

# Measurement and the Study of Change

Michael C. Edwards  
*The Ohio State University*

R. J. Wirth  
*University of North Carolina at Chapel Hill*

Many constructs developmental scientists study cannot be directly observed. In such cases, scales are created that reflect the construct of interest. Observed behaviors are taken as manifestations of an unobserved common cause. As crucial as measurement is to understanding many psychological phenomenon, it is perhaps even more important when the goal of research is to understand how a construct changes over time. In this article we review several approaches to measurement, note features of latent variable measurement models which are ideally suited to the study of change, describe a hypothetical example, and conclude with a discussion of measurement and development.

## INTRODUCTION

Developmental research often focuses on constructs (e.g., depression, anxiety, etc.) that cannot be directly observed. In these instances, it is common to operationalize a construct using a set of items. To the extent that these items are valid indicators of the construct, the scores derived from these items (i.e., the scale) will provide useful information about an individual's level on that construct. These scores are then treated as a manifestation of the unobserved construct. A number of statistical models exist to support this endeavor, ranging from long-standing techniques such as classical test theory (CTT) to more recent models such as factor analysis and item response theory (IRT).

In practice, we recognize that latent variable–based measurement models are more complex to implement than standard CTT-based measurement models. We believe that the more complex models are worth the extra effort but acknowledge that the burden of proof rests on our shoulders to convince developmental researchers that they should put forth the time and effort to use these models in their own research. As we discuss in detail in this article, we feel the benefits of these models far outweigh the costs. Among the benefits we count: the ability to appropriately model changes in constructs that occur over time, the ability to use different sets of items while maintaining comparability of scores, increased likelihood of recovering the true underlying developmental trajectories of a construct, scores that more accurately reflect the operationalized construct, and the possibility of acquiring powerful construct validity evidence.

We begin by discussing the importance of measurement and the additional complexities that arise when studying development. Following this we review two of the more widely used modern measurement frameworks, factor analysis and IRT (known collectively as “item-factor analysis”; see Wirth & Edwards, 2007). Once a general foundation has been established for both modeling frameworks, we turn our attention to features of both frameworks that are relevant to the study of change. These features include item-level weighting, equating, invariance, and their relationships to validity. To help illustrate these ideas we provide a hypothetical example consistent with the repeated assessment of a single construct over three time points. We use the term *time* throughout this article to denote important periods of assessment. Time could easily be replaced with *age*, *period*, *cohort*, or a number of other developmental markers without loss of generality. In our example different combinations of scales are presented at all three time points, yet we demonstrate that the methods presented in this article still accurately model change (e.g., mean increase) in the construct over time. The article concludes with a general discussion of measurement in the context of developmental research.

## THE IMPORTANCE OF MEASUREMENT

The overarching goal of measurement is to differentiate between (and within, in a developmental context) individuals on a given construct. Measurement of this sort has always been a part of the social sciences. Psychophysics, one of the earliest areas of psychological research, used observable measurements to better understand unobservable concepts such as sensation and perception. As the focus of psychology expanded, so did the number of unobservable constructs that were of interest. A prime example is intelligence. It is no accident that some of the first researchers interested in intelligence were also some of

the earliest psychometricians (Galton, 1888; Spearman, 1904; Thomson, 1919; Thurstone, 1934, to name a few). In fact, many of the methods we use today are variations on methods developed by these individuals. It is worth noting that though the ideas for these methods have been around for quite some time it is only in the past decade or two that estimation and software have made them widely accessible.

Intelligence is a useful example to demonstrate why measurement is such an important part of the research process. It is not possible to directly measure a person's intelligence. We must rely on indirect indicators we believe represent intelligence. For any given indicator, it is useful if differential performance on that indicator corresponds to different levels of a construct. In the case of traditional intelligence tests, a "correct" response would indicate a higher level of intelligence than an "incorrect" response. As more indicators are presented to a subject and their responses are observed we gain more evidence about their specific level of intelligence.

The scores that result from the administration of a scale are generally regarded as though they directly represent the construct of interest. However, this belief is only true to the extent that the operationalization of the construct (i.e., the set of items) is valid. For example, scores from a test of addition and subtraction would not be informative about how depressed an individual is. Such a test would not be considered a valid operationalization of the construct of depression. On the other hand, a scale containing items reflecting negative affect as well as changes in an individual's eating and sleeping patterns would reflect many researchers' beliefs about depression. This is an extreme example, but it highlights the critical role items play in the operational definition of a construct.

## MEASUREMENT AND THE STUDY OF CHANGE

If the operationalization of a construct is stable over time, one definition (and set of items) may be a valid representation of the construct regardless of the age of assessment. In other cases, as individuals age the usefulness of a single set of items (i.e., the time- or age-specific operational definition of the construct) may change. There are at least two reasons one may expect behavioral indicators of a construct to change over the course of development: range restriction and differential construct operationalization.

Range restriction (i.e., floor and ceiling effects) occurs when individuals who have different underlying levels of a construct receive the maximum or minimum score on a scale. Although the individuals may be different, the scale in question is unable to detect these differences. In the context of mathematics, this is one

reason items used to assess the ability of an 8-year-old look very different from those used to assess the ability of an 18-year-old. Here the items change not necessarily due to any shift in operationalization, but due to changes in ability levels. The items used to assess 8-year-olds are still relevant for 18-year-olds, but given the hypothesized increase in mathematical ability the items are no longer informative. Range restriction is not limited to the realm of achievement. In the assessment of depression, a scale may be particularly well suited to assessing subclinical levels of depression. However, the same scale may result in every clinically depressed individual receiving the maximum possible score. Although there may be meaningful differences among these clinically depressed individuals, this will not be reflected in their scores. Once again, though the items are relevant, they are not particularly informative.

A second reason indicators may need to change over time is that the operational definition of a construct may change as individuals develop. Psychology has many examples of constructs where the behavioral manifestation changes as a function of the developmental period of the individuals in question (Rutter & Sroufe, 2000). For instance, when assessing psychological constructs in school-aged children, it is very common for a number of the indicators to be related to school-based behaviors. Given how large a role school plays in a child's life, this is an obvious way to gain relevant information. However, a set of items that represent an operationalization of a construct for a school-aged child may be much less relevant as they leave the school environment. One could imagine many school-related items being replaced with work-related items in a scale meant to assess adults. A similar shift would occur once adults reach retirement age. At that point, a new set of items would be required to tap into the elements previously assessed using work-related behaviors. Unlike the case of range restriction, when the operational definition of a construct changes it is possible that items will become more-or-less relevant. This change in relevance can take a number of forms ranging from a slow increase/decrease in relevance over time to a complete change from relevant to irrelevant (or vice versa) at a given time of assessment.

