Psychological Assessment
2007, Vol. 19, No. 1, 88–106

Copyright 2007 by the American Psychological Association
1040-3590/07/$12.00 DOI: 10.1037/1040-3590.19.1.88

Constructing Personality Scales Under the Assumptions of an Ideal Point Response Process: Toward Increasing the Flexibility of Personality Measures

Oleksandr S. Chernyshenko
University of Canterbury

Stephen Stark
University of South Florida

Fritz Drasgow and Brent W. Roberts
University of Illinois at Urbana–Champaign

The main aim of this article is to explicate why a transition to ideal point methods of scale construction is needed to advance the field of personality assessment. The study empirically demonstrated the substantive benefits of ideal point methodology as compared with the dominance framework underlying traditional methods of scale construction. Specifically, using a large, heterogeneous pool of order items, the authors constructed scales using traditional classical test theory, dominance item response theory (IRT), and ideal point IRT methods. The merits of each method were examined in terms of item pool utilization, model–data fit, measurement precision, and construct and criterion-related validity. Results show that adoption of the ideal point approach provided a more flexible platform for creating future personality measures, and this transition did not adversely affect the validity of personality test scores.

Keywords: personality measurement, test construction, item response theory

Measurement issues have been an integral part of personality research over the last several decades. Assertions involving personality test scores were at the heart of the person–situation debate in the 1970s (Kenrick & Funder, 1988; Mischel, 1968), the taxonometric research on the Big Five in the 1980s (i.e., Goldberg, 1993; McCrae & Costa, 1989), and the ongoing polemic concerning the effects of socially desirable responding on the utility of personality measures as predictors of performance in high-stakes settings (Ones, Viswesvaran, & Reiss, 1996; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Conclusions have relied heavily on the use of personality measures, whose quality, or lack thereof, often determines the veracity of research findings.

Although the value of currently available personality measures cannot be overstated, the outlook for their continuing improvement is less certain. To date, most improvements can be traced to expanding, cataloging, and refining the personality construct space; perhaps the best example is Goldberg's (1998) International Personality Item Pool initiative, developed at the Oregon Research Institute. Considerably less progress has been made, however, in the development and application of new measurement technology for personality test construction. For the most part, personality test batteries still consist of many 10- to 15-item scales created using classical test theory methods and scored by summing values for endorsed response options across items (i.e., Likert's approach). More important, because the number of scales is usually large, to achieve reasonable levels of reliability, many items are often administered. This unfortunately increases testing time and, when coupled with rigid test administration procedures (e.g., all items must be administered to each examinee in a predetermined order), makes current personality test batteries cumbersome to use and to maintain.

Note that there certainly has been no shortage of psychometric research in personality assessment. To the contrary, since the early 1990s, Steven Reise, Niels Waller, and their colleagues have published a series of item response theory (IRT) articles (e.g., Reise, 1999; Reise & Waller, 1990) exploring the fit of the one- and two- parameter logistic models (1PLM and 2PLM, respectively) to items of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982). They also designed what is arguably the first computerized adaptive personality test (Waller & Reise, 1989). Subsequently, many researchers have used models discussed in their articles to design better conventional and adaptive personality measures (e.g., Ellis, Becker, & Kimmel, 1993; Reise & Henson, 2000; Rouse, Finger, & Butcher, 1999; Simms & Clark, 2005), to investigate effects of the testing situation on item responding (Stark et al., 2001), and to study context effects (Steinberg, 2001). Whereas this and other research has answered many important substantive and practical questions, there have been minimal changes in the way personality scales are constructed and scored. One possible explanation for this is that the IRT models used in the majority of these studies were created in the context of educational assessment where cognitive ability items having de-

Oleksandr S. Chernyshenko, Department of Psychology, University of Canterbury, Christchurch, New Zealand; Stephen Stark, Department of Psychology, University of South Florida; Fritz Drasgow and Brent W. Roberts, Department of Psychology, University of Illinois at Urbana–Champaign.

Correspondence concerning this article should be addressed to Oleksandr S. Chernyshenko, Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. E-mail: sasha.chernyshenko@canterbury.ac.nz

monstrably correct responses comprise the vast proportion of measures. Although there certainly is nothing wrong with adopting models from other domains, one cannot help but wonder whether this comes with a price. After all, personality questions are very different from cognitive ability questions; they often do not have a "correct" response and, as such, assess one's typical behavior (what one usually does) rather than maximum performance (what one can do).

In this article, we discuss measurement conventions that might be hindering the advancement of personality testing technology, namely, the strict adherence to *dominance* methods for evaluating and scoring items. These methods, which include classical test theory and common factor theory, essentially assume that respondents with higher trait levels exhibit higher item scores than do those with low trait levels and that the probability of observing a high item score increases monotonically as the distance between person and item locations increases (for a more detailed explanation, see the Why Dominance Assumptions May Impede New Developments in Personality Assessment section). Almost all extant personality scales have been created from a dominance perspective. Moreover, the majority of IRT research involving personality data has also used dominance IRT models (i.e., 2PLM or graded response models). Yet, another class of models, known as ideal point models, may provide viable alternatives for personality assessment. Here, we discuss why ideal point methods should be considered for personality test construction and illustrate how to develop scales from this perspective. Specifically, we construct an ideal point scale measuring order, a core facet of conscientiousness (B. Roberts, Chernyshenko, Stark, & Goldberg, 2005), and compare it with scales developed, from the same pool of items, using classical test theory (CTT) and IRT dominance procedures. Finally, we discuss why ideal point approaches to test construction may represent a step forward in personality assessment technology and how this change in thinking might facilitate the transition to more flexible personality tests.

Why Dominance Assumptions May Impede New Developments in Personality Assessment

The majority of personality scales in use today were developed using Likert's (1932) approach, which is rooted in classical test theory and factor-analytic methodologies. In this framework, a large number of homogenous items are first generated with content related to the attribute of interest and administered to a sample of respondents instructed to indicate their level of agreement or disagreement on a scale ranging from, say, 1 to 5. After reverse scoring negatively worded items, response data are used to create scales by selecting items that show moderate to high item–total correlations; those with low correlations are eliminated because they appear not to discriminate well among respondents (a variation of this approach is to drop items that have low loadings on the common factor measured by the scale). As a result, scales developed using Likert's method generally exhibit high internal consistency and reliability and a strong single-factor structure.

Undergirding this process of scale construction is the assumption that item responses follow a *dominance process*; that is, if both persons and items are located on a continuum representing an attribute of interest, then a person will tend to endorse a positively worded item when his or her standing on the latent dimension is more positive than that of the item (the same is true for a negatively worded item after reverse scoring). To illustrate a dominance response process graphically, one can plot the probability of a positive response (item endorsement or agreement) against trait level. The resulting curve is known as an item response function (IRF), an example of which is shown in Figure 1A. The figure shows that the probability of a positive response increases as a person lies increasingly distant and above the item's location. Thus, the IRFs for a dominance response process must be *monotonically* increasing across all trait levels. Note, empirical approaches to personality test construction, typified by the Minnesota Multiphasic Personality Inventory (MMPI; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) and the California Personality Inventory (CPI; Gough & Bradley, 1996), also assume a dominance response process; the only difference is that correlations of item scores with various criteria are used, instead of total score, to determine which items to retain and which to discard.

Figure 1B shows the item information function (IIF) for the same item. In IRT, information is an important conditional statistic. Item information values over a range of trait levels can be displayed graphically using an IIF. As shown in Figure 1B, this

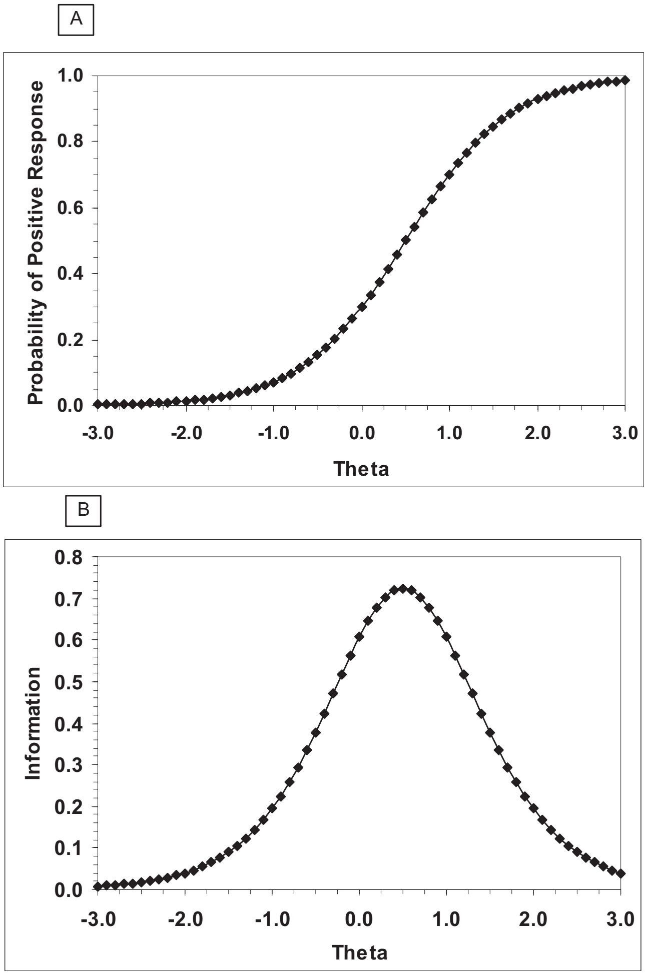

Figure 1. Panel A: Example of dominance item response function. Panel B: Example of the item information function for the same item.

item is most informative for persons having trait levels near 0 to 1 and less informative for examinees at either extreme. Information is inversely related to the standard error of the estimate $\hat{\theta}$ of the latent trait θ $\left(SE(\theta)=1/\sqrt{I(\theta)}\right)$. Where information is high, error is low, and vice versa. Note that if one closely examines Figures 1A and 1B simultaneously, one can see that information is high where the slope of the IRF is steep and relatively low in areas where the IRF is flat. For this reason, such an item would not be very useful for identifying, say, the most emotionally stable candidates for pilot training or the least emotionally stable clients for psychological intervention. This item would be useful, however, for separating respondents into two groups (i.e., stable vs. unstable), and creating a scale by combining many items of this sort would provide effective measurement for persons in the middle of the trait continuum because test (scale) information is computed as the sum of the IIFs. On the other hand, if the goal is to create a scale that measures well at extreme trait levels, items providing high information at θs beyond ± 1 would be needed.

It is important to realize that the decision to discard items exhibiting low item–total correlations, a decision which seems natural because it is ingrained in almost every personality assessment course, is justified only when the condition of monotonicity holds true for all items under consideration. When it does, a low item–total correlation can be viewed as evidence that an item fails to discriminate. However, if in fact *nonmonotonic* items are present in an item pool, then item–total correlations and factor loadings are not appropriate statistics for judging item quality because they inherently assume a monotonic relationship between item score and trait level. For example, in attitude assessment, research has identified discriminating nonmonotonic items that have bell-shaped IRFs (e.g., Andrich, 1996; J. S. Roberts, Laughlin, & Wedell, 1999). The bell-shaped response functions indicate that individuals with low and high standing on the trait continuum have similar expected item scores, which leads to low item–total correlations. Consequently, despite being discriminating, nonmonotonic items will inevitably be judged as “poor” by traditional evaluative criteria.

