RESPONDENT AND TEST DELIVERY CHARACTERISTICS THAT INDUCE ITEM UNFOLDING

Christopher J. Lake

A Thesis

Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

December 2010

Committee:
Michael J. Zickar, Advisor
Scott Highhouse
John Tisak
ABSTRACT

Michael J. Zickar, Advisor

Relatively little is known about the processes that people use when responding to unfolding scale items. In the present study, two types of characteristics (test delivery format and respondent attitude structure) are analyzed to develop an understanding of the processes that induce item unfolding. First, based on the hypothesis that context effects limit unfolding, a scale of attitudes toward workplace drug testing was administered in a computerized item-by-item format ($N = 601$) and compared to a pencil-and-paper block format administration ($N = 566$). Second, the structure of participants’ attitudes was investigated in an attempt to link attitudinal strength, ambivalence, and indifference to specific response patterns that either induce or inhibit item unfolding. Results showed that when scale items were presented item-by-item, respondents’ item-to-item differentiation increased slightly, but this did not systematically lead to better item unfolding or information. With regard to attitude structure, ambivalent respondents endorsed more unfolding items than did univalent or indifferent respondents. The implications of this research are discussed.
To the family and friends who have been so supportive of my academic endeavors: thank you.
ACKNOWLEDGMENTS

I could not have completed this project without the assistance and guidance of several faculty members and students at BGSU. First, I would like to thank my advisor, Mike Zickar, for his encouragement, support, and sense of humor throughout the entirety of this project. Next, thanks to the members of my thesis committee, Scott Highhouse and John Tisak, for providing insights and suggestions; their perspectives have definitely improved the quality of this study. I would additionally like to thank the members of the BGSU Personality Measurement Research Group — Nathan Carter, Dev Dalal, Bing Lin, and Katie Wolford — whose suggestions during our research meetings helped to crystallize this project. I am especially grateful to Nathan Carter who was extremely helpful with regard to many of the analyses required for the results section of this study. Finally, I would like to give a big thanks to Kevin Dawson who single-handedly conducted the questionnaire administration, and was of great assistance with data coding and analysis.
TABLE OF CONTENTS

INTRODUCTION ................................................................................................................. 1

CHAPTER I. THE NATURE OF UNFOLDING SCALE ITEMS .................................................. 4

The Use of Unfolding Items in Scales ........................................................................ 6

Demonstrated Benefits of Unfolding Items ............................................................... 7

Analyzing Attitudes toward Drug Testing ................................................................. 9

The Cognitive Demands of Responding to Unfolding Items ..................................... 11

CHAPTER II. PRESENT STUDY AND HYPOTHESES ...................................................... 15

Reducing Context Effects by Manipulating Test Delivery (Hypotheses 1 to 4) ........ 15

The Effects of Attitudinal Ambivalence and Indifference (Hypotheses 5 to 8) .......... 16

CHAPTER III. METHODS .................................................................................................. 21

Participants and Procedures ....................................................................................... 21

Instruments ................................................................................................................. 22

Measuring Ambivalence ................................................................................ 22

Measuring Attitude Strength .......................................................................... 23

Internal Validity Check: Measuring Attitude Extremity ....................................... 23

CHAPTER IV. RESULTS .................................................................................................... 25

Preliminary Assessments of Fit ............................................................... 25

Option Response Analysis ............................................................................. 25

Item-Level Fit Analyses .................................................................................. 25

Effects of Presenting Items Singly (Hypotheses 1 to 4) ............................................ 26

DIF Analysis .................................................................................................. 27
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of Attitude Structure (Hypotheses 5 to 8)</td>
<td>30</td>
</tr>
<tr>
<td>Person Fit</td>
<td>32</td>
</tr>
<tr>
<td>CHAPTER V. DISCUSSION</td>
<td>34</td>
</tr>
<tr>
<td>Test Delivery Method</td>
<td>34</td>
</tr>
<tr>
<td>Respondent Attitude Structure</td>
<td>36</td>
</tr>
<tr>
<td>Future Research Directions</td>
<td>38</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>40</td>
</tr>
<tr>
<td>TABLES</td>
<td>47</td>
</tr>
<tr>
<td>APPENDIX A. GGUM ITEM PARAMETER ESTIMATES FOR DRUG TESTING SCALE</td>
<td>52</td>
</tr>
<tr>
<td>APPENDIX B. ITEMS MEASURING AMBIVALENCE</td>
<td>55</td>
</tr>
<tr>
<td>APPENDIX C. ITEMS MEASURING ATTITUDE STRENGTH</td>
<td>56</td>
</tr>
<tr>
<td>FIGURES</td>
<td>57</td>
</tr>
</tbody>
</table>
LIST OF TABLES/FIGURES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Defining Characteristics of Unfolding and Traditional (Monotonically-Increasing) Scale Items

Table 2: Likelihood Ratio Item-Level Tests for the Presence of DIF between Items in Experimental and Comparison Samples

Table 3: Comparison of Mean (Standard Deviation) GGUM-Derived Parameter Estimates Obtained from Experimental and Comparison Samples

Table 4: Correlations between Attitude Structure Variables and Number of Unfolding Items Endorsed

Table 5: Linear and Curvilinear Regression Coefficient Estimates used for Predicting Person Mis-fit from Attitude Structure Variables

Figure 1: Comparison of item response functions from a hypothetical monotonically-increasing item and a hypothetical unfolding item

Figure 2: Item response functions from sample Lin et al. (2009) drug testing scale items

Figure 3: Comparison of five respondents’ positive and negative evaluations of drug testing

Figure 4: Comparison of strongly agree and somewhat agree response option functions in three unfolding items

Figure 5: Comparison of strongly agree and somewhat agree response option functions in two extremely-worded items

Figure 6: Item response functions from six of seven items with statistically significant differential item functioning (DIF)

Figure 7: Information functions from four of seven items with statistically significant differential item functioning (DIF)

Figure 8: Scale-level information and standard error function comparisons

Figure 9: Item location estimate dispersion in comparison (left) and experimental (right)

Figure 10: Prediction of person mis-fit from two ambivalence indicator variables
INTRODUCTION

Respondent and Test Delivery Characteristics that Induce Item Unfolding

Self-report questionnaires and scales are commonly used in applied psychological research and practice. A review of studies appearing in top industrial-organizational psychology journals by Sackett and Larson (1990) indicated that more than half of the surveyed studies relied on self-report measures as the main data-collection tool. The most common form of self-report surveys is the Likert-type scale wherein participants are asked to respond to a series of statements by selecting from an ordered set of response options (e.g., strongly disagree - strongly agree). The numeric values associated with each selected response option are then summed and used as a measure of a person’s attribute (e.g., a personality facet or attitude score). However, this ubiquitous form of measurement makes an implicit assumption about a scale’s response options — that they are always strictly ordinal. A response of strongly agree, for instance, always indicates a more extreme attitude than does the response neither agree nor disagree. A different method of measurement, initially proposed by Thurstone and made more practically useful by recent Item Response Theory (IRT) models, allows for nonmonotonicity whereby items can unfold. Although it is uncommon for researchers and practitioners to use unfolding items, some initial research evidence suggest that using unfolding items may provide certain psychometric advantages over the more-popular Likert-type scales.

Item unfolding is a response process whereby a scale item is increasingly endorsed by respondents with higher trait levels, but only to a certain point. At this point of unfolding, the item is then decreasingly endorsed by respondents with higher trait levels. Figure 1 graphically illustrates the difference between an unfolding item and a traditional, monotonically-increasing item. Imagine that the hypothetical unfolding item depicted in Figure 1 reads “genetic
engineering should be allowed if a researcher is attempting to cure deadly human diseases” and that the underlying theta continuum represents attitudes toward genetic engineering. A respondent with a negative attitude toward genetic engineering, say −3, rejects the item with *strongly disagree* because he cannot imagine any scenario in which genetic engineering is acceptable. A respondent with a positive attitude toward genetic engineering, say +3, also rejects the item with *strongly disagree* because she feels that genetic engineering research should not be limited to only deadly human disease. A respondent with a moderate attitude toward genetic engineering, say 0, endorses this item with *strongly agree* because it accurately represents her intermediate feelings toward the topic. Note the non-ordinal nature of the response options in this scenario: people with extremely positive and extremely negative attitudes would select the very same response option. This response pattern would not be possible using Likert-type scales.

Some additional defining characteristics of unfolding items are presented in Table 1, and are referred to throughout the remainder of this paper. The concept of ideal-point measurement, which allows for item unfolding, can be traced back to a model proposed by Coombs (1964). To date, the concept of unfolding has received fairly minimal research attention. Of the available research studies, many focus on broad-level model functioning (e.g., model-data fit; recovery of known parameters) as opposed to individual item unfolding. Although this focus has shown the utility of unfolding and driven innovation in statistical models, it has led to few insights into the characteristics that might cause a particular item to unfold. The present study attempts to clarify the process of item unfolding by demonstrating empirically the importance of specific test delivery and respondent characteristics to the unfolding process.

Data from Lin, Zickar, Carter, Dalal, and Wolford’s (2009) measure of attitudes toward workplace drug testing are used as an example unfolding scale. The Lin et al. measure, which
contains many neutral-tending unfolding items, was developed in accordance with a technique proposed by Chernyshenko, Stark, Drasgow, and Roberts (2007). It is apparent from the Lin et al. study that the bulk of the items written specifically to unfold near the neutral (midpoint) region of the theta continuum did not function as expected. In general, these items exhibited only small amounts of unfolding and the points at which items unfolded were relatively distant from the midpoint. Also concerning was the fact that items near the theta continuum midpoint tended to have severely reduced information functions. A follow-up study using this scale presents an opportunity to gain some insights into the respondent and test delivery characteristics that induce item unfolding. A broad array of research from multiple disciplines will be incorporated into this study such that future researchers studying unfolding should have an increased understanding of the item-level unfolding process — how and why people’s responses might cause an item to unfold. Specifically, research on cognitive information processing, test delivery methods, and attitudinal neutrality seem crucial to this conversation and are thus incorporated into the study. It is hypothesized that manipulating the scale delivery method will result in more accurate responses, especially for neutral-tending unfolding items. Additionally, it is hypothesized that a person’s attitude structure (e.g., ambivalence) plays an important role in how he or she responds to an unfolding item.
CHAPTER I: THE NATURE OF UNFOLDING SCALE ITEMS

Although most attitudinal scale items are written to be extremely positive (e.g., genetic engineering is *always* acceptable) or extremely negative (genetic engineering is *never* acceptable), unfolding items are often written to be relatively neutral (genetic engineering is acceptable in *some circumstances but not others*). Conceptually, then, the degree to which a respondent agrees with a neutral unfolding item should represent how close they are to zero on the underlying trait continuum (Hoijtink, 1991). The more that a respondent disagrees with an unfolding item, the farther from zero they should be. Item agreement/disagreement, in the ideal-point sense, is a function of the distance between a person’s location on theta and an item’s location on theta with increased distance leading to higher probabilities of disagreement (Andrich & Styles, 1998).

