Graduate
Management
Admission
Council®

Adaptive Testing With the Multi-Unidimensional Pairwise Preference Model

Stephen Stark
University of South Florida

and

Oleksandr S. Chernyshenko
University of Canterbury

Presented at the New CAT Models Paper Session, June 8, 2007

Abstract

This paper presents the results of three simulation studies evaluating the feasibility of a multi-unidimensional pairwise preference model (MUPP; Stark, Chernyshenko, & Drasgow, 2005) for administering adaptive, fake-resistant personality tests. The first two studies examined score recovery for 3- and 5- dimensional nonadaptive tests using various test lengths and linking designs. Results indicated that a minimal "circular" linking design was sufficient for recovery of known trait scores, and test length and linking requirements did not pose serious constraints on the development of measures with higher dimensionality. The third study compared estimation accuracy for adaptive and nonadaptive tests involving varying numbers of dimensions (3, 5, 7, 10), items per dimension (5, 10, 20), and proportions of unidimensional pairings (5, 10, 20) needed to identify the latent metric. It was found that adaptive testing with the MUPP model was viable and produced gains in efficiency comparable to those observed with unidimensional CATs created for cognitive ability assessment. Implications of these findings for applied use are discussed.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Stark, S. & Chernyshenko. O. S. (2007). Adaptive testing with the multi-unidimensional pairwise preference model. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

**Stephen Stark, USF Department of Psychology, PCD 4118G,
4202 E. Fowler Ave., Tampa, FL, 33620, sstark@cas.usf.edu.**

Adaptive Testing with the Multi-Unidimensional Pairwise Preference Model

In recent years, there has been a renewed interest among applied psychologists in forced-choice personality tests. This stems primarily from the desire to use personality test scores in high-stakes settings where response distortion (i.e., faking good or faking bad) is a key concern. Multidimensional forced-choice tests, involving items composed of statements similar in desirability but representing different dimensions, are particularly appealing, because they are believed to be more resistant to faking than tests comprising single statement items where the "correct" answers are often easy to discern. However, because of concerns about ipsativity and a lack of formal psychometric models for test construction and scoring, there have been relatively few applications of multidimensional forced-choice tests until recently, and to our knowledge, none involving CAT.

This paper proposes and evaluates a method for administering multidimensional forced-choice tests adaptively using a pairwise preference format. The algorithm is based on the multi-unidimensional pairwise preference item response theory (IRT) model (MUPP; Stark, 2002), which has been used to construct and accurately score one- and two-dimensional forced-choice tests. Two simulation studies are described that were designed to explore the feasibility of this approach with tests involving high dimensionality. First, we examined score recovery for 3- and 5-dimensional nonadaptive tests using various test lengths and linking designs. Then, we compared estimation accuracy for nonadaptive and adaptive tests involving different numbers of dimensions (3, 5, 7, 10), items per dimension (5, 10, 20), and proportions of unidimensional pairings (5, 10, 20) needed to identify the latent metric. Before describing these studies, however, we begin with a brief review of the MUPP model and summarize results from early 1- and 2-dimensional studies that provided a foundation for this work.

Stark's MUPP Model

Stark (2002) proposed an IRT model that can be used to construct and score multidimensional pairwise preference (MDPP) items. In essence, the model assumes that when a respondent is presented with a pair of stimuli (i.e., personality statements), denoted as stimuli *s* and *t*, and is asked to indicate a preference, he/she evaluates each stimulus separately and makes *independent* decisions about stimulus endorsement. If a respondent's endorsement propensity is equal for both stimuli, he/she must reevaluate the stimuli independently until a preference is reached. (Note, this assumption is similar to that of Andrich's (1995) hyperbolic cosine model for unidimensional pairwise preferences.) Thus, the probability of endorsing a stimulus *s* over a stimulus *t* can be formally written as

