MIS Quarterly

RESEARCH COMMENTARY

CONSTRUCT MEASUREMENT AND VALIDATION PROCEDURES IN MIS AND BEHAVIORAL RESEARCH: INTEGRATING NEW AND EXISTING TECHNIQUES[1]

Scott B. MacKenzie
Department of Marketing, Kelley School of Business, Indiana University,
Bloomington, IN 47405 U.S.A. {mackenz@indiana.edu}

Philip M. Podsakoff
Department of Management, Kelley School of Business, Indiana University,
Bloomington, IN 47405 U.S.A. {podsakof@indiana.edu}

Nathan P. Podsakoff
Department of Management and Organizations, Eller College of Management,
University of Arizona, Tucson, AZ 85721 U.S.A. {podsakoff@email.arizona.edu}

Despite the fact that validating the measures of constructs is critical to building cumulative knowledge in MIS and the behavioral sciences, the process of scale development and validation continues to be a challenging activity. Undoubtedly, part of the problem is that many of the scale development procedures advocated in the literature are limited by the fact that they (1) fail to adequately discuss how to develop appropriate conceptual definitions of the focal construct, (2) often fail to properly specify the measurement model that relates the latent construct to its indicators, and (3) underutilize techniques that provide evidence that the set of items used to represent the focal construct actually measures what it purports to measure. Therefore, the purpose of the present paper is to integrate new and existing techniques into a comprehensive set of recommendations that can be used to give researchers in MIS and the behavioral sciences a framework for developing valid measures. First, we briefly elaborate upon some of the limitations of current scale development practices. Following this, we discuss each of the steps in the scale development process while paying particular attention to the differences that are required when one is attempting to develop scales for constructs with formative indicators as opposed to constructs with reflective indicators. Finally, we discuss several things that should be done after the initial development of a scale to examine its generalizability and to enhance its usefulness.

Keywords: Construct validation procedures; Scale development and validation; content, convergent, discriminant and nomological validity; formative and reflective indicator models

[1]Detmar Straub was the accepting senior editor for this paper. Thomas Stafford served as the associate editor.

The appendices for this paper are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

Introduction

It has been over 20 years since Straub (1989) made the following observation about the MIS literature:

> Instrument validation has been inadequately addressed in MIS research. Only a few researchers have devoted serious attention to measurement issues over the last few decades...and while the desirability of verifying findings through internal validity checks has been argued by Jarvenpaa, et al. (1984), the primary and prior value of instrument validation has yet to be widely recognized (p. 147).

Approximately a dozen years later, in a retrospective on the Straub article, Boudreau et al. (2001) surveyed the MIS literature again to assess whether there had been any improvement in the use of construct validation techniques, and concluded that their "findings suggest that the field has advanced in many areas, but, overall, it appears that a majority of published studies are still not sufficiently validating their instruments." (p. 1). Similar concerns regarding the practices used to validate constructs have also been expressed in the field of management by Scandura and Williams (2000), who compared the methodological practices reported in three top journals in two different time periods (1985–1987 and 1995–1997), and concluded that there had actually been a decrease in the proportion of studies that reported information about construct validity and reports of discriminant, convergent, and predictive validity. Therefore, the observation that Bagozzi and Phillips (1982, p. 468) made over 25 years ago still rings true: "Scientists have found it very difficult to translate this seemingly simple notion [of construct validity] into operational terms."

The reason for the apparent lack of progress in this area certainly is not due to a shortage of articles written on the technical procedures that should be used to validate scales (e.g., Anderson and Gerbing 1988; Anderson et al. 1987; Bagozzi et al. 1991; Diamantopoulos and Winklhofer 2001; Edwards 2001; Fornell and Larcker 1981; Gerbing and Anderson 1988; Nunnally and Bernstein 1994; Straub et al. 2004). However, one possibility is that the researchers reading these articles absorb only a portion of what is said because many of these articles are complex and require a fairly well-developed technical knowledge of structural equation modeling procedures. The result is that readers may not understand how to implement the recommendations made in these articles. An even more likely possibility is that there is simply so much work on the topic of scale development and evaluation that it is difficult for researchers to prioritize what needs to be done. Indeed, we believe that one reason Churchill's (1979) seminal article has proven to be so useful to researchers is that he outlined an organized set of activities that set priorities for what needs to be done in the scale development and evaluation process. Therefore, in the spirit of Churchill, the goal of this research is to provide an updated set of recommendations that can be used to give researchers a framework for developing valid scales.

We believe that there are several reasons why an updated set of recommendations would be useful. First, many of the scale development procedures advocated in the literature fail to adequately discuss how to develop appropriate conceptual definitions of a focal construct. Second, many of the recommendations are based on an improper specification of the measurement model[2] that relates the latent variable representing a construct to its measures.[3] Finally, techniques that provide evidence that the scale actually measures what it purports to measure have been underutilized in the management and MIS literatures. In the sections that follow, we will briefly elaborate on each of the limitations identified above. Following this, we will discuss each of the steps in the scale development process while paying particular attention to the differences that are required when one is attempting to develop scales for constructs with formative indicators as opposed to constructs with reflective indicators. Finally, we discuss several steps that should be taken after the initial development of a scale to examine its generalizability and to enhance its usefulness.

Limitations of Current Scale Development Procedures

Failure to Adequately Define the Construct Domain

Even though virtually every discussion of the construct validation or scale development process assumes that it begins with a clear articulation of the construct domain, the existing

[2]For the purposes of our paper, we use the term *measurement model* to refer to a model specifying the relationships between a latent construct and its indicators. Note that some (e.g., Borsboom 2005) prefer to use the term measurement model in a more restricted sense to refer only to instances in which a latent construct has a causal impact on its indicators.

[3]For the purposes of our paper, we use the term *measure* to refer to a standard used to determine or assess the magnitude of an attribute possessed by an entity. This term will sometimes be used interchangeably with the terms *item* and *indicator* depending on the context, because an *item* is a *measure* of an attribute and a response to it can be used as an *indicator* of a latent construct.

literature does not do a very good job of describing the characteristics of a good construct definition and the implications of that definition for measurement model specification. This is important because, as noted by DeVellis (1991, p. 51),

> many researchers *think* they have a clear idea of what they wish to measure, only to find out that their ideas are more vague than they thought. Frequently, this realization occurs after considerable effort has been invested in generating items and collecting data—a time when changes are far more costly than if discovered at the outset of the process.

According to Nunnally and Bernstein (1994, 86-87),

> There are three major aspects of construct validation: (1) specifying the domain of observables related to the construct; (2) determining the extent to which observables tend to measure the same thing ...from empirical research and statistical analyses; and (3) performing subsequent individual difference studies and/or experiments to determine the extent to which supposed measures of the construct are consistent "best guesses" about the construct.

