

AN ITEM RESPONSE THEORY ANALYSIS OF THE SELF-MONITORING SCALE

Edward Burkley¹
Oklahoma State University

ABSTRACT

The Self-Monitoring Scale (SMS) was investigated utilizing item response theory (IRT). First, IRT models that constrained each of the subscale items to have equal discrimination were fitted to the three subscales of the SMS (Acting, Extraversion, and Other-Directedness). These models were then contrasted with separate models that allowed the discriminations to be estimated freely. For all three subscales, model comparison tests of significance indicated that the unconstrained models were a better fit. Thus, the items of each subscale are differentially related to their respective underlying construct. Implications and recommendations are offered for future psychometric development and implementation of the SMS.

Keywords: Self-monitoring, Item Response Theory

INTRODUCTION

The ability of expressive control over one's behavior from situation to situation has important implications for the link between the person and the situation. Variables that moderate this relationship between personality and social behavior are among the most useful constructs in personality and social psychology (Snyder & Ickes, 1985). One such variable that has received a great deal of attention is that of self-monitoring (Snyder, 1974). Since its introduction over 30 years ago, the construct of self-monitoring has spawned a cornucopia of research and has been claimed as "one of the most popular measures of personality to be introduced in recent years" (Briggs & Cheek, 1988, p. 663). Given the importance of a variable like self-monitoring and its many uses in varying research areas, numerous psychometric investigations have been conducted to evaluate the most popular measure of self-monitoring: The Self-Monitoring Scale (SMS; Snyder, 1974). Yet, no detailed item analysis has been performed to improve the psychometric development and implementation

¹ Correspondence concerning this article should be addressed to Edward Burkley, Department of Psychology, Oklahoma State University, 116 N. Murray, Stillwater, OK 74078; email: ed.burkley@okstate.edu.

of the SMS. Thus, one potential avenue of investigation is to utilize the benefits of item response theory techniques (IRT; Embertson & Reise, 2000; Thissen & Wainer, 2001; van der Linden & Hambleton, 1997).

The purpose of the present article is to utilize IRT and offer a more refined item analysis of the SMS than is feasible with other more traditional psychometric techniques. Although there is overlap between classic testing approaches and IRT, IRT does offer some advantages for those interested in personality assessment (Reise & Henson, 2003). For example, by using IRT to analyze the Rosenberg Self-Esteem scale, Gray-Little, Williams, and Hancock (1997) were able to determine that this measure is quite poor at differentiating among high self-esteem individuals. That is, this measure is useful for distinguishing between people low in self-esteem from those that are average, but is not adequate for distinguishing between people who are high in self-esteem. For any researcher interested in using this self-esteem scale, this information may be critical. In another example, Reise and Henson (2000) demonstrated that one item of the Revised NEO-PI Anxiety subscale was nearly four times as informative as the rest of the items. Both of these examples demonstrate situations in which an IRT analysis offered novel evidence above and beyond what could be obtained using a more traditional psychometric approach.

In this article, I will briefly outline the history of self-monitoring assessment. I will then provide an overview of item response theory and explain how I applied this technique to the primary tool for assessing self-monitoring (i.e., the self-monitoring scale and its three subscales). After presenting the results, I will discuss implications and recommendations regarding the future psychometric development and implementation of the self-monitoring scale.

Evolution of the Self-Monitoring Scale

Self-monitoring theory asserts that people vary in the extent to which they regulate their expressive behavior (Snyder, 1974). People who are *high self-monitors* exert a great deal of control over their behaviors out of a concern for situational appropriateness. The behavior of high self-monitors is largely determined by the social constraints of the situation. Consequently, they are often referred to as “social chameleons”. People who are *low self-monitors* do not engage in such self-presentation. Instead, their behavior represents their own inner attitudes and emotions. Accordingly, the behavior of low self-monitors is less influenced by the situation and is more reflective of the individual’s disposition. This distinction between those receptive to the social environment versus those reliant on inner qualities gives self-monitoring the favorable quality of being able to serve as an important moderator of social behavior.

To measure the construct of self-monitoring, Snyder developed the Self-Monitoring Scale (SMS; Snyder, 1974). The scale is comprised of 25 items in a true-false response format. Since the scale’s inception, numerous research investigations have utilized the scale and its importance cannot be overstated. The majority of these investigations have focused on the factor structure of the SMS. Investigators have identified factor solutions ranging from one to four, with three factors being the most widely accepted solution (Briggs, Cheek, & Buss, 1980; Gangestad & Snyder, 1985). These three factors were labeled *Acting*, *Extraversion*, and *Other-Directedness*. However, Snyder and Gangestad (1986) argued that this rotated three-factor solution is not as informative in identifying a single latent construct. To improve the psychometric properties of the scale, Snyder and Gangestad (1986) offered a revised SMS comprised of 18 of the SMS items that had the largest factor loadings. Later investigations

contrasted the 18-item SMS with the original SMS (Briggs & Cheek, 1988; Hoyle & Lennox, 1991; John, Cheek, & Klohnen, 1996), revealing that this revised SMS has its own set of issues. Generally, it has been demonstrated that the deletion of the items from the original SMS increased the scale's reliability and purity of the factors, yet it shifted the scale toward Extraversion and Acting at the expense of weakening the Other-Directedness factor (Hoyle & Lennox, 1991; John et al., 1996). Based on this evidence it has been recommended that researchers restrict investigations of self-monitoring to the use of the original 25-item scale with its three subscales: Acting, Extraversion, and Other-Directedness (e.g., John et al., 1996).