Recent psychometric studies involving the MMPI (Meijer & Baneke, 2004) and the Sixteen Personality Factor Scale (16PF) (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Stark, Chernyshenko, Drasgow, & Williams, 2006) have also revealed the presence of items having nonmonotonic IRFs. Although the number of such items was small in each case, their mere presence calls into question the appropriateness of the nearly universal assumption of a dominance process for personality assessment. As noted by Stark et al. (2006), the number of nonmonotonic items was probably much larger in the original item pools, but requiring all items to have high factor loadings and item–total correlations during scale construction effectively eliminated them from the final measures. This could have unintended consequences in terms of content or measurement precision in particular regions of the trait continuum.

If one examines carefully the items in most currently used personality scales, it becomes immediately apparent that they are relatively positive and negative (e.g., “I like parties and social events”; “I disregard rules and norms”). These items tend to have monotonic IRFs and therefore look good by traditional evaluative criteria. On the other hand, items that describe behaviors tending toward neutrality (e.g., “At parties, I mostly talk with my friends”; “I do pretty standard maintenance of my property and possessions”) are rarely, if ever, observed because they tend to have nonmonotonic IRFs and are judged as poor by traditional criteria (Chernyshenko, 2002).

The decision to retain only items reflecting relative extremity, however, has a number of undesirable psychometric consequences. First, it turns out that the majority of currently used positively and negatively worded items have information functions that are very similar to the ones shown in Figure 1B.[1] Thus, adding such items to a scale does little to improve measurement precision in underrepresented regions of the trait continuum. In this respect, personality scales more closely resemble selection or licensure exams, where items are clustered to provide high information at specific trait levels (i.e., in the middle), rather than diagnostic tools, where the goal typically is to obtain equal measurement precision across the whole trait continuum. Developing the latter requires an item pool with greater variability in terms of the IIFs, but that is difficult to achieve when item–total correlations are used to evaluate item quality. In accordance with the dominance framework, one would need to include more extreme items (e.g., “Under no circumstances would I submit an assignment that I did not check for mistakes at least three times”), but these are rarely endorsed and would also tend to have very low variances and item–total correlations.

Second, item location statistics used in the dominance framework (*p* values or IRT item difficulty [b–] parameters) have no direct correspondence to item content. As is shown later in our empirical example, both positively and negatively worded items often have very similar *p* values. Even more complications arise when negatively worded items are reverse scored (a requirement for nearly all analyses involving the dominance framework) and the results reported. The upshot is that scale developers have little a priori insight as to the kinds of items needed to develop or improve their scales. Instead, one must eagerly await the results of a pretest administration and hope that items with the desired properties are found. This is, again, drastically different from the cognitive ability testing domain, where difficulty parameters are often directly linked to item content and item writers have advanced knowledge of how to construct items with the desired psychometric properties.

Finally, the need to eliminate nonmonotonic items reduces the richness of the pool available for scale construction. Having a large pool of items with a diverse range of *p* values is important for parallel form construction as well as for computerized adaptive testing applications with either unidimensional or multidimensional paired comparison formats (Chernyshenko, Stark, Prewett, Gray, Stilson, & Tuttle, 2006). Forced-choice formats are enjoying a revival in the applied psychology literature, as many organizations want to use fake-resistant personality instruments to predict employee performance and to improve retention (Jackson, Wrobleski, & Ashton, 2000).

[1] Without reverse scoring, a negatively worded item would have a monotonically decreasing IRF rather than an increasing IRF. The difficulty parameter would be the same, and the discrimination parameter would just have the opposite sign, leading to an almost identical item information function.

In short, we argue that the reliance on dominance assumptions and associated scale construction procedures has unnecessarily restricted and perhaps impeded the advancement of technology in personality assessment. In our view, psychometric models that do not impose monotonicity constraints on item response functions are more flexible and can surmount some of these shortcomings. Attractive alternatives include unfolding models capable of scaling both monotonic and nonmonotonic items. These models are based on the notion of an *ideal point* response process and have recently been shown to provide as good or better fit to personality data than do many commonly used dominance models.

Modeling Responses to Personality Items Using an Ideal Point Process

The ideal point process represents a viable alternative to the dominance process. The basic idea can be traced to Thurstone (1928), who provided examples in the context of attitude measurement (the term was proposed formally by Coombs, 1964). Thurstone suggested that a person responds negatively to an item when the attitude or behavior described by the item does not closely reflect the attitude or behavior of the respondent. Such disagreement occurs when the person is located too far above or too far below the item on the trait continuum (the former being contrary to the dominance assumption). On the other hand, people respond positively to items that have locations similar to their own. In probabilistic terms, as the distance between a respondent's location on the trait continuum (called his or her ideal point) and the item's location increases, the probability of endorsing the item decreases. This process implies a single-peaked response function, an example of which is shown in Figure 2A.

Recently, Stark et al. (2006) suggested that, theoretically, an ideal point process better characterizes responses to personality items than does a dominance process, and they tested this idea by comparing the fit of two IRT ideal point models (the generalized graded unfolding model [GGUM; J. S. Roberts, Donoghue, & Laughlin, 2000] and Levine's maximum likelihood formula scoring model [MFSM; Levine, 1984] with ideal point constraints) and two IRT dominance models (2PLM [Birnbaum, 1968] and MFSM with dominance constraints) with dichotomized 16PF data. (A detailed description of these models can be found in Chernyshenko et al.'s, 2001, article.) The fit of each model was examined using graphical and statistical methods. In short, the results indicated that ideal point models provided reasonably good fit to the 16PF data, and the fit tended to improve as the number of items exhibiting nonmononicity increased. The superiority of the ideal point models increased. The authors also stated that, for many personality variables, it is possible to envisage items that measure well in the middle of the trait continuum. Such items would be endorsed primarily by respondents having moderate trait levels, with persons at either extreme tending not to endorse. These findings were recently echoed by Meijer and Baneke (2004), who, using MMPI data, found that dominance constraints were too restrictive. Together, these studies raise interesting questions as to how people actually answer personality items. If indeed a comparison or matching process is involved, wherein respondents endorse only items that are most descriptive of themselves, then ideal point models are theoretically and empirically appropriate. Constructing scales under ideal point assumptions would therefore allow the inclusion of items having a wider range of locations rather than just those tending toward extremes. This, in turn, would improve scale precision, reduce inventory development costs, and offer a relatively straightforward path toward computerized adaptive tests involving forced-choice formats. Importantly, as is shown below, ideal point scale construction and scoring procedures do not require reverse scoring of negatively worded items, and the item location parameters, unlike in a dominance context, do have a direct link to item content. Negative items have negative locations, neutrally worded items have locations near the middle of the trait continuum, and positive items have positive location parameters.

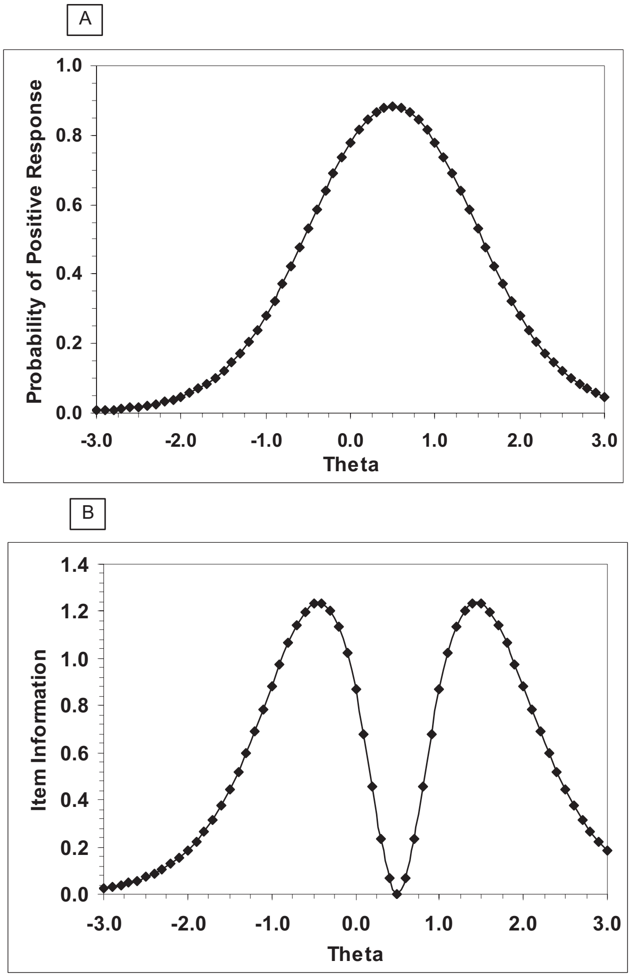

Figure 2. Panel A: Example of ideal point item response function. Panel B: Example of the item information function for the same item.

Scale Construction Using Ideal Point Assumptions

The assumption of an ideal point response process for personality items entails a fundamentally different approach to scale construction and scoring. Scales constructed via an ideal point approach tend to exhibit low item–total correlations, low internal consistency reliability, and a semicircular two-factor structure (Davison, 1977). This occurs because (a) items are intentionally written to reflect the entire range of the trait continuum, including positive, negative, and neutral statements; and (b) responses to negative items are not reverse scored. Importantly, total score, obtained by summing across items, has no meaning in an ideal point framework. One way to calculate a respondent's score is by taking the average (or median) of the location values for the statements he or she endorsed, a procedure first suggested by Thurstone (1928). A better way, however, is to use the general IRT scoring approach in which the presenting behaviors (item responses and their parameters) are used to determine what trait level is most likely. Regardless of the model, the likelihood of an examinee's response pattern can be computed as the product of individual item response probabilities (this stems from the assumption that item responses are locally independent). If one computes the likelihood over a wide range of trait levels, one can thus identify the trait level (theta) where the likelihood is maximized. The result becomes the respondent's score, known formally as the maximum likelihood estimate of theta.[2]

To date, no studies have attempted to design personality instruments under ideal point assumptions. One reason is that there are no simple CTT tools for evaluating the quality of items that reflect neutral or moderate trait levels. Second, Thurstone's approach to scaling items and persons was cumbersome; it required two separate data collections—one to determine item locations and another to score respondents. Because this process was far less attractive than the simple one-step approach proposed by Likert, it eventually fell into obscurity. Today, however, these limitations can be surmounted by using IRT methods, wherein both items and persons can be scaled using a single set of responses; a separate sample of judges is no longer needed. In addition, IRT parameters often have a relatively straightforward interpretation, which allows the evaluation of item quality regardless of extremity. Moreover, because IRT item parameters are subpopulation invariant and person parameters do not depend on the particular set of items used (Hambleton, Swaminathan, & Rogers, 1991), test developers can enjoy greater flexibility in scale construction. First, IRT makes it possible to evaluate the characteristics of items independently of each other so that changes in measurement precision can be examined as items are added and deleted during scale construction. Second, IRT allows test constructors to develop large item banks over time by using linking procedures (e.g., Stocking & Lord, 1983); large item banks are essential for computerized adaptive testing. Finally, "sample-free" item parameters allow for tests of differential functioning at both the item and scale levels, which is important in cross-cultural (e.g., Candell & Hulin, 1986; Ellis et al., 1993) and selection (Stark, Chernyshenko, & Drasgow, 2004) contexts.