Of these aspects, Nunnally and Bernstein argue that specifying the domain of the construct is the most important because

> there is no way to know how to test the adequacy with which a construct is measured without a well-specified domain. In other words, aspect 1 (specifying the domain) is important in telling you what to do in aspect 2 (investigating relations among different proposed measures of a construct (p. 88).

Indeed, we would add that there is no way to know what to do in aspect 3 without a clear conceptual definition of the construct.

Failing to adequately define the conceptual domain of a construct causes several problems (MacKenzie 2003). First, a poor construct definition leads to (1) confusion about what the construct does and does not refer to, and the similarities and differences between it and other constructs that already exist in the field; (2) indicators that may either be *deficient* because the definition of the focal construct is not adequately fleshed out, or *contaminated* because the definition overlaps with other constructs that already exist in the field; and (3) invalid conclusions about relationships with other constructs that later have to be rejected because the indicators of the focal construct are not really capturing what they are intended to capture.

Given the importance of clearly defining the conceptual domain of the construct, it is surprising that so many researchers either neglect this step in the process or fail to properly implement it. One reason this may happen is because it is difficult to do. Writing good construct definitions requires clear conceptual thinking and organization, the lack of which becomes apparent as soon as the researcher tries to write a tight conceptual definition of the construct. In addition, it is hard to find a detailed description of what a researcher needs to do to adequately define a construct. Even those papers that emphasize the importance of developing an adequate conceptual definition do not always tell researchers how to do this. Indeed, as lamented by Nunnally and Bernstein, "no precise method can be stated to outline the domain of variables for a construct properly...the theorizing process is necessarily intuitive" (p. 88). However, even though this may be an intuitive process, we believe that there are ways to structure and guide this theorizing and we discuss this in a later section of the paper.

Failure to Correctly Specify the Measurement Model

Most scale development procedures recommended to-date (e.g., Anastasi and Urbina 1997; Bagozzi et al. 1991; Hinkin 1995; Nunnally and Bernstein 1994; Schwab 1980; Spector 1992) are based on the assumption that a person's score on a measure of a latent construct is a function of his/her true position on the latent construct, plus error. According to this approach to measurement, causality flows from the latent construct to the measures in the sense that each measure is viewed as an imperfect *reflection* of the underlying latent construct (see Bollen 1989; Nunnally and Bernstein 1994). Although this type of measurement model is conceptually appropriate in many instances, Bollen and Lennox (1991) have noted that it does not make sense for all constructs. Indeed, they argue that indicators do not always reflect underlying latent constructs, but sometimes they combine to *form* them. This is consistent with the views of several other researchers (e.g., Blalock 1964; Borsboom 2005; Edwards and Bagozzi 2000; Goertz 2006; Law and Wong 1999; MacCallum and Browne 1993; MacKenzie et al. 2005), who argue that for some latent constructs it makes more sense to view meaning as emanating from the indicators to the construct in a definitional sense, rather than vice versa.

The distinction between *formative* and *reflective* measurement models is critically important for several reasons. First, there

are sound conceptual reasons to believe that many of the most widely used measures in marketing (Jarvis et al. 2003), management (Podsakoff et al. 2003b), and MIS (Petter et al. 2007) should probably be viewed as formative indicators of the constructs they represent, rather than as reflective indicators. Second, recent Monte Carlo simulations reported by Jarvis et al. (2003), MacKenzie et al. (2005), and Petter et al. (2007) suggest that structural parameter estimates can be biased when indicators that should be modeled as having formative relationships with a construct are modeled as having reflective relationships. Finally, the majority of the scale development procedures recommended in the literature only apply to latent constructs with *reflective* indicators, and if they are applied to latent constructs with *formative* indicators, they can undermine construct validity. For example, most articles and texts on scale development processes (see Churchill 1979; DeVellis 1991; Schwab 1980; Spector 1992) recommend that items possessing low item-to-total correlations should be dropped from a scale to enhance internal consistency reliability. Although this recommendation is appropriate in the case of reflective indicators because the items are all assumed to be sampled from the same content domain, if this recommendation is followed for constructs with formative indicators, it may result in the elimination of precisely those items that are most likely to alter the empirical and conceptual meaning of the construct. Thus, as noted by Bollen and Lennox (1991), the conventional wisdom on item selection and scale development and evaluation must be qualified by consideration of the nature of the relationship between the indicators and the latent construct they are intended to represent. In a later section of the paper, we discuss the implications of this distinction for construct validation procedures.

Underutilization of Some Techniques That Are Helpful in Establishing Construct Validity

After a construct has been conceptually defined and tentative measures have been developed, one of the next steps is to test whether the measures behave as one would expect them to if they were valid. Normally, this is evaluated by testing whether the measures of the focal construct relate to measures of other constructs in the nomological network specified by the researcher (Cronbach and Meehl 1955). Although this is certainly one way of assessing the validity of the measures of a construct, the disadvantage of this technique is that it cannot rule out spurious causes as an explanation for the findings. For example, methodological biases cannot be ruled out as a potential explanation, unless the researcher implements procedural or statistical controls (Podsakoff et al. 2003a). Similarly, there may be other constructs of a non-methodological nature that could plausibly account for the observed relationships between the measures of the focal construct and the measures of other constructs included in the nomological network. These disadvantages flow from the fact that the data are correlational in nature. Consequently, one way of eliminating or reducing the plausibility of these rival explanations would be to directly manipulate something that the researcher expects to influence the focal construct in order to see if this affects scores on the measures of the construct. This is consistent with Borsboom's (2005) view that "a test is valid for measuring an attribute [of a construct] if and only if a) the attribute exists, and b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (p. 150). Another way of obtaining evidence of construct validity would be to identify groups known to be high or low on the focal construct and then compare the scores of these groups on the measures of the focal construct you are attempting to evaluate. Although both of these alternative methods of obtaining evidence of construct validity have a long history of use in behavioral research (see Cronbach and Meehl 1955) these techniques are not used very frequently in the MIS and management literatures. Therefore, in this paper we are going to discuss these underutilized techniques and how they compare to the commonly used method of testing nomological validity.

Overview of the Scale Development Process

Figure 1 provides an overview of the steps in the scale development process. As shown in the figure, this process involves a series of steps beginning with construct conceptualization (or reconceptualization of an existing construct) and culminating in the development of norms for the scale. Each of these steps will be discussed in the sections to follow. In our discussion, we will attempt to focus more on the steps that have not been given as much attention in the literature. This does not suggest that the other steps in the validation process are any less important.