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0\}}{P_{st}\{1,0\} + P_{st}\{0,1\}} \approx \frac{P_s\{1\}P_t\{0\}}{P_s\{1\}P_t\{0\} + P_s\{0\}P_t\{1\}}, \tag{1}$$

where

$i =$ index for pairs of stimuli, where $i = 1$ to I (note that an item is defined as a stimulus pair),

$d =$ index for dimensions, where $d = 1, \ldots, D$,

$s, t =$ indices for first and second stimuli, respectively, in a pairing,

$\theta_{d_s}, \theta_{d_t}$ = latent trait values for a respondent on dimensions d_s and d_t respectively,

$P_s\{1\}$, $P_s\{0\}$ = probability of endorsing/not endorsing stimulus *s* at θ_{d_s},

$P_t\{1\}$, $P_t\{0\}$ = probability of endorsing/not endorsing stimulus *t* at θ_{d_t},

$P_{st}\{1,0\}$ = joint probability of endorsing stimulus *s*, and not endorsing stimulus *t* at $(\theta_{d_s}, \theta_{d_t})$,

$P_{st}\{0,1\}$ = joint probability of not endorsing stimulus *s*, and endorsing stimulus *t* at $(\theta_{d_s}, \theta_{d_t})$,

and

$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t})$ = probability of respondent *j* preferring stimulus *s* to stimulus *t* in pairing *i*.

A preference can be represented by the joint outcome {Agree (1), Disagree (0)} or {Disagree (0), Agree (1)}. An outcome of {1,0} indicates that stimulus *s* was preferred to stimulus *t*, and is considered a positive response; an outcome of {0,1} indicates that stimulus *t* was preferred to *s* (a negative response). Thus, the response data for this model are dichotomous. The probability of endorsing a stimulus in a pair depends on θ_{d_s} and θ_{d_t} and also depends fundamentally on the model chosen to characterize the process of single stimulus responding. In principle, any IRT model for unidimensional single stimulus responses could be chosen for computing the $P_s(1)$ and $P_t\{0\}$ terms in Equation 1, but, in our research, we have utilized the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), which has been found to fit data from a variety of personality statements well in recent investigations (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Stark, Chernyshenko, Drasgow & Williams, 2006).

Stark, Chernyshenko, and Drasgow (2005) examined the viability of the proposed approach to test construction and scoring for two-dimensional (2-D) personality tests of 20, 40 and 80 pairwise preference items. They found a .77 correlation between estimated and known (generated) trait scores in the most unfavorable condition (a 20-item test with 10% unidimensional pairings) to .96 in the most favorable (80-item test with 40% unidimensional pairings). The average correlation was about 0.9 for the 40-item tests, regardless of the percentage of unidimensional pairings. This initial study showed good to excellent rank-order correspondence between estimated and known trait scores, thus demonstrating the viability of this approach for, at the very least, tests with low dimensionality.

Purpose

While the two-dimensional simulation study provided evidence that the MUPP approach to test construction and scoring could accurately recover normative trait scores, additional research was needed to investigate the capabilities and limitations of this method under conditions more likely to be encountered in applied settings. First, it was essential to determine how many pairwise preference items would be needed to accurately recover normative trait scores for 3- and 5-dimensional tests and how many unidimensional pairings would be required to identify the latent metric in those cases. If those numbers were unreasonably large, it would make little sense to study tests of 7 dimensions or higher. Second, because requiring all possible dimensional pairings would become prohibitive in terms of test length as dimensionality increased, it was necessary to determine how to most efficiently identify the latent metric for normative score recovery. Specifically, would a minimal or circular linking design suffice, or would all possible pairings of dimensions be necessary? Third, what gains in efficiency might be observed by using adaptive item selection in place of static tests that could be difficult to create because of variations in the multidimensional item information surfaces of multidimensional pairwise preference items? Moreover, with adaptive testing, would it be possible to preserve scoring accuracy with even fewer unidimensional pairs so that tests could, in principle, be made more fake-resistant for applied use?