In general, the task of using IRT methods to create a personality scale involves three basic steps: (a) selecting a model for estimating the parameters of items in the pool, (b) examining model–data fit and eliminating poorly fit items from further consideration, and (c) selecting a subset of items for the final measure, which provides high measurement precision at desired trait levels. However, in the context of ideal point scale development, somewhat different decision rules are used. Because these steps have not been explicated previously in the personality domain, we present them in detail below.

Selecting an IRT Model and Estimating Item Parameters

The last 15 years have seen increased research on the development of parametric probabilistic ideal point models. Several unidimensional IRT models are now available to analyze both dichotomous and polytomous response data. Ideal point IRT models for binary responses include the DeSarbo-Hoffman model (DeSarbo & Hoffman, 1986), the simple single squared logistic model (SSLM; Andrich, 1988), the parallelogram analysis model (PARELLA; Hoijtink, 1991), and the hyperbolic cosine model (HCM; Andrich & Luo, 1993). For polytomous data (graded responses), there is the general hyperbolic cosine model (GHCM; Andrich, 1996), the graded unfolding model (GUM; J. S. Roberts & Laughlin, 1996), and the GGUM (J. S. Roberts et al., 2000). Of the models we have examined, the GGUM appears to be the most suitable for personality scale construction. First, the GGUM can be used with dichotomous and polytomous responses. Second, it does not assume that all items are equally discriminating, and it does not require that all have the same number of response categories. Third, the software for GGUM parameter estimation has been examined in several parameter recovery studies and has been found to perform reasonably well under a variety of conditions (e.g., de la Torre, Stark, & Chernyshenko, 2006; J. S. Roberts, 2001; J. S. Roberts, Donoghue, & Laughlin, 1998, 1999).

In this research, we used the GGUM to model binary (agree/disagree) responses to personality items (readers interested in applying GGUM to polytomous data should refer to J. S. Roberts et al., 2000). For this special case, the model may be written as follows:

$$P[U_i = 1 \mid \theta_j] = \frac{\exp(\alpha_i[(\theta_j - \delta_i) - \tau_{i1}]) + \exp(\alpha_i[2(\theta_j - \delta_i) - \tau_{i1}])}{1 + \exp(\alpha_i[3(\theta_j - \delta_i)]) + \exp(\alpha_i[(\theta_j - \delta_i) - \tau_{i1}]) + \exp(\alpha_i[2(\theta_j - \delta_i) - \tau_{i1}])_i}, \qquad (1)$$

where θ_j is the location of respondent j on the continuum underlying responses, α_i is the discrimination parameter for item i, δ_i is the location of item i on the continuum underlying responses, and τ_{i1} is the location of the subjective response category threshold on the latent continuum.

To better understand this equation, it is useful to examine the IRF presented in Figure 2A, which was obtained by computing the probability of agreeing with the item at discrete values of theta on the interval [–3, 3], using GGUM parameters, $\alpha_i = 2$, $\delta_i = .5$, and $\tau_{i1} = -1$. Note that the function is single peaked and bell shaped and has its maximum at $\delta_i = .5$, which tells us that the item is located close to the middle of the trait continuum. Note also that the probability of item endorsement (a positive response) decreases as the distance between the location of a respondent (theta) and the location of the item (delta) increases in either direction from .5.

[2] IRT-based trait estimation procedures are the same for dominance and ideal point models, with the obvious exception of reverse scoring in the dominance case. IRT dominance models produce latent trait estimates that are nearly monotonically related to the total (summative or "number correct") score (see Hulin, Drasgow, & Parsons, 1983, p. 95), but this is not true for ideal point models because summative scoring is inappropriate.

Figure 2B presents the IIF for the same item. Recall that information is related to the slope of the IRF and indicates the trait regions where an item has the lowest standard error. Because ideal point IRFs are bell shaped, and information is zero when the location of a respondent (theta) equals the item location (delta; see Andrich, 1996, for an explanation), the resulting IIFs are bimodal. In this example, item information is highest on the intervals $[-1.5, 0.5]$ and $[1, 2.5]$. Hence, this relatively neutral item (in terms of location) helps to measure well individuals who are located both above and below the midpoint of the trait continuum. Intuitively, it is easy to see why neutral items like this one are potentially very useful. Most individuals are located in the middle of the trait continuum and would be expected to show positive responses to items such as “My social skills are about average.” Only those with poorly developed social skills (very low on extraversion) and those with well developed social skills (very high on extraversion) would be expected to disagree. Hence, in an ideal point framework, a negative response immediately indicates that a person is either very high or very low on the attribute measured. In contrast, a negative response in a dominance framework suggests that an individual has a lower standing on the latent attribute than is reflected by an item’s location, but how much lower cannot be readily discerned. To do so, one must administer additional, extreme items, which may be difficult to calibrate due to low endorsement rates.

Item parameters for the GGUM ($\alpha_{i,}$ δ_i, τ_{i1}) can be estimated using the marginal maximum likelihood (MML) or Markov chain Monte Carlo (MCMC) approaches (for detailed descriptions, see Bock & Aitkin, 1981, and Johnson & Junker, 2003). The MML procedure is implemented in the GGUM2000 computer program by J. S. Roberts (2001); the MCMC approach is implemented in the GGUM_MCMC program by de la Torre (2004). In this investigation, the GGUM2000 program was used because it is readily available to readers (http://www.education.umd.edu/edms/tutorials/freesoftware.html).

Examining Model–Data Fit

The model–data fit issue in IRT should be addressed in two ways. First, the model assumptions must be consistent with the dimensionality and process of item responding. Second, predictions based on the estimated model should be compared with observed responses using statistical and graphical tests of goodness of fit.

Numerous procedures have been developed to assess unidimensionality (Hattie, 1984, 1985). Examples include modified parallel analysis (Drasgow & Lissak, 1983), dimensionality testing (DIMTEST; Stout, 1987), and a confirmatory factor-analytic method that fits a one-factor model to response data for one scale at a time. However, because all of these procedures assume that a dominance process underlies item responding, they cannot be applied to scales constructed from an ideal point perspective. Thus, only a limited number of strategies are currently available for testing the assumption of unidimensionality in this context.

One such strategy was discussed by Davison (1977), who showed that responses consistent with a unidimensional ideal point (unfolding) model generally display two major principal components and that the component loadings will show a simplex pattern. In connection, J. S. Roberts et al. (2000) suggested that an item can be considered unidimensional if its communality based on the first two principal components is greater than or equal to .3. A limitation of this approach is that we could not find any simulation studies that tested the accuracy of this rule. In our view, more research should be conducted before factor analysis is used to investigate the unidimensionality of ideal point measures. A recent article by Habing, Finch, and Roberts (2005) explores this issue in more detail.

An alternative to factor-analytic tests of unidimensionality is the examination of model–data chi-square fit statistics for item singles, pairs, and triplets. The idea is that if a unidimensional IRT model can fit patterns of responses, then a single latent trait is sufficient to account for item responding (Van den Wollenberg, 1982, and Glas, 1988, showed that fit statistics computed from pairs of items are sensitive to violations of unidimensionality and local independence). If fit statistics are small, then the proposed unidimensional model provides good fit, and there is little reason to suspect that the data are multidimensional. Consequently, in this article, chi-square fit statistics for item singles, pairs, and triplets were used to investigate the fit of the unidimensional GGUM to items measuring the construct of order.

The chi-square statistic for individual dichotomous items is given by

$$\chi_i^2 = \sum_{u=0}^{1} \frac{[O_i(u) - E_i(u)]^2}{E_i(u)}, \tag{2}$$

where i is the item number, $O_i(u)$ is the observed frequency of endorsing option u, and $E_i(u)$ is the expected frequency of option u under a particular IRT model. The expected frequency of respondents selecting option u is given by

$$E_i(u) = N\int P(U_i = u|\theta)^*f(\theta)d\theta, \tag{3}$$

where $f(\theta)$ is a density function, usually a normal, having mean and variance that corresponds to the distribution of person parameters. The expected frequency for a chi-square statistic involving a pair of items, for items i and i', in the $(u,u')^{\text{th}}$ cell of a two-way contingency table, is computed as

$$E_{i,i'}(u,u') = N\int P(U_i = u \mid \theta)P(U_{i'} = u' \mid \theta)^*f(\theta)d\theta. \tag{4}$$

The chi-square statistic for a pair of items can be computed by the usual formula for a two-way table. A similar procedure can be used to calculate chi-square statistics for triplets of items.

To facilitate comparisons of chi-squares based on different sample sizes and models containing different numbers of parameters, we adjusted the chi-squares to a constant sample size of 3,000 and divided them by their degrees of freedom. Previous studies have found that good model–data fit is associated with adjusted χ^2/df of 3 or less. This methodology is implemented in the MODFIT computer program (Stark, 2001).

Selecting Items On the Basis of Their Measurement Precision

The goal of any test construction is to select a subset of items that collectively provides the highest measurement precision at the desired trait levels. Here, we operate under the assumption that most personality scales were intended to measure respondents at all trait levels equally well, so scale developers likely strived to achieve high precision across a broad range of the trait continuum. In IRT, this qualitative description is expressed explicitly by specifying the shape of the test information function (TIF), which is a sum of individual IIFs. Because each item contributes additively to test information, the decision to delete or include an item during scale construction is dictated primarily by its contribution to test information.

To construct a test with a relatively high but flat TIF using an ideal point IRT approach, one should select items with location parameters that are spread evenly across the trait continuum. Neutral items (those with δ_i parameters close to zero) generally help to measure respondents who are above and below average, whereas positive and negative items provide high information in the middle and at extremes. In addition, items with higher discrimination parameters and/or relatively large threshold values should be preferred because they have steeper IRFs and yield more information. These guidelines were used here to construct an ideal point scale representing order.

Method

Item Pool Development

Two subject-matter experts wrote 66 items for the order facet of conscientiousness. Items were in the form of short-sentence statements that described specific behaviors believed to be associated with the facet. Behavioral domains relevant to order were identified by examining the relevant psychological literature as well as the content of items from scales found by B. Roberts et al. (2005) to have high loadings on that facet.

Order is widely considered to be one of the core lower-order traits of conscientiousness. According to the research literature, people who score high on order tend to describe themselves as organized, meticulous, neat, and punctual. Examples of scales measuring this facet are Orderliness and Perfectionism from the Abridged Big Five Dimensional Circumplex Inventory of the International Personality Item Pool (AB5C-IPIP; Hofstee, de Raad, & Goldberg, 1992), the Order scale from the NEO Personality Inventory—Revised (NEO-PI–R; Costa, McCrae, & Dye, 1991), the Perfectionism scale from the 16PF, and the Organization scale from the Jackson Personality Inventory—Revised (JPI–R; Jackson, 1994).