In addition, it is important to keep in mind two caveats regarding the steps we describe in Figure 1. First, we have tried to strike a balance between depth of treatment and breadth of coverage. Undoubtedly, there are readers who will disagree with the tradeoffs that we have made, and we acknowledge that there may be other valuable techniques that could be utilized during some of the steps in the validation process shown in Figure 1 that we are not aware of or chose to exclude. However, we felt that describing every possible technique that might be useful at each step in the construct

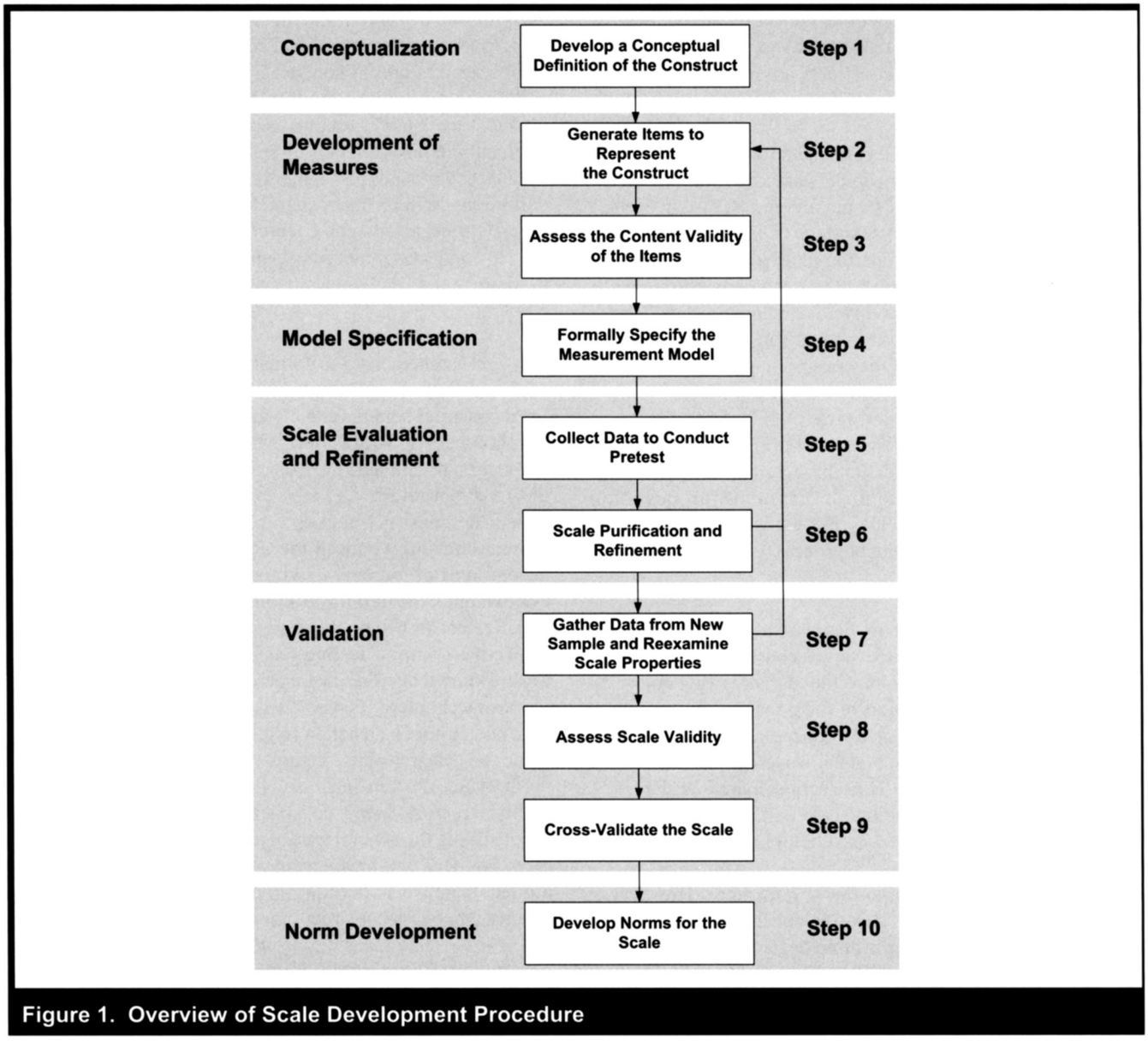


Figure 1. Overview of Scale Development Procedure

validation process would undermine our primary goal of outlining an organized set of activities that set priorities for what needs to be done in the scale development and evaluation process. Second, it is important to bear in mind that there may be practical limitations that prevent researchers from being able to follow all of the recommendations discussed in this paper in a single study, either because of a lack of time or resources, or both. Nevertheless, as noted by Vince Lombardi, the Hall of Fame NFL coach, chasing after perfection serves a useful purpose: "Perfection is not attainable, but if we chase perfection we can catch excellence."

Step 1. Develop a Conceptual Definition of the Construct

According to Nunnally and Bernstein (1994, p. 85),

> To the extent that a variable is abstract and latent rather than concrete and observable (such as the rating itself), it is called a "construct." Such a variable is literally something that scientists "construct" (put together from their own imaginations) and which does not exist as an observable dimension of

> behavior....Nearly all theories concern statements about constructs rather than about specific, observable variables because constructs are more general than specific behaviors by definition.

The first stage of the scale development and validation process involves defining the conceptual domain of the construct. As noted by several authors (Nunnally and Bernstein 1994; Schwab 1980; Spector 1992), this stage of scale development not only requires the identification of what the construct is intended to conceptually represent or capture, but also a discussion of how the construct differs from other related constructs. More specifically, during this stage, the researcher should specify the nature of the construct and its conceptual theme in unambiguous terms and in a manner that is consistent with prior research (MacKenzie 2003). Each of these elements is essential. It is important at this stage of the construct development and validation process for researchers to be as clear and concise in their definition as possible (Churchill 1979; Hinkin 1995). For example, in their discussion of the scale development process, Clark and Watson (1995, p. 310) state that

> A critical first step is to develop a precise and detailed conception of the target construct and its theoretical context. We have found that writing out a brief, formal description of the construct is very useful in crystallizing one's conceptual model... thinking about these theoretical issues prior to the actual process of scale construction increases the likelihood that the resulting scale will make a substantial contribution to the psychological literature.

These points are valid and important to remember. However, even in our own experience, this stage of the construct validation process is the one that is often neglected or dealt with in a superficial manner (e.g., by assuming that labeling or naming the construct is equivalent to defining it). This leads to a significant amount of trouble later in the validation process. Indeed, as noted by MacKenzie (2003, p. 323),

> the failure to adequately specify the conceptual meaning of a study's focal constructs...triggers a sequence of events that undermines construct validity (primarily due to measure deficiency), statistical conclusion validity (due to the biasing effects of measurement model misspecification), and ultimately internal validity (due to a combination of factors).

For this reason, we will briefly elaborate on each of the key factors to keep in mind at the construct conceptualization stage (see Table 1).

As indicated in Table 1, the first thing the researcher needs to do is to examine how the focal construct has been defined in prior research, and to conduct interviews with practitioners and/or subject matter experts. The goal in reviewing the literature is to identify previous uses of the term, rudimentary or dictionary definitions, closely related constructs, etc. Sartori (1984) recommends researchers collect a representative set of definitions, extract their characteristics, and construct matrices that organize such characteristics meaningfully. The goal in conducting interviews with practitioners or experts is to identify the key aspects (or attributes) of the construct's domain.