What is needed in the investigation of the psychometric properties of the SMS is a method 1) that is well suited for the use of dichotomous data (true/false), the typical format used when administering the SMS and 2) that does not make unsound assumptions about item response distributions. Such a resolution is offered by the use of IRT. Not only does IRT fit the above criteria, it also offers a more refined, item-by-item analysis of the scale. First, item response models were originally designed for use on dichotomous data, although more sophisticated models are now available that handle ordered/unordered categorical data (van der Linden & Hambleton, 1997). Second, IRT offers a thorough item-level analysis, providing more information than other traditional techniques. For example, IRT can indicate which items of the SMS do a better job of differentiating between individuals who are low and high self-monitors. Thus, IRT offers a unique contribution to the understanding of the SMS by providing valuable information about the individual items and suggesting possible improvements for the overall scale.

Item Response Theory

Terminology and assumptions. Item Response Theory (IRT) is a collection of statistical methods that assesses the extent to which each item measures the underlying latent construct. This latent construct is generally referred to as theta (θ) and is typically distributed as a z-score that ranges from -3 to +3.

There are two parameters that are central to any IRT model - the threshold (b) and discrimination (a) parameters. The *threshold parameter* (b) locates the level of the underlying construct where there is a .50 probability of endorsing the item. For personality scales, this parameter indicates the level of the personality trait necessary to endorse the item. Thus, SMS items with large positive b values are only endorsed by individuals with very high levels of self-monitoring. These items would distinguish between moderate and high self-monitors. Conversely, SMS items with large negative b values are endorsed by individuals with very low levels of self-monitoring. These items would distinguish between moderate and low self-monitors.

The item *discrimination parameter* (a) quantifies the association between the item and the latent construct. It represents the item's ability to discriminate among people with different levels of the underlying trait. The a parameter typically ranges from 0 to 3. SMS items with large a values do an excellent job of discriminating between various levels of self-monitoring, whereas SMS items with small a values indicate that individuals' responses to these items do not relate to the self-monitoring trait. Thus, the b parameter indicates *where* on the continuum the item is discriminating; whereas, the a parameter indicates *how well* the item is discriminating.

An additional piece of information unique to IRT is a scale's *marginal reliability*. Although reliability is a classic test theory concept, marginal reliability is an analogous index

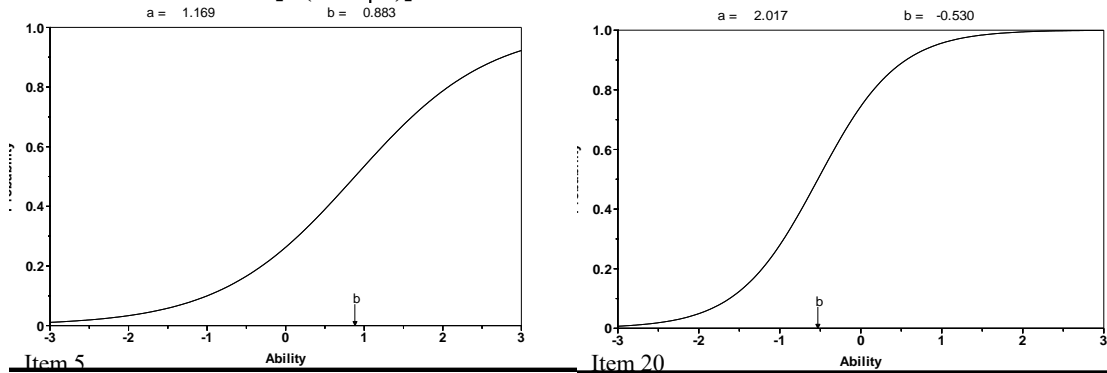
that is often estimated when using IRT. The values for marginal reliability range from 0 to 1 and their interpretation is analogous to Cronbach's alpha (Green, Bock, Humphreys, Linn & Reckase, 1984). This statistic is useful because it provides a single numeric value that summarizes the scale's overall precision (Thissen & Wainer, 2001).

IRT is based on two major assumptions: Unidimensionality and local independence. The assumption of unidimensionality states that all items being analyzed reflect a single continuous latent construct (θ). When a scale consists of multiple subscales, as does the SMS, each subscale is treated as a unidimensional construct. The assumption of *local independence* states the items are independent of each other and that the only connection that does exist between the items is in regards to their relationship with the latent construct. For example, items that are grouped in a testlet would not reflect local independence because there is an additional connection that exists between the items (i.e., grouped items directed toward a single topic). Before conducting an IRT analysis, exploratory or confirmatory factor analysis is typically used to evaluate the unidimensionality and local independence of a particular scale or subscale. In practice, a scale rarely meets these assumptions perfectly; however, IRT is known to be relatively robust to moderate violations of these assumptions (e.g., Drasgow & Parsons, 1983; Hulin, Drasgow, & Parsons, 1983).

Another important note is that IRT relies heavily on graphic representations of item characteristics. The cornerstone of IRT is the *Item Characteristic Curve* (ICC) or trace line, which represents the probability of an item response as a function of the underlying construct. The ICC is central to IRT because it can be used to evaluate the quality of each item. In this graph, the underlying trait comprises the x -axis and the probability of endorsement comprises the y -axis (see Figure 1 for an example). For each item, the slope of the line represents the item discrimination parameter (a), with steeper (i.e., larger) slopes indicating greater discrimination ability.