For this study, items were created to represent the full range of behaviors (negative, positive, and moderate/neutral) associated with order. This was done deliberately because, as was indicated previously, personality test developers traditionally tend to avoid items neutral in content, thus not utilizing fully the universe of available items. We did not know for sure, however, whether dominance models would be inappropriate for such items. Previous researchers (Meijer & Baneke, 2004; Stark et al., 2006) have suggested that possibility, but showing it empirically was crucial. Hence, items were generated in a manner that allowed us to test the main prediction of the study.

The content of the initial item pool was evaluated, and the best items, representing a diverse array of behaviors, were selected. To ensure that items from every region of the trait continuum were present in the pool, two raters were asked to rate on a 7-point scale the content of items in terms of their extremity/location. Items judged to have the content representing the most negative pole of the trait continuum received a rating of "1," items judged to be at the most positive pole received a "7," and items located in the middle of the trait continuum received a rating of "4." For example, the item "I am incapable of planning ahead" was rated 1 by both judges, which meant it was perceived as extremely negative. On the other hand, "I keep detailed notes of important meetings and lectures" received ratings of 6 and 7, respectively, indicating that it is highly positive. Finally, the item "My room neatness is about average" received two ratings of 4, signifying its neutral location. Items showing a marked lack of correspondence between judges' ratings were either rewritten or discarded.

In the final stage of item pool development, 50 items, roughly 7 from each region of the trait continuum, were selected for pretesting. The resulting pool is shown in Table 1. In the table, column 2 presents the item content, and columns 3 and 4 illustrate the perceived location ratings of the two judges. The correlation between the location ratings for the 50 items was .91, indicating a high level of agreement between judges.

After the final pool of items for scale construction was obtained, statements were administered to a large sample of examinees who were asked to indicate their level of agreement using a 4-point response format (1 = *strongly disagree*, 2 = *disagre*e, 3 = *agree*, 4 = *strongly agree*). This format was chosen so that the items could be scored polytomously, at least for the CTT part of this study. Polytomous items violate the normality assumption of confirmatory factor analysis to a lesser degree than do dichotomous items and allow for more accurate item–total correlation estimates. In an IRT context, polytomously scored items are also potentially useful because they provide more psychometric information than do dichotomously scored items, but unfortunately, they require a much larger sample for parameter estimation than was available here. Consequently, for IRT scale development in this investigation, we chose to work with dichotomized responses. Specifically, responses of "strongly disagree" and "disagree" were collapsed and scored as 0, and responses of "strongly agree" and "agree" were recoded as 1. In any case, dichotomization would have been necessary for many extreme items because of infrequent endorsement of the strongly agree/disagree categories.[3]

Participants

Participants were 539 undergraduate students at a large midwestern university in the United States. Four hundred seventy-four participants were recruited from the subject pools of the Psychol-

[3] The decision to dichotomize the data has no effect on the issue of whether an ideal point or dominance response process is appropriate for personality items. Items with neutral content cannot be scaled with either dichotomous or polytomous dominance IRT models if in fact an ideal point process applies. That is because respondents with either very low or very high standings on the order dimension would be expected to select "strongly disagree" or "disagree" response options, whereas those with average standings would be expected to select "strongly agree" or "agree." As a result, near-zero item–total correlations and low dominance IRT item discrimination parameters would be observed for neutral items regardless of whether the item was polytomously or dichotomously scored.

Table 1
Content, Judged Locations, and Item–Total Correlation for the 50 Order Items

Item name	Item content	Item location		Corrected ITC
		Rater 1	Rater 2	
ORD1[a]	I spend a lot of time looking for objects I misplaced.	1	1	.32
ORD2[a]	Usually, my notes are so jumbled, even I have a hard time reading them.	1	1	.37
ORD3[a]	I hate routine or scheduled activities.	1	1	.31
ORD4[a]	I find myself unprepared in most situations.	1	1	.37
ORD5[a]	I am incapable of planning ahead.	1	1	.37
ORD6[a]	Most of the time my room is in complete disarray.	2	1	.60
ORD7[a]	I do not invite people to my home because it is too messy.	1	2	.31
ORD8[a]	I feel comfortable even in very disorganized settings.	3	2	.48
ORD9[a]	Taking care of every detail is a waste of time and effort.	3	2	.37
ORD10[a]	I frequently forget to put things back in their proper place.	2	3	.62
ORD11[a]	I prefer keeping my options open and rarely plan in advance.	3	3	.44
ORD12[a]	I seldom make detailed “to do” lists.	2	3	.47
ORD13[a]	Routines are boring and not for me.	3	3	.34
ORD14[a]	I do not like work spaces that are too clean and tidy.	2	3	.53
ORD15[a]	For me, being organized is unimportant.	2	3	.66
ORD16[a]	It is difficult for me to design a well thought-out plan.	2	3	.21
ORD17[a]	Being neat is not exactly my strength.	3	3	.68
ORD18[a]	I seldom look back to check if I did everything right.	3	3	.24
ORD19[a]	When busy, I spend little time cleaning and organizing things.	3	3	.41
ORD20	I do pretty standard maintenance for my property and possessions.	4	4	.24
ORD21[a]	Although I leave things laying around, I generally remember where most of them are.	3	4	.28
ORD22[a]	Most jobs do not require much planning.	2	4	.28
ORD23[a]	My room neatness is about average.	4	4	−.09
ORD24[a]	Half of the time I do not put things in their proper place.	3	4	.65
ORD25[a]	As long as I have a little bit of clear space on my desk, I am happy to do my work.	3	4	.51
ORD26[a]	My ability to plan is at about average.	4	4	.39
ORD27[a]	My routines are not set in stone. I deviate from them when needed.	4	4	.28
ORD28	I clean my desk about once every two weeks.	5	4	.00
ORD29[a]	Although I try to keep everything in its place, it does not always work for me.	5	4	.33
ORD30[a]	Although I have a daily organizer, I have a hard time keeping it up to date.	3	4	.41
ORD31[a]	Although I pay most of my bills on time, I can occasionally miss a deadline or two.	4	5	.27
ORD32	I try to balance my checkbook at the end of each month.	3	5	.36
ORD33	If I have time, I double check my exam answers before turning them in.	7	6	.17
ORD34	I have a daily routine and stick to it.	7	6	.51
ORD35	I need a neat environment in order to work well.	6	6	.58
ORD36	I dislike doing things without proper planning.	5	6	.40
ORD37	I prefer to do things in a logical order.	5	6	.38
ORD38	Organization is a key component of most things I do.	6	6	.71
ORD39	I rarely deviate from my morning routines.	6	6	.33
ORD40	I am very good at pacing myself to get the job done on time.	4	6	.41
ORD41	I avoid errors by being careful and thorough.	6	6	.39
ORD42	I write notes to myself only if I have too many things to do at once.	5	6	.05
ORD43	I hardly ever lose or misplace things.	6	7	.48
ORD44	I become annoyed when things around me are disorganized.	6	7	.61
ORD45	I usually have a backup plan, in case something goes wrong with the current one.	6	7	.24
ORD46	I keep detailed notes of important meetings and lectures.	6	7	.48
ORD47	I want everything to be exactly the way I plan it.	7	7	.35
ORD48	I hate when people are sloppy.	7	7	.52
ORD49	I try to anticipate problems by planning responses to every possible outcome.	6	7	.17
ORD50	Every item in my room and on my desk has its own designated place.	7	7	.57

Note. $N = 477$. ITC = item–total correlation.
[a] Item was reverse scored.

ogy and Educational Psychology departments. The remaining participants were classroom volunteers from introductory personality and industrial organizational psychology courses.

Demographic information was available for 430 participants. Among those for which demographic data were available, females represented 69% of the sample. Most participants were White (70%), followed by Asians (8%), African Americans (7%), and Hispanics (5%). The mean age of participants was 19.74, because most participants were either college freshmen or sophomores. Eighty-eight percent of the participants had previous work experience, and 48% were employed at the time of the study. Whereas it is true that college samples likely show higher means on order than does the general U.S. population, it is important to note that IRT item parameters are subpopulation invariant (estimates from different samples can be linked by a simple linear transformation; e.g., Embretson & Reise, 2000).

Scale Development

To directly evaluate the advantages of using ideal point methods for personality scale construction and scoring, we created three kinds of order scales from the same item pool. One scale was constructed under ideal point assumptions and is referred to here as the "Ideal Point IRT" scale. The other two were designed under dominance assumptions and are referred to as the "Traditional CTT" and "Dominance IRT" scales. Details of each development process are described below.

Constructing the "Traditional CTT" order scale. Essentially, the standard Likert approach to personality scale construction was followed. First, all negatively worded items in the pool were reverse scored so that only positive item–total correlations were observed. Because the scoring direction of neutral items was often too difficult to determine on the basis of content alone, the correlation between each neutral item and a positively worded subset was examined; if this correlation was negative, then the item was reverse scored. Next, CTT item statistics were computed, and the items were ranked from highest to lowest on the basis of their item–total correlations (Nunnally & Bernstein, 1994). The 20 items having the highest values were selected as candidates for the Traditional CTT order scale, and then some item swapping was done to obtain normally distributed total scores; item difficulty values (i.e., means) were used to make decisions about replacement. In addition, to minimize the influence of response sets (see Jackson, 1994), content balancing was performed by retaining 10 positively worded items and 10 negatively worded items. Responses for the final set of 20 items were then subjected to principal axis factoring to ensure that the resulting scale was unidimensional.

Constructing the "Dominance IRT" order scale. The 2PLM was used as the basis for scale construction. For the 2PLM, the probability of agreeing with item i is written as

$$P(u_i = 1 \mid \theta = t) = \frac{1}{1 + \exp[-1.7a_i(t - b_i)]} \qquad (5)$$

where a_i is the item discrimination parameter and b_i is the item location parameter. A representative 2PLM IRF is presented in Figure 1. It can be seen that individuals with very low trait levels are very unlikely to endorse the item. However, individuals with trait levels greater than the item location parameter ($b_i = .5$) have a greater probability of endorsement, which increases monotonically as the distance between the person and the item increases. (This is consistent with dominance assumptions.) Note also that the item location parameters (b_i) associated with dominance IRT models have a different meaning than do location parameters (δ_i) found in ideal point IRT models. Ideal point locations point out the specific trait level where a positive response is most likely, whereas dominance locations indicate where (in the case of 2PLM) there is a 50% chance of observing such a response. Consequently, the item shown in Figure 1 would not be located in the middle of the trait continuum in an ideal point framework. Instead, extrapolating the bell-shaped curve beyond the normal trait range suggests that a positive response is most likely at around $\theta = 3.5$.

To use the 2PLM, polytomous item scores were dichotomized. Negatively worded items were reverse scored, and 2PLM item parameters were estimated using the BILOG computer program (Mislevy & Bock, 1991). Chi-square fit statistics were then calculated using Stark's (2001) MODFIT computer program, and items showing poor fit were discarded. The remaining items were rank ordered by their contribution to test information (i.e., items with high discrimination parameters received higher ranks), and 20 items were selected in an effort to produce a measure that was informative across a wide range of trait levels. Chi-square fit statistics for pairs and triplets of items in that subset of 20 were then examined to identify possible violations of local independence, and item switching was done as needed to choose 10 negatively worded items and 10 positively worded items for the final measure. Readers interested in learning more about personality scale development using dominance IRT models should refer to Waller, Tellegen, McDonald, and Lykken (1996).