Next, researchers need to formally specify the nature of the construct, including (1) the conceptual domain to which the focal construct belongs and (2) the entity to which it applies. This is consistent with Sartori, who argued that when defining constructs, researchers must specify the phenomena to which the construct refers (i.e., the intension) and the referents to which the construct applies (i.e., the extension). By conceptual domain, we mean the definition should specify the general type of property to which the focal construct refers. For example, the definition should specify whether the construct refers to a thought (e.g., cognition, value, intention, subjective norm), a feeling (e.g., attitude, emotion, attitude toward knowledge sharing), a perception (e.g., perceived ease of use of technology, perceived usefulness of technology, fairness perceptions), an action (e.g., behavior, activity), an outcome (e.g., degree of use, return-on-investment, a stock price, performance), or an intrinsic characteristic (e.g., cognitive ability, structure, speed, conscientiousness). The importance of specifying the general type of property to which a construct refers has been previously recognized by Schwab (1980), who argued that, when defining constructs, it is important to specify whether a construct represents a

> structural property of organizations, perceptions of the property (and if so, by whom), or employee affect toward the property....Much confusion has been created because the construct referent has not been made clear in the definition and/or in moving from definition to measurement (pp. 12-13).

By entity we mean the object to which the property applies (e.g., a person, a task, a process, a relationship, a dyad, a group/team, a network, an organization, a culture). As noted by Kozlowski and Klein (2000, p. 27), the failure to specify the entity to which a construct applies is a common problem:

> This problem, we have noted, once plagued the climate literature. Researchers and critics asked whether climate was to be conceptualized and measured as an organizational (unit) construct or as a

Table 1. Summary of Factors to Consider in Construct Conceptualization

Factor	Considerations
Examine how the focal construct has been used in prior research or by practitioners	• Literature review of previous theoretical and empirical research on the focal construct • Review of literature on the meaning of related constructs • Conduct preliminary research using inductive approach with subject matter experts or practitioners
Specify the nature of the construct's conceptual domain	Identify the type of *property* the construct represents, and the *entity* to which it applies • Job satisfaction: Entity = person; general property = positive feelings about the job • End-user satisfaction: Entity = person; general property = positive feelings about computer technology • Perceived ease of use of technology: Entity = person; general property = perception or belief about the use of technology • IT capabilities: Entity = organization; general property = IT abilities and competencies • Procedural justice: Entity = person; general property = perception of fairness of procedures • Role ambiguity: Entity = person; general property = clarity of perception of role requirements • Fear of technological advances: Entity = person; general property = fear of technological changes • Job performance: Entity = person; general property = job outcomes • Firm performance: Entity = organization; general property = organizational outcomes • Social capital: Entity = organization; general property = resources accruing from network relationships
Specify the conceptual theme of the construct	Describe the necessary and sufficient attributes/characteristics as narrowly as possible • Common attributes/characteristics • Unique attributes/characteristics • Breadth/Inclusiveness Dimensionality • Unidimensional • Multidimensional Stability • Over time • Across situations • Across cases
Define the construct in unambiguous terms	• Provide clear, concise conceptual definition of the construct • Should not be subject to multiple interpretations • Should not be overly technical (technical terms with narrow meanings) • Should define construct positively, not by the denial of other things; negation of one thing does not imply the affirmation of something else • Should not be circular or tautological or self-referential

> psychological (individual) one. Climate researchers resolved this question, differentiating explicitly between a consensual unit climate and its origins in psychological climate. However, the question of level [entity] is often unasked in other research.

Thus, specifying the general type of property to which the focal construct refers and the entity to which it applies is important in the early stage of a construct's conceptualization. For example, according to Doll and Torkzadeh (1988), the definition of end-user satisfaction should focus on a person's (entity) positive feelings about computer technology (general property); and according to Davis (1989) the definition of perceived ease of use of technology should focus on a person's (entity) perception regarding the use of information technology (general property). In contrast, according to

Bharadwaj (2000), IT capabilities is a construct that refers to a firm's (entity) abilities or competencies in the IT area (general property).

Next, the researcher needs to clarify the intension of the focal construct by describing its conceptual theme. The conceptual theme of a construct consists of the set of fundamental attributes/characteristics that are necessary and sufficient for something to be an exemplar of the construct. For example, a submarine might be defined as a boat (1) capable of operation on or below the surface of the water; (2) that can float; (3) possessing an underwater emergency escape trunk; and (4) made of steel. Of these attributes/characteristics, only the first is necessary and sufficient; the others are necessary but not sufficient, sufficient but not necessary, or neither necessary nor sufficient, respectively.

Ideally, each attribute/characteristic specified in the conceptual theme would be common to all exemplars of the construct, and uniquely possessed by exemplars of the construct. However, this often proves to be difficult to do, because it requires a clarity of thought that may be lacking in the early stages of the development of a construct due to ambiguities in the intension and/or extension of the construct. In addition, Goertz (2006) has argued that some constructs conform to a "family resemblance structure," which posits a series of sufficient attributes/characteristics but no necessary ones. Consequently, the conceptual theme is sometimes expressed exclusively in terms of attributes/characteristics that are common but not unique, or exclusively in terms of attributes/characteristics that are unique but not common.

However, the danger of defining a construct solely in terms of common characteristics, ignoring their uniqueness, is that other researchers may falsely believe that all entities possessing those characteristics would qualify as an instance of the construct. This leads to an overly broad definition of the construct. For example, organizational commitment could be defined as a positive feeling about one's organization, but organizational loyalty and organizational involvement are also positive feelings about one's organization. To distinguish between them, their unique characteristics or attributes must also be specified. On the other hand, the danger of defining a construct exclusively in terms of unique characteristics, regardless of commonality, is that other researchers may falsely believe that unless an object possesses this particular characteristic, it cannot qualify as an example of the construct. For example, if the construct of workplace deviance were simply defined as stealing from the company, sexually harassing a coworker, and consuming alcohol or drugs on the job, then others researchers would have no way of knowing whether leaving work early without permission, calling in sick when you are not, or working on a personal matter during work hours are examples of workplace deviance. This is why defining a construct *solely* in terms of examples, without articulating the common characteristics that tie them together, is a problem.

In addition, when specifying the conceptual theme of a construct, it is also important to specify how stable it is expected to be over time, across situations, and across cases (see Chaplin et al. 1988). For example, is the construct expected to be relatively stable over time like a personality trait, or is it expected to vary over time like a mood state? Is the construct expected to apply only in a particular situation and not in others like *task-specific self-efficacy*, or is it expected to be generally applicable across different situations like *generalized self-esteem*? Is the construct expected to be applicable only to specific cases like *military bearing*, or to generalize across cases like *organizational commitment*?