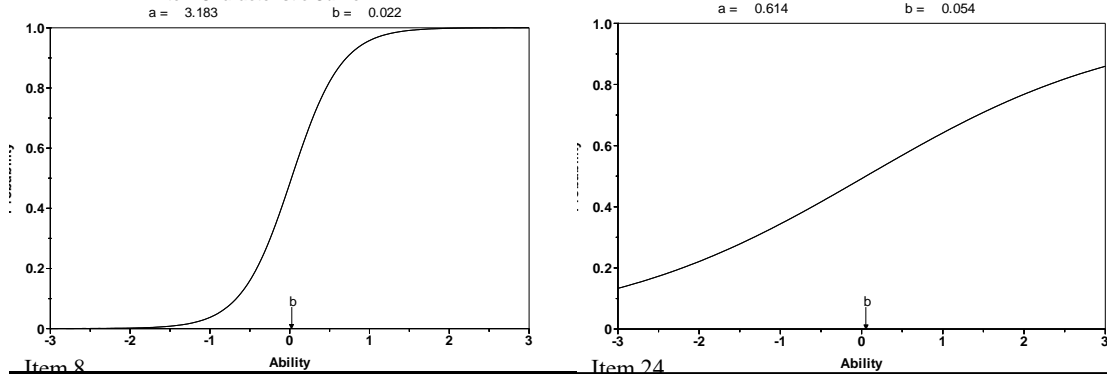
Finally, an important concept in IRT is that of the information function. The *item information function* is an index of how much psychometric information the item provides at each level of the latent construct. In other words, this information function represents how well each item differentiates between people at different levels of the latent construct. More importantly, item information functions are useful because they can be combined to form an overall test information function. The *test information function* is useful in that it indicates how well the entire scale differentiates between people at different levels of the latent construct. For example, a particular self-esteem scale may be more precise at assessing people high in self-esteem; whereas another self-esteem scale may be better equipped at assessing people low in self-esteem. The graphical representation of information functions places the underlying trait on the x -axis and the amount of information on the y -axis (see Figure 4 for example of test information functions). The peak of the information curve indicates where on the theta continuum the scale has the greatest amount of precision, or information. These graphs offer an additional piece of data that is also quite useful, as information increases (solid line), measurement error can be seen to decrease (dotted line), thus offering an assessment of precision and its inverse, error.

Figure 1
Item characteristic curves for the Acting subscale (2PL model). The x-axis represents the latent construct of acting (θ). The y-axis represents the probability of a true response $[T(x = 1|\theta)]$. Note. * indicates reverse scored items.



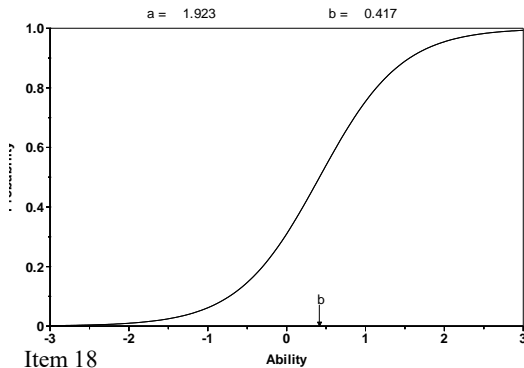
Item 5
 I can make impromptu speeches even on topics about which I have almost no information.

Item 20
 I have never been good at games like charades or improvisational acting.*



Item 8
 I would probably make a good actor.

Item 24
 I can look anyone in the eye and tell a lie with a straight face (if for a right end).



Item 18
 I have considered being an entertainer.

IRT models. Although all IRT analyses assume a single underlying construct, a variety of models can be used to define the causal relationship between this construct and the observed item responses. In essence, these models differ in terms of how many parameters are required to model each item response. A review of all IRT models is beyond the scope of this paper, so instead only the models typically used for scales with dichotomous test items will be described (i.e., logistic models).

One of the simplest IRT models is the *1-parameter logistic model (1PL)*. In this model, only the threshold parameter (b) is required to model the item process. This model is often referred to as a restricted model because it restricts the a parameters to be equal across all items. By imposing this restriction, this model makes the assumption that all the items are equally related to their underlying construct and therefore discriminate equally.

For item j , the 1PL model for the trace line is defined as:

$$T_j(x_j = 1|\theta) = \{1 + \exp[-a(\theta - b_j)]\}^{-1}.$$

In this equation, $T_j(x_j = 1|\theta)$ traces the probability of a “true” response ($x_j = 1$) as a function of θ , b_j represents the threshold parameter for item j and indicates at what level of θ an individual has a 50% chance of giving a “true”(or high trait) response to item j . Finally, a is the item discrimination parameter (slope) and represents the rate of change in the proportion of “true” responses as a function of θ . As stated above, this equation allows the b parameters to be assessed but restricts the a parameters to be equal across all items.

Although the 1PL model is informative, in practice, scale items do not typically discriminate equally. To address this issue, a *2-parameter logistic model (2PL)* is often employed. The 2PL model is almost identical to the 1PL except that the discrimination parameter (a) is allowed to vary across items (i.e., an unrestricted model). This model therefore allows the items to vary in their relation to the underlying construct. The equation for the 2PL model is identical to the one expressed above, with the addition of a j subscript to the a parameter thereby allowing the a parameter to vary across items.

When utilizing IRT models, it is common practice to compare the results of the 1PL and 2PL models to determine if each item is equally related to the latent construct. By examining the difference in fit between the 1PL and the 2PL models, one is able to empirically test the assumption that all scale items relate equally to the underlying trait. The assessment of this assumption is important to test because it is what allows researchers to sum across all items to create a composite score (this is standard procedure for the SMS). If this assumption is violated and items do differ in their relation to the underlying trait, then summed scores are an erroneous practice for that particular scale.

Analyses of the SMS Subscales. Past investigations of the SMS using factor analysis have suggested a multifactorial structure (e.g., Briggs et al., 1980; Hoyle & Lennox, 1991; John et al., 1996). The most often agreed upon solution posits three factors, generally labeled as Acting, Extraversion, and Other-Directedness (Briggs et al., 1980). To examine the psychometric properties of each subscale, separate IRT analyses were conducted on each of the three SMS subscales.

METHOD

Data

The data were collected from introductory psychology students from a large southeastern university. The sample consisted of 581 students who completed the original 25-item SMS in a true-false format. The scale was scored so that true responses received a 1 and false responses received a 0. The items were then recoded so that a 1 represented a “high self-monitoring response” and 0 represented a “low self-monitoring response.”