Constructing the "Ideal Point IRT" order scale. GGUM parameters for the initial set of 50 dichotomous items were estimated using the GGUM2000 computer program (note that reverse scoring is not necessary with ideal point procedures), and chi-square fit statistics were computed as before, using MODFIT. Poorly fit items were removed, and the 20 most discriminating items at negative, neutral, and positive regions of the trait continuum were retained. Again, care was taken to select items that measured well across all trait levels with no obvious violations of local independence. Small chi-square statistics for item pairs and triplets were interpreted as evidence of unidimensionality.

Scale Validation

Because a new type of personality scale was constructed (i.e., ideal point), it was important to show that its scores are comparable to those produced by traditional methods. Here, each scale construction method only used items that were appropriate for that method (i.e., neutral items having low item–total correlations would not be used with CTT procedures) and thus should have yielded rank orderings similar to those of individuals in the total sample.

Because participants answered all 50 items in the initial pool, it was possible to calculate three trait scores for each person on the basis of the respective sets of items retained. For the Traditional CTT scale, trait estimates were calculated by summing item scores. For the Dominance IRT and Ideal Point IRT scales, trait estimates were computed using expected a posteriori (EAP) estimation (see Thissen & Wainer, 2001, for details). The resulting three order scores were then correlated with a widely used Big Five measure to examine convergent and discriminant validity. In addition, the order scores were correlated with scales from the Study Behavior Questionnaire (SBQ; B. Roberts, 2001) and the Health Behavior Checklist (HBCL; Vickers, Conway, & Hervig, 1990) to show that the new scales performed equally well in predicting behavioral criteria.

Big Five measures. The Big Five dimensions of personality were assessed using the 35-item version of Goldberg's (1992) Big Five Adjective Markers. The measure was composed of bipolar adjectives (seven per Big Five scale) presented using a 9-point format as suggested by Goldberg (1992). Data for this 35-item scale were available for 185 of the 539 respondents. The reliability of the Big Five scales ranged from .78 to .91. It was expected that the order scales would correlate higher with the global Conscientiousness scale than with any of the other Big Five scales.

Health-related behaviors. Participants' health-related behaviors were assessed using the shortened version of the HBCL. The shortened HBCL is a 35-item measure that consists of four subscales: Preventative Health Behaviors, Accident Control Behaviors, Traffic Risk, and Substance Risk. For this study, the Preventative Health Behaviors subscale contained 16 items such as "I exercise to stay healthy" and "I see a doctor for regular checkups." The Accident Control Behaviors subscale had 7 items such as "I have a first aid kit in my home" and "I fix broken things around the home right away." The Traffic Risk subscale consisted of 8 items such as "I speed while driving" and "I cross busy streets in the middle of the block." Items were scored so that higher values indicated higher risk taking. Finally, the Substance Risk subscale contained 4 items such as "I do not drink alcohol" and "I don't smoke." Unlike the Traffic Risk subscale, these items were scored such that higher values indicated substance avoidance. All HBCL items were administered using a 5-point format ranging from 1 (*very uncharacteristic of me*) to 5 (*very characteristic of me*). Reliability estimates across the four subscales varied from .65 to .72.

Previous research has clearly shown that conscientiousness is one of the strongest dispositional predictors of health behaviors. Conscientiousness measures have been linked to lower tobacco and alcohol consumption (Shedler & Block, 1990; Sher & Trull, 1994; Watson & Clark, 1993), less risky sexual and driving behavior (Caspi et al., 1997; Clark & Watson, 1999; Martin & Boomsma, 1989; White & Johnson, 1988), and fewer problems with obesity (Chalmers, Bowyer, & Olenick, 1990). Consequently, it was expected that the new order scales would be positively related to the HBCL Preventative Health Behaviors, Accident Control Behaviors, and Substance Risk Behavior subscales but negatively related to the Risk Taking Behavior subscale.

SBQ

The SBQ (B. Roberts, 2001) contains 12 items that ask respondents how often they attend classes, take notes, or turn in assignments late. A 5-point response format is used, where 1 = *never*, 2 = *seldom*, 3 = *sometimes*, 4 = *often*, and 5 = *always*. The reliability of the SBQ in the present sample (N = 180) was .80. It was expected that all three order scales would have positive relationships with SBQ scores.

Results

Traditional CTT Order Scale

Table 1 (column 5) presents item–total correlations for the pool of 50 order items after reverse scoring. Thirteen of 50 items had correlations lower than .3, which is the cutoff suggested by Nunnally & Bernstein (1994) for item retention. Interestingly, the majority of items with low correlations were those judged by the independent raters in the pool development phase of this study to be in the middle of the trait continuum (see Table 1, columns 3 and 4). For example, the item "Although I pay most of my bills on time, I can occasionally miss a deadline or two" was rated 4 and 5, respectively, by the two raters and had an item–total correlation of .27. Even more dramatic was the result observed for the item "My room neatness is about average." It was rated 4 by both judges and had an item–total correlation of −.09. Overall, these findings support the earlier assertion that, if an ideal point response process applies, items located in the middle of the trait continuum would have low item–total correlations and tend to be eliminated during scale construction.

A set of 10 positively worded items and 10 negatively worded items with the highest item–total correlations was then selected from the remaining pool of 37, and principal axis factor analysis was used to examine dimensionality. Table 2 presents CTT item statistics (means, standard deviations, and item–total correlations) as well as factor loadings for the resulting 20-item subset. It can be seen that the item–total correlations were reasonably high, ranging from .34 to .71, and as expected, slightly higher values were observed for the factor loadings because factor scores do not contain measurement error. Overall, the results indicate that a single dominant factor was present (the ratio of the first to second eigenvalue was $>$ 5.0) and that the total score distribution was very close to normal (see Figure 3). The scale mean was 55.6, with a standard deviation of 9.2 and an internal consistency reliability of .91.

Dominance IRT Order Scale

For the IRT analyses, data from the 50 reverse-scored order items were dichotomized, and 2PLM item parameters were estimated, using BILOG. Item ORD23 ("My room neatness is about average") had to be deleted for BILOG to converge. Chi-square to degrees-of-freedom ratio fit statistics for individual items were then computed to screen out those that were not fit well by 2PLM. As a result, Item ORD4 ("I find myself unprepared in most situations") was eliminated from further analyses because its adjusted χ^2/df statistic was equal to 9.96, which is considerably larger than the recommended value of 3 (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Of the remaining items, 12 had low discrimination parameters ($a < .40$) and were deemed to be of "poor quality" for scale development purposes. Items with low discrimination parameters provide little information and thus contribute little to reducing the error in trait estimates. Altogether, 14 of the 50 items in the initial pool had to be screened out prior to the Dominance IRT scale assembly.

Next, fit statistics for item pairs and triplets were examined. This analysis revealed several combinations of items that seemed to violate the local independence assumption. As expected, the problematic combinations involved items similar in content. For example, Items ORD10 ("I frequently forget to put things back in their proper place") and ORD43 ("I hardly ever lose or misplace things") had an adjusted χ^2/df of 27.5. Similarly, Items ORD34 ("I have a daily routine and stick to it") and ORD39 ("I rarely deviate from my morning routines") had an adjusted χ^2/df of 110.1 (!). Hence, despite the fact that these items were adequately discriminating, only one item from each pair could be included in the final measure (see Embretson and Reise, 2000, for a discussion of local dependence and its consequences). With this in mind, a 20-item Dominance IRT scale was created. The resulting fit statistics for item pairs and triplets were 1.26 and 2.16, respectively, and no pair had an adjusted χ^2/df larger than 15.

Table 2
CTT Item Statistics and Factor Loadings for the Dominance CTT Order Scale

Item name	Item content	*M*	*SD*	Corrected ITC	Factor loading
ORD6[a]	Most of the time my room is in complete disarray.	3.16	.76	.66	.71
ORD8[a]	I feel comfortable even in very disorganized settings.	2.72	.73	.54	.57
ORD10[a]	I frequently forget to put things back in their proper place.	2.76	.72	.60	.63
ORD11[a]	I prefer keeping my options open and rarely plan in advance.	2.59	.76	.40	.39
ORD12[a]	I seldom make detailed "to do" lists.	2.96	.90	.40	.41
ORD14[a]	I do not like work spaces that are too clean and tidy.	3.17	.68	.54	.57
ORD15[a]	For me, being organized is unimportant.	3.13	.69	.67	.70
ORD17[a]	Being neat is not exactly my strength.	2.73	.90	.73	.78
ORD24[a]	Half of the time I do not put things in their proper place.	2.75	.78	.67	.71
ORD25[a]	As long as I have a little bit of clear space on my desk, I am happy to do my work.	2.52	.81	.59	.62
ORD34	I have a daily routine and stick to it.	2.46	.66	.45	.45
ORD35	I need a neat environment in order to work well.	2.73	.78	.67	.72
ORD36	I dislike doing things without proper planning.	2.61	.67	.35	.36
ORD37	I prefer to do things in a logical order.	3.08	.58	.34	.35
ORD38	Organization is a key component of most things I do.	2.85	.74	.69	.71
ORD43	I hardly ever lose or misplace things.	2.44	.79	.43	.46
ORD44	I become annoyed when things around me are disorganized.	2.88	.73	.67	.71
ORD46	I keep detailed notes of important meetings and lectures.	2.73	.73	.41	.42
ORD48	I hate when people are sloppy.	2.83	.75	.58	.62
ORD50	Every item in my room and on my desk has its own designated place.	2.52	.86	.62	.66

Note. $N = 501$. CTT = classical test theory; ITC = item–total correlation.
[a] Item was reverse scored.

Before presenting the content of the resulting Dominance IRT scale, a slight digression is in order. The problem of locally dependent items illustrates an important difference between CTT and IRT approaches to scale construction, which emerges primarily in noncognitive domains. CTT methods encourage the creation of scales having high internal consistency reliability, which is easiest to achieve by selecting items very similar in content (these will have very high interitem correlations). Whereas this poses no psychometric problem in the cognitive domain, repetition of content in noncognitive measures seems to induce response sets or memory effects that lead to violations of local independence, in the sense that answers to "repeated" items are influenced not only by one's standing on the latent trait measured by a scale (e.g., Order) but also by the recollection of answers to previous items. (In factor-analytic terminology, such violations of local independence would manifest in items having common specific variance.) IRT is able to identify such dependencies and either remove or combine offending items prior to scoring (e.g., Orlando & Thissen, 2000; Thissen, Steinberg, & Mooney, 1989). This is the main reason why the CTT and IRT Dominance order scales developed here had only 13 items in common. If all items could have been used in scale construction (i.e., no dependencies found), the two methods would have produced nearly identical results, because items having high 2PLM discrimination parameters would also have high item–total correlations. Examination of item fit, however, allowed us to detect locally dependent items and thus constitutes an important advantage of IRT scale construction.