Finally, all of this should be done using language that is clear and concise, that is not subject to multiple interpretations, and that is not overly technical. In addition, it is important that the definition (1) is not tautological or self-referential, and (2) describes the construct positively in terms of what it is (and not exclusively by the denial of other things, or in terms of its antecedents and/or consequences).

Construct Dimensionality

Once the construct has been carefully defined, it is important to step back and evaluate whether there are multiple sub-dimensions of the focal construct and how they relate to the focal construct and to each other. In this section, we will explore each of these issues.

The first issue is whether there are multiple sub-dimensions of the focal construct, or to put it another way, does the construct have more than one conceptually distinguishable facet, aspect, or sub-dimension? Many constructs are defined as having multiple, distinct sub-dimensions. For example, trustworthiness has been defined by Serva et al. (2005) as having three distinct sub-dimensions (integrity, benevolence, and ability); firm performance has been defined by Rai et al. (2006) as a function of operational excellence, customer relationships, and revenue growth; and Yi and Davis (2003) have defined observational learning processes as having several distinct facets (attention, retention, production, and motivation). If a construct is multidimensional, then it is important to define each of the sub-dimensions with the same care that was used in the case of the focal construct itself.

In order to determine whether the focal construct is multidimensional, we have found it useful to list the essential characteristics of the construct and answer the following questions:

(1) How distinctive are the essential characteristics from each other (apart from their common theme)?

(2) Would eliminating any one of them restrict the domain of the construct in a significant or important way?

If the essential characteristics have no unique aspects, and eliminating any one of them would not restrict the conceptual domain of the construct, then the construct is unidimensional from a conceptual perspective. In contrast, if the essential characteristics describe relatively unique aspects of the construct, and eliminating any of them would restrict the conceptual domain of the construct, then the construct is multidimensional from a conceptual perspective. For example, Bagozzi et al. (1979) have noted that a tripartite conceptualization of a person's attitude toward an object views the focal construct as being multidimensional and consisting of affective, cognitive, and conative components; whereas a unidimensional conceptualization of attitude views this construct as consisting of affect only.

If the construct is multidimensional, a second conceptual question that should be considered is the nature of the relationship between the sub-dimensions and the higher-order (more general) construct. As noted by Edwards (2001),

> The relationships between a multidimensional construct and its [sub-]dimensions are not causal forces linking separate conceptual entities, but instead represent associations between a general concept and the [sub-]dimensions that represent or constitute the construct (p. 146).

When making the decision about how the sub-dimensions relate to the more general focal construct, several authors (Bollen and Lennox 1991; Jarvis et al. 2003; Wong et al. 2008) have noted that it is helpful to ask

(1) Are the sub-dimensions viewed as manifestations of the focal construct or as defining characteristics of it?

(2) Does the focal construct exist separately at a deeper and more embedded level than its sub-dimensions, or is the focal construct a function of its sub-dimensions?

(3) Would a change in the focal construct be associated with changes in all of the sub-dimensions, or is it possible for a change in the focal construct to be associated with a change in only one of the sub-dimensions (but not the others)?

If the sub-dimensions are viewed as defining characteristics, the focal construct is a function of its sub-dimensions, and a change in only one of the sub-dimensions could be associated with a change in the focal construct, then the sub-dimensions are best thought of as *formative* indicators of the second-order focal construct. For example, transformational leadership is often conceptualized (see Avolio et al. 1999; Howell and Hall-Merenda 1999) as having multiple behavioral sub-dimensions (e.g., individualized consideration, idealized influence, intellectual stimulation, etc.) that together define what it means to be a transformational leader and determine a leader's level of transformational leadership. These are viewed as formative indicators, rather than as reflective indicators, because it seems reasonable that an increase in the level of a leader's individualized consideration behavior might be associated with an increase his/her level of transformational leadership, without necessarily being associated with any changes in the leader's intellectual stimulation behavior or idealized influence behavior.

In contrast, if the sub-dimensions are viewed as manifestations of a focal construct, the focal construct exists separately at a deeper and more embedded level than its sub-dimensions, and a change in the focal construct would be expected to produce a change in all of its sub-dimensions, then the sub-dimensions are best thought of as *reflective* of the second-order focal construct. For example, a leader's *general tendency* to exhibit contingent reward behavior toward his or her subordinates might be measured by asking a leader's subordinates to respond at several different points in time to the following types of items: "my supervisor provides positive feedback when I do my job well," "my supervisor praises me when my work is especially good," and so on. The responses to these items are reflective indicators of the leader's tendency to exhibit this form of behavior at a particular point in time, and the leader's tendencies at these specific points in time are themselves reflective of his/her general tendency to exhibit this form of behavior. More generally, a second-order measurement model with multiple first-order sub-dimensions as reflective indicators might be appropriate when a researcher (1) is interested in measuring a stable focal construct (e.g., an individual difference variable) over time or across situations, or (2) has several randomly selected parcels of items each of which is reflective of a focal construct. Note, however, that the latter procedure is not without limitations (see Bandalos 2002), as we discuss later.

For constructs with multiple sub-dimensions as formative indicators, a third question that needs to be considered is how

the sub-dimensions combine to form the focal construct. In all of these models, the focal construct is a function of the sub-dimensions that jointly define it. The question is, what type of function? Is it an additive or multiplicative one? Goertz (2006) argues that "concepts have causal theories embedded in them" (p. 12) in the sense that "the core attributes of a concept constitute a theory of the ontology of the phenomenon under consideration." (p. 27). In his view, an essential part of a construct's conceptualization is the specification of the manner in which the sub-dimensions combine to give the construct its meaning.

For some constructs, the sub-dimensions combine in a compensatory fashion to produce the meaning of the focal construct in such a way that the effect of each sub-dimension on the focal construct is independent of the effects of the other sub-dimensions. Implicitly, this structure suggests that a change in each individual sub-dimension is sufficient (but not necessary) to produce a change in the meaning of the focal construct. This structure might be appropriate for a construct like job performance (e.g., Rotundo and Sackett 2001), if one hypothesizes that each of its sub-dimensions (task performance, organizational citizenship behavior, and workplace deviance behaviors) contribute to changes in job performance, the magnitude of the effect of each sub-dimension is unrelated to the effects of any of the other sub-dimensions, and the sub-dimensions are substitutable in the sense that one might compensate for another. In this type of model, the sub-dimensions are added together to form the focal construct. For example, an employee can increase his/her job performance either by increasing task performance or increasing organizational citizenship behavior or by decreasing workplace deviance behavior. Conceptually, this means that the focal construct represents the *union* of its sub-dimensions.