Because unfolding items are difficult to interpret and score with Classical Test Theory methods, they are typically avoided in scale construction (Chernyshenko et al., 2007). This difficulty stems from the fact that there are two types of disagreement with an unfolding item: *disagreement from above* and *disagreement from below* (Andrich & Styles, 1998; Roberts & Laughlin, 1996). Disagreement from above occurs when someone is high enough on the theta continuum ($\theta = +3$, say) that a neutral unfolding item is not strong enough for them to endorse. Disagreement from below occurs when someone is low enough on the theta continuum ($\theta = -3$, say) that a neutral unfolding item is too strong for them to endorse. Although this would be impossible to score in a Likert scale, certain IRT models allow for such analysis.

The GGUM IRT model (Roberts et al., 2000), as an example, is designed specifically to model unfolding item response patterns. The GGUM is expressed as:
which describes the probability of an observed response \( (Z_i) \) to item \( i \) given person \( j \)'s location on the latent trait continuum, \( \theta \). The \( \alpha \) and \( \delta \) parameters refer to the discrimination and location of item \( i \). The \( \tau \) parameter denotes the location of every \( k \) response option boundary. \( C \) is the value assigned to the highest response option, \( M \) is the number of subjective response functions, defined as \( M = 2C + 1 \), and \( w \) is a response option index variable. The probability of a person selecting an observable response such as strongly agree, is the summed probabilities of strongly agreeing from above and strongly agreeing from below a particular item. The model does this by calculating the distance between the person and item locations on \( \theta (\delta_i - \theta_j) \). A negative distance \((\delta_i < \theta_j)\) indicates that a person is agreeing/disagreeing from above; a positive distance \((\delta_i > \theta_j)\) indicates that a person is agreeing/disagreeing from below. In short, the GGUM provides the flexibility necessary to precisely assess the functioning of unfolding items in addition to traditionally-written items.

At a broader level, using neutral-tending unfolding items in a scale represents a Thurstonian method of measurement (Andrich, 1988; Drasgow, 2009). Thurstone (1928) proposed a method of measurement wherein items of varying strengths (including neutral-tending items) and respondents reside on the same scale. This is in contrast to the ubiquitous Likert (1932) method of measurement where people’s responses to extremely-worded items are additively compiled for comparison with one another. If one were to place Thurstonian scale items onto a latent trait continuum, they would appear relatively dispersed across the entirety of the continuum, including the center of the continuum. However, if one were to place Likert
scale items onto a latent trait continuum, they would appear densely concentrated around the poles of the trait continuum and there would be few or no items in center of the continuum (Andrich & Styles, 1998).

The Use of Unfolding Items in Scales

Given recent studies (Chernyshenko et al., 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006) that show the benefits of using the GGUM IRT model to analyze traditionally-designed (i.e., using the Likert paradigm) personality scales, Chernyshenko et al. suggest that even greater benefits would be derived from using an unfolding IRT model to analyze a scale created with the ideal-point concept in mind from the beginning. In an ideal-point (Thurstonian) scale, items are written to cover the entire range of the latent trait rather than just the extreme positive/negative regions. Embretson and Reise (2000, pp. 270-271) seem to endorse this notion, suggesting that, in order to create a scale with the most precision across all levels of theta, researchers should include items that discriminate at different points across a broad range of theta levels.

The Likert scale development process tends to exclude neutral-tending items. Generating scale items according to the Likert paradigm, therefore, may limit test information (Chernyshenko et al., 2007) which in turn limits measurement precision. Test information describes the cumulative ability of an item pool to discriminate between gradations of theta at different points along the continuum. The greater information a scale has at a given point on theta, the greater the scale’s ability to differentiate between finer gradations in attitudes. Overall, analyzing items via ideal-point modeling potentially allows for a richer, more accurate understanding of attitudes. Using an ideal-point model allows an item to be understood in an unrestrained way, without the assumption of monotonically-increasing item functioning and
Use of unfolding items in attitudinal scales is still quite rare despite advances in statistical modeling. As researchers have developed increasingly sophisticated models, the models are often evaluated initially with simulated data and via re-analysis of pre-existing attitudinal or personality scale data. In this way, researchers have been able to demonstrate the benefits of allowing items to unfold.

Andrich (1988), for example, was able to compare his new unfolding model to a traditionally-made Thurstone scale. Attitudes toward capital punishment were analyzed with both the unfolding model and a Thurstonian judgment group. As expected, the item location estimates derived from the unfolding model were quite similar to Thurstone scale values as determined by a judgment group, thus establishing a solid link between modern IRT models and Thurstone’s concept of measurement.

A study by Hoijtink (1991) used the same scale of attitudes toward capital punishment and found support for a second unfolding model. This model was then used to analyze a scale of people’s attitudes toward nuclear energy. Hoijtink found that the item location estimates, including those for neutral-tending unfolding items, were ordered as expected. Attitudes ranged from the desire to keep nuclear energy infrastructure in place to the desire to close every nuclear power plant worldwide. Of specific note are two near-neutral items in the scale written to assess ambivalent (simultaneous positive and negative) attitudes. The item location estimates for these two items were conceptually sound ($\theta = .02; \theta = .65$), furthering the prospect of using unfolding items to measure ambivalence.

Rost and Luo (1997) made a contribution to the research in their analysis of adolescents’
attitudes toward adults using a traditional and unfolding IRT model. They found that unfolding items, especially those with location estimates near the poles were being markedly skewed by the assumption of monotonicity. Four of 10 items changed signs, moving from negative to positive, when allowed to unfold. All but one item location estimate changed by at least .75 units when items unfolded.

Roberts et al. (2000) were the first to model attitudes using the GGUM model that allows for polytomous, rather than dichotomous (agree/disagree) responses. An earlier form of their model was, like previous models, tested with a scale of attitudes toward capital punishment (Roberts & Laughlin, 1996). The generalized form of the model was then used to analyze attitudes toward abortion. Estimated model parameters were able to predict expected attitude scores reasonably accurately; $r_s = .61$ to .85 between expected and observed responses across 20 items. The abortion scale items covered the latent continuum quite well. Many neutral-tending items showed definite unfolding, and were discriminating enough to add useful amounts of information to the overall scale. The authors concluded that, to the extent that unfolding items are (intentionally or unintentionally) included in a scale, a dominance IRT model will inaccurately estimate the location of people toward the continuum’s extremes. The Roberts et al. study demonstrated that positive, negative, and neutral unfolding items can (1) coexist and contribute to the overall functioning of an attitudinal scale, and (2) form a scale that spans all regions of theta, including the scale’s midpoint.

It seems clear that using neutral-tending unfolding items, in concordance with an unfolding IRT model, can accurately model response data. Most recently, Chernyshenko et al. (2007) advocated the construction of scales that fully utilize unfolding items. Subsequently, Lin et al. (2009) followed the advice set forth by Chernyshenko et al. when constructing their scale
measuring attitudes toward workplace drug testing, described in the following section. Despite the promise of increased information and improved precision of measurement achieved with the use of unfolding items, data from Lin et al. study show that incorporating meaningful unfolding items may be difficult.

**Analyzing Attitudes toward Drug Testing**

Scale items with corresponding GGUM-derived (Roberts, Fang, Cui, & Wang, 2006) location estimates \(\delta_i\) calibrated using the sample from Lin et al. (2009) are presented in Appendix A, ordered from lowest to highest location estimate. Responses were collected using five response categories: 1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, and 5=strongly agree. At first glance, it may appear as though the items are relatively spread out over the continuum. Notice, however, that no items are located between \(-.50\) and \(+1.61\) on \(\theta\), creating a gap of 2.11 units near the continuum midpoint. Andrich and Styles (1998) suggest that such a gap is typical of scales created via the Likert paradigm. This is quite concerning given that many items were generated specifically to unfold at a near-neutral point on the continuum. What follows is an explanation of why it is believed that biased participant responses caused unreasonable parameter estimates in this scale.

Evidence gathered from Lin et al. (2009) suggests that items having a location \(\delta_i\) near the continuum midpoint provide very little information. Figure 2 shows the item characteristic curves and corresponding item information functions for three of the drug testing scale items. Part A of Figure 2 depicts the neutral-tending unfolding item located closest to zero on theta. Part B of the figure depicts a second neutral-tending unfolding item, this one located about 2 units away from zero. Notice that this second item shows more marked unfolding (peakedness) and provides more information about participant responses. Finally, Part C of Figure 2 depicts a
traditional, extremely-worded item. The extremely-worded item provides the most information about participant responses.

The trend seen in Figure 2 is the same trend seen in the scale as a whole: items with location estimates near zero on theta provide only limited amounts of information. That is, the neutral-tending unfolding items in this scale do not contribute much to the scale’s ability to differentiate between gradations in respondent attitudes. Item information in the GGUM is jointly determined by two parameters: $\alpha$ and $\psi$ (Roberts et al., 2000). The alpha ($\alpha$) parameter is a measure of how well an item’s response options discriminate between people with various latent trait scores. A high alpha estimate indicates that a person’s response to that item is very useful in determining that person’s attitude toward drug testing. The second parameter, psi ($\psi$), indicates the degree to which an item’s response options are spread out across theta. Specifically, $\psi$ is mean distance between the thresholds ($\tau$s) that separate an item’s response options. A high psi estimate indicates that response options are widely dispersed across theta. Because the vast majority of participants are between +3 and −3 on theta, an item provides the most information when psi is low and response option thresholds are relatively condensed. In sum, maximum information between ± 3 units is achieved when $\alpha$ is large and $\psi$ is small.

