scores.

For the personality scales, the Cronbach's α of Order, Dominance, and Curiosity scales were .81, .90, and .76, respectively, suggesting that the scales containing only positive and negative items showed satisfactory reliabilities. Similar results were found for the vocational interest scales, with Cronbach's α of .87, .93, .85, and .77 for the Artistic, Investigative, Realistic, and Social scales.

The corrected item-total correlations (ITC_c) are displayed for each item in Table 1 through Table 7. For both personality and vocational interest scales, almost all of the positive and negative items show high ITC_c. In traditional CTT analysis, these items would be considered well-performing. Theoretically, intermediate items are expected to show low positive or negative item-total correlations. In our results, however, the intermediate items did not consistently exhibit the hypothesized ITC_c pattern. By adopting $|\text{ITC}_c| < .3$ as a rule-of-thumb cutoff value for assessing low ITC_c, we found that only a portion of items were classified as low, and the results varied across different intermediate domains. In particular, the Average personality domain had the most items classified as low ITC$_c$ (8 out of 9 items), followed by Transition (6 out of 9) and Condition (4 out of 9). The Frequency domain had the least items with low ITC$_c$ (2 out of 9). For vocational interest scales, however, almost all intermediate items showed high ITC$_c$, except for three items in the Social scale.

A possible explanation of the above results is that the intermediate items generated from different domains vary in their properties. Another explanation is that the ITC$_c$ is not an appropriate statistic for identifying intermediate items, as it is sensitive to the empirical trait distribution of the sample. For example, the intermediate items will show low item-total correlations only when the sample is large and the latent trait distribution is broad and symmetric. However, when the trait distribution is positively skewed, the ITC$_c$ will be strongly influenced by the large number of people who have low trait values, and we will find that they tend to reject the intermediate items. In that case, the ITC$_c$ may appear large even though the items are truly intermediate. Because our sample was selected from a university subject pool, the distribution of vocational interests can be restricted and therefore skewed in some facets. Moreover, a low ITC$_c$ does not necessarily mean that the item is intermediate, because that item can simply be a bad item that does not measure the target trait. Therefore, we essentially need to look at the results of the IRT analysis, which is sample invariant, and provides more solid evidence on whether the items are intermediate or not.

Unidimentionality

Both the GGUM and the 2PL model require the items to be unidimensional. Thus, we examined the unidimensionality assumption of all personality and vocational interest scales before proceeding to the IRT analysis. Exploratory factor analysis (EFA) with principal axis factoring was conducted on each scale to examine whether there existed a dominant factor. As suggested by previous research, the first factor should account for at least 20% of the total variance to obtain stable item parameter estimations (Reckase, 1979). All scales satisfied the EFA criterion, despite the problems associated with factoring ideal point items (Davison, 1977), indicating that it was appropriate to conduct unidimentional IRT analysis on those scales.

IRT Analysis of Personality Scales

Item parameters

GGUM and 2PL parameter estimates of the personality items are displayed in Table 1 through Table 3. The 2PL parameters of Item 17 from the Dominance scale could not be estimated by BILOG as the "initial slope was less than -0.15". Examination of the item parameters revealed some important psychometric properties of the intermediate items. Specifically, we first focused on the discrimination parameters, which are the α-parameters in GGUM and the a-parameters in 2PL. For the 2PL model, the average estimated a-parameters of intermediate items were 0.33, 0.48, and 0.28 for the Order, Dominance, and Curiosity scales, respectively. Such a-parameters are typically considered as low, and they are much lower than the discrimination parameters of extreme items (averages of 0.91, 1.73, and 0.87, respectively for the Order, Dominance, and Curiosity scales). Low 2PL a-parameters suggest that the intermediate items are poor in discriminating trait values. However, when analyzed with the ideal point GGUM model, the α-parameters of intermediate items were excellent, with averages of 1.47, 1.23, and 0.91 respectively for the Order, Dominance, and Curiosity scales. Thus, with the ideal point model, the intermediate items would not be labeled as poor.

Next, we looked at the location parameters (i.e. the δ-parameters in GGUM), which denote the trait level where the probability of endorsing that item is maximized. As suggested by Roberts and Shim (2008), an item with a δ-parameter lying between the 10th to the 90th percentiles of the estimated θ distribution can be considered as exhibiting unfolding. Otherwise, the item essentially acts as an extreme item and exhibits dominance-like properties. The computed unfolding θ range was (-1.40, 1.32) for Order, (-1.42, 1.16) for Dominance, and (-1.17, 1.04) for Curiosity. As shown in Table 1 through Table 3, almost all positive and negative items had δ-parameters lying outside the unfolding ranges, except for Item 2 in Dominance, as well as Item 1 and 3 in Curiosity. For the intermediate items, although the δ-parameters show large variations across items and scales, most of them reside within the unfolding range. The Order scale had the most intermediate items identified as exhibiting unfolding (11 out of 12), followed by the Dominance scale (8 out of 12). The Curiosity scale, however, only had 5 out 12 intermediate items with δ-parameters lying within the unfolding range.

Model fits

The model fits of GGUM and 2PL were examined by calculating the adjusted χ^2/df ratios. As chi-squares of single items are insensitive to various misfits when assessed in the same sample used for parameter estimations, we focus on the chi-squares of item pairs (i.e., doublets) and item triads (i.e., triplets), because they can better detect violations of local independence (Drasgow et al., 1995). As presented in Table 8, the average adjusted χ^2/df ratios of item doublets and item triplets are all below 3.0 for GGUM, indicating that the GGUM exhibited satisfactory fit for all three personality dimensions. On the other hand, the 2PL model generated considerably larger chi-square values for item doublets and triplets, suggesting that the 2PL did not fit the data well. The results supported our Hypothesis 1 and 2 in that when intermediate items are included in the scale, only an ideal point model is able to fit the data.

The fit plots illustrate the difference in model fits between extreme and intermediate items. For example, Figures 3a and 3b present the GGUM and 2PL fit plots of the positive item "*I plan my time very carefully*" from the Order scale. We can see that in both models, the empirical response curve lies almost exactly on top of the expected item response curve, suggesting that both models fit the item well. Specifically, the GGUM empirical response curve is similar to a dominance-based response curve, which is monotonically increasing. The results show that the ideal point model is flexible in fitting extreme items, as it can generate extreme location parameters for those items such that there will be a monotonic relation between the expected response and the latent trait for trait values with nontrivial frequencies.

The fit plots for intermediate items show very different patterns for the GGUM and the 2PL model. For example, the fit plots of Item 10 "*I try to keep my room clean and tidy, but I don't always have time to do so*" from the Order scale are shown in Figures 4a and 4b. The GGUM fit plot shows bell-shaped curve with unfolding around the ideal point, and the empirical proportions are similar to the estimated item response function. However, the IRF generated by the 2PL model is almost a straight line, whereas the empirical response curve is somewhat twisted and does not fit the IRF. We found that many intermediate items demonstrated this pattern, indicating that even though the dominance model can fit the extreme items well, it is not able to provide satisfactory descriptions of intermediate items. The adjusted χ^2/df ratios also supported this notion, as the large chi-square values were always those of item doublets and triplets that contained intermediate items.

Intermediate items by domains

A research question of substantial importance concerns the performance of items from the four intermediate domains (FACT). To address this question, we first created four sets of items for each scale, such that each set contained all 8 extreme items but only one domain of intermediate items. We then separately estimated the GGUM parameters for items in each set and computed model fit.

Results are presented in Table 9. As shown in the table, all item sets demonstrate satisfactory model fit. Interestingly, the results indicated that not all items constructed in the FACT domains behaved like intermediate items. Based on the values of δ-parameters, we found that the Average domain performed the best among FACT, with 8 out of 9 items lying within the unfolding range. Researchers have speculated that respondents would be confused by comparing-to-average items, as they might not be able determine to whom they were comparing themselves (Brown & Maydeu-Olivares, 2010). In our study, however, we found that this was not a problem. Furthermore, we found that the Frequency domain also enabled us to successfully construct ideal intermediate items, as it had 7 items showing unfolding. Thus, it appears that the respondents pay attention to the frequency modifiers when they answered items, and perceived those items as representing a moderate extremity. The other two domains did not work as well as Average and Frequency, with 5 unfolding items for Transition and 4 for Condition showing unfolding. Items in those two domains would usually be referred to as "double-barreled" items. Our results showed that although those "double-barreled" items may somewhat tap into the moderate extremities, the respondents may have difficulties understanding the items and making their choice. This also coincides with the informal feedback provided by some of our respondents, who said that they found "some items have two parts which are confusing to answer", as they may "agree with one part but disagree with the other part".