Based on previous SMS research (Briggs et al., 1980), the subscales were defined as follows: The Acting subscale consisted of items 5, 8, 18, 20, and 24 (see Figures 1, 2, and 3 for items), the Extraversion subscale consisted of items 12, 14, 20, 21, 22, and 23, the Other-Directedness subscale consisted of items 2, 3, 6, 7, 13, 15, 16, 17, 19, 23, and 25. Items 1 and 4 were not included in any of the subscales (Briggs et al., 1980).

Models

The three subscales of the SMS (Acting, Extraversion, and Other-Directedness) were analyzed using binary IRT models in the Multilog computer program (Thissen, 1991; Thissen, Chen, & Bock, 2003). First a 1PL model was fit to the data, restricting all items to be equally related to their underlying construct. Next, a 2PL model was fit to the data, allowing the items to vary in their relation to the underlying construct.

RESULTS

Test of Assumptions

A confirmatory factor analysis with three uncorrelated factors representing each subscale was conducted to assess unidimensionality and local independence. Because the data were dichotomous, a robust weighted least squares (WLSMV) estimation was conducted in Mplus (Muthén & Muthén, 2000). The *RMSEA* demonstrated good fit (*RMSEA* = .06) but the *TLI* fell just short of the desirable cutoff (*TLI* = .85). Although these results are not conclusive, there is a great deal of research demonstrating evidence for the prominent 3-factor solution of the SMS and there is also evidence that IRT is relatively robust to moderate violations of its assumptions (e.g., Drasgow & Parsons, 1983; Hulin et al., 1983); thus, IRT analyses were pursued with an assessment of each of the SMS subscales.

Acting Subscale

Descriptive statistics. The average summed score on the Acting subscale was 2.32 (*SD* = 1.58), with scores ranging from 0 to 5. The means for the acting items were .31, .49, .38, .66, and .49, respectively. The coefficient alpha was .66 and the mean item-to-total correlation (r_{ϕ}) was .67. The interitem correlations ranged from $r_{\phi 20,24} = .10$ to $r_{\phi 8,20} = .46$ and the item-total correlations ranged from $r_{\phi T,24} = .55$ to $r_{\phi T,8} = .73$.

IRT analysis. First, the 1PL model was applied to the data. This restricted model estimated 6 parameters. The $-2\log(\text{likelihood})$ computed for this model was -3868.10 and the marginal reliability was .63. Next, the 2PL model was applied to the data. This unrestricted model estimated 10 parameters. The $-2\log(\text{likelihood})$ was -3923.30 and the marginal reliability was

.66. The relative fit of the unrestricted and restricted models was then assessed for the Acting subscale,

$$G_{\text{diff}}^2 = -3868.10 - (-3923.30) = 55.20 \text{ (} df = 4, p < .001 \text{)}.$$

The results indicated that the 2PL model fit the data significantly better than the 1PL model. Thus, one can conclude that the discrimination parameters are not equal across items and that these subscale items differ in their relation to acting.

Table 1
Parameter Estimates for the Self-Monitoring Subscales (2PL model)

Subscale		<i>a</i> (SE)	<i>b</i> (SE)
Acting	Item 5	1.17 (0.16)	0.88 (0.14)
	Item 8	3.18 (0.36)	0.02 (0.05)
	Item 18	1.92 (0.22)	0.42 (0.08)
	Item 20	2.02 (0.22)	-0.53 (0.07)
	Item 24	0.61 (0.12)	0.05 (0.18)
Extraversion	Item 12	1.86 (0.20)	-0.06 (0.07)
	Item 14	1.36 (0.18)	-1.33 (0.16)
	Item 20	0.96 (0.14)	-0.80 (0.15)
	Item 21	0.59 (0.13)	-1.58 (0.36)
	Item 22	1.88 (0.20)	-0.04 (0.07)
	Item 23	1.73 (0.20)	-0.64 (0.09)
Other-Directedness	Item 2	0.87 (0.16)	1.78 (0.30)
	Item 3	0.57 (0.14)	-2.29 (0.73)
	Item 6	0.90 (0.14)	0.56 (0.35)
	Item 7	0.47 (0.15)	-3.82 (1.59)
	Item 13	1.30 (0.16)	-0.48 (0.10)
	Item 15	0.93 (0.14)	-0.15 (0.13)
	Item 16	1.29 (0.20)	-0.76 (0.12)
	Item 17	0.41 (0.11)	1.30 (0.42)
	Item 19	1.48 (0.23)	1.26 (0.14)
	Item 23	0.07 (0.12)	-10.69 (* *)
	Item 25	0.75 (0.13)	-0.71 (0.19)

* * Unable to obtain estimate of standard error, value too large.

The threshold and discrimination parameters for the 2PL model are displayed in Table 1 and the ICCs for each item are shown in Figure 1. The threshold parameters ranged from $b_{20} = -.53$ to $b_5 = .88$. All of the items' threshold values were close to zero, where zero on the theta continuum represents the unobserved population mean. This suggests that this subscale allows fine distinctions only among individuals with moderate levels of the acting construct. These items are only providing information about moderate levels of acting and do not supply much information regarding the lower and upper ends of the acting continuum. This indicates that the use of this subscale does not allow fine distinctions among individuals with low or high levels of acting. This pattern can also be seen graphically in the test information function (see Figure 4), with the greatest amount of measurement precision, and the least amount of error, occurring in the center of the continuum.

The item discriminations ranged from a poor $a_{24} = .61$ to a very strong $a_8 = 3.18$. This variability can also be seen in the ICCs for these items. Items 8, 18, and 20 show strong slopes, with item 8 displaying the steepest slope. Thus, items 8, 18, and 20 show the strongest relation to acting. Conversely, item 24 shows a weaker relationship to acting, as indicated by its flatter slope.