The “comparison” rows in Table 3 show the mean item parameter estimates from the Lin et al. sample for positively-worded, negatively-worded, and unfolding items. The positive, negative, and unfolding item-type categories were assigned during the item-writing phase of scale development. These categories may be somewhat artificial in that they represent more of a continuum than discrete categorical differences. Nonetheless, comparing items in this manner provides an excellent comparison of item functioning by item type. Notice that the unfolding items provide the lowest alpha estimates (and thus the least amount of discrimination). Based on
this pattern of data, there is reason to believe that unfolding items provided less information due to their diminished capacity to discriminate between gradations in people’s theta levels, compared to other item types. Unfolding item response options did not appear to be more spread out than other item types, indicating that the psi parameter is probably not the cause of unfolding items’ reduced information.

Another problem is that several unfolding scale items do not logically fall into place on the latent trait continuum, as compared to surrounding items. It defies logic, for instance, that the item “I am somewhat opposed to drug testing” ($\delta_i = -4.06$) is estimated to be more extreme than the item “drug testing is a violation of a person’s privacy” ($\delta_i = -3.57$). It is possible that this is simply estimation error, but it is also possible that this suggests some problems with the scale.

The evidence presented so far suggests that something is flawed in the way participants responded to the unfolding items in the drug testing scale. It is at this point that research from other disciplines becomes extremely helpful. What follows is a broad discussion of the cognitive processes participants use to respond to scale items, how these processes are influenced by characteristics of test delivery, and, finally, how attitude structure should play a role in item response patterns if unfolding items function as intended.

*The Cognitive Demands of Responding to Unfolding Items*

The section focuses on the cognitive processes used by participants to respond to survey items, with the goal of using this research to better understand the participant response processes used when completing the drug testing scale. Tourangeau (1984) proposed a four-step process by which participants respond to items. Participants (1) interpret the question, (2) retrieve a response from memory, (3) form a judgment on the information retrieved, and (4) report this
judged response. Krosnick (1991; 1999) added to this theory, stating that there are differences in the degree to which a respondent thoughtfully and carefully goes through each of these stages. Respondents who put forth adequate cognitive effort when responding to scale items are said to be *optimizing*; those who use mental shortcuts or otherwise fail to exert much cognitive effort are said to be *satisficing*. For example, Krosnick (1991; 1998) has shown that participants who satisfice are more likely to use neutral response options (e.g. *no opinion* or *neither agree nor disagree*) and less likely to differentiate between individual scale items.

There are a few factors that are likely to determine how much mental effort a respondent uses when filling out a survey. For example, it is likely that individual characteristics (e.g., need for cognition) have some bearing on the amount of effort expelled (Petty & Jarvis, 1996). More relevant to this study, however, is the finding that satisficing increases when participants are asked to perform cognitively difficult tasks (Krosnick, 1991). Krosnick (p. 548) indicated that “task difficulty is a function of the difficulty of interpreting the meaning of a question and response choices.” It seems apparent that interpreting an unfolding item (e.g. “drug tests are okay for entry-level positions, but not management-level positions”) is more difficult than interpreting a traditional, extremely-worded scale item. Even a simply-worded unfolding item such as “employer drug testing is wrong in certain situations,” requires simultaneous evaluation of situations in which drug testing is, and is not, acceptable.

A likely reason that respondents experience unfolding items as cognitively demanding is that they are composed of two disparate (positive and negative) components that must be mentally *decomposed*, or broken down into smaller steps, and comparatively weighed. Some unfolding items are double-barreled and overtly ask a person to consider positive and negative components of their attitude; other unfolding items covertly assess positive and negative
components by asking a person to state how neutral they feel toward a topic. Evidence suggests that this decomposition process is both difficult and prone to error. In a study by Armstrong, Denniston, and Gordon (1975) participants were asked to either (1) mentally decompose and calculate or (2) write out a series of decomposed steps and then mentally calculate, to solve multi-part problems. They found that participants who took the time to think through and write out problems’ components step-by-step were significantly more accurate in their responses. Further evidence shows that evaluations requiring integration of both positive and negative facets require more mental calculation time than evaluations of purely positive or purely negative facets (van Harreveld, van der Pligt, de Vries, Wenneker, & Verhue, 2004). These findings suggest that mentally decomposing and calculating a response to a neutral-tending unfolding item is cognitively difficult. Kuncel (1973) notes that it is more cognitively demanding for a person to select the appropriate response option to an item that is proximal, rather than distal, to the person’s location on the attitude being measured. Therefore, the cognitive demands associated with answering neutral-tending unfolding items may be magnified for someone who is, themselves, nearly neutral (zero) on the attitude being measured.

The drug testing scale with its mixture of positive, negative, and neutral-tending unfolding items, presents a rather difficult set of cognitive challenges to the respondent. It therefore seems likely that many respondents satisficed, using mental shortcuts rather than deliberate consideration to respond to items. A case will now be made for one particular mental shortcut implicated in the lackluster results of the drug testing scale. It is believed that a high number of participants relied on simplistic context effects to respond to scale items.

Context effects arise when respondents focus not on what an individual item is asking, but rather on surrounding (contextual) information (Strack & Martin, 1987; Tourangeau &
Rasinski, 1988). Rather than taking the time to interpret every single item, respondents make judgments about the meaning of the scale and rely on those simplistic, overall meanings when responding to items. Context effects are a type of error generated by the measurement instrument (also known as method variance; see Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Stated explicitly, context effects create biased item response data.

Participants are known to extract contextual information from many different sources. Research shows that scale items listed before (Schwarz, Strack, Hippler, & Bishop, 1991; Strack, 1992) and after (Schwarz & Hippler, 1995) the current item of interest can bias responses. Research also shows that participants garner contextual information from the numeric rating scale used (Schwarz & Hippler, 1991). Evidence indicates that people absorb and accumulate contextual information as they complete scale items (Krosnick & Alwin, 1987) leading to increased response consistency (Knowles, 1988). Furthermore, there is evidence demonstrating that participants are more likely to use contextual information when they encounter ambiguous, as opposed to unambiguous, items (Strack, Schwarz, & Wänke, 1991).

Taken collectively, evidence on context effects suggests that, when respondents encounter an ambiguous item (i.e. an unfolding item), they look to other nearby sources for help. Unsure how to respond to an unfolding item, a respondent may look to the previous or subsequent item(s) and assume that a response of “1” means disfavoring drug testing whereas “5” means favoring it. Moreover, exposure to many unfolding items in a single questionnaire may induce more satisficing, leading to an increased reliance on simplistic contextual information.
Reducing Context Effects by Manipulating Test Delivery (Hypotheses 1 to 4)

Research has shown how much information a respondent can absorb (and subsequently use) when scale items are presented together on a single sheet of paper or on a single computer screen. Budd (1987) showed that three scales had drastically different coefficient alphas when whole scales were presented sequentially (.84, .79, and .78) versus when scale items were intermixed (.47, .52, and .50). Other studies (e.g., Harrison & McLaughlin, 1996) show less drastic, but uniformly lower coefficient alpha estimates when items from multiple scales are intermixed. Additionally, consider what happens when individual scale items are removed from their context (other scale items) and delivered to participants one at a time. Tourangeau, Couper, and Conrad (2004) showed that coefficient alpha for an eight-item scale significantly decreased (.62 to .51) when items were presented individually. When scale items were presented individually, participants showed increased variability in responses such that they give their most-common (modal) response less frequently. In other words, intra-individual variability in responding seemed to increase when items were presented one at a time. Thus, presenting items singly seems likely to induce thoughtful responses to individual items thereby reducing the context bias presented by the presence of other scale items.

**Hypothesis 1:** Participants who are administered drug testing scale items singly, as opposed to together, will optimize response behaviors and demonstrate less context effect bias, resulting in increased intra-individual response variability.

To the extent that participants optimize response behaviors and differentiate between items, the scale’s neutral-tending unfolding items should show more marked unfolding.

**Hypothesis 2:** When drug testing scale items are administered singly, the scale’s
neutral-tending unfolding items will show more marked unfolding.

Not only should the scale’s neutral-tending unfolding items show more marked unfolding, but other item and scale properties should improve as well. If the manipulation decreases the amount of error present in the responses to unfolding items, those items should then provide greater information about a person’s attitude. If neutral-tending unfolding items show more marked unfolding, the item locations should become closer to the midpoint

**Hypothesis 3:** When drug testing scale items are administered singly, unfolding items will show increased information relative to when items were administered together.

**Hypothesis 4:** When drug testing scale items are administered singly, unfolding items will shift toward the center of the latent trait continuum, creating more uniform coverage between +2 and −2.

*The Effects of Attitudinal Ambivalence and Indifference (Hypotheses 5 - 8)*

The use of unfolding items to investigate the structure of attitudes is important. Coombs, who developed the ideal-point theory of data, pointed out just how integral positive (approach) and negative (avoidance) aspects of attitudes are to the unfolding process (Coombs & Avrunin, 1977). Whereas a Likert scale’s monotonic items only allow for bipolar measurement along a favorable-disfavorable continuum, unfolding items allow for simultaneous evaluation of favorable and disfavorable aspects of an attitude.

Consider the meaning assigned to Likert scale scores: high scores indicate that someone favors a particular object and low scores indicate that someone opposes that same object. It is impossible to tell what an average score indicates, though. It is therefore difficult to describe a middling respondent. An average score could mean a person has no feelings, positive or
negative, about the object. Alternatively, the score could mean the person has a variety of positive and negative feelings about the object which cancel one another out. The former case refers to respondents who are indifferent toward an object; the latter case refers to respondents who are ambivalent toward an object. Unfortunately, the bipolar nature of Likert scales cannot satisfactorily distinguish ambivalent respondents from indifferent respondents (Cacioppo, Gardner & Berntson, 1997). Hypotheses 5 through 7, below, are proposed in an attempt to disentangle ambivalence and indifference, and determine how these attitudinal constructs relate to item unfolding.