## MEASUREMENT MODELS

Measurement models are the statistical representation of how items (or scales) are related to the operationalized construct. In this section we review three of the more widely used measurement models: CTT, factor analysis, and IRT. A plethora of course- and book-length treatments exist on all three models. What follows are very brief introductions with an emphasis on the features of the models that most affect the study of change.

## Classical Test Theory

One early and still popular measurement model is true score theory, which is a fundamental element of CTT. The true score model is typically written as

$$X_{ikt} = \tau_{ikt} + e_{ikt}, \quad (1)$$

where  $X_{ikt}$  is the observed score of person  $i$  on exam  $k$  at time  $t$ ,  $\tau_{ikt}$  is the “true score” of person  $i$  on exam  $k$  at time  $t$ , and  $e_{ikt}$  is the error term that is assumed to have an average of zero and be uncorrelated with the true score. There are several ways to understand the true score model. The true score can be thought of as the score we would observe if we could measure the construct without error, although technically it is an expected value. That is, it is the observed score we would expect a person to obtain on a given scale if that individual completed the scale infinitely many times, and we averaged over those repeated assessments.

Although not typically regarded as such, true score theory can be considered a latent variable model (LVM). There are a number of ways to define a latent variable (see Bollen, 2002), but one popular mode of thought views latent variables as a distribution of scores on an unobserved factor or construct. The CTT true scores described above fall into this category, as they are distributions of scores which are unobservable. However, the true scores of CTT are not the same as latent scores hypothesized in other LVMs. In CTT, a true score exists for each individual for each scale—the score is scale dependent. The  $i$  and  $k$  subscripts on the  $\tau$  in Equation 1 above highlight this fact. The implication is that there is no score unique to the person that exists outside of a given test. Based on CTT, any change to a scale results in changes to the individual’s true score. These changes can include even minor modifications to the items contained on a scale. More details are provided in later sections about how the inability to easily compare scores across different versions of a scale (or different scales) can complicate the study of development. We next turn our attention to two modeling frameworks that are more flexible (impose fewer constraints) than the true score model presented above.

## Factor Analysis

Similar to true score theory, factor analysis assumes that individuals possess some true unobservable level of a construct. Conceptually, factor analysis models the relationship between a latent distribution of scores and the observable manifestations of those scores (the item responses). Technically, factor analysis models the common variance among a set of items. The common variance represents the true variance of the construct. An important consequence of the factor analysis method is that the model is item independent (Widaman, Cudeck, &

MacCallum, 2007). Items that are valid measures of the same construct can be added or removed from a factor analysis model without changing the definition of, or person's latent score on, the construct.

Factor analytic methods are item independent regardless of whether we are interested in conducting an exploratory or confirmatory factor analysis (EFA and CFA, respectively). EFA is useful when a researcher has no *a priori* hypotheses about which items define which constructs or when examining an item set for possible violations of unidimensionality (e.g., a set of items may be measuring more than the intended construct). Although we find this latter use of EFA quite helpful, it is rare to find circumstances where no *a priori* beliefs exist regarding the factor structure of an item set. Unlike EFA, CFA offers a way to formally test the structure of an item set. Along with theory supporting the item content, a CFA model that can account for the observed data lends further support to assertions regarding validity.

Factor analysis involving categorical outcomes (e.g., Likert-type item responses) poses a number of technical challenges (e.g., nonlinear parameters, alternative estimation techniques, etc.), which have been addressed extensively elsewhere (see, e.g., Bartholomew, Steele, Moustaki, & Galbraith, 2002; Mislavy, 1986; Takane & de Leeuw, 1987; Wirth & Edwards, 2007). Here we focus only on key concepts and parameters as well as their role in modeling developmentally relevant constructs.

The measurement characteristics of an item can be described within the factor analysis framework using two parameters,  $\tau$  and  $\lambda$ . The categorical item response ( $x$ ) is assumed to be a discrete manifestation of an underlying, continuous response ( $x^*$ ). The  $\tau$  parameter (i.e., threshold) helps to link these two representations. The  $\lambda$  parameter (i.e., factor loading) describes the strength of the relationship between  $x^*$  and the construct. Note that the  $\tau$  presented here is not the same as the  $\tau$  presented in the true-score model (Equation 1). Although the similarity in notation can be confusing, we have remained consistent with the notation used in the extant literature.

The threshold parameters can be defined as

$$x_{ijt} = c, \text{ if } \tau_{cjt} < x_{ijt}^* < \tau_{(c+1)jt}, \quad (2)$$

where  $\tau_{cjt}$  defines the threshold or "cut-point" on a latent response distribution between response category  $c$  and category  $c + 1$  for the observed response  $x$  to item  $j$  at time  $t$ . The first threshold for a given item denotes the  $z$ -score on the standard normal distribution<sup>1</sup> that has to the left of it an area under the curve

---

<sup>1</sup>The standard normal assumption for the latent response distribution is not required (Pearson, 1913). However, at the time of this publication all factor analytic software makes this assumption.

equal to the proportion of individuals who endorsed the first category (e.g., *strongly disagree*). The second threshold denotes the  $z$ -score that has to the left of it an area under the curve equal to the proportion of individuals who endorsed the first two categories (e.g., *strongly disagree* and *disagree*). There are  $C - 1$  thresholds for an item with  $C$  categories (e.g., there are four thresholds for an item with five response options). The threshold parameters play an important role in understanding the measurement properties of an item. In the dichotomous case, a high threshold suggests fewer people endorsed the item and thus individuals must possess more of the latent construct to endorse the item.