Extraversion Subscale

Descriptive statistics. The average summed score on the Extraversion subscale was 3.86 ($SD = 1.68$), with scores ranging from 0 to 6. The means for the extraversion items were .52, .80, .66, .70, .51, and .67, respectively. The coefficient alpha was .64 and the mean item-to-total correlation (r_{ϕ}) was .61. The interitem correlations ranged from $r_{14,21} = .07$ to $r_{12,22} = .40$ and the item-total correlations ranged from $r_{T,21} = .48$ to $r_{T,(12,22)} = .66$.

IRT analysis. First, the 1PL model was applied to the data. This restricted model estimated 7 parameters. The $-2\log(\text{likelihood})$ computed for this model was -3310.80 and the marginal reliability was .61. Next, the 2PL model was applied to the data. This unrestricted model estimated 12 parameters. The $-2\log(\text{likelihood})$ computed for this model was -3344.60 and the marginal reliability was .63. The relative fit of the unrestricted and restricted models was then assessed for the Extraversion subscale,

$$G_{\text{diff}}^2 = -3310.80 - (-3344.60) = 33.80 \quad (df = 5, p < .001).$$

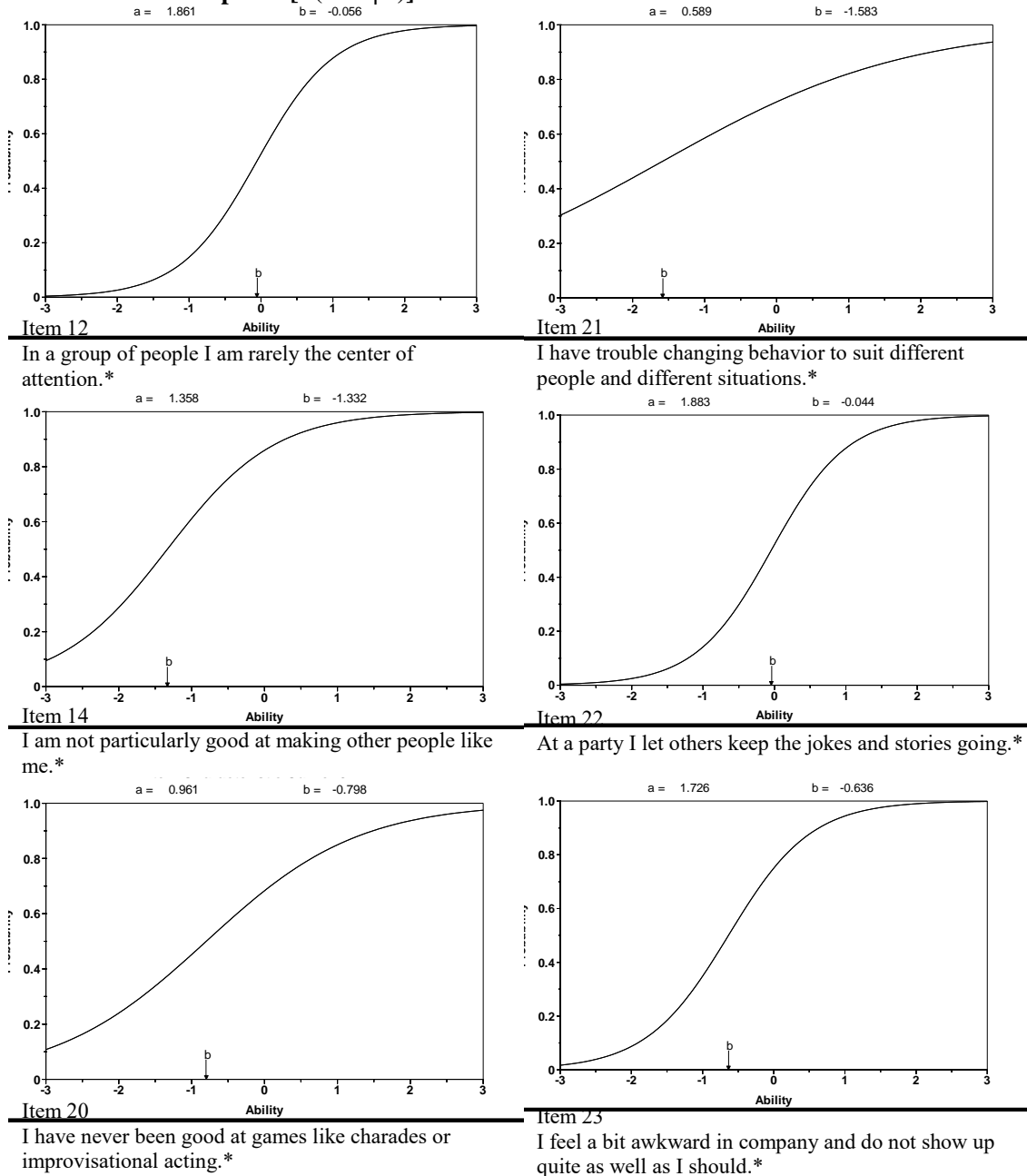
The results indicated that the 2PL model fit the data significantly better than the 1PL model. Thus, one can conclude that the discrimination parameters are not equal across items and that these subscale items differ in their relation to extraversion.

The threshold and discrimination parameters for the 2PL model are displayed in Table 1 and the ICCs for each item are shown in Figure 2. The threshold parameters ranged from $b_{21} = -1.58$ to $b_{22} = -.04$. All of the items' threshold values were negative, indicating that this subscale allows fine distinctions only among individuals with low levels of extraversion and that none of the items are adequately identifying those individuals high in extraversion. This pattern is reiterated graphically in the test information function (see Figure 4), with the greatest amount of measurement precision occurring in the center and lower end of the continuum.

The item discriminations ranged from $a_{21} = .59$ to $a_{22} = 1.88$. As shown in the ICCs, items 12, 22, and 23 show the strongest slopes, and thus are most related to the construct of extraversion. Item 21 has the weakest relation to extraversion, as is evident by its flatter slope.