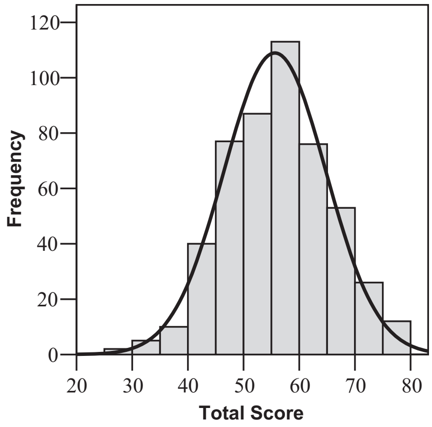

Figure 3. Total score distribution for the Traditional CTT order scale. CTT = classical test theory.

Table 3 presents item content and 2PLM parameters for the final 20-item Dominance IRT order scale. The average item discrimination parameter was .82, which is similar to what is often found in personality scales like the 16PF. Figure 4 shows IRFs for a negatively worded item (ORD8: "I feel comfortable even in very disorganized settings") and a positively worded item (ORD48: "I hate when people are sloppy"). Note that both items have monotonically increasing response functions, which are required under the 2PLM. Also, note that despite one item being positive and the other being negative, the two IRFs are nearly identical, because they have very similar item parameters. This peculiarity resulted from reverse scoring the negatively worded item, so its IRF is monotonically increasing and not decreasing (the IRF shown represents response probabilities associated with not endorsing the item rather than endorsing the item). However, this reverse scoring did not affect the location of the IRF because the trait location

Table 3
2PLM Item Parameters for the Dominance IRT Order Scale

		2PLM item parameter	
Item name	Item content	a	b
ORD6[a]	Most of the time my room is in complete disarray.	1.31	−1.22
ORD8[a]	I feel comfortable even in very disorganized settings.	0.82	−0.53
ORD15[a]	For me, being organized is unimportant.	1.04	−1.41
ORD17[a]	Being neat is not exactly my strength.	1.58	−0.34
ORD19[a]	When busy, I spend little time cleaning and organizing things.	0.70	0.53
ORD21[a]	Although I leave things laying around, I generally remember where most of them are.	0.54	1.14
ORD24[a]	Half of the time I do not put things in their proper place.	1.14	−0.39
ORD25[a]	As long as I have a little bit of clear space on my desk, I am happy to do my work.	0.92	0.06
ORD27[a]	My routines are not set in stone. I deviate from them when needed.	0.46	3.04
ORD29[a]	Although I try to keep everything in its place, it does not always work for me.	0.54	1.14
ORD32	I try to balance my checkbook at the end of each month.	0.51	0.05
ORD35	I need a neat environment in order to work well.	1.25	−0.44
ORD36	I dislike doing things without proper planning.	0.42	−0.30
ORD38	Organization is a key component of most things I do.	1.22	−0.66
ORD39	I rarely deviate from my morning routines.	0.43	−0.01
ORD41	I avoid errors by being careful and thorough.	0.49	−1.95
ORD43	I hardly ever lose or misplace things.	0.61	0.18
ORD46	I keep detailed notes of important meetings and lectures.	0.60	−0.74
ORD48	I hate when people are sloppy.	0.86	−0.67
ORD50	Every item in my room and on my desk has its own designated place.	1.07	−0.02

Note. $N = 539$. IRT = item response theory; 2PLM = two-parameter logistic model.
[a] Item was reverse scored.

where a person is expected to endorse the item with a .5 probability is, by definition, also the location where he or she is expected to not endorse that item with that probability. Also, because 2PLM IRFs for reverse-scored and non-reverse-scored items with similar location parameters are basically mirror images centered on item location, the shapes of the IIFs are very similar. The conclusion one should draw from this discussion is that 2PLM item location parameters (or CTT p values) are not related to the content of the items. In fact, the correlation between the average perceived item extremity/location (see location ratings in Table 1) and the 2PLM b parameters was −.07.

The dashed line in Figure 5 presents the 2PLM TIF; these functions are important because the standard error of a latent trait estimate equals $1/\sqrt{\text{Test Information}}$. As expected, the Dominance IRT scale measured well in the middle of the trait continuum but poorly at the extremes. Information is particularly low at high trait levels because of a lack of discriminating items with positive location parameters (i.e., with b parameters > 0). Tests with such TIFs are appropriate for personnel screening because they have low measurement error in the region [−1.0, +1.0], but they are probably less suitable for clinical and counseling purposes, where there is also interest in measuring at extremes.

Ideal Point IRT Order Scale

With ideal point scale development, the reverse scoring of negatively worded items is not necessary. Consequently, we analyzed the original dichotomous responses to the 50 order items using the GGUM2000 computer program. As before, adjusted chi-square to degrees-of-freedom ratio fit statistics, provided by MODFIT, were used to identify items that were not fit well by the GGUM. None were found. Importantly, because only 2 items in

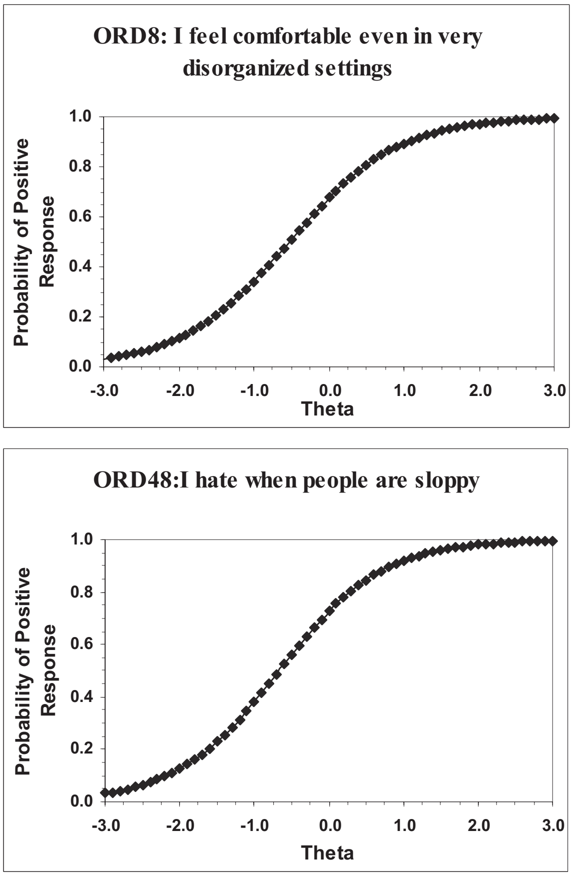

Figure 4. Top panel: Two-parameter logistic model item response functions for Item ORD8. Bottom panel: Two-parameter logistic model item response functions for Item ORD48.

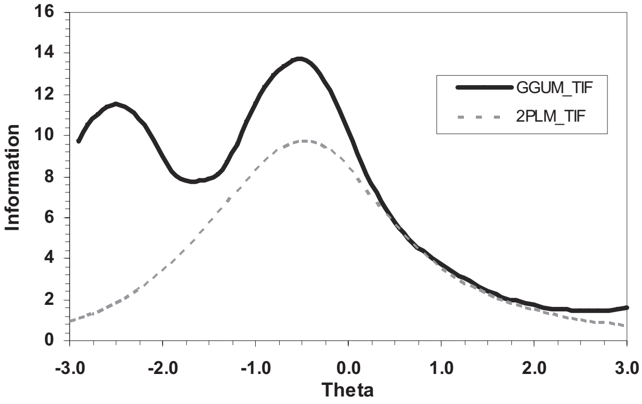

Figure 5. Comparison of test information functions for the Dominance IRT and Ideal Point IRT order scales. IRT = item response theory; GGUM = generalized graded unfolding model; TIF = test information function; 2PLM = two-parameter logistic model.

the initial pool had low (< .4) item discrimination parameters, 48 of 50 could be used for scale assembly (as compared with 36 with 2PLM). Interestingly, Item ORD23, which had to be deleted from the 2PLM analysis for BILOG to converge, was not only fit well by the GGUM but also had a neutral location parameter and a bell-shaped item response function. Overall, 20 of 48 items exhibited nonmonotonic IRFs over the range –3 to +3. Therefore, it is not surprising that many of these items were dropped from the pool during dominance scale construction.

Next, the 48 items remaining in the pool were ranked by their discrimination parameters and sorted in ascending order by location. Twenty highly discriminating items with location parameters distributed over a wide range of trait levels were selected to produce a scale with an approximately flat TIF. Fit statistics for item pairs and triplets were examined to identify possible violations of local independence. Items ORD44 ("I become annoyed when things around me are disorganized") and ORD48 ("I hate when people are sloppy") were flagged as dependent (adjusted χ^2/df for the pair was > 25); thus, the latter item was replaced with ORD37 ("I prefer to do things in a logical order"), resulting in improved fit.

Table 4 presents items and their GGUM parameters for the Ideal Point IRT order scale. In this set, 11 of 20 items exhibited nonmonotonicity. The most striking example was Item ORD23 ("My room neatness is about average"), whose IRF is presented in Figure 6. The GGUM parameters for this item were $\alpha = 1.46$, $\delta = .23$, and $\tau_1 = -1.34$. It is evident that the neutral item location (δ) positioned the peak of the IRF in the middle of the trait continuum. Respondents who were about average in terms of order tended to endorse the item, whereas those who were very disorganized or very organized tended not to endorse it. It is not surprising, therefore, that the item–total correlation for this item was near zero, and it was discarded from the pool in both instances of dominance scale construction. The remaining 9 of 20 items had relatively extreme location parameters, and folding (a downward turn in an IRF) would have been observed only at trait levels beyond ±3. Because very few respondents lie in those regions of the trait continuum, those items were also fit well by dominance models, and in fact, many appeared in the Dominance CTT and Dominance IRT scales.