However, this model is not appropriate for all constructs. For some constructs, the sub-dimensions represent attributes/characteristics that are necessary and jointly sufficient for the meaning of the construct. This concept structure suggests the focal construct represents the *intersection* of sub-dimension 1 *and* sub-dimension 2 *and* sub-dimension 3, etc. Practically speaking, this type of structure can be represented by a multiplicative interaction among the sub-dimensions. This is consistent with Goertz (2006, p. 7), who noted that this type of model

> goes back to Aristotle and builds concepts using the structure of necessary and sufficient conditions. In classic philosophical logic, to define a concept is to give the conditions necessary and sufficient for something to fit into the category. Each of these necessary conditions is a [sub-dimension]: the structural glue that binds the [sub-dimensions] together to form the basic level is the mathematics of necessary and sufficient conditions.

For example, for over 50 years, social scientists have conceptualized source credibility as requiring both expertise and trustworthiness (Hovland et al. 1953). This implies that a source *must* have some minimal level of both attributes to possess credibility. In other words, a source that possesses expertise, but that is not trustworthy, is not credible. Likewise, a source that is trustworthy but has no expertise is also not credible. Thus, when a construct is comprised of several necessary attributes, the construct should be viewed as being formed by the interaction among these attributes, and should be modeled in a fashion consistent with this logic. It is important to note that, conceptually, source credibility is not a distinct construct that is *caused by* trustworthiness and expertise; rather it is *defined as being* the product of trustworthiness and expertise. Another example of a construct in which multiple, distinct sub-dimensions interact to form a focal construct might be Vroom's (1964) force to perform an act construct (which is viewed as the valence of the outcome times the strength of the expectancy that a specific act will be followed by the attainment of that outcome). Although we are not aware of any specific examples of the use of this multiplicative structure for a measurement model, we do believe that this type of measurement model is appropriate for some constructs and should be explored in future research.

Constructs Are Not Inherently Formative or Reflective

It is important to note that the terms *formative* and *reflective* describe the relationship between an indicator and the latent construct with which it is associated. Constructs are not inherently formative or reflective in nature, and most can be modeled as having either formative or reflective indicators depending upon the researcher's theoretical expectations about how they should be related based on the conceptual definition of the construct. For example, job satisfaction has been conceptualized as both a unidimensional construct (Cammann et al. 1983) and a multidimensional construct with several distinct facets (Smith et al. 1969). In the first case, job satisfaction is measured with three reflective indicators; whereas in the second case, job satisfaction has multiple formative indicators, each of which represents one of the facets (e.g., Law and Wong 1999). The key point is that the way in which the construct and the indicators are linked depends on the content of the indicator and how the construct is conceptualized by the researcher. This is consistent with

Borsboom (2005), who argued that even a construct like socio-economic status (SES), which is frequently used as a prototypical example of a construct with formative indicators, can be measured with reflective indicators:

> One may also imagine that there could be procedures to measure constructs like SES reflectively—for example, through a series of questions like "how high are you up the social ladder?" Thus, the fact that attributes [constructs] like SES are typically addressed with formative models does not mean that they could not be assessed reflectively (p. 169).

Ontological Issues

A final point worth noting is that formative and reflective measurement models have different ontological assumptions that rarely have been discussed in the literature, but nevertheless have important implications. As noted by Borsboom (2005, p. 63), latent variable theory is ontologically ambiguous depending upon whether a reflective or formative indicator model is assumed: "the realist interpretation of a latent variable implies a reflective model, whereas constructivist, operationalist, or instrumentalist interpretations are more compatible with a formative model."

Several authors (e.g., Borsboom 2005; Howell et al. 2007b) have reasoned that measurement models with reflective indicators imply that the latent construct (1) is a *real entity* that exists independently of a person and the way in which s/he finds out about it, and (2) *causes* the observed variation in the responses to the items used to measure it. Although we believe that a realist interpretation is reasonable for many constructs represented by reflective indicator measurement models, we suspect that some researchers would be reluctant to endorse this strict interpretation in the case of latent factors "discovered" through exploratory factor analysis, and perhaps in the case of certain types of psychological constructs. For example, Bagozzi has recently questioned both of these two widely held assumptions (i.e., the assumption that a latent construct must always represent a *real, mind-independent entity*, and the assumption that it *causes* a person's responses to its measures):

> A strict realist conception of latent variables becomes less tenable when we consider such abstract or theoretical concepts as attitudes, attributions, beliefs, desires, emotions, goals, intentions, motivation, and personality traits. These mental events and states are widely studied and likely constitute the majority of applications of structural equation models (SEMs) by psychologists. It has not been possible to relate such mental events and states closely with physical processes in the brain or to make claims about how the mental events or states function in a physical sense. Indeed, it is an open question whether this will ever be possible, if one assumes that there is something more to attitudes, intentions, and so on than the firing of neurons and other physical processes (Bagozzi 2007, p. 230).

> Now to the claim that the relationship between a latent variable and its manifest or measured variables is causal....It seems to me that the relationship in question is not causal, per se, but rather one of hypothetical measurement. That is, the relationship is between an abstract, unobserved concept and a concrete, observed measurement hypothesized to measure the concept; the relationship is part logical, part empirical, and part theoretical (conceptual), with the inferred factor loading representing, in and of itself, only part of empirical meaning of the relationship (Bagozzi 2010, p. 210).

In contrast, measurement models with formative indicators need not assume that the composite latent construct is a real entity. Typically, constructs with formative indicators are seen as theoretical constructions (rather than real entities) that summarize (and therefore depend upon) people's responses to the items used to represent the construct. This point has been well articulated by Borsboom (2005, p. 62), who has noted that

> Latent variables of the formative kind are not conceptualized as determining our measurements, but as a summary of these measurements. These measurements may very well be thought to be determined by a number of underlying latent variables (which would give rise to the spurious model with multiple common causes of Edwards and Bagozzi 2000), but we are not forced in any way to make such an assumption. Now, if we wanted to know how to weight the relative importance of each of the measurements comprising SES in predicting, say, health, we could use a formative model....In such a model, we could also test whether SES acts as a single variable in predicting health. In fact, this predictive value would be the main motivation for conceptualizing SES as a single latent variable. However, nowhere in this development have we been forced to admit that SES exists independent of our measurements....The formative model thus does not necessarily require a realist interpretation of the latent

variable that it invokes. In fact, if a realist interpretation were to be given, it would be natural to conceptualize this as a spurious model with multiple common causes in the sense of Edwards and Bagozzi (2000). This would again introduce a reflective model part in the model, which would correspond to that part of the model that has a realist interpretation.

Step 2: Generate Items to Represent the Construct

Once the focal construct has been conceptually defined, the next step in the process is to generate a set of items that fully represents the conceptual domain of the construct. These items may come from a variety of sources (see Churchill 1979; Haynes et al. 1995; Nunnally and Bernstein 1994), including reviews of the literature, deduction from the theoretical definition of the construct, previous theoretical and empirical research on the focal construct, suggestions from experts in the field, interviews or focus group discussions with representatives of the population(s) to which the focal construct is expected to generalize, and an examination of other measures of the construct that already exist.