Measurement accuracy

As discussed in the Introduction section, an important property of intermediate ideal point items is that they can provide more information for respondents who have relatively extreme latent traits. Figure 5 shows the item information function (IIF) of Item 3, which is a positive item in the Order scale. We can see that the IIF curves generated by GGUM and 2PL are close to each other, except the GGUM provides more information around θ = 3.0. Because θ = 3.0 is outside the θ range where most respondents are located, it appears that the IIFs of extreme items are similar for ideal point and dominance models. However, Figure 6, which displays the IIFs of Item 10 in the Order scale, shows quite distinctive item information patterns. This typical intermediate item has a 2PL IIF that is almost a straight with low values, whereas the GGUM IIF is a bimodal curve which peaks at around θ = -2.0 and 1.5, with much higher information values than the 2PL curve. Thus, the misspecification of the 2PL leads to the incorrect conclusion that such items provide trivial information, but fitting an appropriate model leads to the reverse conclusion.

Figures 7a through 7c display the *test information functions* (TIF), which are computed as the summations of IIFs across all items within a scale. In general, the GGUM test information functions provide more information than the 2PL ones when the trait levels approach the positive and negative ends of the distribution. To compare the different patterns in information between extreme and intermediate items in general, we computed cumulative GGUM information separately for extreme items and for intermediate items within a scale. Although the absolute values of cumulated information are not comparable between extreme and intermediate items (because the total number of items is not equal), the difference in where they show maximized information can be observed. For example, Figure 8 shows the cumulative information functions for the Dominance scale. The plot clearly indicates that the extreme items provide more information for the middle-ranged traits, whereas the intermediate items provide more information at the positive and negative ends. Thus Hypothesis 3 was fully supported.

IRT Analysis of Vocational Interest Scales

GGUM and 2PL parameters of the vocational interest items are presented in Table 4 through Table 7, except for the 2PL parameters of Item 10 in the Social scale, as the parameters cannot be estimated by BILOG because the “initial slope was less than -0.15”. Examinations of discrimination parameters showed parallel results to what we obtained from the personality scales: The GGUM on average produced higher discrimination parameters than the 2PL did, and the intermediate items were shown to be more discriminating for the GGUM than the 2PL model.

Using the same criteria (central 80 percent) as we adopted for the personality scales to label items as truly intermediate, we found that 3 out 5 items showed unfolding for the Artistic and Investigative scales. Note that those two scales were constructed in an *item triad* format, for the items to be considered as intermediate, their location parameters also need to be between the location parameters of positive and negative items. Thus, Item 5 in the Artistic scale was ruled out as it had a higher location parameter than the corresponding positive item. For the remaining 5 unfolding items, the GGUM successfully recovered the structures of the item triads by producing δ-parameters with values arranged in the same order as the item content. The above results suggested that it is possible to use the FACT domains to construct intermediate items for the vocational interest. For the Social and Realistic scales, however, the "central 80 percent" method failed to contain items designed to be intermediate. We found all 3 negative items from the Realistic scale to lie within the unfolding range, and 4 out of 5 positive items in the Social scale were within that range. As our respondents were recruited from the Psychology Subject Pool with many psychology majors having strong interests in Social but not in Realistic activities, the trait distributions of the two scales can be rather skewed. Thus, the 10th and 90th percentiles may not work as reasonable cutoff values for classifying items as intermediate, as extreme items may be mislabeled as intermediate. When the trait distribution is skewed, we may need to look at the model fit and the empirical item response curve of individual items to accurately identify intermediate items.

The scale level model fit indices are presented in Table 8. In terms of adjusted χ^2/df ratios for item doublets and triplets, we found that all but the Artistic scale fitted the GGUM well. A closer examination of the Artistic scale model fit showed that large chi-squares were obtained for item doubles and triples when items in the same triad were included. The Investigative scale, which was also constructed with item triads, exhibited the same problem, though not as severe as with the Artistic scale. The results reflect violations of the local independence assumption for items in the same triad. Given that those items describe the same activity, it is likely that individuals' responses to one item were too highly correlated with their responses to another item in that triad. Interestingly, the 2PL model fit the Investigative and the Realistic scales well, even though those scales contained intermediate items. By looking at the fit plots of intermediate items, we found 10 intermediate items empirically showing unfolding, including Item 2, 11, 14 in Artistic and Investigate, Item 8 in Realistic, and Item 8, 9 , and 10 in Social. Out of these 10 items, 4 items were in the Condition domain, 4 were in Average, and the other two were in Frequency and Transition. The results were to some extent different from what was found in the personality scales, where the Average and Frequency domains worked the best among all FACT domains. It seems that respondents still perceive "comparing to average" as describing intermediate preferences to certain activities reflecting vocational interest, but they had more difficulty comprehending what was meant by "*Sometimes I like to a play a musical instrument*". However, with "*only when I have time*", respondents apparently perceived a moderate preference.

CHAPTER 4
DISCUSSION

Ever since the dominance-based Likert scale became the most prevalent scaling method for non-cognitive assessments, researchers have avoided intermediate items, as they were believed to be “double-barreled”, and therefore ambiguous. Recent research on response processes underlying non-cognitive measures have shown importance of intermediate items. Intermediate items, which served as crucial components in Thurstone scaling 80 years ago (Thurstone, 1931), can now be analyzed with ideal point IRT models, and make important contributions to measurement.

This study provided strong evidence to support the use of the ideal point model for personality and vocational interest measures by showing that intermediate items can only be appropriately described by an ideal point model, but not by the dominance model. Firstly, we found that the ITC_C, which is an index for item discrimination based on dominance models, failed to appropriately characterize the usefulness of many items. Similarly, 2PL model, as a dominance-based IRT model, generally produced low discrimination parameters for intermediate items. However, in the ideal-point-based GGUM, these intermediate items were found to have good to excellent discrimination parameters and to often have location parameters that reside in the middle range of the trait continuum. Model fit indices showed that the ideal point model fit scales with intermediate items, but these scales were not adequately fit the dominance model. By examining the fit plots of individual items, we found that the IRFs of intermediate items based on the ideal point model were usually bell-shaped curves with unfolding above the ideal points. The IRFs generated by the dominance model, however, were usually flat curves and did not fit the empirical response curves very well. In sum, the study has shown that personality and vocational interest scales with intermediate items should be analyzed with an ideal point model. If the intermediate items are well constructed, they will show satisfactory discriminating power and exhibit unfolding properties.

Furthermore, we empirically validated that intermediate items provided more information than the extreme items did for respondents with relatively extreme traits. In general, we found that the GGUM test information functions were higher than the 2PL ones at extreme trait levels. The findings replicated what Chernyshenko et al. (2007) found in their personality scales constructed based on the ideal point approach. We also separately examined the item information and cumulative information of intermediate and extreme items. Extreme items generally exhibited a unimodal information curve with maximum around the middle of the trait continuum, whereas the information functions of intermediate items were usually bimodal. These results suggest that intermediate items are valuable because they enhance trait estimation at low and high values, whereas the extreme items did not provide satisfactory measurement.

In practice, our study provides a guidance on how to construct well-performed intermediate items. The majority of the intermediate items constructed from our FACT domains exhibited unfolding and fit the ideal point model well, indicating that by simply adopting one strategy from the FACT domains, we were able to convert traditional non-cognitive items to intermediate items. We also found diversity in performance among the FACT domains. The Average domain consistently worked well for both personality and vocational interest tests. Adding modifiers denoting the frequencies also influenced people’s perceptions of personality items and made them show unfolding, as has been found in emotional intelligence items (Nye et al., 2010). However, this approach did not work well for vocational interest tests. Interestingly, the Condition domain performed well for vocational interest items, but did not show as much unfolding as items in Average and Frequency domains. Not surprisingly, the Transition domain items exhibited the poorest performance among all FACT domains, suggesting that the "double-barreled" item structure could sometimes cause confusion. We also noticed that item wording and content were also important for intermediate items to perform well. For example, we found that the specific conditions we used in the intermediate items would affect the performance of the item. When the conditions were common and were expressed in a mild way (e.g. "*I do not mind trying new things when there are not many choices*"), items tended to show unfolding. However, when the conditions were restricted and emphasized as uncommon (e.g. "*I am open to new concepts only if they are not hard to understand*"), individuals were inclined to perceive that item as more extreme.