The factor loading ( $\lambda_{ij}$ ) for item  $j$  at time  $t$  is a measure of the relationship between the factor (i.e., construct) and the item at a given point in time. A larger  $\lambda$  indicates that an item is a more reliable indicator of the construct compared to an item with a smaller (in absolute magnitude)  $\lambda$ . Formally the relationship between an item and the latent factor at a given time point can be described as<sup>2</sup>

$$x^*_{ijt} = \lambda_{jt}\xi_{it} + v_{ijt} + \varepsilon_{ijt}, \quad (3)$$

In this formulation  $x^*_{ijt}$  denotes individual  $i$ 's latent response to item  $j$  at time  $t$ . The response is a function of individual  $i$ 's latent score on construct  $\xi_{it}$  at time  $t$  weighted by the relationship ( $\lambda_{jt}$ ) between item  $j$  and the construct ( $\xi_t$ ) at time  $t$ . There are two residual variances:  $v_{ijt}$  and  $\varepsilon_{ijt}$ .  $v_{ijt}$  denotes individual  $i$ 's item-specific deviation from the mean of  $\xi_t$  at time  $t$  and  $\varepsilon_{ijt}$  denotes measurement error. Both are assumed to have a mean of zero and be uncorrelated with each other and with the latent factor. Although the subscripts can be difficult to follow, they highlight that each item can have a unique relationship to the construct and that these relationships can differ over time.

This introduction of the categorical factor model focuses on three important measurement characteristics that are not found in CTT's true score model. First, factor analysis allows each item to have a unique factor loading and set of thresholds, which in turn allows items to be differentially weighted. Beyond different weights being applied to different items within a single time point, this enables us to model situations where the same item behaves differently over time. Second, given item- and time-specific weighting, factor analysis has the ability to partition observed variability to provide unbiased estimates of the

---

<sup>2</sup>There are several ways this model can be presented. Although Equation 3 does not include a threshold or intercept term, the information from the thresholds contributes to the estimation of the correlations among the latent response distributions (known as polychoric correlations).

true variability in the latent construct (Bollen & Lennox, 1991). When a factor analysis model is defined over multiple time points, these two features allow for an item's relationship to the construct to change over time while (given certain constraints) continuing to provide an unbiased estimate of the variability of the latent construct over time. And third, because of the partitioning of variances, the factor model is item independent. This means that an item used at one time of assessment does not necessarily have to be used at all times of assessment.

These three features offer great flexibility in measuring constructs in a developmental context. Before exploring the impact of these model characteristics on the measurement of constructs over time, we will introduce IRT, a closely related latent variable measurement model.

### Item Response Theory

The name item response theory is somewhat misleading, as it is really a collection of many different measurement models. One widely used IRT model is the two-parameter logistic model (2PLM), which was developed explicitly for dichotomous item responses. In the 2PLM, the probability of a particular individual endorsing an item at a given time period can be expressed as

$$P(x_{ijt} = 1 | \theta_{it}) = \frac{1}{1 + \exp[-a_{jt}(\theta_{it} - b_{jt})]}, \quad (4)$$

where  $x_{ijt}$  is an observed response from individual  $i$  to item  $j$  at time  $t$ ,  $\theta_{it}$  is a latent score for person  $i$  at time  $t$ ,  $a_{jt}$  is a slope parameter describing the strength of the relationship between item  $j$  and the latent factor at time  $t$ , and  $b_{jt}$  is a severity or threshold parameter describing the amount of the latent construct someone must possess to have a 50% probability of endorsing item  $j$  at time  $t$ .

One beneficial feature of IRT models is how readily they lend themselves to graphical displays. These are called trace lines or item characteristic curves (ICCs). Figure 1 contains trace lines of four items that differ both in terms of their slopes ( $a$ ) and their severity parameters ( $b$ ). The items in the top panel have the same slope, but different severity parameters. An item with a higher severity parameter requires a higher level of the construct for there to be a 50% chance of the item being endorsed. The items in the bottom panel have different slopes, but the same severity parameters. As the slope increases, the steepness of the resulting logistic curve increases. This corresponds to a faster change in the predicted probability of endorsement as the level of the latent construct increases.

Another widely used IRT model in the social sciences is the graded response model (GRM; Samejima, 1969). The GRM is appropriate when there are more

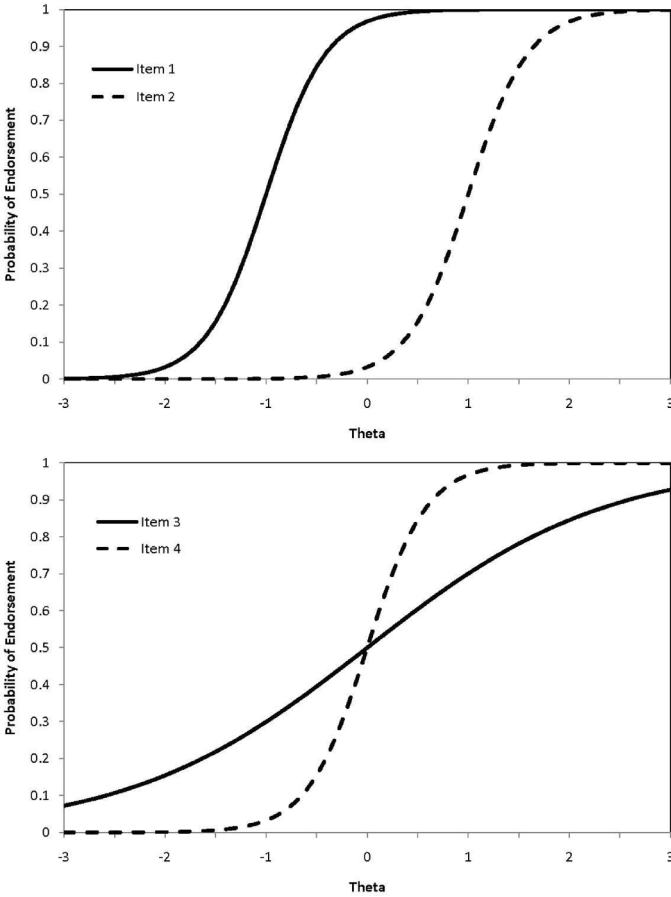


FIGURE 1 2PLM trace lines for four dichotomous items.