Regardless of whether the focal construct is unidimensional or multidimensional, the ultimate goal of the item generation process is to produce a set of items that fully captures all of the essential aspects of the domain of the focal construct, while minimizing the extent to which the items tap concepts outside of the domain of the focal construct. For multidimensional constructs, that would mean developing a set of items for each individual sub-dimension, while also making sure that the sub-dimensions comprise all essential aspects of the focal construct's definition. Importantly, this is true regardless of how the sub-dimensions relate to the focal construct (i.e., as formative or reflective indicators). In both cases, the ultimate objective in generating the initial set of items is the same. Indeed, as noted by Diamantopoulos and Sigauw (2006, p. 267),

> In short, according to the extant literature, there appears to be no compelling reason as to why the *initial* item pool would differ *purely* because of the choice of measurement perspective. Assuming that literature guidelines on comprehensiveness and inclusiveness are diligently followed [e.g., Bollen and Lennox 1991; DeVellis 1991; Diamantopoulos and Winklhofer 2001; Spector 1992], item generation under each perspective would not be expected to result in widely divergent item pools.

In addition, there are several other considerations regarding the items that should be taken into account at this stage of the scale development process. One consideration relates to the manner in which the items are written (Peterson 2000; Podsakoff et al. 2003a; Spector 1992; Torangeau et al. 2000). Generally speaking, each item should be written so that its wording is as simple and precise as possible. Double-barreled items (e.g., "Credible speakers possess expertise and are trustworthy") should be split into two single-idea statements, and if that proves impossible, the item should be eliminated altogether. Items that contain ambiguous or unfamiliar terms should be clarified, and items that possess complicated syntax should be simplified and made more specific and concise. Finally, efforts should also be made to refine or remove items that contain obvious social desirability (see Nederhof 1985).

Step 3: Assess the Content Validity of the Items

Once items have been generated for representing the focal construct, they should be evaluated for their content validity. According to Straub et al. (2004, p. 424), content validity concerns "the degree to which items in an instrument reflect the content universe to which the instrument will be generalized." Similarly, Kerlinger (1973, p. 459), defines content validity as "the 'representativeness' or 'sampling adequacy' of the content—the substance, the matter, the topics—of a measuring instrument." Thus, two related judgments must be made when assessing content validity:

(1) Is the individual item representative of an aspect of the content domain of the construct?

(2) Are the items as a set collectively representative of the entire content domain of the construct?

Although there are a variety of methods that have been developed to assess the content adequacy of new measures (Anderson and Gerbing 1991; Hinkin and Tracey 1999; Lawshe 1975; Schriesheim et al. 1999; Schriesheim et al. 1993), we recommend the procedure suggested by Hinkin and Tracey (1999) as illustrated by Yao et al. (2008). To our knowledge, this technique has not been used in the MIS literature. In this procedure, the researcher constructs a matrix in which definitions of different aspects of the construct domain are listed at the top of the columns and the items are listed in the rows (see Table 2). Next, raters are asked to rate the extent to which each item captures each aspect of the construct domain using a five point Likert-type scale ranging from 1 (not at all) to 5 (completely). This information in Table 2 can be transposed to produce a matrix

Table 2. Hypothetical Example of Item Rating Task to Assess Content Adequacy

Rater Number = 001 Trustworthiness Scale Items†	***Benevolence*** is the degree to which the trustor believes that the trustee has goodwill or positive intentions toward the trustor (Serva et al. 2005, p. 630).	The other party's ***ability*** to accomplish a task important to the trustor, where ***ability*** is the set of skills or attributes that enable the trustee to have influence (Serva et al. 2005, pp. 629-630).	***Integrity*** is a trustor's perception that the trustee adheres to acceptable values, which could include issues such as consistency, honesty, and fairness (Serva et al. 2005, p. 630).
1. The management team really looked out for what was important to our development team.	4	2	1
2. Our development team felt that the management team was very capable of performing its job.	1	5	2
3. Our development team believed that the management team tried to be fair in dealings with others.	1	1	5
4. Our development team's needs and desires were very important to the management team.	5	1	2
5. Our development team had confidence in the skills of the management team.	1	5	2
6. The management team had a strong sense of justice.	2	1	4
7. The management team went out of its way to help our development team.	5	2	2
8. Our development team believed that the management team was well qualified.	1	5	1
9. Our development team liked the values of the management team.	2	1	5

†The trustworthiness items used for illustration purposes were taken from Serva et al. (2005).

Table 3. Example of a Data Array for Content Adequacy Scores

Rater Number	Aspects of Trustworthiness	Item #1	Item #2	Item #3	Item #4	Item #5	Item #6	Item #7	Item #8	Item #9
001	Benevolence	4	1	1	5	1	2	5	1	2
	Ability	2	5	1	1	5	1	2	5	1
	Integrity	1	2	5	2	2	4	2	1	5
002	Benevolence	5	2	1	5	2	1	5	1	1
	Ability	1	5	1	1	4	1	1	5	2
	Integrity	1	1	5	1	2	5	2	1	4
n…	Benevolence									
	Ability									
	Integrity									

of ratings data similar to the one shown in Table 3. In this table, each case consists of multiple lines of data (one for each aspect of the construct domain). The first line of each case consists of the rater's ratings of each of the items on the first aspect of the construct domain (i.e., benevolence), the second line contains the rater's ratings of each of the items on the second aspect of the construct domain (i.e., ability), and so on. A one-way repeated measures ANOVA is then used to assess whether an item's mean rating on one aspect of the construct's domain differs from its ratings on other aspects of the construct's domain. Assuming the F-statistic is significant, a planned contrast is conducted to test whether the mean of the rating for the item on the hypothesized aspect of the construct domain is higher than the mean rating for this item on all other aspects of the construct domain.

When using the technique described above, it is important to keep several things in mind. First, because each rater makes multiple ratings for each item, it is essential to use a one-way repeated measures ANOVA, so that an adjustment is made to the error term (see Winer 1971, p. 270). It would only be appropriate to use a one-way between-subjects ANOVA to analyze the data if the ratings of each item on each aspect of the construct domain were provided by different raters. The disadvantages of a between-subjects approach is that it would require substantially more subjects and the test of the item rating differences across aspects of the construct domain would be less powerful because individual differences across raters would be lumped into the error term.

Second, it is important to remember that the effectiveness of any content adequacy assessment technique is only as good as the definitions of the construct (and the items) that are developed by the researcher in the first place. Thus, as we have noted earlier, good construct definitions are a critical element of the construct validation process.

Third, when selecting people to serve as raters, it is important to make sure that they have sufficient intellectual ability to rate the correspondence between items and the theoretical definitions. Anderson and Gerbing (1991) have argued that, in addition, it is also important for the raters to be representative of the main population of interest; whereas Schriesheim et al. (1993) and Hinkin and Tracey (1999) do not feel that this is a critical issue and that college students can be used for this task. In principle, we tend to agree with Anderson and Gerbing's position, because we believe it is important to develop items that are conceptually understood by the population of interest. However, we recognize that in many instances, college educated students may be representative of the population to which one desires to generalize.