Study 1

Method

A simulation study was conducted using GGUM parameters for personality statements measuring three lower-order facets of the "big five" factor Conscientiousness (Order, Traditionalism, and Responsibility) and two lower-order facets of Extraversion (Energy and Dominance). To assess the potential effects of differences in statement quality across facets, we first conducted one-dimensional (1-D) simulations to establish a baseline for comparing estimation accuracy subsequently using three-dimensional (3-D) and five-dimensional (5-D) tests. 10- and 20-item tests were first constructed for each facet in a manner similar to that described by Stark (2002). However, unlike those initial simulations, where scoring accuracy was examined by comparing estimated scores for 50 to 100 simulated examinees at discrete points on either 1-D or 2-D grids reflecting combinations of θ (trait levels), a sampling approach was used instead. Specifically, for each 1-D test, 1,000 θs were sampled randomly from independent standard normal distributions; responses were generated based on the MUPP model; and response patterns were scored using a Visual Basic for Applications computer program. The quality of normative score recovery in each condition was then assessed using the correlation between the estimated and known θ values.

A similar approach was then used to develop tests involving three and five personality dimensions. For the 3-D simulations, we created items by pairing statements representing Order, Traditionalism, and Energy. For the 5-D simulations, we also included Responsibility and Dominance. Building on the 1-D studies, where either 10 or 20 items per dimension were used, we created items for the 3-D- and 5-D tests by pairing statements representing all possible combinations of facets, while also including either 10% or 20% of unidimensional pairings (e.g., 1-1, 2-2, 3-3) to identify the latent metric. The exact types and numbers of pairings (i.e., items) in each condition of these "complete linking" studies are shown in Table 1 (in the column of Table 1 entitled "Facet Pairings", 1 = Order, 2 = Traditionalism, 3 = Energy, 4 = Responsibility, and 5 = Dominance).

Table 1. Numbers and Types of Items in the 3-D and 5-D Tests Using a Complete Linking Design With 10% and 20% Unidimensional Pairings

Facet Pairings	3-Dimensional Tests			
	30 items		60 items	
	10%	20%	10%	20%
1-1	1	2	2	4
2-2	1	2	2	4
3-3	1	2	2	4
1-2	9	8	18	16
1-3	9	8	18	16
2-3	9	8	18	16
Facet Pairings	**5-Dimensional Tests**			
	50 items		100 items	
	10%	20%	10%	20%
1-1	1	2	2	4
2-2	1	2	2	4
3-3	1	2	2	4
4-4	1	2	2	4
5-5	1	2	2	4
1-2	5	4	9	8
1-3	4	4	9	8
1-4	5	4	9	8
1-5	4	4	9	8
2-3	5	4	9	8
2-4	4	4	9	8
2-5	4	4	9	8
3-4	5	4	9	8
3-5	4	4	9	8
4-5	5	4	9	8

In an effort to adequately cover the many possible combinations of θs in the 3-D and 5-D studies, 3,000 (3-D) and 5,000 (5-D) θs were sampled for the facets in each study from independent standard normal distributions, and the rank-order correspondence of estimated and known θs was examined using Pearson correlations. Then, to obtain a single index of recovery for each experimental condition, the correlations were averaged across dimensions.

Results

Table 2 presents the correlations between known and estimated θs across each personality facet and test type. For example, the .95 in the first row of the last column represents the correlation between the estimated and sampled θs for the facet of Order, as measured by the 100-item 5-D test having 20% unidimensional pairings. As in the 2-D studies reported by Stark et al. (2005), there was little, if any, effect for the percent of unidimensional pairings, which suggests that 10% unidimensional pairings is all that is required with a complete linking design. In addition, the correlations between estimated and known θ s were high even for the short tests in each condition (.88), and they improved with increases in test length. In terms of the number of pairings involving each facet, the MUPP scoring algorithm seemed to perform better for the 3-D and 5-D tests than with1-D tests of comparable length, suggesting that the recovery of normative scores might actually improve with higher dimensionality, provided that enough combinations of facets are represented by the multidimensional pairings. Based on these results, we concluded that 15 items per dimension, with 10% of those being unidimensional, would be sufficient to yield good rank-order θ recovery with nonadaptive measures.