*Note.* Item 1 has a slope of 2 and a severity of  $-1$  whereas Item 2 has a slope of 2 and a severity of 1. Item 3 has a slope of 0.5 and a severity of 0 whereas Item 4 has a slope of 2 and a severity of 0. The y-axis is the probability of endorsement, and the x-axis is the level of the latent construct (which is assumed to follow a standard normal distribution).

than two ordered responses. The GRM expresses the probability of choosing a particular response category as

$$P(x_{ijt} = c | \theta_{it}) = \frac{1}{1 + \exp[-a_{jt}(\theta_{it} - b_{cjt})]} - \frac{1}{1 + \exp[-a_{jt}(\theta_{it} - b_{(c+1)jt})]}, \quad (5)$$

The definitions for  $a$  and  $\theta$  are the same as in the 2PLM (Equation 4). Unlike the 2PLM, where there is only one severity parameter, in the GRM there is a set of

severity parameters. There are  $C - 1$  severity parameters where  $C$  is the number of response alternatives for a given item. Expressed verbally, Equation 5 states that the probability of choosing a particular category ( $c$ ) is the difference between choosing that category or higher and the probability of choosing the next category ( $c + 1$ ) or higher.

As with the 2PLM, it is also possible to plot trace lines for the GRM. Trace lines for four 5-category items are shown in the four panels of Figure 2. Items in the same row have the same slope parameter ( $a = 2$  in the top row and  $a = 1$  in the bottom row), and items in the same column have the same set of severity parameters. A higher slope suggests that an item is more discriminating than an item with a lower slope. The more discriminating an item is, the more distinct the probabilities will be for choosing a particular category at a particular level of the latent construct. In the GRM, higher slopes result in more peaked response functions, as can be seen by comparing the top row to the bottom row. The set of severity parameters for the right column items in Figure 2 are higher than those for the items in the left column. The right column items are more severe. This is represented in the trace lines in the right column being shifted to the right. In terms of response behavior, the different severity parameters mean that choosing response category two (for example)

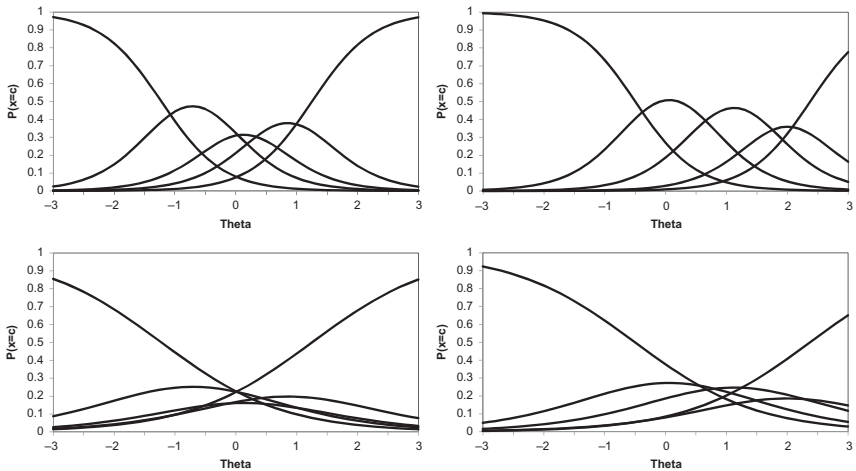


FIGURE 2 GRM trace lines for four 5-category items.

*Note.* The items in the top row have slopes of 2 and the items in the bottom row have slopes of 1. The items in the left column have severity parameters of  $-1.22$ ,  $-0.19$ ,  $0.46$ , and  $1.25$ , and the items in the right column have severity parameters  $-0.5$ ,  $0.62$ ,  $1.63$ , and  $2.38$ . In all four plots the y-axis is the probability of choosing a particular category and the x-axis is the level of the latent construct (which is assumed to follow a standard normal distribution). GRM = graded response model.

requires higher levels of the construct in the two items on the right than it does for the two items on the left.

The categorical factor analysis model presented in the previous section and the IRT models just described are essentially different parameterizations of the same latent variable measurement model (Takane & de Leeuw, 1987). Although estimation and implementation differences exist (see Wirth & Edwards, 2007), the measurement characteristics described for the factor analytic model (e.g., item-level weighting, item independence) also hold for IRT. We now turn our attention to how the flexibility of factor analysis and IRT offers greater freedom in measuring constructs over time.

## STUDYING CHANGE WITH MODERN MEASUREMENT MODELS

This section focuses on the aspects of factor analysis and IRT that are advantageous when studying development. We begin with a discussion of item-level weighting and how this affects the creation of individual scores as well as scale construction. After this we discuss the concept of equating, with an emphasis on how equating scales gives researchers greater flexibility when modeling a construct that is operationalized differently over developmental periods. We end this section with a discussion on measurement invariance and how violations of measurement invariance, traditionally viewed as problems to be solved, may in fact represent extremely strong tests of construct validity.

### Item-Level Weighting

As opposed to CTT models, which assume unit weighting for items, factor analysis and IRT allow the items to contribute differentially to an individual's score. Although many researchers view CTT as a simpler model than factor analysis or IRT, it is in several senses a constrained version of either model. To make a factor analysis or IRT model correspond more closely with CTT, one would impose constraints such that all items had the same factor loading (or slope) and threshold(s)<sup>3</sup> (or severity parameter[s]). In this light, CTT can be viewed as a nested model within the more general factor analytic or IRT frameworks. Despite the widespread popularity of CTT, in our experience these constraints diminish the ability of the model to account for the observed data. There is indeed something odd about the common practice of using factor analysis to establish the

---

<sup>3</sup>It is unusual to consider constraining thresholds in the factor analytic framework. However, by referring to any of the equations relating IRT parameters to factor analysis parameters (for instance, Equations 5 and 6 in Wirth and Edwards, 2007) it can be shown that there is a direct relationship between constraining severity parameters in IRT and thresholds in factor analysis.

dimensionality of a scale but then ignoring the parameter estimates themselves when creating scale scores. Statements about the adequacy of a model from a factor analytic standpoint may not apply when the parameters from that model are ignored.