Attitudes that simultaneously take into account positive and negative aspects are known as ambivalent (Kaplan, 1972). Tetlock (1986) proposed a model of ambivalent attitudes based upon a process of value pluralism. That is, when people mentally derive an attitude, they compare relative weights placed upon competing values. When considering whether or not they favor workplace drug testing, people may experience a mental tradeoff between competing positive and negative values. On the one hand, testing may facilitate a safe work environment and reduce employee drug use. On the other hand, testing may infringe upon a person’s right to privacy and individual freedoms.

Figure 3 illustrates five hypothetical respondents’ positive and negative evaluations of workplace drug testing. Person A and Person B have essentially univalent (one-sided) attitudes toward drug testing. The strength of Person A’s positive evaluations, for example, greatly outweigh the strength of his negative evaluations, resulting in an overall favorable attitude toward drug testing. Person C and D, are ambivalent, rather than univalent, because both have at least a moderate amount of positive and negative evaluations and the strength of their positive and negative evaluations are roughly equal. It is fair to say that Person C and D have “mixed
feelings” toward drug testing.

Theoretically, an unfolding item is most likely to be endorsed by a respondent when that respondent harbors a mixture of positive and negative feelings about an object. Stated simply, ambivalent people should endorse unfolding items. Consider how two people might respond to the unfolding item “drug tests are okay for entry-level positions but not management-level positions.” A univalent person who strongly favors drug testing should respond *strongly disagree* because they feel that the item is not strong enough for them to endorse. This person thinks that drug tests are okay for *all* employees, including management-level positions, and thus disagrees with the item. An ambivalent person, however, should respond *strongly agree* because they have the type of mixed feelings that are elicited by the item. This logic applies to all unfolding items, and it is thus expected that ambivalent, but not univalent, participants will endorse unfolding items.

**Hypothesis 5:** Participants with ambivalent attitudes will endorse more unfolding items than participants with univalent (one-sided) attitudes.

As shown in Figure 3, there are different levels of ambivalence. At the extreme, Person C has high-strength positive and negative evaluations, indicating that the person feels strongly about the costs and benefits of drug testing, yet remains torn on the issue. Person D has moderate-strength evaluations, indicating that the person feels somewhat less strong about the costs and benefits of drug testing. As an ambivalent respondent’s attitude strength increases, it seems reasonable to assume that the number and intensity of mixed feelings experienced should also increase.

**Hypothesis 6:** For ambivalent respondents, there will be a positive linear relationship between attitude strength and number of unfolding items endorsed.
Participants who report feeling much ambivalence about workplace drug testing will endorse more unfolding items than those who report feeling little ambivalence.

At this point it is necessary to distinguish between the interrelated concepts of ambivalence and indifference. As stated, ambivalence occurs when a person has simultaneous positive and negative evaluations. Indifference, on the other hand, occurs when a respondent simply has no discernable opinion, positive or negative, about a target object. Whereas ambivalent people have feelings (albeit mixed feelings) about a topic, indifferent people are unlikely to care about the topic at all (Fazio, 1995). Person E in Figure 3 appears to be indifferent toward drug testing because she lacks any real attitude strength.

Because neutral-tending unfolding items elicit mixed feelings, they should not be endorsed by indifferent respondents who hold no opinions on drug testing. Having said that, it is admittedly quite debatable how an indifferent participant will respond to unfolding items. It seems logical, though, that an indifferent person would not endorse the item. They may respond with Neither Agree nor Disagree or Strongly Disagree to indicate that they do not have the mixed feelings elicited by the item. The claim made here is that ambivalent participants, as a group, should respond homogenously to neutral-tending unfolding items (by endorsing them) whereas indifferent participants will be erratic in their responses to these items.

**Hypothesis 7:** Participants who are indifferent about workplace drug testing will be less likely to endorse neutral-tending unfolding items than those who are ambivalent.

*Internal Validity Check: Attitude Extremity (Hypothesis 8)*

In order to ensure that participants are accurately being measured by the drug testing
scale, an internal validity check is in order. This is especially warranted given that the proposed manipulation is supposed to elicit “true” responses to scale items by removing context error. The test delivery manipulation should not create unusual responses; rather, it should increase the accuracy and validity of responses, increasing the model’s ability to estimate item and person parameters. Attitude extremity is, very simply, the intensity of a person’s attitude. If all possible attitudes are laid out onto a continuum, extremity refers to the absolute value of the distance away from the scale’s midpoint (Abelson, 1995).

**Hypothesis 8:** There will be a positive linear relationship between participant attitude extremity and GGUM-derived person location estimates. Participants with extreme attitudes will have person locations toward the poles of the latent trait continuum; participants with weak attitudes will have person location estimates near the scale midpoint.
CHAPTER III: METHODS

Participants and Procedures

Two samples were used in the present study. One sample, referred to as the comparison sample, was collected by Lin et al. (2009) and serves primarily as a point-of-comparison for an experimental sample. The comparison sample consists of $N = 566$ university undergraduate students who were given extra credit or class credit for participation. The mean respondent age was 18.9, $SD = 1.3$. Eighty-two percent of respondents were Caucasian and 11% were African-American. This sample was administered the workplace drug testing scale in its original pencil-and-paper block format.

The second sample, referred to as the experimental sample, was collected by this study’s author. The experimental sample consists of $N = 601$ university undergraduate students from the same university as the first sample, and, in the same manner as the comparison group, participants were given extra credit or class credit for participation. The mean participant age was 19.3, $SD = 2.4$. Eighty-nine percent were Caucasian and seven percent were African-American. Rather than responding to the scale items while being able to see the rest of the scale, as was done in Lin et al., participants in this sample were administered the workplace drug testing items via computer, with each item displayed singly, on its own screen. Further, participants had to wait six seconds before they could respond to each item. This was meant to slow participants down and create thoughtful, rather than hurried, responding. The survey program did not allow participants to go back to previous screens. After responding to the drug testing scale items, participants in this sample were asked to respond to a series of scales measuring their ambivalence, indifference, and attitude extremity.
Instruments

The first two clusters of instruments described here were used to identify people with ambivalent and indifferent attitudes. Because no single instrument is capable of making the distinction between ambivalence and indifference, the two constructs are measured separately with different scales. Measurement is complicated by the fact that neither construct has a single, widely-accepted method of measurement. In fact, most published research seems to rely on the combined results of multiple instruments to assess ambivalence or indifference. Likewise, the present study uses two measures of experienced ambivalence as indicators of ambivalence, and three measures of attitude strength as indicators of indifference.

Measuring Ambivalence. There are two methods typically used to assess a person’s level of attitudinal ambivalence: meta-attitudinal and formula-based scores (Bassili, 1996; Jonas, Broemer, & Diehl, 2000). A meta-attitudinal measure assumes a person has insights into the amount of “mixed feelings” they have about a target object. A formula-based measure, on the other hand, uses a split semantic differential technique where positive and negative evaluations are assessed separately and compared (Kaplan, 1972). Both of the above techniques were used because research indicates that they account for unique variance (Bassili, 1996).

The items used to measure ambivalence are listed in Appendix B. Eight items from a study by Cacioppo, Gardner, and Berntson (1997) asked participants to report how much ambivalence they feel when thinking about workplace drug testing (the meta-attitudinal measure). When a participant reports feeling many ambivalent thoughts, this suggests that they are ambivalent, rather than indifferent, about drug testing. Two items adapted from Kaplan (1972; see also Bassili, 1996) were used to separately measure participants’ positive and negative evaluations toward workplace drug testing (the formula-based measure). As shown in Figure 3,
when a participant has non-zero positive and negative evaluations that are nearly equal, the person is said to be ambivalent rather than indifferent (Jonas et al., 2000).

Measuring Attitude Strength. Attitude strength is the degree to which someone knows how they feel about a target object. Recall from Figure 3 that indifference, by definition, means having little or no attitude strength. Scale items (listed in Appendix C) from three attitude strength components were used to measure attitude strength: self-reported cognitive effort, attitude clarity, and attitude correctness.

First, the measure of self-reported cognitive effort asked participants to report how much effort they put forth when they responded to survey items. Self-reported cognitive effort is a mixture of two scales as described in Wegener, Downing, Krosnick, and Petty (1995). Participants were asked how many thoughts they generated about scale items following a procedure from Barta and Ray (1986). Participants were also asked how much effort they put into answering the scale items following a procedure from Petty, Harkins, Williams, and Latané (1977). Because participants who are indifferent need not engage in mental decomposition and weighting as other participants would, they should report having used relatively little cognitive effort.

A measure of attitude clarity asked participants how sure they are that they know what their attitude toward a target object truly is (Petrocelli, Tormala, & Rucker, 2007). A related measure, attitude correctness, asked participants how sure they are that their attitude is valid (Petrocelli et al.). To the extent that someone does not know their attitude and does not have confidence in their attitude, they are thought to be indifferent toward workplace drug testing.

Internal Validity Check: Measuring Attitude Extremity. To measure attitude extremity, participants were asked to consider and then rate their overall attitude toward workplace drug
testing. Because extremity is a simple construct, it was measured with a single item: “overall, how do you feel about workplace drug testing?” To allow for fine gradations in favor or opposition toward drug testing, responses were collected on an 11-point scale.
CHAPTER IV: RESULTS

Preliminary Assessments of Fit

Option Response Analysis. An initial review of the option response and item response functions was performed before conducting any analyses. This review revealed a particular pattern in option responses specific to neutral-tending unfolding items. The strongly agree response option seemed to be ineffective at capturing endorsement of unfolding items. That is, very few respondents chose this option as their response to the unfolding items. Figure 4 shows three representative neutral-tending unfolding items. Notice that, even at the higher points of the theta continuum (i.e. + 2.5), a respondent is more likely to select somewhat agree than strongly agree. Compare this to two option response functions from extremely-worded positive items, shown in Figure 5. In the typical extremely-worded item, the strongly agree/disagree option almost always becomes the most probable response in the higher ranges of theta, yet this does not seem to be the case with neutral-tending unfolding items. In many unfolding items, strongly agree was completely subsumed by somewhat agree. Santor and Ramsey (1998) call this particular condition “option ineffectiveness” and suggest that the situation be remedied by collapsing response options such that the ineffective option is merged with the subsuming option. Strongly agree and somewhat agree were thus collapsed into a single response option.