Figure 2

Item characteristic curves for the Extraversion subscale (2PL model). The x-axis represents the latent construct of extraversion (θ). The y-axis represents the probability of a true response $[T(x = 1|\theta)]$. Note. * indicates reverse scored items.



Other-Directedness Subscale

Descriptive statistics. The average summed score on the Other-Directedness subscale was 5.92 ($SD = 2.03$), with scores ranging from 0 to 11. The means for the other-directedness items were .21, .77, .39, .85, .62, .53, .68, .38, .20, .67, and .62, respectively. The coefficient alpha was .50 and the mean item-to-total correlation (r_ϕ) was .45. The interitem correlations ranged from $r_{7,23} = .001$ to $r_{13,16} = .28$ and the item-total correlations ranged from $r_{T,23} = .10$ to $r_{T,13} = .52$.

IRT analysis. First, the 1PL model was applied to the data. This restricted model estimated 12 parameters. The $-2\log(\text{likelihood})$ was 117.40 and the marginal reliability was .50. Next, the 2PL model was applied to the data. This unrestricted model estimated 22 parameters. The $-2\log(\text{likelihood})$ computed for this model was 7.6 and the marginal reliability was .60. The relative fit of the unrestricted and restricted models was then assessed for the Other-Directedness subscale,

$$G_{\text{diff}}^2 = 117.40 - 7.6 = 109.80 \quad (df = 10, p < .001).$$

Once again, the results indicated that the 2PL model fit the data significantly better than the 1PL model. Thus, the discrimination parameters are not equal for all items, indicating that these subscale items differ in their relation to other-directedness.

The threshold and discrimination parameters for the 2PL model are displayed in Table 1 and the ICCs for each item are shown in Figure 3. The threshold parameters ranged from $b_2 = 1.78$ to $b_{23} = -10.69$. The majority of the items had negative threshold values, indicating that this subscale allows for fine distinctions only among individuals with low levels of other-directedness. Two of the items (7 and 23) showed extremely negative values, indicating that only a very low level of other-directedness would be needed to endorse these items. Items such as these are less informative because most individuals are likely to endorse them and therefore they do not distinguish among different levels of other-directedness.

The item discriminations ranged from a very poor $a_{23} = .07$ to a moderate $a_{19} = 1.48$. This wide variability can also be seen in the ICCs for these items. None of the items show strong slopes. Three of the items (13, 16, and 19) show moderate slopes, with item 19 displaying the steepest slope in the subscale. Thus, item 19 shows the strongest relation to other-directedness, albeit only modestly. Conversely, items 3, 7, 17, and 23 show an extremely weak relationship to other-directedness, as indicated by their very flat slopes. This finding suggests that these items are poor measures of other-directedness and may be related to a different construct. The test information function (see Figure 4) also suggests that this is the case, given that the information curve does not appear to peak and the error curve remains high across the theta continuum.

Figure 3
Item characteristic curves for the Other-Directedness subscale (2PL model). The x-axis represents the latent construct of other-directedness (θ). The y-axis represents the probability of a true response [$T(x = 1|\theta)$]. Note. * indicates reverse scored items.