Another practical implication is that by acquiring a better understanding of intermediate items, we can facilitate the development of *computerized adaptive tests* (CATs) based on the ideal point model. With the development of IRT models, there is a growing trend of using CATs to measure individual differences, especially personality (see Drasgow & Olson-Buchanan, 1999). An integral part of applying CATs to personality tests is to generate accurate estimates of item information functions, so that items with maximal information can be selected (Reise & Henson, 2000). If the scale consists of only extreme items, then few items will show large information at the extremes of the trait continuum, making it difficult to select appropriate items for some respondents. The inclusion of intermediate items provides a solution to this problem, as they can aid in the estimation of the traits of respondents at the extremes.

Admittedly, the design and results of our study are limited in several ways. Firstly, even though we found that intermediate items worked better for the ideal point model in personality tests, we did not obtain consistent results in vocational interest tests. Surprisingly, two of the vocational interest tests showed satisfactory model fits for the dominance model, though there were intermediate items exhibiting unfolding in those scales. Moreover, although we used two different approaches (normal vs. item triads) in constructing vocational interest scales, we did not find consistent evidence supporting either approach. Considering that vocational interest items ask about preferences and thus are different from personality items, in the future we should consider different strategies for developing intermediate vocational interest items. Secondly, our sample was collected in a Psychology Subject Pool, which limited the generalizability of this study. Moreover, the sample also restricted our examination of Realistic and Social vocational interest scales because of the restricted trait distributions. According to Tay et al. (2009), a sample from a restricted range of interest will lead to relatively monotonic curves. In this study, we found that for Social and Realistic scales, the “central 80 percentile” method we used to determine intermediate items did not function well. Thirdly, because the respondents were college students, we were not able to collect information on organizational job criteria to test the validity of our scales. Note, however, that Chernyshenko et al.’s (2007) found that personality scales constructed on an ideal-point-based approach showed satisfactory predictive validities for predicting student and health behaviors, we believe that the inclusion of intermediate items will not compromise the validity of the scales.

Given that we have validated the importance of intermediate items, in the future researchers should focus on investigating what factors may affect the performance of intermediate items. For example, the number of response categories may affect how respondents answer intermediate items. In the current study, we used the 4-point Likert scale as the response format, and then dichotomized the responses for IRT analysis. As Likert scales with polytomous response options are so prevalent for personality measures, most respondents are quite familiar with specifying their trait levels based on the corresponding scale response options, thus are likely to ignore the intermediate modifiers within the items. If this is true, the performance of intermediate items might be improved if we simply use a “yes” or “no” response format. Another factor that may affect the performance of intermediate items is the testing environments. O’Brien and LaHuis (2011) found that respondents may not utilize an ideal point process when answering personality items in a high-stake environment. However, the scales they used to test their hypotheses were constructed based on the dominance approach. It is intriguing to see if the conclusions of that paper would be replicated if future researchers can construct intermediate items based on the FACT domains and examine the performance of intermediate items in a personnel selection environment.

Another interesting topic involves the possibility of detecting *differential item functioning* (DIF) for the intermediate items among different samples, especially across different cultures. Culture has been shown to be an influential factor that can affect people’s responses to self-report measures (Hui & Triandis, 1989). For instance, people from Asian cultures tend to avoid choosing extreme options on a Likert-scale. Perhaps this middle response style may lead to over-endorsement of intermediate items. Thus, it is important to test whether there are any cultural differences in people’s response styles for measures developed with an ideal point approach. Recently, researchers have developed different approaches to detecting DIF based on the ideal point model (e.g., Carter & Zickar, 2011; Wang, Tay, & Drasgow, 2013), making it possible to compare cultural differences in people’s responses to intermediate items.

CHAPTER 5

CONCLUSION

The current study empirically showed that intermediate items, which were an essential part of Thurstone scaling, can also be used in a Likert scale format without sacrificing performance. With the inclusion of intermediate items, personality and vocational interest scales not only exhibited better model fit with the ideal point model than with a dominance model, but also provided more information for individuals with extreme latent traits. The study also investigated different approaches to constructing intermediate items, and found that comparing-to-average and specifying-frequencies strategies can be used to write intermediate items.

TABLES

Table 1

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Personality Items in the Order Scale

					GGUM			**2PL**	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	***α***	***δ***	***τ***	***a***	***b***
1	Positive	I wish that everyone was as organized as me.		.57	1.52	1.62	-1.64	0.91	-0.05
2	Positive	Organizing and arranging things is extremely fulfilling.		.51	1.22	1.71	-2.67	0.79	-0.96
3	Positive	I plan my time very carefully.		.57	1.21	1.89	-2.06	0.88	-0.21
4	Positive	I follow a strict daily schedule.		.48	0.92	1.46	-1.01	0.63	0.27
5	Negative	Organizing things is a waste of time.		.51	1.67	-3.11	-0.79	0.99	-2.28
6	Negative	I prefer not to plan ahead and instead take life as it comes.		.42	1.13	-1.71	-0.47	0.62	-1.22
7	Negative	I am an unorganized person.		.61	2.15	-1.99	-0.94	1.23	-1.02
8	Negative	It's hard for me to keep things in order.		.57	2.33	-1.83	-0.82	1.21	-1.01
9	Intermediate	I try to keep track of my bills, but I'm not too accurate	Transition	-.26	1.03	-1.22	-0.29	0.12	4.13
10	Intermediate	I try to keep my room clean and tidy, but I don't always have time to do so.	Transition	-.11	1.99	-0.32	-1.68	0.27	-3.08
11	Intermediate	I can ignore a mess for a long time, but eventually I clean it up.	Transition	-.47	1.68	-1.28	-1.64	0.09	-2.71
12	Intermediate	Occasionally I miss a deadline or two.	Frequency	-.33	1.02	-1.37	-0.76	0.10	2.63
13	Intermediate	Sometimes I do not put things in their proper place.	Frequency	-.54	2.14	-0.96	-1.96	0.12	-6.16
14	Intermediate	Sometimes I can tolerate the messiness of my room.	Frequency	-.44	1.68	-0.81	-1.90	0.14	-5.18
15	Intermediate	I spend time cleaning and organizing things when I am not busy.	Condition	.58	1.15	1.69	-2.29	0.80	-0.60
16	Intermediate	I deviate from my routines when needed.	Condition	-.08	1.06	-0.39	-3.15	0.26	-5.53
17	Intermediate	When my desk gets too messy, I will clean it up.	Condition	.45	2.06	0.69	-2.47	1.11	-1.72
18	Intermediate	I am about average in regard to details.	Average	-.14	0.69	-0.65	-0.74	0.17	-0.14
19	Intermediate	My room neatness is about average.	Average	.01	1.93	-0.05	-1.23	0.35	-1.39
20	Intermediate	I consider myself as organized as most other people.	Average	.21	1.24	0.27	-1.26	0.48	-0.88

Note. *N* = 355. ITCc = Corrected item-total correlation.

Table 2

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Personality Items in the Dominance Scale

					GGUM			**2PL**	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	α	δ	τ	a	b
1	Positive	I enjoy being in a position of power.		.74	2.55	1.61	-2.34	1.50	-0.71
2	Positive	I can give orders when I am in a leadership position.		.59	1.53	0.66	-2.89	0.86	-2.22
3	Positive	I enjoy leading a group.		.78	3.26	1.45	-2.40	2.11	-0.89
4	Positive	I want to take charge of everything I do.		.40	0.73	1.94	-1.43	0.43	0.24
5	Negative	I am not a leader.		.75	3.14	-2.56	-1.23	1.99	-1.26
6	Negative	I dislike taking the responsibility of leading a group.		.71	2.48	-2.25	-1.32	1.90	-0.82
7	Negative	I try to avoid leadership roles.		.80	3.98	-2.37	-1.49	2.96	-0.80
8	Negative	Given a choice of being a follower or a leader, I would always choose to be a follower.		.73	3.34	-2.43	-1.44	2.07	-0.92
9	Intermediate	I am usually not vocal about my opinions, but I will speak up when needed.	Transition	-.32	0.97	-1.53	-2.13	0.11	-3.19
10	Intermediate	I am a forceful person, but I also feel comfortable making compromises.	Transition	.39	0.73	1.47	-1.83	0.43	-0.60
11	Intermediate	I like to be a leader, but I also enjoy being a follower.	Transition	.11	1.52	-0.23	-1.90	0.35	-2.69
12	Intermediate	Sometimes I can persuade my friends to do things in my way.	Frequency	.28	0.83	1.12	-4.10	0.50	-3.02
13	Intermediate	Sometimes I feel comfortable leading a group.	Frequency	.58	1.65	0.87	-2.65	1.01	-1.72
14	Intermediate	Occasionally I speak up to influence others' decisions.	Frequency	.41	1.14	0.66	-3.17	0.65	-2.56
15	Intermediate	I will lead a group only when I'm interested in getting the task done.	Condition	-.27	0.93	-1.42	-1.72	0.11	-1.77
16	Intermediate	I do not mind taking the leadership position if nobody else in the group would like to.	Condition	.65	1.83	1.37	-3.03	1.22	-1.56
17	Intermediate	I will take charge only when I feel it is necessary.	Condition	-.41	1.43	-1.82	-2.06	##	##
18	Intermediate	Compared to my friends, I am about average in showing dominance over others.	Average	.13	1.16	-0.03	-1.28	0.34	-1.31
19	Intermediate	I am as dominant as other people on average.	Average	.17	1.08	0.05	-1.15	0.35	-0.92
20	Intermediate	My desire to lead a group is about average.	Average	-.01	1.47	-0.37	-1.26	0.23	-1.66

Note. N = 355. ITCc = Corrected item-total correlation. ## = Item parameters cannot be estimated by BILOG program.