Beyond model fit, the issue of item-level weighting manifests itself in at least two additional facets of measurement. First, there are implications for scoring. The goal of measurement is to assign a value to each individual that can serve as an estimate of their level of the construct being measured. If items are differentially related to that construct, it is important to have a system of scoring which incorporates these differences. Factor loadings and slopes contain information about how strongly items are related to the construct. Items that have a stronger relationship to the construct are essentially more reliable indicators of that construct. Factor analysis and IRT are able to differentially weight items relative to the amount of information they provide. There are also the threshold (or severity) parameters to consider. These too play an important part in the item-level weighting accomplished by factor analysis and IRT. A brief example should highlight how important it is to consider an item's severity level in scoring. Imagine a 10-item depression scale that includes as items "I felt sad" and "I tried to kill myself." Suppose that one individual endorsed the "I felt sad" item (but no others) and another individual endorsed the "I tried to kill myself" item (but no others). In CTT, without some sort of external (and typically ad hoc) weighting system, both subjects would receive the same score. However, it seems unlikely that anyone observing the pattern of responses would assess these two individuals as possessing the same level of depression. By incorporating the severity of the "I tried to kill myself" and "I felt sad" items, factor analysis and IRT are able to provide scores which go beyond the number of items endorsed to consider which items were endorsed. These IRT or factor scores can then serve as dependent variables in other analyses of the sort described by Grimm and Ram (2009), Hoffman and Stawski (2009), Selig and Preacher (2009).

The second way in which the item-level weighting affects measurement is on the construction of scales themselves. Once differential item weighting is available, it becomes possible to develop scales based on who the scale is meant to assess. For instance, a scale meant to track change on a highly variable construct would ideally have items spaced along the entire continuum of the construct being assessed (i.e., items with a wide range of severity). On the other hand, a scale designed to judge whether a person was above or below some diagnostic criterion would be most useful if it contained items with severity parameters in close proximity to the criterion level. By focusing on the latent cut-point of interest, such items would be the most informative regarding whether a particular individual was above or below the criterion.

Another validity-based advantage to the differential item weighting is that the scores from such a system will more accurately reflect the operationalization of

the construct. Most symptoms or behaviors are not viewed as equally important or equally severe in their relation to a construct, and factor analysis and IRT provide this level of differentiation. Indeed, with careful planning, empirically derived item weights can provide evidence of construct validity.

*Item-level weighting as construct validity.* In *Test Scoring*, Thissen and Wainer (2001) stated that “In casual terms, we can define validity as measuring the right thing, and reliability as measuring the thing right” (p. 11). The process of validating a scale is an ongoing accumulation of evidence that supports the assertion that the scale measures the construct it purports to measure. As with scientific theories in general, the more nuanced the predictions, the more strength they lend to the argument if they turn out to be correct. Within the domain of measurement, predictions about how each item is related to a construct (even relative to the other items) is a very precise prediction and would therefore lend strong support to the claims that a scale does indeed “measure the right thing.”

Although it may be difficult in all applications to explicitly predict the relationship between each item and the construct of interest, doing so offers a unique opportunity to provide falsifiable hypotheses regarding construct validity. In our example above we presented two hypothetical depression items. In an analysis including these two items, the “suicide” item would be hypothesized to be more severe than the “sad” item and likely more severe than all other depression items. Should such predictions be correct, the results should be taken as, at least in part, evidence of construct validity (see Shadish, Cook, & Campbell, 2002, for a discussion of the many aspects of construct validity).

## Equating

An important outcome of item-level weighting is that it becomes possible to relate scores from one set of items to a different set of items (e.g., two different scales) that assess the same construct. This is only part of the necessary model features to enable such a comparison—there must also be an invariant person score. As described earlier, true scores in CTT are test specific and do not generalize beyond a given person/test interaction. In this setting, there is no easy basis on which to relate scores from one scale to those on another. In factor analysis and IRT, in contrast, it is assumed that individuals possess an internal (but unobservable) level of the latent construct that exists independently from the scale used.

The existence of the latent person score means that two different scales assessing the same operationalized construct could realistically be viewed as attempting to estimate the same quantity. However, without additional work, there is no guarantee that the scores they produce will be directly comparable. This is where the item-level

weighting again proves invaluable. By taking into account the different properties of the scales, it is possible to perform an equating procedure that will make the scores from the two scales directly comparable. Equating procedures range widely in difficulty of implementation, but given preplanning it is often possible to equate scales with little effort. One widely used form of equating, known as common item equating, uses a set of invariant common items (also known as an anchor) to insure that any noncommon items are kept in the same metric. For an excellent overview of equating issues in the IRT framework see Kolen and Brennan (2004).<sup>4</sup>

The use of two different scales is just one of the possible reasons respondents may see different sets of items. Another reason, which is particularly salient to the study of change, relates to our earlier point about items changing to reflect the differential operationalization of a construct over time. If the operational definition of a construct changes over time, it is not hard to understand the need for different items. By using equating, it is possible to use different sets of developmentally appropriate items but still obtain scores which are directly comparable.

Up to this point we have focused on the role equating can play when the desire is to create comparable scores between different sets of items. However, it is also possible for items to become more or less relevant over time, or more or less severe. This represents changes in the item-level weights that, as far as the statistical models are concerned, render items “different” over time. Fortunately, equating can play a role in this case as well by allowing the changing nature of an item’s relationship to the construct to be modeled in a way that maintains comparability of scores over time. This moves us toward the intersection of equating and measurement invariance, the topic we turn to next.

## Invariance

Whether dealing with  $\lambda$  and  $\tau$  parameters in factor analysis or  $a$  and  $b$  parameters in an IRT model, *invariance* refers to the equality and stability of the item parameters across groups and/or over time. For example, suppose a researcher assesses a group of individuals on three separate occasions using a delinquency measure with 10 dichotomous items. Each individual was assessed once when they were 8, 12, and 16 years old. For full item invariance to hold over time, the way in which each item is related to delinquency must remain constant across all

---

<sup>4</sup>Equating is one example of differing terminology and emphasis between the IRT and factor analysis literatures. Equating is not often discussed in factor analysis, and when it is discussed it is typically presented in the context of scaling and invariance. However, the concepts covered in the IRT literature translate directly to factor analysis.

time points. In a factor analysis model this would require that all  $\lambda$  and  $\tau$  parameters remain equal (within item) across all time points. In an IRT model this would require that all  $a$  and  $b$  parameters remain equal (within item) across all time points. Note that this does not imply that levels of delinquency are constant over time. Indeed, we would expect the frequency of endorsements to increase if a construct is increasing over time. With invariance, any change over time in observed response frequencies can be accounted for by changes in the latent construct. In the event that an item is non-invariant, the observed response frequencies are changing more (or less) than would be predicted by changes in the latent construct.