Table 2. Correlations Between Estimated and Known θs for 1-D, 3-D, and 5-D Tests in the Complete Linking Simulation Studies for 10% and 20% Unidimensional Pairings

Temperament Facet	1-D Tests		3-D Tests				5-D Tests			
	10 items	20 items	30 items		60 items		50 items		100 items	
			10%	20%	10%	20%	10%	20%	10%	20%
Order	.88	.93	.91	.91	.95	.95	.91	.91	.95	.95
Traditionalism	.92	.95	.89	.90	.93	.94	.89	.89	.94	.94
Energy	.85	.91	.86	.85	.92	.92	.85	.85	.92	.92
Responsibility	.85	.93	*	*	*	*	.86	.85	.92	.92
Dominance	.90	.93	*	*	*	*	.90	.90	.95	.95
Average	.88	.93	.89	.89	.93	.94	.88	.88	.94	.94

*Data not simulated for this facet.

Study 2: Examining Minimal Cross-Dimensional Linking with 5-Dimensional Tests

Method

Although the results of the complete linking studies with 3-D and 5-D tests indicated excellent θ recovery, practical concerns dictated the exploration of alternative linking methods, because requiring all possible pairings would become prohibitive as test dimensionality increased (i.e., test length would increase exponentially). From a test administration standpoint, it would be preferable to have minimal constraints on the types of pairings presented to an examinee, so that an adaptive algorithm could take full advantage of the available statements in a testing pool. One of the simplest minimal linking designs is *circular* linking, which pairs statements representing facets adjacent to each other when arranged in a circle. For example, with a 5-D test, we could pair dimensions 1-2, 2-3, 3-4, 4-5, and 5-1, while also including a proportion (10%) of unidimensional pairings (1-1, 2-2, 3-3, 4-4, and 5-5) to identify the latent metric. To explore the efficacy of this linking strategy, we conducted simulations involving 50- and 100-item tests with 10% and 20% unidimensional pairings. As before, θs were sampled for 5,000 examinees from five independent standard normal distributions and Pearson correlations were used to assess the correspondence between known and estimated θs.

Results

The results of this study are presented in Table 3. As shown in the table, the accuracy of θ estimation in the circular linking conditions was virtually identical to the accuracy in the complete linking conditions. In fact, the average correlations in the corresponding conditions were identical. This result is important for applied purposes because it indicates that circular linking is sufficient for accurate rank-order recovery of trait scores with the MUPP procedure.

Therefore, complete linking is unnecessary and linking requirements do not pose serious constraints on the development of tests with higher dimensionality.

Table 3. Comparison of Correlations Between Estimated and Known θs for Circular and Complete Linking Designs With 10% and 20% Unidimensional Pairings

Temperament Facet	5-D Tests with Circular Linking				5-D Tests with Complete Linking			
	50 items		100 items		50 items		100 items	
	10%	20%	10%	20%	10%	20%	10%	20%
Order	.91	.92	.96	.95	.91	.91	.95	.95
Traditionalism	.89	.89	.93	.94	.89	.89	.94	.94
Energy	.87	.86	.92	.92	.85	.85	.92	.92
Responsibility	.86	.85	.93	.92	.86	.85	.92	.92
Dominance	.89	.89	.95	.95	.90	.90	.95	.95
Average	.88	.88	.94	.94	.88	.88	.94	.94

Study 3: Comparing the Accuracy of Nonadaptive and Adaptive Tests

Adaptive personality testing with the MUPP model in applied settings must address three issues. First, one must determine the number of personality dimensions that will be assessed and develop pools of statements that vary adequately in terms of location and social desirability. Second, constraints must be implemented to pair statements in a way that will not only identify the latent metric, but will also enhance their resistance to faking. Third, one must decide whether to terminate testing based on estimated standard errors of θ estimates or using a fixed number of items. With both administration time and perceived fairness in mind, we developed the "fixed length" adaptive algorithm described below.