A more detailed examination of the GGUM parameters in Table 4 reveals a fairly close correspondence between item content and the respective location parameters (δ). For example, Item ORD2 ("Usually, my notes are so jumbled, even I have a hard time reading them") had the second lowest GGUM location rating (−3.74) and was judged by both independent raters as extremely

Table 4
GGUM Item Parameters for the Ideal Point IRT Order Scale

		GGUM item parameter		
Item name	Item content	α	δ	τ
ORD2	Usually, my notes are so jumbled, even I have a hard time reading them.	0.82	−3.74	−1.26
ORD6	Most of the time my room is in complete disarray.	2.91	−1.89	−0.76
ORD10	I frequently forget to put things back in their proper place.	3.30	−1.30	−0.87
ORD14	I do not like work spaces that are too clean and tidy.	1.22	−5.24	−3.25
ORD15	For me, being organized is unimportant.	1.68	−2.29	−0.78
ORD17	Being neat is not exactly my strength.	2.72	−1.60	−1.28
ORD20	I do pretty standard maintenance for my property and possessions.	0.87	0.27	−3.24
ORD23	My room neatness is about average.	1.46	0.23	−1.34
ORD24	Half of the time I do not put things in their proper place.	2.81	−1.48	−1.14
ORD26	My ability to plan is at about average.	0.75	−1.06	−1.04
ORD29	Although I try to keep everything in its place, it does not always work for me.	1.85	−0.61	−1.39
ORD30	Although I have a daily organizer, I have a hard time keeping it up to date.	0.88	−1.56	−0.90
ORD34	I have a daily routine and stick to it.	0.84	2.82	−2.49
ORD35	I need a neat environment in order to work well.	1.81	1.57	−2.00
ORD37	I prefer to do things in a logical order.	0.87	3.46	−5.96
ORD38	Organization is a key component of most things I do.	2.15	2.26	−2.89
ORD42	I write notes to myself only if I have too many things to do at once.	0.54	−0.26	−2.00
ORD44	I become annoyed when things around me are disorganized.	1.94	2.69	−3.45
ORD46	I keep detailed notes of important meetings and lectures.	0.93	2.03	−2.68
ORD50	Every item in my room and on my desk has its own designated place.	1.76	2.62	−2.62

Note. GGUM = generalized graded unfolding model; IRT = item response theory.

negative (rating of “1”). Conversely, Item ORD50 (“Every item in my room and on my desk has its own designated place”) had one of the highest GGUM locations (2.62) as well as ratings of “7” by both judges. Items with GGUM locations in the middle of the trait continuum were also judged relatively neutral in their content. The correlation between the GGUM location parameters and the average of the rater judgments was .89, whereas the same correlation for 2PLM location parameters was −.07. Clearly, it is much easier to anticipate the sign and the extremity of location parameters in an ideal point context than in a dominance context, a finding which we believe many scale developers will find useful. At this point, the same cannot be said about item discrimination parameters. Writing highly discriminating items proved difficult in both contexts, and research studies examining this issue are clearly needed.

The solid line in Figure 5 presents the TIF for the Ideal Point order scale. A comparison with the dominance TIF reveals substantial differences in information at extreme negative trait levels (i.e., the Ideal Point order scale has much less error in this region). At the low end of the trait continuum, the Ideal Point IRT scale performed considerably better than did its dominance counterpart, but neither provided much information at the upper end due to a lack of discriminating items in that region. To improve measurement at the positive end of the trait continuum, items with locations in the region [+0.5, +1.5] should have been included in the ideal point scale, but in this instance, there were no such items in the pool.

Validity of the New Order Scales

Table 5 presents correlations among the three order scales, the Big Five scales, and two criterion measures. The most important result in this table is that the Ideal Point IRT order scores correlated highly with the Dominance IRT and Traditional CTT scores (.92 and .88, respectively), and hence, the validity was not reduced by incorporating neutral items into the order scale. As can be seen from the table, all three order scales showed evidence of convergent and discriminant validity—they correlated highly with the Conscientiousness scale of Goldberg’s Adjective Markers and did not correlate with the other Big Five markers. Also, as expected, the order scores correlated highly with scores on the SBQ and moderately with the HBCL scores.

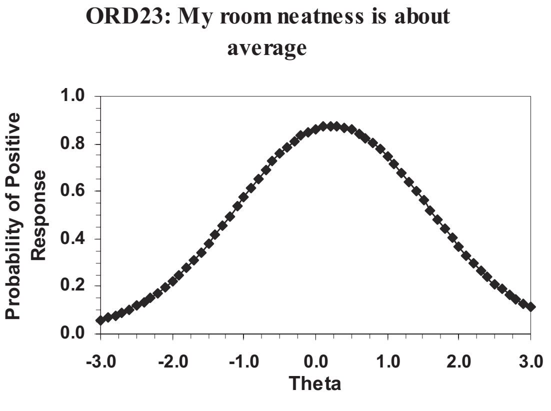

Figure 6. Generalized graded unfolding model item response function for Item ORD23.

Table 5

Convergent, Discriminant, and Criterion-Related Validities of Three Order Scales

Criterion and indicator	N	Ideal Point IRT	Dominance IRT	Traditional CTT
Order				
Ideal Point IRT	539	1.00		
Dominance IRT	539	.92	1.00	
Traditional CTT	501	.88	.91	1.00
Big Five Adjective Markers				
Conscientiousness	177	.49	.51	.54
Extraversion	179	−.02	.00	.01
Agreeableness	180	.04	.07	.06
Emotional Stability	179	−.11	−.09	−.10
Intellect	179	−.14	−.10	−.10
Study Behavior	181	.40	.39	.39
Preventative Health Behaviors	174	.22	.17	.19
Accident Control Behaviors	181	.17	.15	.13
Health Behaviors				
Traffic Risk	180	−.21	−.27	−.27
Avoiding Substance Risk	180	.14	.14	.16

Note. Correlations greater than or equal to ± .14 are significant. IRT = item response theory; CTT = classical test theory.

These criterion validity results indicate that the rank ordering of individuals remained roughly the same regardless of the measurement strategy used. This is not particularly surprising given that each scale was composed of items meeting the constraints of the various models.[4] However, the flexibility of using an ideal point model can be evinced by using it to score the Dominance IRT order scale. As shown by the scatter plot in Figure 7B, the original 2PLM scores correspond rather closely to the new scores obtained by fitting the GGUM to the same data. In fact, the correlation between the two sets of scores for the total sample was .97. Although there were some differences in rank order for those scoring in the lowest and highest regions of the trait continuum, the correlations in each trait region remained relatively high; they were .57 for those scoring at the lower end of the trait continuum [−3.0, −1.0], .94 for those scoring in the middle of the trait continuum [−1.0, +1.0], and .67 for those scoring at the upper end [+1.0, +3.0]. In contrast, when the 2PLM model was used to score the Ideal Point order scale, the resulting scores did not correspond well to the original GGUM scores (see Figure 7A). This was especially true in the lowest (i.e., [−3.0, −1.0]) and highest (i.e., [+1.0, +3.0]) regions of the trait continuum, where correlations between sets of scores were only .33 and .21, respectively. Moreover, the overall criterion-related validity of scores dropped in every case when 2PLM was used to score the Ideal Point order items but remained the same when GGUM was used to score the Dominance IRT order items. The main conclusion here is that the GGUM is flexible enough to accurately score any set of

[4] Also, because correlations are known to be insensitive to minor changes in rank order (Drasgow & Kang, 1984), gains in the measurement precision that the Ideal Point order scale seemed to achieve for low-scoring individuals could not be adequately expressed by means of the correlation index or any other index.

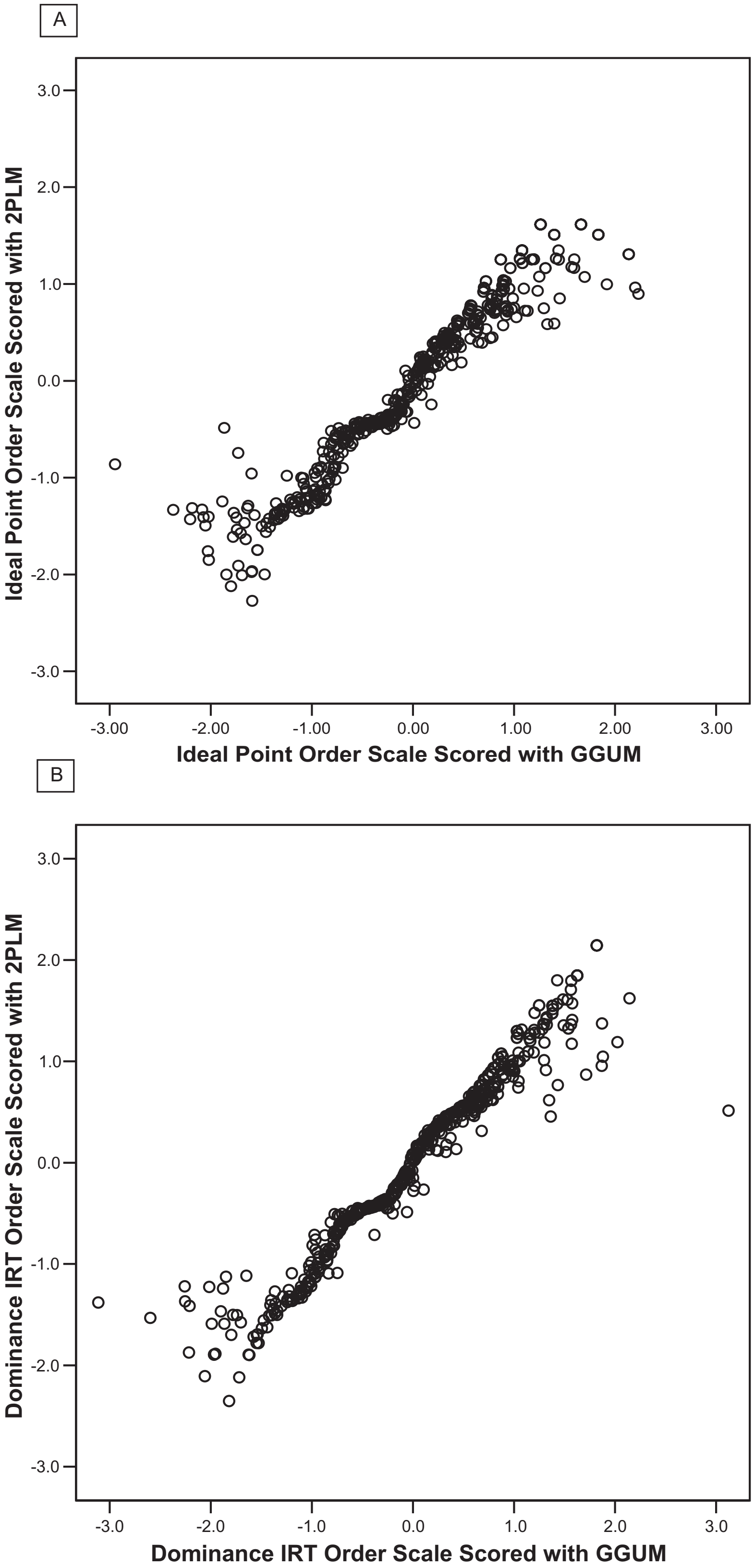

Figure 7. Panel A: Scatter-plot comparison of two-parameter logistic model (2PLM) and generalized graded unfolding model (GGUM) scores computed using responses to items included in the Ideal Point order scale. Panel B: Scatter-plot comparison of 2PLM and GGUM scores computed using responses to items included in the Dominance IRT order scale. IRT = item response theory.

personality items, whereas the 2PLM can be used only to score items with IRFs that are monotonic on the typically observed range of trait levels.

Discussion

The main aim of this article was to explicate why a transition to ideal point methods of scale construction is needed to advance the field of personality assessment. Whereas previous research has only suggested that an ideal point response process is more appropriate for personality items than is the dominance framework underlying traditional methods of scale construction (Stark et al., 2006), this study has empirically demonstrated the substantive benefits of ideal point methodology in the personality domain (see also Chernyshenko, 2002). Here, we started with a large, heterogeneous pool of order items and constructed scales utilizing traditional CTT, dominance IRT, and ideal point IRT methods. The virtues of each method were examined in terms of item pool utilization, model–data fit, measurement precision, and construct and criterion-related validity.