Table 3

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Personality Items in the Curiosity Scale

					GGUM			**2PL**	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	***α***	***δ***	***τ***	***a***	***b***
1	Positive	I am excited about new knowledge.		.67	2.32	0.95	-2.64	1.43	-1.61
2	Positive	I am fascinated by science.		.37	0.73	1.42	-2.20	0.58	-0.84
3	Positive	I like to learn new things whenever I have time.		.57	1.31	1.02	-2.50	1.07	-1.31
4	Positive	I am always intrigued by what I learn in classes.		.36	0.81	1.20	-1.05	0.69	-0.03
5	Negative	I am not curious about the things that I don’t know.		.40	1.08	-2.69	-0.57	0.73	-1.89
6	Negative	I learn new things only when I have to.		.48	1.83	-2.70	-1.38	0.82	-1.49
7	Negative	I am not interested in learning new things.		.45	1.17	-3.27	-0.61	0.64	-2.75
8	Negative	I would prefer a job where I don't have to learn anything new.		.50	1.88	-2.77	-1.33	0.98	-1.50
9	Intermediate	I can be persuaded to try some new things, but most of the time I am reluctant to do so.	Transition	-.30	0.86	-1.81	-0.76	0.14	2.92
10	Intermediate	I like to experience new things, but seldom have time.	Transition	.02	0.65	-0.74	-1.15	0.24	-0.76
11	Intermediate	I am not excited about new technology, but I become interested when others show me how to use it.	Transition	-.20	0.75	-1.63	-0.84	0.15	1.52
12	Intermediate	Sometimes I read non-fiction books to learn something new.	Frequency	.38	0.65	1.86	-1.03	0.48	0.42
13	Intermediate	At times I prefer to try new things rather than stick to old choices.	Frequency	.26	0.85	1.09	-2.59	0.46	-1.79
14	Intermediate	Occasionally I find myself interested in information that I really don't need.	Frequency	.32	1.06	0.95	-3.14	0.65	-2.23
15	Intermediate	I am open to new concepts only if they are not hard to understand.	Condition	-.42	1.69	-1.79	-1.41	0.09	2.65
16	Intermediate	I try new restaurants only when other people recommend them.	Condition	-.17	0.77	-1.74	-0.77	0.16	1.99
17	Intermediate	I do not mind trying new things when there are not many choices.	Condition	.10	0.97	-0.18	-3.03	0.32	-4.28
18	Intermediate	I am about as curious as my friends.	Average	-.07	0.79	-0.29	-2.16	0.24	-2.99
19	Intermediate	I am about average in curiosity about new knowledge.	Average	-.35	1.05	-1.26	-1.61	0.12	-1.93
20	Intermediate	I have a moderate interest in learning new skills.	Average	.11	0.86	-0.52	-2.79	0.28	-3.76

Note. *N* = 355. ITCc = Corrected item-total correlation.

Table 4

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Vocational Interest Items in the Realistic Scale

					GGUM			2PL	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	***α***	***δ***	***τ***	***a***	***b***
1	Positive	I enjoy things like laying brick or tile very much.		.68	1.81	2.56	-0.81	1.18	1.63
2	Positive	I like to work on an offshore oil-drilling rig more than most people.		.42	1.44	3.17	-0.84	0.95	2.14
3	Positive	I have a passion on setting up and operating machines to make products.		.62	2.14	2.00	-0.50	1.33	1.47
4	Positive	I am always willing to repair household appliances.		.66	2.31	1.97	-1.30	1.46	0.66
5	Positive	I always feel excited about fixing things around the house.		.58	1.70	1.69	-1.22	1.04	0.44
6	Intermediate	I like fixing a broken faucet only when nobody else can.	Condition	.55	1.17	1.82	-0.65	0.73	1.06
7	Intermediate	Sometimes I like to fix mechanical things for fun.	Frequency	.77	3.01	2.25	-1.52	1.90	0.73
8	Intermediate	My interest in installing flooring in houses is about average.	Average	.54	1.62	1.52	-0.83	0.86	0.72
9	Intermediate	I would like to operate a machine on a production line, but I would soon get bored.	Transition	.49	1.17	1.93	-0.92	0.76	0.87
10	Intermediate	I have a moderate interest in repairing and installing locks.	Average	.68	1.97	2.06	-0.73	1.27	1.26
11	Negative	I don't think it interesting to operate a grinding machine in a factory.		.40	1.07	-0.69	-2.25	0.59	1.64
12	Negative	I don't like building kitchen cabinets.		.64	2.39	-0.84	-1.55	1.25	0.69
13	Negative	I have no interest in building a brick walkway.		.67	2.17	-0.91	-2.03	1.26	1.11

Note. *N* = 355. ITCc = Corrected item-total correlation.

Table 5

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Vocational Interest Items in the Social Scale

					GGUM			2PL	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	α	δ	τ	a	b
1	Positive	I love giving career advice to people.		.23	0.58	1.15	-1.18	0.36	-0.32
2	Positive	I always enjoy helping elderly people with their daily activities.		.43	1.33	0.90	-1.16	0.65	-0.27
3	Positive	I am more interested in teaching an elementary school class than most other people.		.57	1.14	1.63	-1.34	0.79	0.16
4	Positive	I would always love to work with mentally disabled children.		.65	3.12	1.00	-1.04	1.53	0.03
5	Positive	I would love to have the opportunity to teach disabled people work and living skills.		.67	3.48	1.03	-1.24	1.95	-0.16
6	Intermediate	I would help people with family-related problems only if I was paid.	Condition	-.13	0.45	-2.99	-0.53	0.15	3.42
7	Intermediate	Sometimes I like to teach children how to read.	Frequency	.61	1.56	1.63	-2.38	0.98	-0.74
8	Intermediate	I like to help my neighbors only when I have time.	Condition	-.04	0.60	-0.71	-1.04	0.15	-0.98
9	Intermediate	My interest in taking care of children is about average.	Average	.00	0.62	-0.21	-1.26	0.18	-1.60
10	Intermediate	Although I like to volunteer in charities, I am usually not motivated.	Transition	-.35	1.30	-1.09	-0.50	##	##
11	Negative	I don't find teaching a high-school class attractive to me.		.33	0.60	-1.09	-0.84	0.34	0.01
12	Negative	I have no interest in helping people with drug or alcohol problems.		.30	1.01	-1.81	-0.44	0.51	-1.41
13	Negative	I don't like supervising the activities of children at a camp.		.55	1.48	-2.05	-0.91	0.93	-1.04

Note. *N* = 355. ITCc = Corrected item-total correlation. ## = Item parameters cannot be estimated by BILOG program.