Item-level Fit Analyses. To assess item fit, response data from the experimental and comparison samples were calibrated separately in GGUM 2004 (Roberts, Fang, Cui, & Wang, 2006). The obtained item parameter estimates were used to generate expected response patterns which were compared to the observed response patterns using the computer program MODFIT (Stark, 2001). A chi-square based index of fit was used, with higher values being indicative of divergence between the expected and observed response pattern for a particular item. Other
researchers (e.g., Drasgow, Chernyshenko, & Stark, in press) have argued that a fit value greater than 3 represents a meaningful amount of mis-fit. The mean item fit was .29 in the comparison group and slightly higher, .36, in the experimental group, with no values greater than 2 in either condition. In addition to testing each item individually, MODFIT also analyzes the fit of two- and three-item sets, known as doubles and triples. The mean fit value of item doubles was 1.66 in the comparison group and 1.18 in the experimental group. Six item doubles in the comparison group were greater than 3, whereas one item in the experimental condition was greater than 3. The mean fit value of item triples was 1.89 in the comparison group and 1.38 in the experimental group. Three item triples in the comparison group were greater than 3, whereas two items in the experimental group were greater than 3. The relatively small mean fit values, and the small number of item sets (doubles or triples) exceeding 3 indicates a good degree of fit between expected and observed responses in the vast majority of the 48 items. No particular items displayed abnormally high mis-fit and therefore all items were included in subsequent analyses.

Effects of Presenting Items Singly (Hypotheses 1 to 4)

The results of Hypotheses 1 through 4 compared data from the experimental sample to that from the comparison sample to determine the effectiveness of the test delivery manipulation.

Hypothesis 1. To test whether respondents in the experimental condition exhibited more intra-individual variability in responses than those in the comparison condition, nondifferentiation scores were calculated. Tourangeau, Couper, and Conrad’s (2004) measure of respondent nondifferentiation was used to measure of intra-individual variability. A nondifferentiation score is calculated by first determining a given respondent’s modal response, and then calculating the proportion of items to which that respondent give his or her modal response. With 48 items and 5 response options, nondifferentiation scores had a possible range
of .23 (modal response was observed in 11 of 48 items) to 1.00 (modal response was observed in 48 of 48 items). In the experimental condition, the mean nondifferentiation score was .40 ($SD = .08$). In the comparison condition, the mean nondifferentiation score was .41 ($SD = .11$). Although the magnitude of the difference between conditions is quite small ($d = .10$), an independent samples t-test showed that scores were significantly different between samples, $t(1,061) = 1.96, p < .05)$. Hypothesis 1 was supported.

**DIF Analysis.** Hypothesis 2 and 3 predicted varying degrees of differential item functioning (DIF) in the neutral-tending unfolding items. DIF was expected when comparing the same item as calibrated in the experimental versus comparison groups. Before proceeding with these item-level analyses, an omnibus test of DIF was conducted, as suggested by Carter, Dalal, and Zickar (2010) and Roberts and Gordon (2008). Roberts and Gordon suggest using a likelihood ratio that compares marginal maximum likelihood estimates of model-fit when item parameters are constrained to be equal between samples and when those parameters are allowed to vary across samples. Fourteen items – those that were constructed to be neutral-tending unfolding – were suspected to have DIF and were therefore incorporated into the omnibus test. The omnibus test was significant, $G^2 = 375.88, df = 70, p < .001$. Because the results suggested that there was DIF between the models, follow-up item-level DIF analyses were conducted. Results of post-hoc item-level analyses, shown in Table 2, indicate that seven of 14 suspected items did indeed exhibit a significant amount of DIF. Taken alone, the presence of DIF neither proves nor disproves the any particular hypotheses; it simply indicates which items have parameters that significantly differ between the experimental and comparison samples.

**Hypothesis 2.** To determine whether the seven significant-DIF items exhibited unfolding characteristics, item response functions were plotted from $+3$ to $-3$ across the theta continuum.
Item response functions for all seven items showed at least a minimal amount of non-
monotonicity and item locations that were between +2 and −2 on theta, thus indicating that these
items are indeed neutral-tending unfolding items (refer to Table 1). Of interest here is how item
characteristics differ between the experimental and comparison conditions.

Figure 6 shows the comparative item response functions of significant-DIF items as
calibrated in both conditions. Two items in the experimental condition show greater unfolding
(Figure 6, Part A) and two items show less (Figure 6, Part B) unfolding, as compared to the
comparison condition. Three items (two of which are presented in Figure 6, Part C) do not show
marked unfolding in either condition. These three items are characterized by relatively
horizontal item response functions that do not seem to discriminate very well. In sum, two items
supported the hypothesis, two items contradicted the hypothesis, and three items neither
supported nor contradicted the hypothesis.

Because item response function analyses were inconclusive, the overall trends in
parameter estimates between groups were analyzed. Item parameters from the experimental
condition were equated to those from the comparison condition so that all parameters reside on
the same metric, allowing for meaningful comparison. The GGUMLINK program was used to
equate parameters using an item response function comparison technique (see Roberts & Huang,
2003). Roberts, Donoghue, and Laughlin (2000) note that unfolding increases as (1) alpha
increases and (2) psi decreases. If, across most unfolding items, alpha increased and psi
decreased in the unfolding condition, then it seems reasonable to conclude that unfolding
generally increased. Table 3 compares the parameter estimates obtained from the experimental
and comparison groups. With regard to unfolding items, there were no significant differences in
parameters between the experimental and comparison conditions. The inconclusive results
shown by the item response functions and the lack of difference between the parameters in each group indicates that there is little support for this hypothesis. Hypothesis 2 was not supported.

Hypothesis 3. To determine if item information is greater in the experimental group than the comparison group, item information functions from both groups were plotted for the seven significant-DIF items, as shown in Figure 7. Two of seven items showed increased information in the experimental condition (Part A of Figure 7) and five items showed decreased information in the experimental condition (two of which are presented in Part B of Figure 7), relative to the comparison.

The cumulative test information function was also examined to see if, on the whole, information was greater in the experimental condition. Figure 8 shows that there was a noticeable increase in the overall amount of information that the questionnaire provided in the experimental condition. The standard error of measurement, which is inversely related to test information, also shown in Figure 8, appears to be lower in the experimental condition. In sum, only two of the significant-DIF items showed improved information. In spite of that, the scale taken as a whole appeared to provide more information in the experimental condition. However, because cumulative test information is influenced by positive and negative (i.e. extremely-worded) items as well as unfolding items, the test information function does not provide strong enough evidence to support the hypothesis. Hypothesis 3 was not supported.

Hypothesis 4. Hypothesis 4 proposed that when the experimental group was used to calibrate the item parameters, item location parameters would shift toward the center of the theta continuum, creating more uniform coverage of the range from $+2$ to $-2$. In the experimental condition, 29 items have delta (item location) estimates between $+2$ and $-2$, which is more than the 21 in the comparison condition. Figure 9 shows the distribution of the delta parameters in
each condition. In both conditions, there is a gap present in the center of the continuum. Note, however, that the gap is smaller by .20 units in the experimental condition than in the comparison condition. Finally, as seen in Table 3 there was a general trend for items of all three types to have less polar delta estimates in the experimental condition than in the comparison condition. Hypothesis 4 was supported.

Effect of Attitude Structure (Hypotheses 5 to 8)

The measures used to assess attitude structure were not administered to the comparison sample, and thus the results of Hypotheses 5 through 8 rely solely on data from the experimental sample.

Some initial coding of the semantic differential variable was required before any analyses could be performed. The semantic differential is comprised of each respondent’s separately-assessed positive and negative feelings about drug testing, and these two dimensions served as the basis of an ambivalence measure and a strength measure. Semantic differential ambivalence was coded as follows: smaller dimension - larger dimension. This produced a metric with scores ranging from 0 (respondent is entirely univalent) to 3 (respondent is ambivalent). Semantic differential strength was coded as follows: positive dimension + negative dimension, producing a metric ranging from 0 (no strength whatsoever) to 6 (maximum strength).

To determine how attitude structure influences responses to unfolding items, many indicators of ambivalence and attitude strength were used in this study. A correlation matrix showing the relationship between these indicators and number of unfolding items endorsed is presented in Table 4. The coefficient alphas shown in the table, where applicable, indicate that the attitude structure measures had acceptable internal consistency.

Hypothesis 5. To determine whether ambivalent respondents endorsed more unfolding items than univalent respondents, the pattern of correlations in Table 4 was examined. Number
of unfolding items endorsed was significantly related to participants’ meta-attitudinal experienced ambivalence score \( (r = .14, p < .01) \) and formula-based semantic differential ambivalence score \( (r = .17, p < .01) \). Additionally, the number of unfolding items a respondent endorsed was significantly and inversely related to participants’ attitude clarity \( (r = - .10, p < .05) \) and extremity \( (r = - .16, p < .01) \). This pattern supports the proposition that ambivalent respondents were more likely than univalent respondents to endorse unfolding items. Hypothesis 5 was supported.

**Hypothesis 6.** To determine if high-strength ambivalent people endorsed more unfolding items than low-strength ambivalent people, experienced ambivalence was hypothesized to have a direct effect on number of unfolding items endorsed, with the semantic differential attitude strength score acting as a mediator of this relationship. The Sobel (1982) test of mediation was conducted using the unstandardized regression coefficients of experienced ambivalence \( (b = .01, SE = .005, p < .05) \) and semantic differential strength \( (b = .36, SE = .08, p < .01) \) scores on number of unfolding items endorsed. The resulting two-tailed Sobel test of mediation was significant, \( Z = 2.09, p < .05 \). Hypothesis 6 was supported.

**Hypothesis 7.** Multiple attitude structure indicators were use to determine if ambivalent respondents endorsed more unfolding items than did indifferent respondents. First, attitude extremity was used to differentiate people who were univalent from those who were not. Those with an attitude extremity score greater than 3 on a scale of 0 (least extreme) to 5 (most extreme) were deemed univalent and thus excluded from this analysis. The remaining group, then, should be comprised of people who are a mixture of ambivalent and indifferent. Three variables should, in theory, be able to distinguish indifference from ambivalence: mental effort required, number of thoughts generated, and experienced ambivalence. Bivariate correlations were run to test the
relationship between these three indicators of ambivalence and the number of unfolding items endorsed. There were small but significant relationships for two of these indicators — mental effort \((r = .11, p < .05)\) and thoughts generated \((r = .13, p < .01)\), and a non-significant relationship for the third indicator — experienced ambivalence \((r = .05, ns)\). Given that two of three ambivalence indicators show a slightly positive, significant relationship with the number of neutral-tending unfolding items endorsed, hypothesis 7 was supported.