Our results show that the ideal point approach was more advantageous than either dominance method in several ways. First, we showed that it is possible to create discriminating neutral (moderate) items that produce nonmonotonic IRFs when scaled with an ideal point model. Whereas these items contributed substantially to measurement precision, they were identified as “poor” and eliminated by dominance methods due to low item–total correlations and small a parameters. Second, the greater flexibility of the ideal point model allowed for better utilization of the original 50-item pool. When item pools contain statements describing a wide range of behaviors (negative, neutral, and positive), ideal point models should clearly be preferred over dominance models. Third, the item content (as judged by raters) was congruent with the ordering of location parameters on the trait continuum in an ideal point framework, but this was not the case with the dominance models. Congruence of content and location parameters facilitates scale development by helping item writers generate statements with desired properties prior to pretesting. This flexibility makes ideal point approaches particularly attractive for large-scale testing applications, which require large item pools with substantial variation in location parameters. Fourth, utilizing neutral items here allowed for the creation of a scale that measured well at the extremes of the trait continuum, an objective that is sometimes difficult to achieve when dominance approaches are used for dichotomous scales.[5]

Whereas all the results mentioned above are certainly appealing, ideal point scale construction and scoring methods have limitations. One important issue is test scoring. Many researchers and practitioners are accustomed to using sum scores, which are not appropriate for ideal point contexts; instead, more computationally intensive IRT scoring methods are required (i.e., searching for a trait score that maximizes the observed response pattern). This limitation, in our view, is likely to become less salient, however, as more personality testing programs move toward computerized test administration. Programming maximum likelihood or expected a posteriori estimation is straightforward, and this extra effort is well justified by benefits inherent in IRT scores. In fact, many cognitive ability testing programs that rely on dominance models now prefer IRT scoring to total scores, even though the two sets of scores are usually highly correlated.

The second important issue is the type of items that test developers desire for their scales. As mentioned previously, response functions for moderately extreme items (e.g., “I like order”) are virtually the same over the observed theta range for dominance and ideal point models (this is true for both dichotomous and polytomous cases). Therefore, if one has such extreme items only, there would be little reason to use ideal point approaches, and a dominance model would be sufficient.

Note also that examinations of correlations between ideal point and dominance order scales with external variables indicated that construct and criterion-related validities did not decrease when neutral items were used for scale construction. Consequently, the adoption of ideal point approaches in personality scale development would not negate past research findings but would provide a more flexible platform for creating a new generation of personality measures. The use of ideal point methods, however, is unlikely to yield increases in criterion-related validities. Even a simple trichotomous scoring of all respondents in this sample (where low-scoring individuals are recoded as 1, those in the middle as 2, and those scoring highest as 3) would have produced correlations similar to those observed in Table 5. In our view, the gains associated with transitioning to ideal point methods will be found in better measurement of high/low scoring individuals and, hence, improved diagnostic, selection, and classification decisions. This argument is similar to the one used in ability testing, where short (10–15 items), general cognitive ability measures yield criterion-related validities that are almost as high as much longer multiple-aptitude test batteries. Yet, in that context, very few researchers or practitioners would advocate making college admission, licensure, or diagnostic decisions using a single score from a short test; typically, at least two or more dimensions (quantitative, analytical, and verbal) are used in addition to other measures of knowledge or skill, such as subject tests and grade-point average. Finally, having a model that fits data well is crucial for the accurate identification of differentially functioning items and tests. This is particularly important in applied settings, where secondary dimensions could unwittingly influence item responding and thus affect hiring and promotion decisions (see Bolt, 2002; Stark et al., 2004, 2006).

An interesting question is whether the assumption of an ideal point response process is reasonable for all personality dimensions or whether some personality traits, like cognitive ability, require dominance assumptions. For example, is it possible to write neutral items measuring, say, negative affect? To answer such questions definitively, more research is needed. However, in our view, if one can envision a neutral item that someone with very high negative affect would choose not to endorse, then an ideal point response process would still apply (e.g., “I only get upset when really bad things happen to me”). To date, we have found nonmonotonic items in scales measuring each of the six lower-order facets of conscientiousness: order, self-control, traditionalism, in-

[5] The issue of insufficient measurement precision at the extreme trait levels would of course be less pertinent for polytomous response formats. The use of multiple response options would substantially increase measurement efficiency (see Reise & Henson, 2000) as long as a well fitting IRT model can be found.

dustriousness, virtue, and responsibility (Chernyshenko, 2002), as well as in three facets of extraversion (sociability, dominance, and energy). So, at this point, our findings do support the general notion of ideal point responding in personality assessment, but we will continue to explore this issue with facets from the other factors of the Big Five.

Along these lines, a reviewer of this article suggested that one could easily transform a neutral item that might be fit well by an ideal point model (e.g., "My social skills are about average") into a dominance item (e.g., "My social skills are at least as good as those of an average person") and, hence, avoid dealing with nonmonotonicity altogether. This is an interesting idea, but we don't see a strong rationale for doing so. The first item is parsimonious and relatively straightforward, but it will almost surely appear poor by traditional (dominance) psychometric criteria. The main aim of this article is to encourage readers to consider more flexible methods of developing and scaling items that are consistent with, for example, common dialogue, rather than dropping items to fit the requirements of our current models. Discarding or transforming items that are otherwise usable, in our view, is no longer necessary.

Finally, identifying psychometric models that more closely correspond to the way people answer items provides a better understanding of the psychology of personality assessment. We believe that dominance models provide a misleading characterization of the response process: People do not endorse an item such as "My room neatness is about average" when their standing on the order latent trait is very high. Instead, when confronted with an item, people ask, "Does this item describe me?" Clearly, the response process is ideal point (Stark et al., 2006), and research can only benefit by the use of psychometric models that appropriately mirror this process.

References

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12,* 33–51.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49,* 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17,* 253–276.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15,* 113–141.

Butcher, J. N., Dahlstrom, W. G., Graham, I. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Candell, G. L., & Hulin, C. L. (1986). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17,* 417–440.

Caspi, A., Begg, D., Dickson, N., Harrington, H., Langley, J., Moffitt, T. E., et al. (1997). Personality differences predict health-risk behaviors in young adulthood: Evidence from a longitudinal study. *Journal of Personality and Social Psychology, 73,* 1052–1063.

Chalmers, D. K., Bowyer, C. A., & Olenick, N. L. (1990). Problem drinking and obesity: A comparison in personality patterns and life-style. *International Journal of the Addictions, 25,* 803–817.

Chernyshenko, O. S. (2002). *Applications of ideal point approaches to scale construction and scoring in personality measurement: The development of a six-faceted measure of conscientiousness.* Unpublished doctoral dissertation. University of Illinois at Urbana–Champaign.

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., Prewett, M., Gray, A., Stilson, R., & Tuttle, M. (2006, April). *Normative score comparisons from single stimulus, unidimensional forced choice, and multidimensional forced choice personality measures using item response theory.* Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Clark, L. A., & Watson, D. (1999). Temperament: A new paradigm for trait psychology. In L. Pervin & O. John (Eds.), *Handbook of personality research and theory* (Vol. 2). New York: Guilford Press.

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

Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences, 12,* 887–898.

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

de la Torre, J. (2004). *GGUM_MCMC: Markov chain Monte Carlo estimation of parameters in the generalized graded unfolding model.* Unpublished manuscript. Rutgers, The State University of New Jersey, University College—New Brunswick.

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30,* 1–17.

DeSarbo, W. S., & Hoffman, D. L. (1986). Simple and weighted unfolding threshold models for the spatial representation of binary choice data. *Applied Psychological Measurement, 10,* 247–264.

Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction analyses for detecting measurement nonequivalence. *Journal of Applied Psychology, 69,* 498–508.

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

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

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory. *Journal of Cross-Cultural Psychology, 2,* 133–148.

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

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment, 4,* 26–42.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Goldberg, L. R. (1998, March 18). *International Personality Item Pool: A scientific collaboratory for the development of advanced measures of* *personality and other individual differences.* Retrieved May 20, 2001, from http://ipip.ori.org/ipip

Gough, H. G., & Bradley, P. (1996). *CPI Manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Habing, B., Finch, H., & Roberts, J. (2005). A Q-sub-3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement, 29,* 457–471.

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

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49–78.

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

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63,* 146–163.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15,* 153–169.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood, IL: Dow Jones–Irwin.

Jackson, D. N. (1994). *Jackson Personality Inventory–Revised manual.* Port Huron, MI: Sigma Assessment Systems, Inc.

Jackson, D. N., Wrobleski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13,* 371–388.

Johnson, M. S., & Junker, B. W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding item response models. *Journal of Educational and Behavioral Statistics, 28,* 195–230.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist, 43,* 23–34.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Series No. 84–4). Arlington, VA: Office of Naval Research, Personnel and Training Research Programs.

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

Martin, N. G., & Boomsma, D. I. (1989). Willingness to drive when drunk and personality: A twin study. *Behavior Genetics, 19,* 97–111.

McCrae, R. R., & Costa, P. T., Jr. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56,* 586–595.

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

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

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

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 219–242). Mahwah, NJ: Erlbaum.

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

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

Roberts, B. (2001). *Study Behavior Questionnaire.* Unpublished manuscript. University of Illinois at Urbana–Champaign, Department of Psychology.

Roberts, B., Chernyshenko, O. S., Stark, S., & Goldberg, L. (2005). The construct of conscientiousness: The convergence between lexical models and scales drawn from six major personality questionnaires. *Personnel Psychology, 58,* 103–139.

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

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1998). *The generalized graded unfolding model: A general parametric item response model for unfolding graded responses* (RR-98–32). Princeton, NJ: Educational Testing Service.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (1999, April). *Estimating parameters in the generalized graded unfolding model: Sensitivity to the prior distribution assumption and the number of quadrature points used.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Québec, Canada.

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., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20,* 231–255.

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

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,* 282–307.

Shedler, J., & Block, J. (1990). Adolescent drug use and psychological health: A longitudinal inquiry. *American Psychologist, 45,* 612–630.

Sher, K. J., & Trull, T. J. (1994). Personality and disinhibitory psychopathology: Alcoholism and antisocial personality disorder. *Journal of Abnormal Psychology, 103,* 92–102.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological assessment, 17,* 28–43.

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. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86,* 943–953.

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

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. 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,* 25–39.

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Psychological Bulletin, 99,* 118−128.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 55,* 293–325.

Tellegen, A. (1982). *A brief manual for the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247–260.

Thissen, D., & Wainer, H. (Eds). (2001). *Test scoring.* Mahwah, NJ: Erlbaum.

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

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

Vickers, R. R., Jr., Conway, T. L., & Hervig, L. K. (1990). Demonstration of replicable dimensions of health behaviors. *Preventive Medicine, 19,* 377–401.

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

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,* 545–576.

Watson, D., & Clark, L. A. (1993). Behavioral disinhibition versus constraint: A dispositional perspective. In D. M. Wegner & J. W. Pennebaker (Eds.), *Handbook of mental control* (pp. 506–527). Upper Saddle River, NJ: Prentice Hall.

White, H. R., & Johnson, V. (1988). Risk taking as a predictor of adolescent sexual activity and use of contraception. *Journal of Adolescent Research, 3,* 317–331.

Received June 3, 2005
Revision received July 31, 2006
Accepted August 22, 2006 ■