Table 6

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Vocational Interest Items in the Artistic Scale

					GGUM			**2PL**	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	α	δ	τ	a	b
1	Positive	I love drawing pictures.		.69	2.70	1.95	-1.69	1.62	0.20
2	Intermediate	I like to draw pictures only when I have time.	Condition	.58	1.97	1.07	-1.00	0.90	0.06
3	Negative	I have no interest in drawing.		.68	2.64	-1.77	-1.75	1.60	-0.04
4	Positive	I always have a strong interest in playing a musical instrument.		.36	0.81	0.82	-0.94	0.28	-1.38
5	Intermediate	Sometimes I like to play a musical instrument.	Frequency	.50	1.01	0.88	-1.58	0.65	-0.06
6	Negative	I don't think playing musical instruments is interesting.		.39	0.86	-2.80	-0.95	0.48	-2.21
7	Positive	I always dream of acting in a play.		.46	0.97	1.96	-0.35	0.67	1.43
8	Intermediate	I have a moderate interest in acting compared to people on average.	Average	.46	0.73	2.26	-1.09	0.52	0.75
9	Negative	I don't like acting at all.		.46	0.88	-1.95	-1.53	0.61	-0.33
10	Positive	I love taking Art courses.		.75	3.65	1.98	-1.54	2.38	0.43
11	Intermediate	I have a moderate interest in taking Art courses.	Average	.68	2.56	1.30	-1.27	1.53	0.00
12	Negative	I don't find Art courses are interesting.		.73	2.40	-1.95	-1.76	1.60	-0.22
13	Positive	I have a passion on designing artwork.		.70	2.31	2.27	-1.24	0.94	0.92
14	Intermediate	I like designing artwork only when I have time.	Condition	.63	3.18	1.02	-0.75	1.28	0.36
15	Negative	I have no interest in designing artwork.		.73	3.23	-1.38	-1.62	1.84	0.24

Note. *N* = 355. ITCc = Corrected item-total correlation.

Table 7

Item Types, Contents, Domains, ITC$_c$, GGUM and 2PL Parameters of Vocational Interest Items in the Investigative Scale

					GGUM			**2PL**	
Item	**Type**	**Contents**	**Domain**	**ITC$_c$**	***α***	***δ***	***τ***	***a***	***b***
1	Positive	I have a strong interest in reading science-related articles		.80	2.52	1.89	-1.47	1.63	0.40
2	Intermediate	My interest in reading scientific articles is about average.	Average	.30	1.31	0.51	-1.19	0.46	-0.65
3	Negative	I don't find it interesting to read science-related articles.		.68	2.28	-1.79	-1.28	1.47	-0.44
4	Positive	I always feel excited about conducting research in a lab.		.64	1.27	2.09	-1.60	0.82	0.39
5	Intermediate	l like to working in a lab only if it is not time consuming.	Condition	.36	1.10	0.62	-0.23	0.39	1.01
6	Negative	I have no interest in doing research in a lab.		.64	1.62	-1.96	-1.33	1.06	-0.53
7	Positive	I think it is very fascinating to study scientific theories.		.79	2.84	1.88	-1.79	1.83	0.11
8	Intermediate	Sometimes I like to study scientific theories, but not always.	Frequency	.51	2.50	0.59	-1.10	0.80	-0.35
9	Negative	I hate studying a scientific theory.		.67	2.62	-1.63	-1.22	1.63	-0.37
10	Positive	I am very interested in science-related courses.		.80	3.37	1.55	-1.62	2.17	-0.04
11	Intermediate	From time to time I find science-related courses are interesting.	Frequency	.67	1.99	1.41	-2.64	1.26	-1.16
12	Negative	I think courses on science are boring.		.78	3.21	-2.10	-1.57	2.11	-0.48
13	Positive	I would love to have the opportunity to work on a scientific project.		.82	3.73	1.78	-1.83	2.51	-0.02
14	Intermediate	I have a moderate interest in working on scientific projects.	Average	.76	2.32	1.70	-1.96	1.49	-0.23
15	Negative	I dislike working on any scientific project.		.70	2.60	-1.58	-0.99	1.50	-0.57

Note. *N* = 355. ITCc = Corrected item-total correlation.

Table 8

Means and Standard Deviations of the Adjusted χ^2/df Ratios of GGUM and 2PL Models for Each Scale

Scale	**Number of Items**	**GGUM Mean Adjusted χ^2/df**			**2PL Mean Adjusted χ^2/df**		
		Singlets	**Doublets**	**Triplets**	**Singlets**	**Doublets**	**Triplets**
Personality Scales							
Order	20	0.00 (0.00)	1.33 (5.45)	1.68 (4.24)	0.00 (0.00)	19.92 (32.77)	27.86 (31.33)
Dominance	20	0.00 (0.00)	1.28 (5.17)	1.41 (3.78)	0.00 (0.00)	7.13 (18.23)	9.15 (14.38)
Curiosity	20	0.00 (0.00)	1.07 (3.84)	1.22 (3.26)	0.00 (0.00)	10.22 (21.79)	12.83 (18.65)
Vocational Interest Scales							
Artistic	15	0.00 (0.00)	10.02 (40.75)	13.34 (31.50)	32.38 (71.32)	33.97 (62.20)	31.14 (44.05)
Investigative	15	0.04 (0.15)	2.12 (6.64)	2.16 (4.73)	0.00 (0.00)	1.34 (5.63)	3.04 (6.21)
Realistic	13	0.00 (0.00)	1.04 (4.32)	1.05 (2.84)	0.00 (0.00)	0.74 (3.33)	0.96 (2.78)
Social	13	0.00 (0.00)	2.29 (8.07)	3.12 (6.92)	0.00 (0.00)	5.36 (10.15)	6.91 (7.91)

Note. Values in the parentheses denote the standard deviations.

Table 9

Means and Standard Deviations of the Adjusted χ^2/df Ratios of GGUM and 2PL Models for Personality Scales by Intermediate Domains

Scale	**Number of Items**	**GGUM Mean Adjusted χ^2/df**		
		Singlets	**Doublets**	**Triplets**
Order				
Transition	11	0.00 (0.00)	1.33 (5.37)	1.60 (3.84)
Frequency	11	0.00 (0.00)	0.81 (3.14)	0.98 (2.35)
Condition	11	0.00 (0.00)	1.03 (4.38)	1.35 (3.59)
Average	11	0.00 (0.00)	0.81 (3.91)	0.88 (2.64)
Dominance				
Transition	11	0.00 (0.00)	0.71 (3.24)	1.06 (2.55)
Frequency	11	0.00 (0.00)	1.22 (3.79)	1.47 (3.36)
Condition	11	0.00 (0.00)	0.78 (3.41)	0.83 (2.40)
Average	11	0.00 (0.00)	1.56 (6.47)	1.72 (4.11)
Curiosity				
Transition	11	0.00 (0.00)	0.38 (1.58)	0.35 (1.17)
Frequency	11	0.00 (0.00)	0.52 (2.32)	0.33 (1.27)
Condition	11	0.00 (0.00)	0.44 (2.28)	0.38 (1.43)
Average	11	0.00 (0.00)	0.46 (2.11)	0.24 (1.32)

Note. Values in the parentheses denote the standard deviations.

FIGURES

Figure 1. Example of an item response function (IRF) based on the dominance model.