**Hypothesis 8.** Hypothesis 8 was meant to serve as an internal validity check to ensure that respondent attitudes were being accurately assessed by the statistical model. Indeed, participant self-rated overall attitude extremity was strongly correlated with GGUM-derived theta estimates, \(r = .83, p < .001\). Hypothesis 8 was supported.

**Person Fit.** Person fit is the degree to which an individual respondent’s observed responses equal the responses predicted by the GGUM IRT model. Person fit is a chi-square metric with higher values indicating higher degrees of mis-fit. Results from the above hypotheses indicate that individual differences in attitude structure can play a role in individuals’ responses to unfolding items. It is of practical importance to know which, if any, of these attitude structure variables lead to response patterns that do not fit the model.

Although there are multiple measures of person mis-fit provided by the GGUM 2004 (Roberts, Fang, Cui, & Wang, 2006) computer program, all indices were very highly correlated in this sample, a finding that has been corroborated by Lin et al. (2009). A measure of *infit* was arbitrarily chosen to represent the degree of person mis-fit in these analyses, but the results do not differ based on the specific index chosen. Table 5 shows the regression coefficient estimates used to predict person mis-fit. Note that three attitude variables had significant curvilinear regression terms, perhaps due to the exponential nature of mis-fit values. Several of these
variables predicted person mis-fit fairly well, with the highest *multiple R* being .20. Five of the eight variables examined here showed significant relationships to person mis-fit. Figure 10 graphically depicts the relationship of semantic differential ambivalence and mental effort — the two strongest predictors — with person mis-fit.

Based on the person mis-fit analysis, it appears that high scores on ambivalence indicators (i.e., number of unfolding items endorsed, semantic differential ambivalence, required cognitive effort, and number of thoughts generated) were positively associated with person mis-fit. On the other hand, a strength indicator (i.e., extremity) was negatively related to mis-fit. Ambivalent people, it seems, were more likely than univalent people to exhibit response patterns that were discrepant from those predicted by the statistical model.
CHAPTER V: DISCUSSION

Unfolding scales incorporate both extremely-worded and neutral items into a single scale, analogous to the technique proposed by Thurstone (1928). Analysis and interpretation of unfolding scales has been aided by advanced IRT models such as the GGUM (Roberts, Donoghue, & Laughlin, 2000). Some previous research suggests that unfolding scales have the potential to offer increased information and precision of measurement compared to Likert-type scales. However, research on item-level unfolding is scarce. This research contributes to the literature by attempting to elucidate some of the psychological mechanisms — test delivery and respondent characteristics — that underlie unfolding response patterns.

Test Delivery Method. Based on the proposition that neutral-tending unfolding items present a cognitive challenge to respondents, thereby resulting in across-item context bias, a test delivery manipulation presented scale items to respondents one-by-one with a small time delay before an item response could be entered. Although this manipulation slightly increased intra-individual response variability, it did not induce a response pattern that led to uniformly more marked item unfolding. The results of this manipulation were quite mixed: some neutral-tending items showed more marked unfolding in the experimental condition (as compared to the comparison condition), but other items showed markedly less unfolding in the experimental condition. Similarly, some neutral-tending items showed increased information in the experimental condition, but other items showed decreased information. The hypothesized effects of the manipulation were generally not supported by the data.

There are a couple possible reasons for the lackluster results of this portion of the study. First, it is possible that the manipulation failed to stop respondents from satisficing. Respondents may have continued to satisfice in the experimental condition, perhaps using the mental heuristic
that 1 = disfavoring drug testing and 5 = favoring drug testing to respond to all items. It is possible that such a heuristic is formed after seeing just a few scale items, and thus presenting one item at a time would not have prevented the use of this heuristic. Second, it is possible that the neutral-tending unfolding items in this particular scale are simply ineffective at discriminating between people with different attitude levels. Items that are inherently non-discriminating would not likely be improved by any sort of test delivery manipulation.

A third possible reason that the manipulation failed has to do with a methodological weakness associated with the test delivery manipulation: the absence of a true control group. Respondents in the experimental group were compared to respondents in a comparison group, comprised of data collected earlier by other researchers. Respondents were not randomly assigned to experimental and control conditions, leaving the possibility that the results were influenced by differing characteristics between the samples. Although there is no reason to suspect any key characteristic differences between samples, given that there was no random assignment, this cannot be ruled out.

There were, however, two potentially useful findings that arose from this manipulation. First is the finding that scale-level information was greater across the majority of the theta continuum in the experimental condition than in the comparison condition. Presenting items one-by-one with a small time delay seemed to produce responses that were less contaminated by error. Given the available data, it is not possible to tell whether this is the result of reduced context bias from presenting items singly, or whether it is the result of increased mindfulness from forcing respondents to wait a few seconds before entering a response to each item. Nonetheless, researchers or practitioners may consider using similar test delivery protocols when response accuracy is of critical importance.
The second potentially useful finding is that the gap in item locations that existed in the center of the theta continuum in the comparison condition was decreased in the experimental condition. The experimental manipulation seemed to move item locations slightly toward the center of the theta continuum. Interestingly, a similar result — less extreme item location estimates — is obtained when items contain vague quantifiers such as the word “sometimes” (Carter, Dalal, & Zickar, 2010). The presence of vague quantifiers in an item would seemingly require more mental calculation time from respondents than if the quantifier was absent (see van Harreveld et al., 2004). Thus, one interpretation of the similar results in this study and Carter, Dalal, and Zickar is that the presence of vague quantifiers causes participants to spend more time on the item, creating a time-delay somewhat similar to the forced six-second delay in this study. This seems to lend some support to the proposition made by Drasgow, Chernyshenko, and Stark (in press) that introspection plays a critical role in the unfolding response process.

**Respondent Attitude Structure.** The data generally supported the hypotheses made with regard to the relationship between attitude structure and responses to unfolding items. First, ambivalent respondents endorsed more unfolding items than did univalent respondents. Further, strongly ambivalent respondents (those with a lot of attitude strength) endorsed more unfolding items than did weakly ambivalent respondents (those with low attitude strength). Finally, ambivalent respondents endorsed more unfolding items than did indifferent respondents. On the whole, these results seem to provide some good preliminary evidence that attitude structure plays an important role in determining responses to neutral-tending unfolding items. That said, there are a few areas of concern regarding these results.

One readily apparent issue with these findings is the small effect sizes that characterize most of the relationships — $r$s generally less than .20 between attitudinal variables and outcomes
such as number of unfolding items endorsed or person fit (see Column 1 in Table 4; Table 5). One possible explanation for these small relationships is the existence of different types of unfolding items. Edwards (1946) performed an analysis of neutral Thurstonian scale items and found that neutral items can have different characteristics. He discovered that certain neutral items tap into people’s ambivalent attitudes whereas others tap into people’s indifference. Based on this finding, one might expect a respondent attitude structure × item-type interaction effect: ambivalent items might elicit endorsement from ambivalent respondents but not indifferent respondents, and vice versa. The small effect sizes seen in this study, then, may be the result of having grouped together somewhat dissimilar types of unfolding items.

One unexpected finding was with regard to person fit. Ambivalence indicators were positively associated with mis-fit, whereas a strength indicator was negatively associated with mis-fit. Furthermore, the number of unfolding items endorsed was predictive of mis-fit. This is unexpected because, theoretically, an unfolding scale which includes neutral-tending unfolding items, should allow ambivalent respondents to appropriately express their ambivalent feelings, as compared to a traditional Likert scale which would not. One possible explanation for this finding is that the large majority of respondents had well-defined, univalent attitudes toward drug testing. Analysis of the 11-point attitude extremity measure (1 = strongly disfavor drug testing; 11 = strongly favor drug testing) indicates that over 50% of respondents self-reported with a response of 9 or greater. An IRT model constructs predicted response patterns empirically, by estimating the item response functions that produce the least amount of error. If univalent respondents were numerous enough to dominate the scale and impose a strictly monotonic-type response pattern (feasible given the response option ineffectiveness problem noted in the results section), then item response functions would be representative of the majority (univalent
respondents), and the minority (ambivalent respondents) would represent deviation away from the typical item response function.

*Future Research Directions*. The first future research direction stems directly from one overall weakness of the present study. The results of this study hinge upon the characteristics of a specific measurement instrument that focuses on attitudes toward drug testing. The generalizability of these results are not known. Due to a limited amount of research on item-level unfolding processes, it is difficult to draw general conclusions. Thus, this stream of research would benefit from some replication studies. The preliminary and positive findings regarding the importance of attitude structure, for example, should be assessed using a different scale, focusing on a different attitude, to determine if the results are replicable and generalizable.

The next direction stems from the converging evidence from this study and Carter, Dalal, and Zickar (2010) that mental calculation time (introspection) may play a role in the unfolding response process. Many studies use the internet as a medium to connect with respondents and collect data. Many survey administration programs offer a means of assessing the amount of time a person takes to respond to a given item, known as response latency. The latency of extremely-worded items could be compared to the latency from unfolding items to see if unfolding items require more participant response time. Further, one could assess latency differences amongst different unfolding items to see if there is a link between the degree of unfolding (e.g., peakedness; point of unfolding) of an item and response time.

The final suggested research direction is regarding unfolding item “types.” In light of Edwards’ (1946) analysis of neutral items, it may not make sense to treat all unfolding items as though they were equivalent. A future study could attempt to categorize unfolding items into ambivalent-type or indifferent-type items. If the results of the present study are an indication,
then ambivalent-type items are likely to show marked unfolding whereas indifferent items would not.