While the factor analysis and IRT literatures rely on different language, the underlying theory of invariance (or differential item functioning in IRT) remains unchanged. When the parameter values of a measurement model remain constant over time, there is evidence that the assessment tool is measuring the same construct at each assessment (Meredith & Horn, 2001). Moreover, when invariance holds, scores derived from measurement models at two or more time points are on the same scale (i.e., have been equated). Scores must be on the same scale to compare scores within or between individuals (Drasgow, 1984).

Traditional views of invariance, especially within the factor analysis literature, regard item-level noninvariance (i.e., changing item properties) as problematic. We view invariance (and noninvariance) in a different light; one much more consistent with the IRT literature. When a subset of items are invariant (called partial invariance), factor analysis and IRT models can be parameterized such that all scores remain on the same scale. If at least a subset of items is shown to be invariant, researchers retain the ability to accurately differentiate individuals. Of course, if some of the items have changing parameter values over time, the question of whether or not the same construct is being measured is a valid one. We argue that *a priori* hypotheses about item characteristics, including the presence of noninvariance (see, Horn, McArdle, & Mason, 1983), are not just testable within the factor analysis and IRT frameworks but offer a unique opportunity to assess construct validity. Thus, with at least partial invariance and strong *a priori* hypotheses about the invariance of the items, valid scores can be obtained while simultaneously strengthening the case for the scores' construct validity.

*Invariance as construct validity.* Cronbach and Meehl (1955) once wrote "Whether a high degree of stability is encouraging or discouraging for the proposed interpretation depends upon the theory defining the construct" (p. 288). Anxiety (Ferdinand, Dieleman, Ormel, & Verhulst, 2007), temperament (Durbin, Hayden, Klein, & Olino, 2007), reading ability (Fergusson, Horwood, Caspi, Moffitt, & Silvia, 1996), various aspects of personality (Caspi & Roberts, 2001; Kagan, 1980), and antisocial behaviors (Moffitt, 1993; Pajter, 1998) have all been

suggested to change in their manifestation over time. If a theory predicts change in the manifestation of a construct over time, this strongly suggests changes in the construct's measurement over time. Theory should predict which items remain invariant and which ones increase or decrease in the strength of their relationship to the construct or change in their relative severity.

This way of thinking about measurement invariance as evidence for construct validity stands in stark contrast to the predominant views of measurement invariance as a hurdle to be overcome. It is very common to see what we have termed "defensive" measurement invariance work that is conducted with the hopes of finding complete measurement invariance over time (or groups). In some cases, this may be all that is possible. Even if noninvariance is detected, not all noninvariance will be valuable in contributing construct validity evidence. One example of noninvariance that would not necessarily inform construct validity would be differences in item parameters due to a change in mode of administration (e.g., from paper-and-pencil survey to telephone survey). However, we believe theory often suggests that over time we should expect a dynamic relationship between some of the items and the operationalized construct. In either case, whether defensive or constructive, the ability of factor analysis and IRT to account for noninvariance is a powerful tool.

The previous sections have focused on demonstrating what researchers who are interested in studying change have to gain from more complex latent variable-based measurement models. From greater correspondence between measurement model and operationalized construct to opportunities to provide strong construct validity evidence, factor analysis and IRT provide significant advantages over CTT. In the next section, we use a simulated example to help show researchers how they can apply these models in their own work.

## A HYPOTHETICAL EXAMPLE

To provide a concrete illustration of some of the points raised throughout this article, we have created a simulated example. We do not provide complete details on all the generating values, but these (along with the generated data and software syntax) can be downloaded from the first author's Website.<sup>5</sup> In what follows we use the construct of delinquency to provide a context for the simulated example as well as to illustrate how one would interpret the results in a statistical sense. Neither author is an expert in delinquency, and we do not mean to suggest that these would be the results if a study like the one described here

---

<sup>5</sup><http://faculty.psy.ohio-state.edu/edwards>.

were actually conducted. Rather, our intention is to provide a statistical interpretation of one possible pattern of results.

The simulated example represents a situation where 3,000 individuals are assessed over three time points. At each time point individuals respond to two 10-item delinquency scales. For the purposes of this example, imagine that these assessments are conducted at three developmentally distinct junctures (at ages 8, 12, and 16), such that the scales that best measure the time-specific operationalization of delinquency change. A total of four scales are used across the three time points. Scale A (appropriate for ages 4–8) is used only at Time 1. Scale B (appropriate for ages 8–12) is used at Time 1 and Time 2. Scale C (appropriate for ages 12–16) is used at Time 2 and Time 3. Last, Scale D (appropriate for ages 16–20) is used only at Time 3. This structure is represented in Figure 3. We note that to use this sort of model in practice, one would need to provide evidence that the two scales used at any time point could be adequately accounted for by a

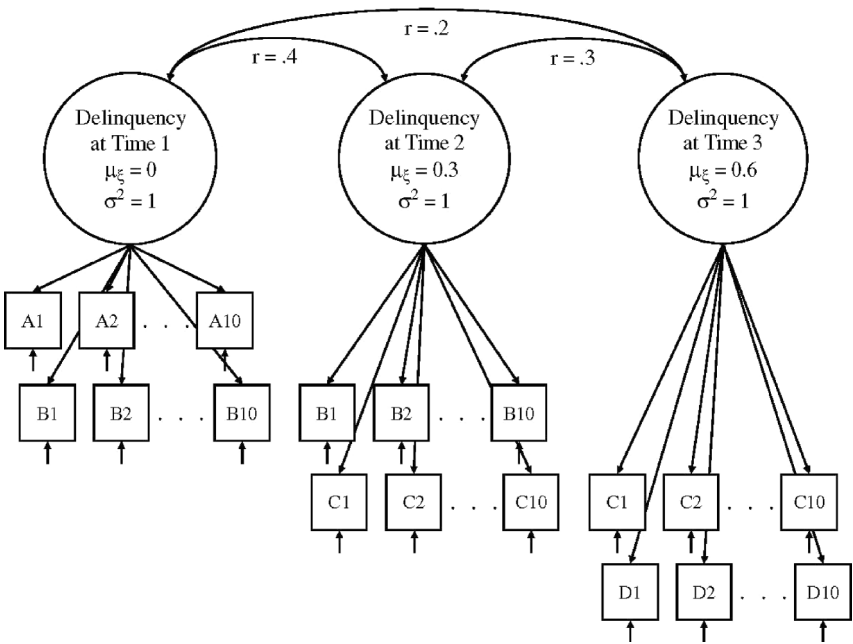


FIGURE 3 Path diagram of simulated example.