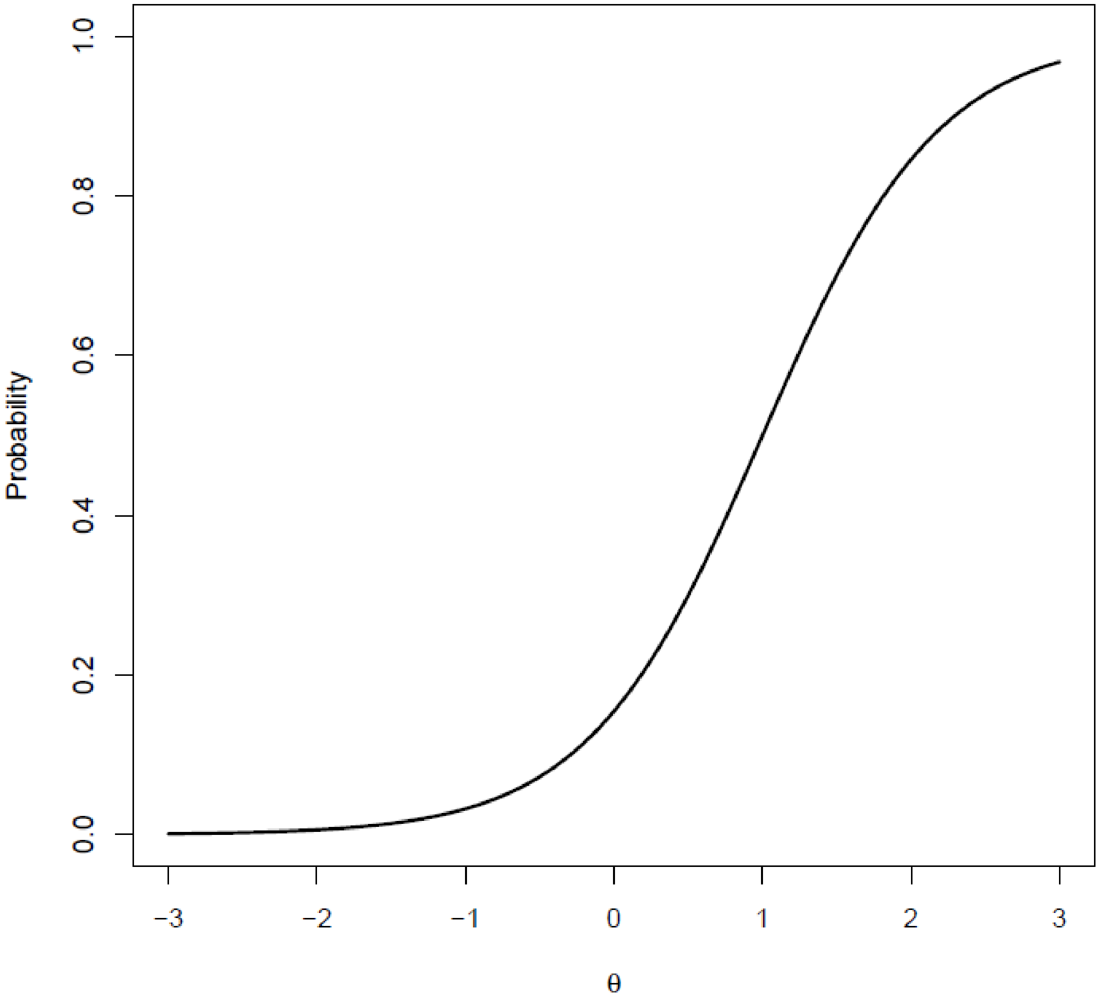


Figure 2. Example of an item response function (IRF) based on the ideal point model.

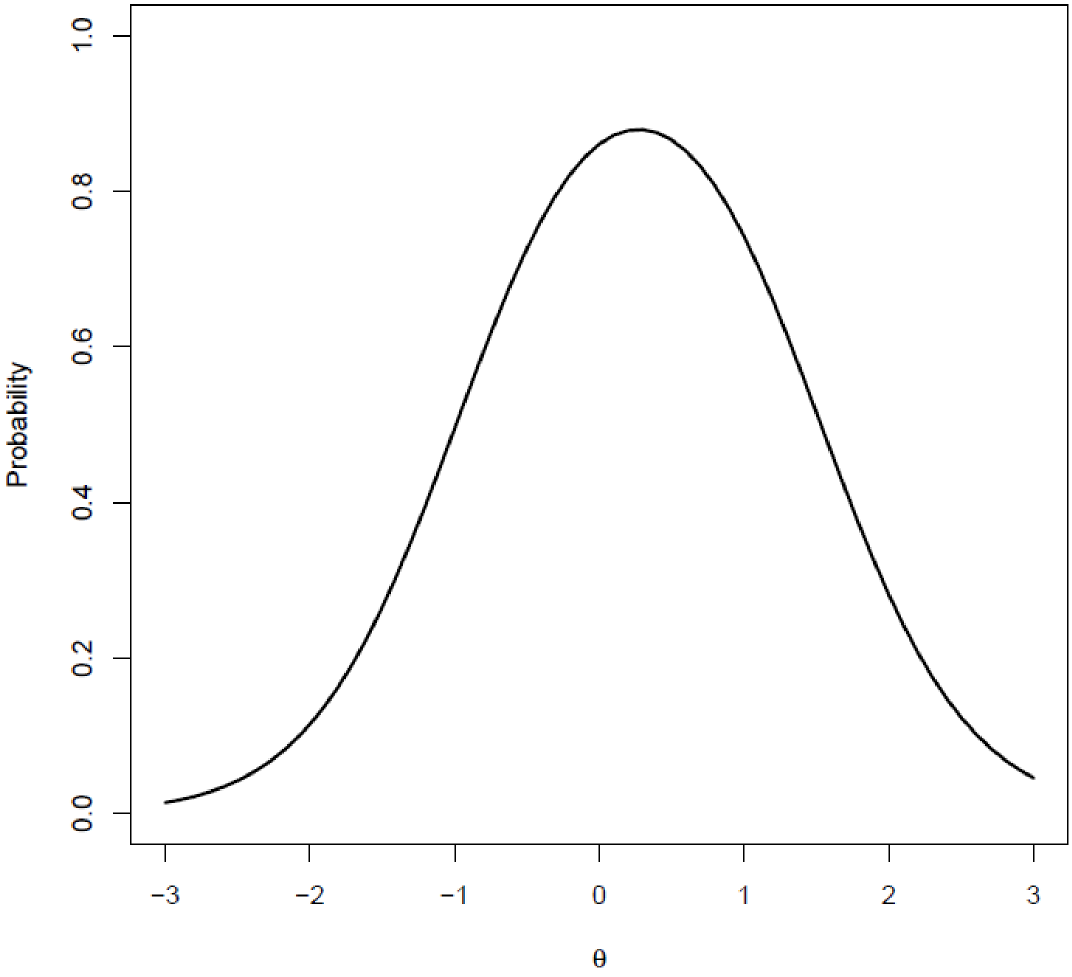


Figure 3a. 2PL Fit plots for Item 3 of the Order scale generated by MODFIT 3.0 (Stark, 2001). IRF stands for the item response function predicted by the model, and EMP stands for the empirical response function. The error bars represent the 95% confidence intervals for the empirical points.

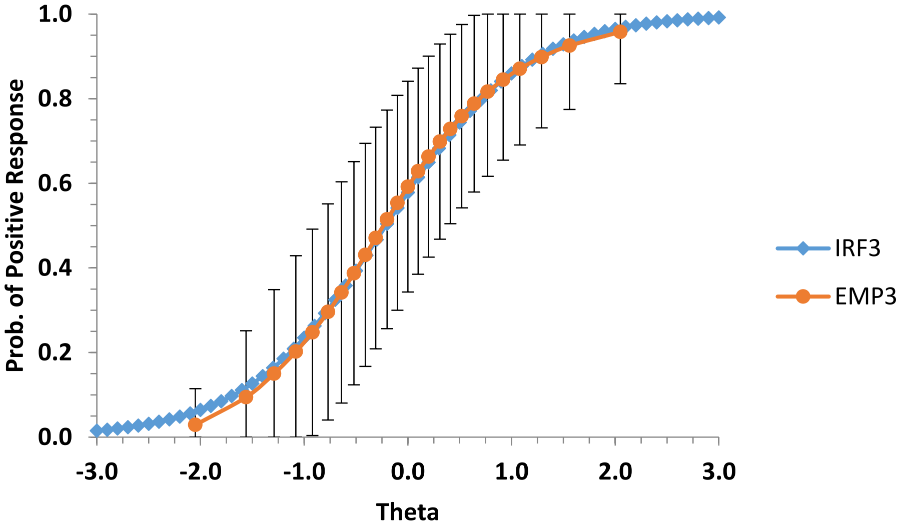


Figure 3b. GGUM Fit plots for Item 3 of the Order scale by MODFIT 3.0 (Stark, 2001). ORF stands for the option response function of option 1 predicted by the model, and EMP stands for the empirical response function. The error bars represent the 95% confidence intervals for the empirical points.

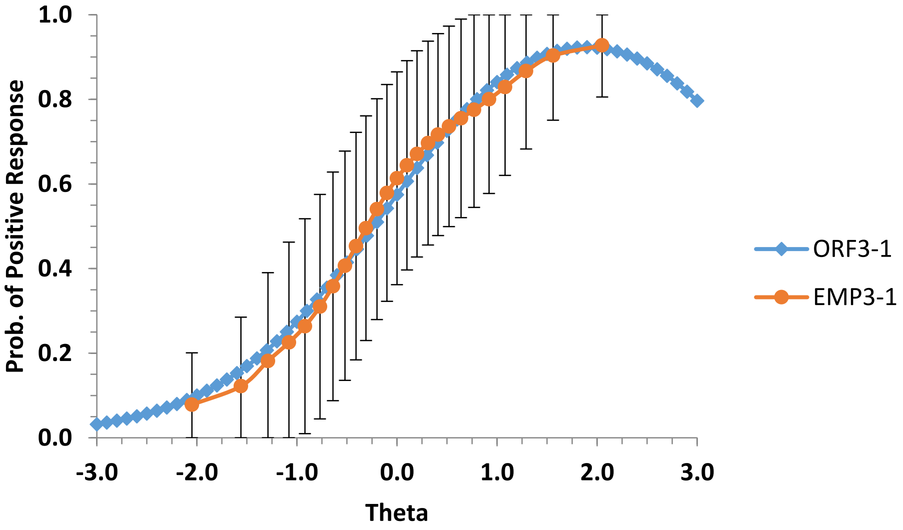


Figure 4a. 2PL Fit plots for Item 10 of the Order scale by MODFIT 3.0 (Stark, 2001). IRF stands for the item response function predicted by the model, and EMP stands for the empirical response function. The error bars represent the 95% confidence intervals for the empirical points.