1. Specify the number of dimensions to assess and the number of items to administer "per dimension". These choices determine the total test length. For example, one might choose to assess three dimensions with ten items per dimension, thus yielding a test composed of 30 pairwise preference items.
2. Create and store content codes representing all unique multidimensional and unidimensional pairings (e.g., for a 3-D test, 1-1, 2-2, 3-3, 1-2, 1-3, 2-3).
3. Specify what percent of unidimensional pairings (i.e., items) will be used to identify the latent metric. (Previous research suggested that 10% unidimensional pairings would be adequate with nonadaptive tests.)
4. Create and store an item dimensionality sequence for the whole test to ensure that the metric is identified and content is balanced.

a. Create a "linking group" to enable trait score estimation as soon as possible using a circular linking design.

b. Then, use a complete linking design to cycle through the remaining dimensional combinations adaptively after the first actual trait scores have been estimated.

5. Create the linking group items by randomly pairing statements or by using item information, assuming average (zero) initial trait scores for a respondent on all dimensions. Administer the linking group.
6. Estimate trait scores and standard errors using the Bayesian modal scoring algorithm developed by Stark (2002).
7. Select and administer subsequently items adaptively based on information provided at the estimated trait scores until the desired number of items has been reached. Create items representing the dimensional combinations as follows:
    a. For *unidimensional* items needed to identify the metric, pair statements that are similar in desirability but different in location to the greatest extent possible.
    b. For multidimensional items, needed to enhance resistance to faking, pair statements that are similar in desirability and location.

Method

A Visual Basic.NET program was used to compare the efficacy of the adaptive algorithm described above with nonadaptive tests of the same length and dimensionality. Estimation accuracy was examined using a fully crossed design involving varying numbers of dimensions (3, 5, 7, 10), items per dimension (5, 10, 20), and proportions of unidimensional pairings (5, 10, 20) needed to identify the latent metric. In each condition, data were generated by sampling 1,000 θs from independent standard normal distributions, and θ recovery was assessed using correlations and bias statistics. To increase the realism of the simulations, we utilized 40 personality statements for each dimension having parameters and social desirability ratings derived from actual examinee data collected in previous studies. However, some minor changes to the parameters were made, as needed, to balance discrimination and location across dimensions.

Results

Tables 4 and 5 present the correlations and error statistics for the nonadaptive and adaptive test simulations. In each table, the first column shows the percentage of unidimensional items used to identify the metric. The second column indicates the number of items per construct; for example, a 3-D test involving 5 items per construct would comprise 15 items. The remaining columns show the results for the nonadaptive and adaptive conditions respectively. In each case, the value shown in a cell represents the average across dimensions.

As can be seen in Table 4, the average correlations between estimated and known θs increased as test length increased from 5 to 20 items per dimension, and there was virtually no effect for the percent of unidimensional pairings. These results are consistent with the 1-D and 2-D results reported by Stark et al. (2005). Moreover, the fact that 5% unidimensional pairings was adequate for attaining good rank-order recovery of θs, regardless of how many dimensions were assessed, is important from an applied perspective because such items are arguably less resistant to faking than MDPP items. Finally, note the striking gains in efficiency that were obtained by going from nonadaptive to adaptive item selection. Adaptive tests yielded approximately the same correlations as nonadaptive tests that were nearly twice as long, a finding which is consistent with a multitude of studies involving unidimensional cognitive ability CATs.