*Note.* Circles denote latent variables. Squares denote manifest variables (items). Single-headed arrows between two variables denote directional effects. Single-headed arrows with no variable of origin denote unique variances, and dual-headed arrows denote correlations. The mean ( $\mu_{\xi}$ ), variance ( $\sigma^2$ ), and correlations ( $r$ ) among the latent variables are provided. Factor loading and thresholds values are not provided.

single factor (i.e., are unidimensional). Rather than scoring each scale separately, a more precise score can be achieved by accumulating data from all available items administered at a given time point. The structure here is somewhat complex, although not unrealistically so for situations encountered by developmental researchers.

In this example delinquency increases over time, starting with a mean of zero (in the standard normal metric) at Time 1 and increasing by 0.3 each subsequent time point. For simplicity we have left the variance of the latent factors at one across all three time points, although this is not necessary. The data have been generated such that the scores across time are correlated as follows:  $r = 0.4$  for Time 1 and Time 2,  $r = 0.2$  for Time 1 and Time 3, and  $r = 0.3$  for Time 2 and Time 3. The data were simulated in the IRT framework, but the resulting data were analyzed in the IRT and factor analytic frameworks.

Due to differences in available software, the way in which the analyses are conducted is slightly different. Despite these differences, the results from the analyses are comparable. To highlight this comparability, the estimated latent means and covariances provided in Table 1 are based on the latent delinquency scores calculated from each set of results. As the results show, despite the fact that the items used to assess delinquency at each time point change, the means and covariances are recovered. Both modeling frameworks were able to recover the generating distributional properties of delinquency over time even though no item was seen at all three time points.

Relying on CTT results in a very different idea about how delinquency changes over time. Earlier, we mentioned the fact that in CTT, any values (such as person means or item endorsement rates) only have meaning for a single item set by person interaction. In the example we present here, there are three

TABLE 1  
Correlation and Mean Estimates over Time  
from IRT Scale Scores and Regression-Based Factor Scores

	<i>Time 1</i>	<i>Time 2</i>	<i>Time 3</i>
	<i>Correlations</i>		
Time 1	—	0.43	0.23
Time 2	0.38	—	0.35
Time 3	0.20	0.30	—
	<i>Means</i>		
IRT $\hat{\mu}$	-0.04	0.24	0.57
FA $\hat{\mu}$	0.00	0.29	0.63

*Note.* IRT = item response theory; FA = factor analysis.

IRT results are in the lower triangle, FA results on the upper. First three rows are correlation estimates, last two rows are mean estimates.

occasions of measurement each using a unique item set. From a CTT standpoint, this example yields three sets of nonequivalent values. To illustrate why this can be problematic, the 10 items constituting Scale D (used at Time 3) are more severe indicators of delinquency (harder to endorse) than the other items in the analysis. Although delinquency is increasing linearly over time, using CTT the observed mean scores are 1.99, 2.17, and 2.19, suggesting a more complex curvilinear trend. This is a direct result of CTT having no capacity to relate the severity of the items at the third time point to any of the earlier time points (Wirth, 2008). IRT and factor analysis, by virtue of item-level weighting and equating, “know” that the items at the third time point are more severe and weight those responses accordingly. Given this ability, both methods capture the linear trend over time (see Table 1) that generated the data. We do not present individual-level results, but the macro-level results described in this section are a direct result of similar trends in the underlying data.

The sample size used here is quite large by most standards, but sample sizes of this magnitude are not required to complete an IRT or factor analysis. As is always the case, with smaller sample sizes there is less information with which to estimate model parameters. We believe that at some point a sample becomes so small that the complexity of the model overwhelms the data’s ability to provide information. In these cases, it can be safer to use a more constrained model that requires less of the data. We tend to use a sample size of 200 as an approximate lower bound. However, we must stress that this is a very general guideline based on our personal experiences. Both authors have seen smaller sample sizes lead to excellent solutions and larger sample sizes yield inadmissible solutions. One hope we have for the future is that social scientists may begin to adopt an IRT-based item bank view of their scales. In such a world, great efforts would be made during the validation of a scale to obtain a large and representative sample of the population for which the scale was designed. If item parameters can be obtained from this large sample, they can be applied (without the need for re-estimation) to any subsequent samples from that population. To illustrate this point we examined scores for 300 new simulated respondents to see if the trends observed above still held. The observed mean scores were 2.01, 2.14, and 2.18 over the three time points and the observed IRT scores were  $-0.01$ ,  $0.2$ , and  $0.55$ . Although the latent means are not quite as well recovered as when the sample size is 3,000, the overall trend is still recovered reasonably well with the IRT scores.

## Invariance

There are many different ways to examine a set of items for invariance. For an excellent overview of methods for assessing measurement invariance we recommend Teresi (2006). This article is the lead article in a set of articles that describe various methods to assess measurement invariance. A number of these articles use

different techniques to assess the same scale (the Mini-Mental State Examination), which is particularly informative to understand how the various methods compare to one another. Because we used simulated data, we did not conduct any invariance tests. When we do conduct these tests, we prefer using the IRT-based likelihood-ratio differential item functioning (DIF) procedure (Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Thissen, Steinberg, & Wainer, 1993). However, either the IRT-based or factor analysis-based tests of invariance would have worked in the hypothetical example.

In the hypothetical results presented here, the first item on Scale B (B1) was generated to be non-invariant from the first to the second time period. Between the first and second time periods Item B1 becomes a weaker indicator of delinquency (slope decreases) and less severe (thresholds decrease). In terms of item content, such a shift in item parameters could be expected when a particular symptom or behavior becomes less relevant over time perhaps due to an increase in the frequency of the symptom/behavior that is attributable to some construct other than that which the scale is intended to assess. For example, suppose the item in question is related to defying parents. In the realm of delinquency, the parameter changes above imply that: (1) Between the ages of 8 and 12 the amount of defiance increases more than would be expected based on the increase in overall levels of delinquency in the population and (2) defiance as a behavior is not as strongly linked to delinquency at age 12 as it was at age 8.