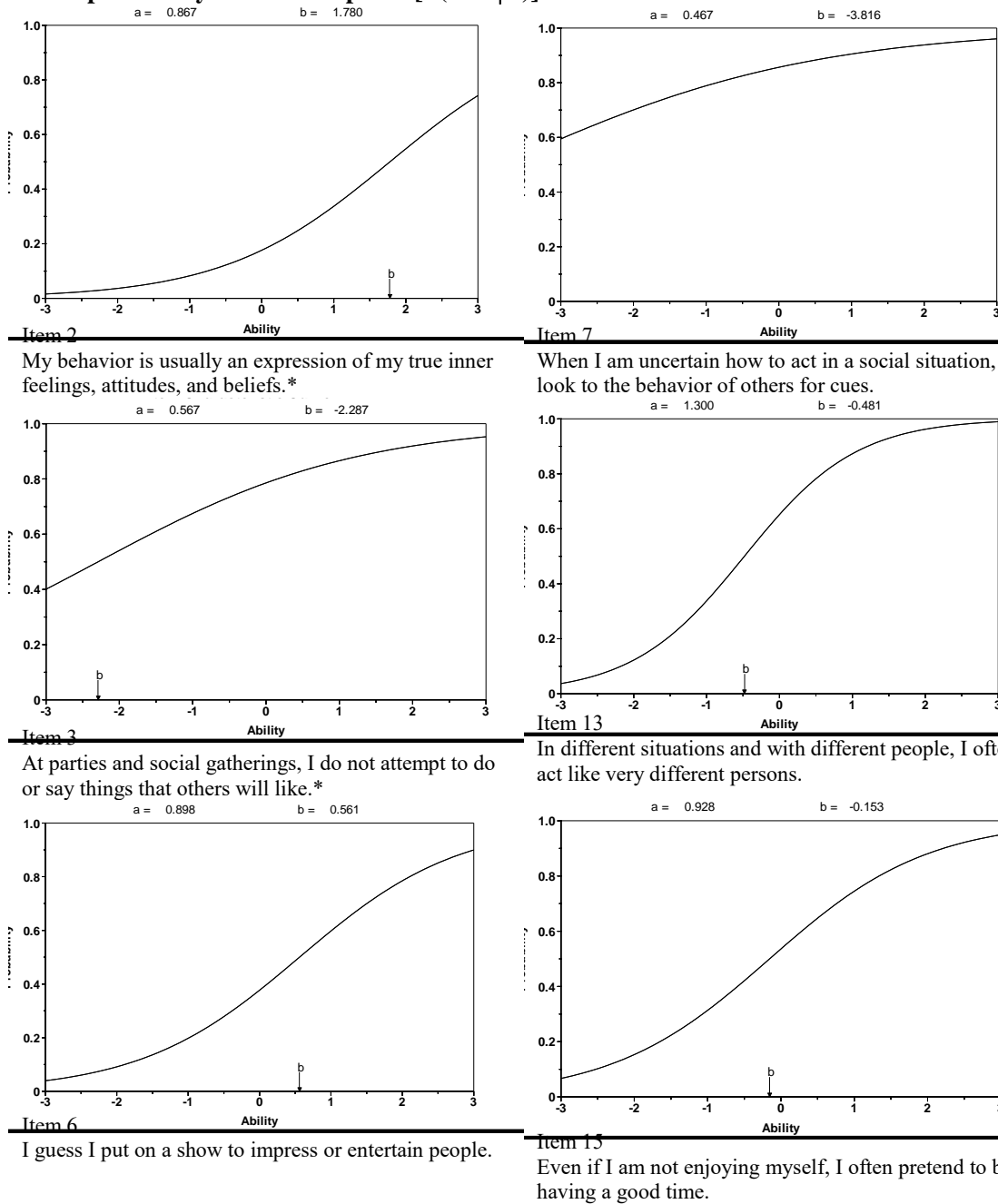
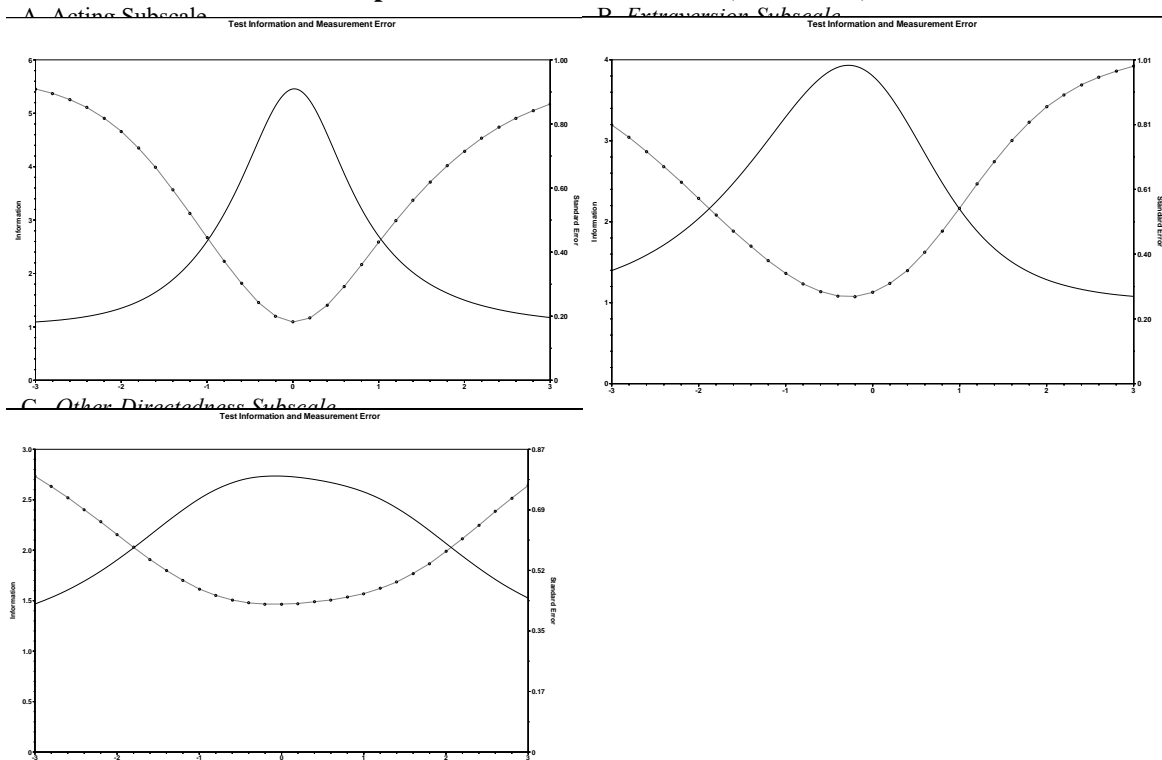


Figure 4
Test information functions for the Acting, Extraversion and Other-Directedness subscales (2PL models). The x-axis represents the latent construct for each subscale (θ). The left y-axis represents the amount of information provided by each subscale (solid line) and the right y-axis represents the amount of error (dotted line).



DISCUSSION

Summary

This paper represents the first item response theory analysis of the SMS subscales and thus offers an important contribution to the understanding of the Self-Monitoring Scale's psychometric properties. Importantly, the IRT analyses revealed aspects about the scale that mirror previous findings yet it also offered new information, clarification, and suggestions about its use.

Given that some researchers have suggested solely using the SMS subscales (e.g., John et al., 1996), separate IRT analyses were conducted for each subscale. For all three subscales, the IRT analyses showed that the unrestricted model provided a superior fit to the data. Importantly, this indicates that the subscale items are not equal in their relationship to their respective underlying latent construct. Further inspection of the item parameters demonstrated that some subscales displayed greater variability than others, with the Other-Directedness subscale displaying the greatest amount of variability.

The Acting subscale possessed the best psychometric properties. Three of the five items showed very large slopes, indicating that the majority of the items are highly related to the construct of acting. The IRT analysis revealed that items 8, 18, and 20 best captured the acting construct. Further, the thresholds of the items were all close to zero, indicating that these items are differentiating between individuals in the middle of the acting continuum. This property may be desirable when the construct is thought to be normally distributed in the population.

Next, the Extraversion subscale showed reasonable psychometric properties. Three of the six items showed strong slopes, indicating that half of the items are highly related to the construct of extraversion. Items 12, 22, and 23 best captured the extraversion construct. The thresholds of the items were all negative, indicating that the items are primarily identifying individuals low in extraversion. That is, the Extraversion subscale is adequate for identifying people who are low in extraversion. However, this subscale is inadequate for researchers interested in distinguishing among people who are high in extraversion. This information is critical for researchers interested in studying highly extraverted individuals using the SMS.