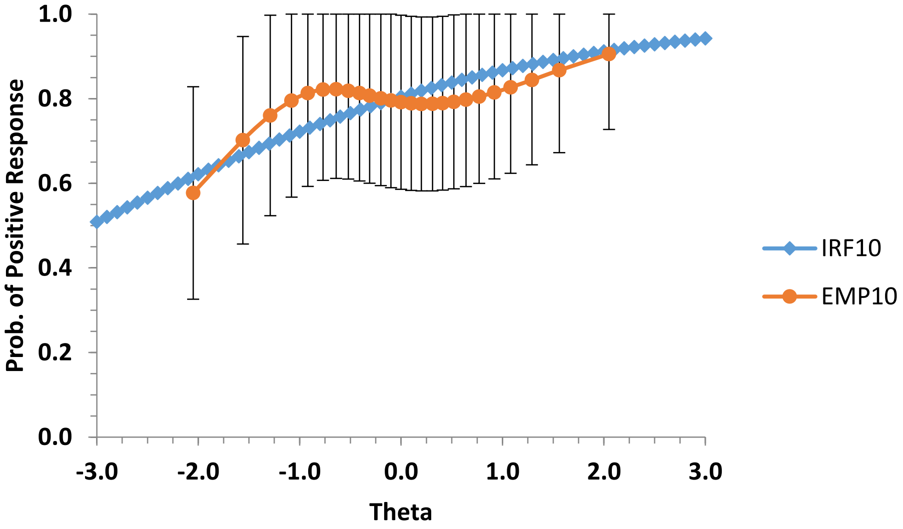


Figure 4b. GGUM Fit plots for Item 10 of the Order scale by MODFIT 3.0 (Stark, 2001). ORF stands for the option response function of option 1 predicted by the model, and EMP stands for the empirical response function. The error bars represent the 95% confidence intervals for the empirical points.

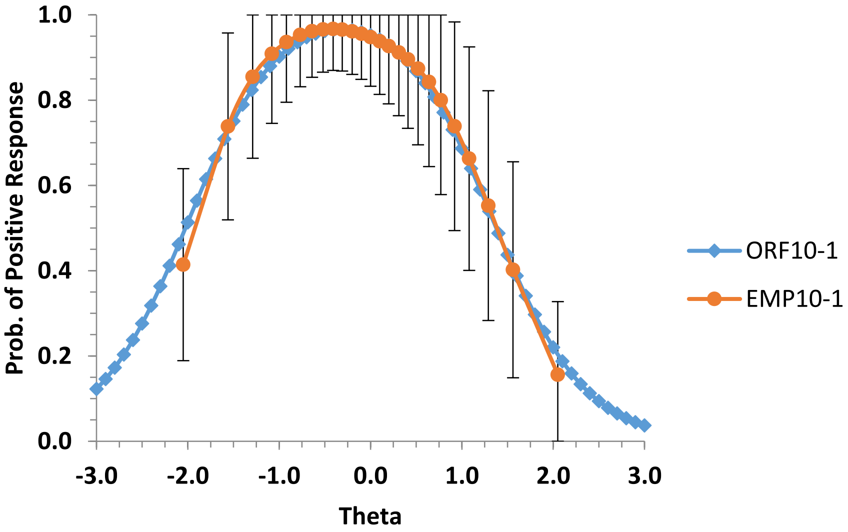


Figure 5. Comparison of item information functions for Item 3 in the Order scale between the GGUM and the 2PL model.

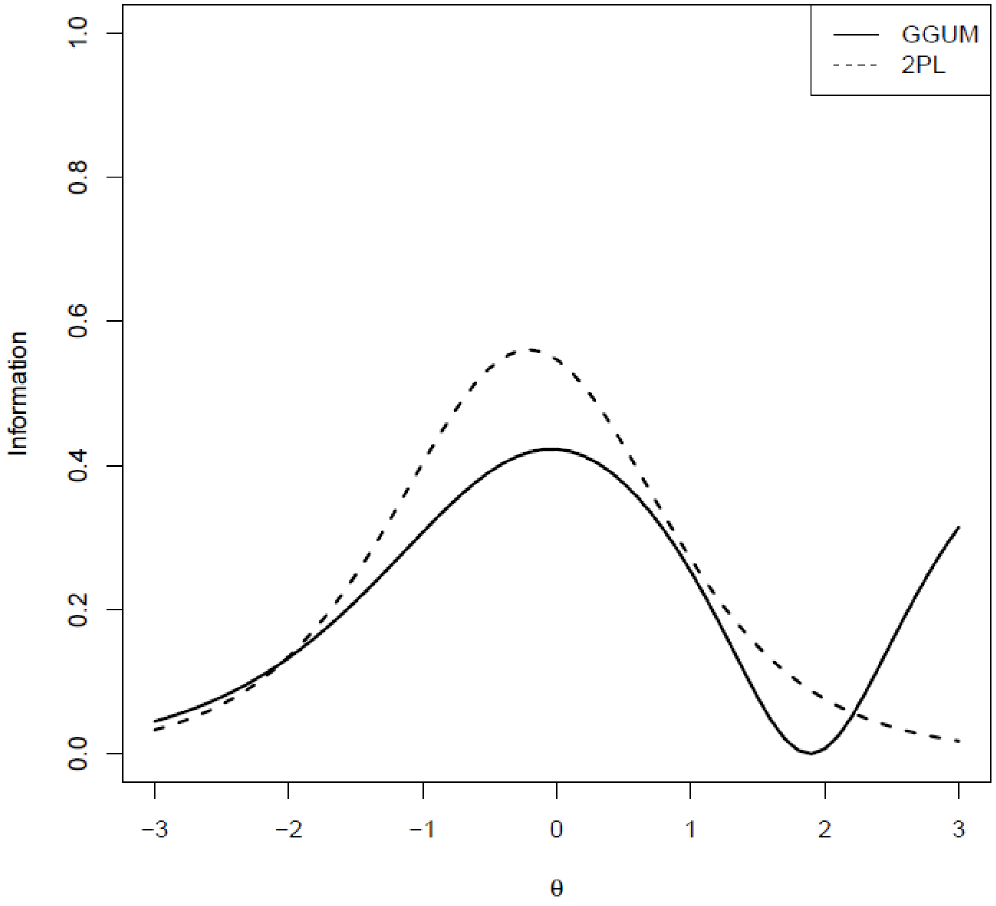


Figure 6. Comparison of item information functions for Item 10 in the Order scale between the GGUM and the 2PL model.

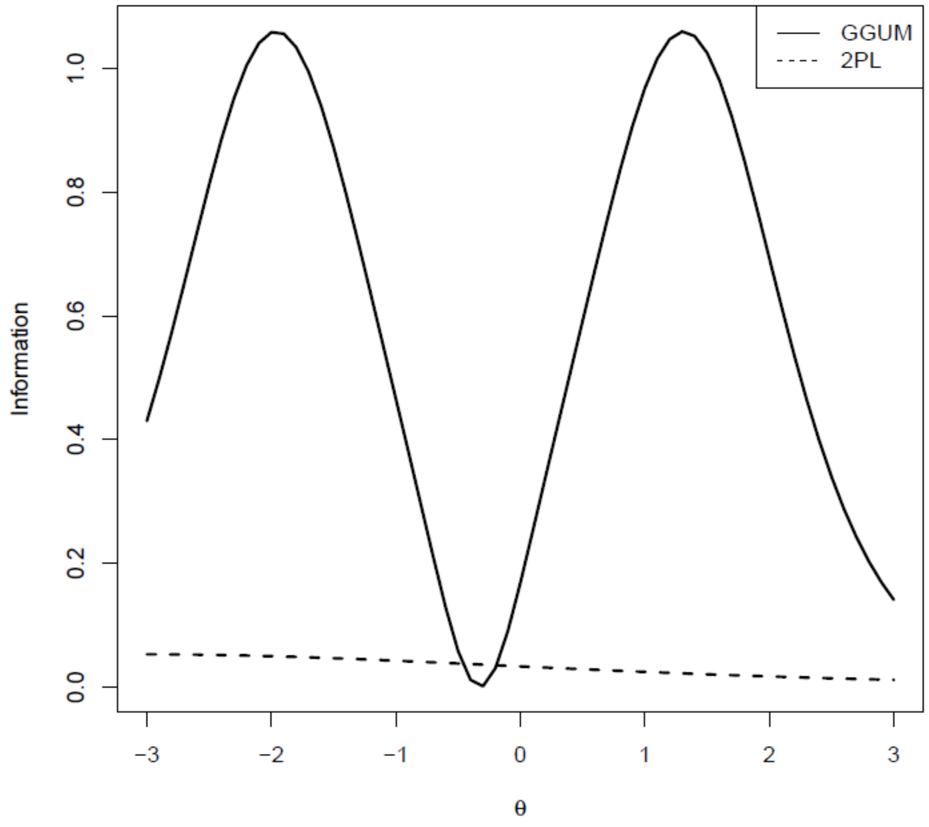


Figure 7a. Comparison of GGUM and 2PL test information functions for the Order scale.