The inclusion of neutral-tending unfolding items in scales can potentially lead to a richer and more accurate assessment of people’s attitudes. However, there is still much to be learned about the functionality and utility of unfolding items. It remains debatable whether altering the test delivery method can improve the functionality of an unfolding (Thurstonian) scale; the results presented here did not strictly support or disconfirm this proposition. The results of this study provide initial support for the proposition that attitude structure helps drive unfolding response processes. Future studies will hopefully expand upon both of the topics explored here — test delivery and respondent characteristics — so that researchers and practitioners will eventually be able to rely upon studies such as this for guidance in writing, analyzing, and interpreting unfolding scale items.
REFERENCES


Drasgow, F. (2009). Assessing personality 75 years after Likert: Thurstone was right! Presentation at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Drasgow, F., Chernyshenko, O.S., & Stark, S. (in press). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice*.


Management, 22, 313-338


LA.


Table 1

Defining Characteristics of Unfolding and Traditional (Monotonically-Increasing) Scale Items

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Unfolding Item</th>
<th>Traditional Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wording</td>
<td>Moderately-worded (not extremely positive or extremely negative)</td>
<td>Extremely-worded</td>
</tr>
<tr>
<td>Scale Development Paradigm</td>
<td>Included in Thurstone-type scales</td>
<td>Included in both Thurstone-type scales and Likert-type scales</td>
</tr>
<tr>
<td>Technical characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shape of Item Response Function</td>
<td>Non-monotonic item response function with a noticeable “peak” falling relatively close to the theta midpoint</td>
<td>Monotonically-increasing item response function</td>
</tr>
<tr>
<td>Item Location/Point of Unfolding</td>
<td>Between −2 and +2, approximately</td>
<td>Less than −2 or greater than +2, approximately</td>
</tr>
</tbody>
</table>
Table 2

Likelihood Ratio Item-Level Tests for the Presence of DIF between Items in Experimental and Comparison Samples

<table>
<thead>
<tr>
<th>Suspect Item</th>
<th>AIC</th>
<th>Log Likelihood</th>
<th>df</th>
<th>$G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 7</td>
<td>89,182.28</td>
<td>-44,286.14</td>
<td>5</td>
<td>7.16</td>
</tr>
<tr>
<td>Item 16</td>
<td>89,192.11</td>
<td>-44,291.06</td>
<td>5</td>
<td>17.00</td>
</tr>
<tr>
<td>Item 17</td>
<td>89,198.95</td>
<td>-44,294.47</td>
<td>5</td>
<td>23.82*</td>
</tr>
<tr>
<td>Item 23</td>
<td>89,209.24</td>
<td>-44,299.62</td>
<td>5</td>
<td>34.12*</td>
</tr>
<tr>
<td>Item 24</td>
<td>89,181.49</td>
<td>-44,285.74</td>
<td>5</td>
<td>6.36</td>
</tr>
<tr>
<td>Item 30</td>
<td>89,230.57</td>
<td>-44,310.28</td>
<td>5</td>
<td>55.44*</td>
</tr>
<tr>
<td>Item 32</td>
<td>89,212.82</td>
<td>-44,301.41</td>
<td>5</td>
<td>37.70*</td>
</tr>
<tr>
<td>Item 35</td>
<td>89,217.79</td>
<td>-44,303.89</td>
<td>5</td>
<td>42.66*</td>
</tr>
<tr>
<td>Item 38</td>
<td>89,162.21</td>
<td>-44,276.56</td>
<td>5</td>
<td>0.00</td>
</tr>
<tr>
<td>Item 40</td>
<td>89,273.80</td>
<td>-44,331.90</td>
<td>5</td>
<td>98.68*</td>
</tr>
<tr>
<td>Item 43</td>
<td>89,195.63</td>
<td>-44,292.82</td>
<td>5</td>
<td>10.26</td>
</tr>
<tr>
<td>Item 44</td>
<td>89,184.29</td>
<td>-44,287.15</td>
<td>5</td>
<td>9.18</td>
</tr>
<tr>
<td>Item 45</td>
<td>89,191.62</td>
<td>-44,290.81</td>
<td>5</td>
<td>16.50</td>
</tr>
<tr>
<td>Item 47</td>
<td>89,193.12</td>
<td>-44,291.56</td>
<td>5</td>
<td>18.00*</td>
</tr>
</tbody>
</table>

Note: Bonferroni correction for multiple comparisons was utilized; values with a corresponding $p < .0036$ – the corrected alpha level – are denoted with an asterisk (*).
Table 3

*Comparison of Mean (Standard Deviation) GGUM-Derived Parameter Estimates Obtained from Experimental and Comparison Samples†*

<table>
<thead>
<tr>
<th>Item Category</th>
<th>Mean delta</th>
<th>Mean alpha</th>
<th>Mean psi</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>1.50 (.18)</td>
<td>1.64 (1.18)</td>
<td>.59 (.20)</td>
</tr>
<tr>
<td>Comparison</td>
<td>2.41 (.72)</td>
<td>1.35 (.75)</td>
<td>.61 (.27)</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>−2.56 (1.19)</td>
<td>1.20 (.47)</td>
<td>.91 (.23)</td>
</tr>
<tr>
<td>Comparison</td>
<td>−3.40 (1.05)</td>
<td>1.24 (.41)</td>
<td>.79 (.19)</td>
</tr>
<tr>
<td><strong>Unfolding</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>.02 (1.41)</td>
<td>.92 (.47)</td>
<td>.83 (.33)</td>
</tr>
<tr>
<td>Comparison</td>
<td>−.13 (1.40)</td>
<td>.86 (.50)</td>
<td>.76 (.33)</td>
</tr>
</tbody>
</table>

†Several items were excluded from this analysis due to poorly-estimated item location estimates, which become artificially large.
Table 4

*Correlations between Attitude Structure Variables and Number of Unfolding Items Endorsed*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of unfolding items endorsed†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Experienced Ambivalence</td>
<td>.14**</td>
<td>(.85)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Attitude Clarity</td>
<td>-.10*</td>
<td>-.33**</td>
<td>(.89)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Attitude Correctness</td>
<td>-.06</td>
<td>-.27**</td>
<td>.64**</td>
<td>(.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Semantic Differential</td>
<td>.17**</td>
<td>.28**</td>
<td>-.24**</td>
<td>-.24**</td>
<td>(.74)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambivalence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Required Cognitive Effort</td>
<td>.06</td>
<td>.09*</td>
<td>.14**</td>
<td>.08</td>
<td>-.06</td>
<td>(N/A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Number of Thoughts Generated</td>
<td>.05</td>
<td>.06</td>
<td>.17**</td>
<td>.09*</td>
<td>-.05*</td>
<td>.29**</td>
<td>(.80)</td>
<td></td>
</tr>
<tr>
<td>8. Attitude Extremity</td>
<td>-.16**</td>
<td>-.38</td>
<td>.48**</td>
<td>.43**</td>
<td>-.42**</td>
<td>.03</td>
<td>.08*</td>
<td>(N/A)</td>
</tr>
</tbody>
</table>

** p < .01 and * p < .05

†After statistically controlling for the relationship between each variable and the number of somewhat agree responses given. Coefficient alpha estimates for multi-item variables are presented in the diagonal.
Table 5

*Linear and Curvilinear Regression Coefficient Estimates used for Predicting Person Mis-fit from Attitude Structure Variables*

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Linear term, $b$</th>
<th>Quadratic term, $b^2$</th>
<th>Multiple $R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of unfolding items endorsed</td>
<td>.86**</td>
<td>.02*</td>
<td></td>
<td>.08</td>
</tr>
<tr>
<td>2. Experienced Ambivalence</td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>3. Attitude Clarity</td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>4. Attitude Correctness</td>
<td></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>5. Semantic Differential Ambivalence</td>
<td>1.26**</td>
<td>−.29**</td>
<td>.07**</td>
<td>.17</td>
</tr>
<tr>
<td>6. Required Cognitive Effort</td>
<td>1.26**</td>
<td>−.32*</td>
<td>.07**</td>
<td>.20</td>
</tr>
<tr>
<td>7. Number of Thoughts Generated</td>
<td>1.69**</td>
<td>−.23**</td>
<td>.02**</td>
<td>.18</td>
</tr>
<tr>
<td>8. Attitude Extremity</td>
<td>−.86**</td>
<td>.05**</td>
<td></td>
<td>.18</td>
</tr>
</tbody>
</table>

*Coefficient estimates significant, $p < .05$, ** Coefficient estimates significant, $p < .001$.
Blank spots indicate non-significant coefficients.
For all regression models with at least one significant term, the overall model was significant, $p < .01$.
APPENDIX A: GGUM ITEM PARAMETER ESTIMATES FOR DRUG TESTING SCALE