Table 4. Comparison of Correlations Between Estimated and Known θs or Nonadaptive and Adaptive MDPP Tests

		Average Correlation Across Dimensions							
		Nonadaptive				Adaptive			
% Unidim.	Items Per Dimension	3-D	5-D	7-D	10-D	3-D	5-D	7-D	10-D
	5	.73	.72	.76	.76	.87	.85	.86	.87
5	10	.85	.87	.87	.86	.93	.93	.93	.93
	20	.93	.93	.93	.94	.96	.96	.96	.96
	5	.73	.74	.75	.75	.87	.87	.85	.88
10	10	.85	.85	.86	.87	.92	.93	.93	.93
	20	.93	.93	.94	.94	.96	.96	.96	.96
	5	.74	.74	.74	.75	.87	.84	.86	.87
20	10	.85	.85	.87	.86	.92	.90	.93	.93
	20	.92	.93	.93	.94	.96	.96	.96	.96

Table 5 presents the average absolute bias statistics for the nonadaptive and adaptive tests. These findings mirror those for the correlations shown in Table 4. Specifically, there appeared to be no effect for the percent of unidimensional pairings, which suggests that 5%, rather than 10%, unidimensional pairings is sufficient to identify the latent metric. Second, as before, test length had the primary effect on accuracy, and test length can be reduced by 50% using adaptive item selection. Perhaps most importantly, these results show that the estimated θs were not only useful in terms of rank order, but in terms of accuracy, even with tests of high dimensionality. For example, the average absolute bias was just .27 with 10-D tests involving 100 items, only 5 of which were unidimensional.

Table 5. Comparison of Absolute Bias for Nonadaptive and Adaptive MDPP Tests

		Average Absolute Bias Across Dimensions							
		Nonadaptive				Adaptive			
% Unidim.	Items Per Dimension	3-D	5-D	7-D	10-D	3-D	5-D	7-D	10-D
	5	.53	.53	.51	.49	.38	.39	.38	.37
5	10	.40	.39	.38	.38	.28	.28	.27	.27
	20	.29	.29	.39	.27	.22	.21	.21	.27
	5	.52	.52	.51	.50	.38	.38	.39	.36
10	10	.41	.40	.39	.38	.29	.28	.27	.27
	20	.29	.29	.28	.27	.22	.21	.20	.20
	5	.52	.52	.52	.50	.39	.40	.37	.37
20	10	.42	.40	.38	.38	.29	.31	.35	.27
	20	.30	.29	.28	.27	.22	.21	.29	.20

Summary and Conclusions

In applied psychology, there has been a resurgence of interest in alternative formats for assessing noncognitive constructs, particularly in the personality domain. This interest has been driven largely by concerns about response sets or biases, such as faking, and their effects on the quality of personnel decisions (e.g., Schmitt & Oswald, 2007). Of primary interest are multidimensional forced-choice items involving pairs or tetrads of statements that are balanced in terms of social desirability, because they seem to offer some promise in terms of their resistance to faking (Jackson, Wrobleski, & Ashton, 2000). However, although enthusiasm for using forced-choice formats in applied settings runs high, relatively little is known about the psychometric properties of the models used for constructing and scoring such measures. This paper attempted to bridge that gap by focusing on multidimensional pairwise preference tests and specifically the MUPP IRT approach to test construction and scoring (Stark, 2000l Stark et al., 2005) in a CAT framework.

We conducted two simulation studies that addressed issues relevant to the design of a CAT algorithm based on the MUPP model. In the first study, we showed that good recovery of trait scores could be achieved using tests of up to five dimensions with as few as 10% of the items being unidimensional. We then examined whether test length could be held to reasonable levels in settings involving high dimensionality by using a circular (minimal) linking design, as opposed to full linking, and we found no difference in terms of the average correlations between the estimated and known trait scores. Finally, we proposed and tested a CAT algorithm that used both circular and full linking to administer fixed length MDPP tests scored via Bayesian modal trait estimation (see Stark et al., 2005). It was found, for example, with a 7-D test involving just 70 items, 7 being unidimensional, the average correlations across dimensions between known and estimated θs were .83 and .91 for nonadaptive and adaptive tests, respectively. With 10-D adaptive tests, correlations as high as .94 were observed, thus demonstrating the viability of this approach in situations commonly encountered in organizations.

References

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19,* 269-290.

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

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.

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.

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology, 91,* 613-621.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment* [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement, 29,* 184-201.

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