The first item on Scale C (C1) was also generated to be non-invariant between Time 2 and Time 3, with the slope and threshold values increasing. Controlling for changes at the level of the latent variable, Item C1 is more strongly related to delinquency and the same response (i.e., a response of *agree* on a Likert-type scale) indicates a higher level of delinquency at Time 3 than it does at Time 2. Consider what this would mean if Item C1 was about hitting another person. We observe that the level of delinquency increases by 0.3 from Time 1 to Time 2 and again from Time 2 to Time 3. If the item remained invariant, we would expect to see hitting increase over time based on its particular set of item parameters. A lack of invariance means that at some point, the observed responses deviate from this expected trend. For the item in this example we would interpret the parameter values to mean that: (1) hitting becomes a stronger indicator of delinquency from age 12 to age 16 and (2) fewer instances of hitting occur at 16 than the model would predict. This shift in the threshold leads to the same level of hitting over both time periods to be treated as more severe at age 16 than at age 12.

From a modeling perspective, there is no difference between two different items at two time points and a single item that has a different relationship to the construct at two time points. As long as the changing relationship is modeled, as it is in the above example, it does not have a disruptive effect on measurement at any particular time point or on the ability to recover trends over time. This is again due to the ability of IRT and factor analysis to incorporate time-specific,

item-level weighting. As long as some invariant item set exists between two time points (i.e., an anchor), the mechanics of equating can be used to provide valid scores despite other changes to the item set (e.g., adding/deleting items or noninvariance). In general, the larger the anchor, the more stable the equating procedure will be.

## CONCLUSION

We began this article by highlighting the important role of item content in the operationalization of a construct. The results illustrated in the previous section demonstrate that, by adopting measurement models such as IRT and factor analysis, researchers are no longer required to choose one set of items to assess a construct when studying change. This flexibility enables researchers to respond to changes in the operationalization of a construct over time. The ability to insure that the measurement instrument remains in close agreement with the operational definition of a construct is a tremendous asset when building the case for the validity of a set of scores. Added to this is the idea of constructive noninvariance, which provides an extremely detailed way to support arguments that one is “measuring the right thing.” Viewed as a whole, the measurement models described here (i.e., IRT and factor analysis) represent a significant step forward for the measurement of constructs and the ability to understand how those constructs develop.

## ACKNOWLEDGMENT

The authors would like to thank Bill Gardner and Li Cai for helpful comments on earlier drafts of this manuscript. We would also like to express our appreciation to Denis Gerstorff, Nilam Ram, and three anonymous reviewers. The resulting paper is stronger for the input of all seven individuals. Any deficiencies that remain are, of course, our own.

## REFERENCES

- Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The analysis and interpretation of multivariate data for social scientists*. New York: Chapman & Hall/CRC.
- Bollen, K. A. (2002). Latent variables in psychology and social sciences. *Annual Review of Psychology*, 53, 605–634.
- Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation prospective. *Psychological Bulletin*, 110, 305–314.
- Caspi, A., & Roberts, B. W. (2001). Personality development across the life course: The argument for change and continuity. *Psychological Inquiry*, 12, 49–66.

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
- Dragow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*, 134–135.
- Durbin, C. E., Hayden, E. P., Klein, D. N., & Olino, T. M. (2007). Stability of laboratory-assessed temperamental emotional traits from ages 3 to 7. *Emotion*, *7*, 388–399.
- Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. *Medical Care*, *44*, S134–S142.
- Ferdinand, R. F., Dieleman, G., Ormel, J., & Verhulst, F. C. (2007). Homotypic versus heterotypic continuity of anxiety symptoms in young adolescents: Evidence for distinctions between DSM-IV subtypes. *Journal of Abnormal Child Psychology*, *35*, 325–333.
- Fergusson, D. M., Horwood, L. J., Caspi, A., Moffitt, T. E., & Silvia, P. A. (1996). The (artificial) remission of reading disability: Psychometric lessons in the study of stability and change in behavioral development. *Developmental Psychology*, *32*, 132–140.
- Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, *45*, 135–145.
- Grimm, K. J., & Ram, N. (2009). A second-order growth mixture model for developmental research. *Research in Human Development*, *6*(2–3), 121–143.
- Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*(2–3), 97–120.
- Horn, J. L., McArdle, J., & Mason, R. (1983). When is invariance not invariance: A practical scientists look at the ethereal concept of factor invariance. *Southern Psychologist*, *1*, 179–188.
- Kagan, J. (1980). Four questions in psychological development. *International Journal of Behavioral Development*, *3*, 231–241.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.
- Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association.
- Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*, 3–31.
- Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*, 674–701.
- Pajer, K. (1998). What happens to “bad” girls? A review of the adult outcomes of antisocial adolescent girls. *American Journal of Psychiatry*, *155*, 862–870.
- Pearson, K. (1913). On the measurement of the influence of “broad categories” on correlation. *Biometrika*, *9*, 116–139.
- Rutter, M., & Sroufe, L. A. (2000). Developmental psychopathology: Concepts and challenges. *Development and Psychopathology*, *12*, 265–296.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*.
- Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research in Human Development*, *6*(2–3), 144–164.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Spearman, C. (1904). “General intelligence” objectively determined and measured. *American Journal of Psychology*, *16*, 201–293.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

- Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health outcomes. *Medical Care, 44*, S39–S49.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Hillsdale, NJ: Erlbaum.
- Thissen, D., & Wainer, H. (2001). An overview of test scoring. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 1–19). Mahwah, NJ: Erlbaum.
- Thomson, G. H. (1919). The proof and disproof of the existence of general ability. *British Journal of Psychology, 9*, 321–336.
- Thurstone, L. L. (1934). The vectors of mind. *Psychological Review, 41*, 1–32.
- Widaman, K. F., Cudeck, R., & MacCallum, R. C. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Erlbaum.
- Wirth, R. J. (2008). *The effects of measurement non-invariance on parameter estimation in latent growth models*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.