GGUM Item Parameter Estimates for Drug Testing Scale

<table>
<thead>
<tr>
<th>Statement</th>
<th>$\hat{\delta}_i$</th>
<th>$\hat{\alpha}_i$</th>
<th>$r_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug testing will not solve all of an organization’s problems</td>
<td>−12.99</td>
<td>0.19</td>
<td>−0.72</td>
</tr>
<tr>
<td>Drug testing probably won’t help improve the workplace</td>
<td>−4.56</td>
<td>1.44</td>
<td>−0.41</td>
</tr>
<tr>
<td>Testing employees for drugs is a waste of time</td>
<td>−4.34</td>
<td>1.76</td>
<td>−0.49</td>
</tr>
<tr>
<td>Employee drug-testing is never a good idea</td>
<td>−4.30</td>
<td>1.75</td>
<td>−0.26</td>
</tr>
<tr>
<td>Employers should never be allowed to test their employees for drug use</td>
<td>−4.13</td>
<td>1.68</td>
<td>−0.10</td>
</tr>
<tr>
<td>I am somewhat opposed to employment drug testing</td>
<td>−4.06</td>
<td>2.04</td>
<td>−0.41</td>
</tr>
<tr>
<td>Drug-testing is a violation of a person’s privacy</td>
<td>−3.57</td>
<td>1.62</td>
<td>−0.18</td>
</tr>
<tr>
<td>No one should have to take a drug test to get a job</td>
<td>−2.50</td>
<td>1.26</td>
<td>0.14</td>
</tr>
<tr>
<td>Employers are hardly ever justified in drug testing employees for drug use</td>
<td>−2.16</td>
<td>0.93</td>
<td>0.34</td>
</tr>
<tr>
<td>I would NOT work for an organization with a drug-testing program</td>
<td>−2.07</td>
<td>1.31</td>
<td>0.39</td>
</tr>
<tr>
<td>Employers should give ample warning to their employees before administering a drug test</td>
<td>−2.03</td>
<td>0.95</td>
<td>0.21</td>
</tr>
<tr>
<td>Random drug tests are unfair because they are not related to job performance</td>
<td>−2.03</td>
<td>1.56</td>
<td>0.15</td>
</tr>
<tr>
<td>Drug-testing is unnecessary for most jobs</td>
<td>−2.01</td>
<td>1.24</td>
<td>0.35</td>
</tr>
<tr>
<td>Drug testing might alienate some employees</td>
<td>−1.88</td>
<td>0.24</td>
<td>0.49</td>
</tr>
<tr>
<td>Employers should be allowed to test their employees for hard drug use (e.g. heroin, cocaine, etc.) but not for softer drugs (e.g. marijuana)</td>
<td>−1.79</td>
<td>0.50</td>
<td>0.35</td>
</tr>
<tr>
<td>Statement</td>
<td>Score 1</td>
<td>Score 2</td>
<td>Score 3</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Employer drug testing is wrong in certain situations</td>
<td>-1.73</td>
<td>0.70</td>
<td>0.33</td>
</tr>
<tr>
<td>Companies should provide their reasons for testing employees for drug use</td>
<td>-1.56</td>
<td>0.32</td>
<td>0.46</td>
</tr>
<tr>
<td>The only reason employers should use drug tests is for issues related to performance on the job</td>
<td>-1.48</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>Employment drug-testing should only be allowed before someone is hired</td>
<td>-0.99</td>
<td>0.36</td>
<td>0.50</td>
</tr>
<tr>
<td>Drug tests are okay for entry-level positions but not management-level positions</td>
<td>-0.75</td>
<td>1.00</td>
<td>0.25</td>
</tr>
<tr>
<td>Employment drug testing should be allowed in special circumstances</td>
<td>-0.60</td>
<td>0.19</td>
<td>0.46</td>
</tr>
<tr>
<td>Employers should be allowed to use urine tests to test employees for drugs, but they should not be allowed to conduct hair follicle tests</td>
<td>-0.50</td>
<td>0.42</td>
<td>0.39</td>
</tr>
<tr>
<td>Sometimes an employer is justified in testing employees for drug use</td>
<td>1.61</td>
<td>0.71</td>
<td>-0.45</td>
</tr>
<tr>
<td>The government should regulate employers’ drug-testing practices</td>
<td>1.79</td>
<td>0.24</td>
<td>-0.55</td>
</tr>
<tr>
<td>Employee drug testing is acceptable if privacy and confidentiality can be guaranteed</td>
<td>1.87</td>
<td>1.11</td>
<td>-0.22</td>
</tr>
<tr>
<td>Some organizations would benefit from a drug-testing policy</td>
<td>2.07</td>
<td>1.30</td>
<td>-0.54</td>
</tr>
<tr>
<td>Employers should have the right to test for drug use</td>
<td>2.32</td>
<td>2.44</td>
<td>-0.09</td>
</tr>
<tr>
<td>Employers should be able to drug-test their employees at any time</td>
<td>2.33</td>
<td>2.49</td>
<td>-0.06</td>
</tr>
<tr>
<td>Employers have the right to test their employees for drug use</td>
<td>2.34</td>
<td>2.58</td>
<td>-0.09</td>
</tr>
<tr>
<td>If given probable cause, organizations should be allowed to test employees for drug use</td>
<td>2.37</td>
<td>1.00</td>
<td>-0.42</td>
</tr>
<tr>
<td>Drug-testing is the employer’s right</td>
<td>2.50</td>
<td>1.39</td>
<td>-0.10</td>
</tr>
<tr>
<td>Employers should drug-test their employees</td>
<td>2.59</td>
<td>1.27</td>
<td>-0.16</td>
</tr>
<tr>
<td>Employers should use drug tests that can detect drug use over the</td>
<td>2.62</td>
<td>0.96</td>
<td>-0.29</td>
</tr>
</tbody>
</table>
last month

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>Parameter Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employers should give no warning to employees before they are drug-tested</td>
<td>2.64</td>
<td>1.61</td>
<td>-0.04</td>
</tr>
<tr>
<td>Employees should be tested if there is a reason to believe they are on drugs</td>
<td>2.66</td>
<td>0.70</td>
<td>-0.22</td>
</tr>
<tr>
<td>An employer is always justified in drug-testing employees</td>
<td>2.68</td>
<td>0.88</td>
<td>-0.17</td>
</tr>
<tr>
<td>All employers should drug test their employees</td>
<td>2.76</td>
<td>1.46</td>
<td>0.00</td>
</tr>
<tr>
<td>I think drug-testing in the workplace is a good thing</td>
<td>2.82</td>
<td>3.83</td>
<td>0.03</td>
</tr>
<tr>
<td>Employers should be allowed to drug-test randomly selected groups of employees</td>
<td>2.84</td>
<td>1.13</td>
<td>0.06</td>
</tr>
<tr>
<td>I would consider working for an organization with a drug-testing program</td>
<td>3.03</td>
<td>1.22</td>
<td>0.24</td>
</tr>
<tr>
<td>People who work dangerous jobs should be drug-tested</td>
<td>3.18</td>
<td>0.90</td>
<td>0.39</td>
</tr>
<tr>
<td>An employer should be allowed to drug-test employees for any reason, at any time.</td>
<td>3.25</td>
<td>1.38</td>
<td>0.04</td>
</tr>
<tr>
<td>Employers should give drug tests to all employees despite their level of job performance</td>
<td>3.32</td>
<td>1.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Drug tests should be used to measure what employees do during their free time, even though they may not be impaired when they go to work</td>
<td>3.35</td>
<td>0.69</td>
<td>-0.07</td>
</tr>
<tr>
<td>Drug-testing is a great way for employers to screen employees</td>
<td>3.37</td>
<td>1.36</td>
<td>0.06</td>
</tr>
<tr>
<td>An employer should be allowed to drug-test a person who has an accident on the job</td>
<td>4.71</td>
<td>0.41</td>
<td>0.46</td>
</tr>
<tr>
<td>A positive drug test is proof that drugs contributed to the accident</td>
<td>6.29</td>
<td>0.32</td>
<td>0.63</td>
</tr>
<tr>
<td>An employer should be allowed to drug-test an employee to decide whether they receive pay while on leave for a work-related accident</td>
<td>6.56</td>
<td>0.27</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Parameter estimates from Lin et al. (2009).
APPENDIX B: ITEMS MEASURING AMBIVALENCE

Items measuring ambivalence

Instructions: Consider the positive and negative aspects of drug testing. Please indicate the extent to which each of the following words describes how you feel about drug testing in the workplace.

Response options: Strongly disagree - strongly agree

1. Muddled
2. Divided
3. Tense
4. Contradictory
5. Jumbled
6. Conflicted
7. Consistent (reverse-keyed)
8. Uniform (reverse-keyed)

9. Considering only the positive qualities of workplace drug testing and ignoring their negative qualities, would you say that the positive qualities of workplace drug testing are: extremely positive; quite positive; slightly positive; or not at all positive?

Response options: Extremely positive; quite positive; slightly positive; not at all positive

10. Considering only the negative qualities of workplace drug testing and ignoring their positive qualities, would you say that the negative qualities of workplace drug testing are: extremely negative; quite negative; slightly negative; or not at all negative?

Response options: Extremely negative; quite negative, slightly negative; not at all negative
APPENDIX C: ITEMS MEASURING ATTITUDE STRENGTH

Items measuring attitude strength

Instructions: Please keep the drug testing questionnaire in mind as you respond to the following statements.

1. How much mental effort did you put into evaluating the drug testing items?
   Response options: No effort at all; a small amount of effort; moderate effort; a lot of effort

2. When completing the drug testing items, I had many thoughts about workplace drug testing.
3. When completing the drug testing items, I had few thoughts about workplace drug testing. (reverse-keyed)
   Response options: Strongly disagree - strongly agree

Instructions: Please think about your attitude toward drug testing.

4. How certain are you that you know what your true attitude on workplace drug testing really is?
5. How certain are you that the attitude you expressed toward workplace drug testing really reflects your true thoughts and feelings?
6. To what extent is your true attitude toward workplace drug testing clear in your mind?
7. How certain are you that the attitude you just expressed toward workplace drug testing is really the attitude you have?
8. How certain are you that your attitude toward workplace drug testing is the correct attitude to have?
9. To what extent do you think other people should have the same attitude as you on this issue?
10. How certain are you that, of all possible attitudes one might have toward workplace drug testing, your attitude reflects the right way to think and feel about the issue?
   Response options: Very uncertain, somewhat uncertain, somewhat certain, very certain
Figure 1. Comparison of item response functions from a hypothetical monotonically-increasing item and a hypothetical unfolding item.
A) Item: “Employers should be allowed to use urine tests for drugs, but they should not be allowed to conduct hair follicle tests.”

B) Item: “I would NOT work for an organization with a drug-testing program”

C) Item: “Employers should have the right to test for drug use”

Figure 2. Item response functions from sample Lin et al. (2009) drug testing scale items.
Figure 3. Comparison of five respondents’ positive and negative evaluations of drug testing.
Figure 4. Comparison of *strongly agree* and *somewhat agree* response option functions in three unfolding items.
Figure 5. Comparison of *strongly agree* and *somewhat agree* response option functions in two extremely-worded items.
A) Experimental condition shows more marked unfolding than comparison

B) Comparison condition shows more marked unfolding than experimental

C) Neither condition shows marked unfolding

Figure 6. Item response functions from six of seven items with statistically significant differential item functioning (DIF).
A) Experimental condition provides more information than comparison

B) Comparison condition provides more information than experimental

Figure 7. Information functions from four of seven items with statistically significant differential item functioning (DIF).
A) Test information functions for experimental and comparison groups

![Graph showing test information functions for experimental and comparison groups.](image)

B) Standard error functions for experimental and comparison groups

![Graph showing standard error functions for experimental and comparison groups.](image)

Figure 8. Scale-level information and standard error function comparisons.
Figure 9. Item location estimate dispersion in comparison (left) and experimental (right).
Figure 10. Prediction of person mis-fit from two ambivalence indicator variables.