Finally, the Other-Directedness subscale demonstrated poor psychometric properties. None of the items had strong slopes. Out of the eleven items, only three showed moderate slopes, four of the items had extremely weak slopes, and the rest were in between this range. This suggests that none of these items allow for fine distinctions within the construct of other-directedness. It is important to note that the item with the strongest relation (item 19) was deleted from the original 25-item SMS to create the revised 18-item scale (Snyder & Gangestad, 1986). As previously stated, many argue that this revision shifted the SMS toward Extraversion and Acting and weakened the Other-Directedness construct (Hoyle & Lennox, 1991; John et al., 1996, Briggs & Cheek, 1988). The present analyses reveal that this is most likely because the revised SMS removed the strongest item of the Other-Directedness subscale. Finally, the thresholds of these items were mostly negative, with two items having extremely negative thresholds. Because many of the items do not strongly relate to the construct, their thresholds are more variable and less informative. Upon examination of the few items that did show a moderate relation to other-directedness (i.e., items 13, 16, and 19), one can see that two of the thresholds are negative and one is positive. Thus, together these items are differentiating the upper and lower ends of the continuum.

In sum, the present IRT analysis suggest that the Acting subscale has good psychometric properties, the Extraversion subscale is only useful in identifying Introverts, and the Other-Directedness subscale is a poor measure of its construct. Researchers interested in using the SMS should be aware of these limitations so that they can decide if this measure is appropriate for their particular research needs.

Implications and Recommendations

These results offer some pragmatic suggestions for researchers utilizing the SMS. First, in accordance with other researchers (John et al., 1996), it is recommended that the original 25-item SMS be used over the 18-item scale. This is primarily because the 18-item scale deletes the strongest item from the Other-Directedness subscale, making a valid and reliable measure of this construct tenuous. Furthermore, the present analyses found the unrestricted models for all three subscales provided a superior fit to the data, indicating that the SMS subscale items are not equal in their relationship to their respective underlying construct. This calls into question the common practice of creating a summed score for each subscale. Instead, researchers may want to adopt a weighted scoring system such that items with stronger slopes are weighted

more heavily. Alternatively, researchers could omit the weakest items from the subscale before computing a summed score.

Finally, the results offer suggestions for the use of each subscale. For the Acting subscale, the thresholds all approximated zero. This implies that the subscale is useful for researchers interested in identifying two broad groups of individuals, those high and low in acting. However, if the researcher's goal is to make more finite distinctions, such as identifying those very high/low, then this subscale is unsuited for their task. Instead, one would need to create Acting subscale items that have greater variability in their threshold parameters and thus discriminate among varying levels of acting. For the Extraversion subscale, the thresholds were all negative. This implies that the subscale is only useful for researchers interested in identifying individuals low in extraversion. Only a small level of extraversion is required for endorsement of these items; therefore, they do not distinguish between moderate and high extraverts. If the researcher's goal is to separate those individuals who are low and high in extraversion, this subscale is not adequate for this purpose. Instead, one would need to create Extraversion subscale items that reflect the upper end of the continuum by having large positive threshold parameters. For the Other-Directedness subscale, the thresholds were mostly negative, with some extremely low thresholds. First, it is clear that none of the items are adequately assessing the other-directedness construct. Furthermore, several of the items are extremely poor and should possibly be omitted. For example, item 23 is the poorest item in the Other-Directedness subscale, but it is among the strongest items identified in the Extraversion subscale. It is therefore recommended that this item be removed from the Other-Directedness subscale and retained solely as an indication of extraversion. Also, more items should be developed that more strongly relate to other-directedness and that differentiate at the upper levels of this construct.

CONCLUSION

Item response theory is a useful psychometric tool and personality researchers can greatly benefit from IRT applications. Unlike more traditional approaches, IRT provides a detailed item-by-item analysis and allows researchers to identify when a particular scale is most useful and when it is an inappropriate application (for review of the use of IRT in personality, see Reise & Henson, 2003; Rouse, Finger & Butcher, 1999). The current IRT application provides useful information to anyone interested in studying the construct of self-monitoring. Armed with this information, researchers will be better able to understand what the SMS does and does not measure.

REFERENCES

- Briggs, S. R., & Cheek, J. M. (1988). On the nature of self-monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology, 54*, 663-678.
- Briggs, S. R., & Cheek, J. M., & Buss, A. H. (1980). An analysis of the Self-Monitoring Scale. *Journal of Personality and Social Psychology, 38*, 679-686.
- Dragow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

- Gangestad, S. W., & Snyder, M. (1985). To carve nature at its joints: On the existence of discrete classes in personality. *Psychological Review*, *92*, 317-349.
- Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item-response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*, 443-451.
- Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.
- Hoyle, R. H., & Lennox, R. D. (1991). Latent structure of self-monitoring. *Multivariate Behavioral Research*, *26*, 511-540.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- John, O. P., Cheek, J. M., & Klohnen, E. C. (1996). On the nature of self-monitoring: Construct explication with Q-sort ratings. *Journal of Personality and Social Psychology*, *4*, 763-776.
- Muthén, L. K., & Muthén, B. O. (2000). *Mplus User's Guide*. Second Edition. Los Angeles, CA: Muthén & Muthén.
- Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO-PI-R. *Assessment*, *7*, 347-364.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*, 93-103.
- Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *72*, 282-307.
- Snyder, M., & Ickes, W. (1985). Personality and social behavior. In S. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol 3, pp. 883-946). New York: Random House.
- Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, *30*, 526-537.
- Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology*, *51*, 125-139.
- Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item and test scoring using item response theory*. Chicago: Scientific Software.
- Thissen, D., Chen, W-H., & Bock, R. D. (2003). *Multilog (version 7)* [Computer software]. Lincolnwood, IL: Scientific Software International.
- Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. Mahwah, New Jersey: Lawrence Erlbaum.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.