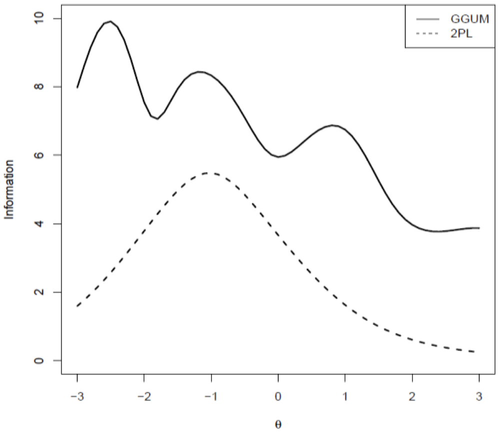


Figure 7b. Comparison of GGUM and 2PL test information functions for the Dominance scale.

.

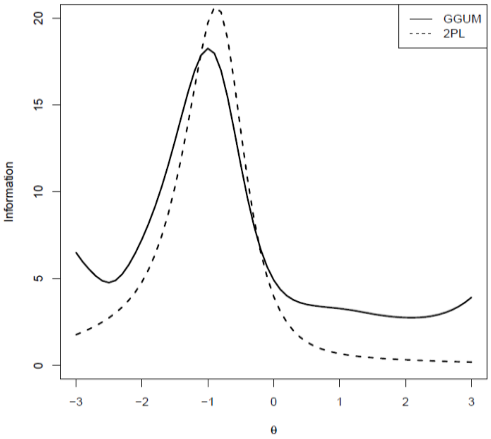


Figure 7c. Comparison of GGUM and 2PL test information functions for the Curiosity scale.

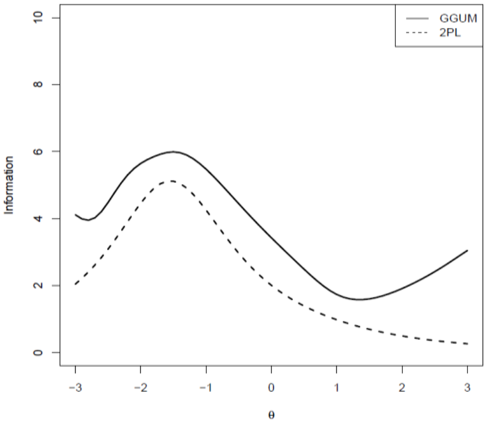
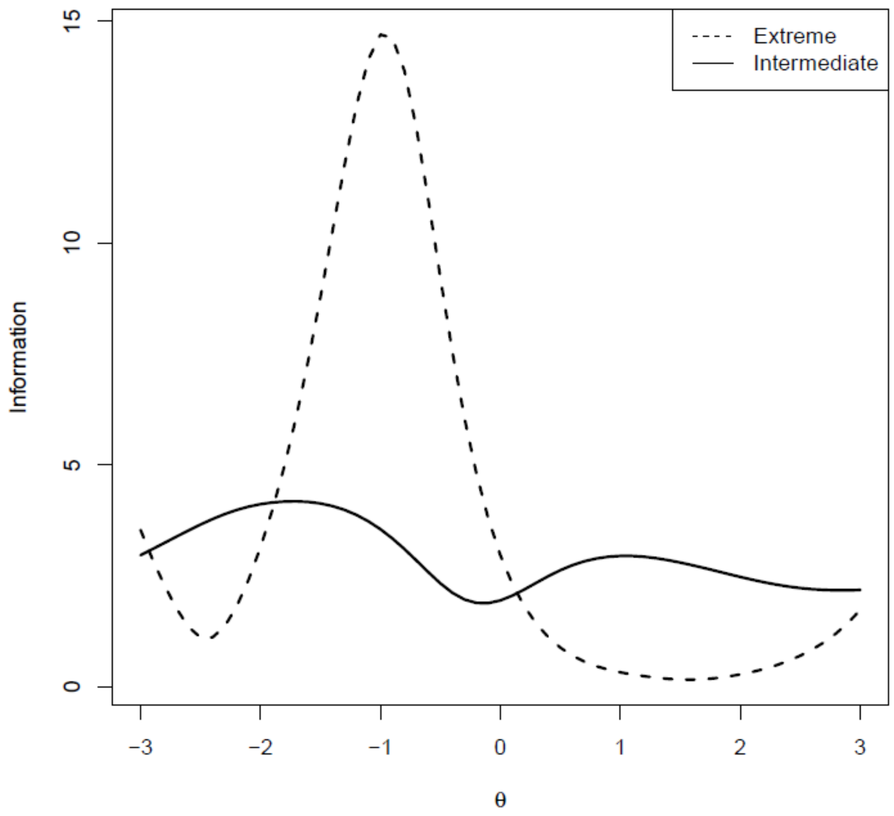


Figure 8. Comparison of cumulative information functions between all extreme items and all intermediate items in the Dominance scale based on the GGUM.



REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Sage.

Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology: Perspectives on Science and Practice,* 3, 489–493.

Carter, N. T., Lake, C. J., & Zickar, M. J. (2010). Toward understanding the psychology of unfolding. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 3, 511–514.

Carter, N. T., & Zickar, M. J. (2011). A comparison of the LR and DFIT frameworks of differential functioning applied to the generalized graded unfolding model. *Applied Psychological Measurement*, *35*(8), 623-642.

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

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

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

Cronbach, L. J. (1949). Essentials of psychological testing. New York: Harper.

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika,* 42, 523–548.

Deng, C.-P., Armstrong, P. I., & Rounds, J. (2007). The fit of Holland's RIASEC model to U.S. occupations. *Journal of Vocational Behavior,* 71, 1–22.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial & Organizational Psychology: Perspectives on Science and Practice*, 3(4), 465-476.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response models to multiple-choice tests. *Applied Psychological Measurement*, 19, 145-165.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Psychology Press.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1, 104-121.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296-309.

Liao, H-Y., Armstrong, P. I., Rounds, J., & Su, R. (2007). *Interest Item Pool (IIP).* Retrieved February 23, 2013 from http://jrounds.weebly.com.

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

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

Mislevy, R. J., & Bock, R. D. (1991). *BILOG user's guide*. Chicago, IL: Scientific Software.

Nye, C. D., Newman, D. A., & Joseph, D. L. (2010). Never say ‘‘always’’? Extreme item wording effects on scalar invariance and item response curves. *Organizational Research Methods*, *13*(4), 806-830.

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207-230.

Reise, S. P. (2010). Thurstone might have been right about attitudes, but Drasgow, Chernyshenko, and Stark fail to make the case for personality. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 3, 485–488.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, *7*(4), 347-364.

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

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, 24, 3–32.

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2004). GGUM2004: A Windows based program to estimate parameters in the generalized graded unfolding model. *Applied Psychological Measurement*, 30, 64–65.

Roberts, J. S., & Shim, H. S. (2008). *GGUM2004 Technical Reference Manual (v1.1)*. Atlanta, GA: Georgia Polytechnic University.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Scherbaum, C. A., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *The Leadership Quarterly*, 17, 366-386.

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

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91, 25–39.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement*, *35*(4), 280-295.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology*, 94, 1287-1304.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.

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

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review*, 36, 222–241.

Thurstone, L. L. (1931). The measurement of change in social attitude. *The Journal of Social Psychology*, 2, 230-245.

Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement*.

Waples, C. J., Weyhrauch, W. S., Connell, A. R., & Culbertson, S. S. (2010). Questionable defeats and discounted victories for Likert rating scales. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 3, 477–480.

Zampetakis, L. A. (2011). The measurement of trait emotional intelligence with TEIQue-SF: An analysis based on unfolding item response theory models. *Research on Emotion in Organizations*, 7, 289-315.

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