Fourth, it is also important to avoid overburdening the raters. Based on evidence from psychophysical judgment research, Schreisheim et al. (1993) suggest that raters can reliably distinguish between a maximum of only eight to ten aspects of the content domain at a time. However, even this may be too much when the number of items is also large. In such cases, it may be necessary to have raters assess only a subset of the items to reduce the overall burdens of the task.

Fifth, the procedure described above can be used to not only determine whether the items capture the hypothesized construct, but also whether they capture unintended constructs as well (i.e., whether the items are contaminated). This can be done by including definitions of other constructs in the column headings of Table 2 and asking the raters to rate the extent to which the items represent these other constructs. Indeed, as noted by Schreisheim et al. (1993, p. 406), when constructing new scales,

> the developer might be well-advised to employ not only the categories and definitions of the scales or subscales which are being developed, but also those from related constructs or from constructs which have been problematic sources of content confounding in the past....The use of these additional content categories should help ensure that any new measures which are developed have item content which is free of items from troublesome and/or extraneous domains.

Finally, it is important to recognize that this technique does not make any implicit assumptions about the direction of the relationship between the items and their corresponding factors or about the correlations between the items themselves. Because of this, it can be used to assess the content validity of either formative or reflective indicators. This is a key advantage of this technique relative to the Q-method approach to content assessment discussed by Schriesheim et al. (1993), and it is particularly important because Petter et al. (2007) have noted that a lack of content validity is a particularly serious problem for constructs with formative indicators.

Step 4: Formally Specify the Measurement Model

Once a content valid set of items has been generated, the next step is to formally specify a measurement model that captures the expected relationships between the indicators and the focal construct and/or sub-dimension they are intended to

represent.[4] This is complicated by the need to set the scale of measurement and to ensure that the parameters of the model are all identified. The scale of measurement for a first-order construct with multiple reflective or formative indicators can be set (1) by fixing a path between the latent construct and one of its indicators at some nonzero value or (2) by fixing the variance of the construct at some nonzero value (Bollen 1989; MacCallum and Browne 1993). In both instances, the nonzero value is usually 1.0 to aid in interpretation. Either of these two solutions is acceptable. It is also necessary to set the scale of measurement for a second-order construct with multiple reflective or formative first-order sub-dimensions as indicators. This can be done (1) by fixing a path between the second-order construct and one of its sub-dimensions at some nonzero value (usually 1.0) or (2) by fixing the variance of the second-order construct at some nonzero value (again, usually 1.0). Once again, either of these solutions is acceptable. The advantage of fixing the path at 1.0 is that it aids interpretation by causing the scale of measurement for the second-order construct to be the same as one of its sub-dimensions. The advantage of fixing the variance at 1.0 is that it aids interpretation by standardizing the construct.

A second important issue that complicates the specification of constructs with formative indicators is that, depending on the model structure, it is not always possible to identify the construct-level error term (Bollen and Davis 2009; MacCallum and Browne 1993). In rare instances, it might be appropriate to resolve this indeterminancy by fixing the construct-level error term (ζ) at zero. For example, in the case of an "exogenous" second-order construct with multiple first-order sub-dimensions as formative indicators, and with multiple reflective indicators of each first-order sub-dimension, it might be appropriate to fix the error term associated with the second-order latent construct at zero provided that one is confident that (1) the first-order indicators of the second-order composite latent construct are free of measurement error, (2) all of the essential sub-dimensions of the second-order construct are represented, and (3) the sub-dimensions do not interact with each other. With respect to the first point, in this type of model it may be reasonable to assume that the first-order indicators are free of measurement error since random measurement error has been partialled out of these indicators. With respect to the second and third points, if the procedure outlined in step 1 of our scale development process has been followed, there may be instances where one might be willing to assume that all of the essential sub-dimensions of the constructs are represented and the sub-dimensions independently influence the composite latent construct. Under these circumstances, it may make sense to fix the error term associated with the second-order composite latent construct at zero. Indeed, as noted by Diamantopoulos (2006, p. 11),

> It will be recalled from the previous discussion that the error term in a formative measurement model represents the impact of all remaining causes other than those represented by the indicators included in the model....Given this interpretation of the error term, it becomes apparent that it would be legitimate to set $\zeta = 0$ as long as all possible causes on the construct are included as indicators in the model. This is not as far-fetched as it may initially sound. Consider, for example, Carlson and Grossbart's (1988) television coviewing measure which indicates the frequency of watching television with one's children (1 – "very seldom" 5 – "very often"). This measure contains three items (weekdays, Saturdays, and Sundays, respectively) which collectively exhaust all possibilities for viewing television in a week. If we were to apply the model...to this measure, then ζ would have to be set to zero as no additional occasions could be specified to capture coviewing behaviour. Thus, in some instances, the best way of dealing with the error term would be to simply exclude it from the model.

However, an error term is necessary whenever the composite latent construct is in the endogenous position, or when the formative indicators do not fully represent the construct domain, as is often the case. The latter might occur when the formative indicators consist of several *sufficient* determinants of the composite latent construct (Goertz 2006), but there are other conceptually appropriate determinants of the composite latent construct that are not included in the set of indicators. Bollen and Davis (2009, p. 503) have shown that for the construct-level error term to be identified, "every latent variable with an unrestricted variance (or error variance) must emit at least two directed paths to variables when these latter variables have unrestricted error variances." Practically speaking, this condition is satisfied if the latent construct emits paths to (1) at least two theoretically appropriate reflective indicators, (2) at least two other latent constructs with reflective indicators, or (3) one reflective indicator and at least one other latent construct with reflective indicators. However, because satisfying this requirement is necessary but not

[4]Note that our discussion here focuses on the use of covariance-based SEM techniques rather than components-based techniques for specifying and testing measurement models (for a discussion of the merits of each approach, see Diamantopoulos 2011). In addition, our discussion does not apply to multidimensional constructs that are specified as "profiles" or discrete combinations of various levels of their dimensions (for a more extensive discussion of multidimensional profile models, see Law et al. 1998).

sufficient for identification, additional conditions must also be met (e.g., satisfying the scaling rule, the t-rule, and certain restrictions on the error covariances).[5]

We have always recommended (see Jarvis et al. 2003; MacKenzie et al. 2005) that researchers resolve this identification problem by including two global reflective indicators of the composite latent construct, along with the formative indicators. For example, consistent with multidimensional conceptualizations of job satisfaction (Law and Wong 1999; Smith et al. 1969; Spector 1997), let us assume that job satisfaction is defined as a composite latent construct comprised of three facets (e.g., pay satisfaction, work satisfaction, and promotion satisfaction) and these three facets were measured with the following items: "I am very satisfied with my pay" (X1), "I am very satisfied with the nature of my work" (X2), and "I am very satisfied with my opportunities for promotion" (X3), respectively. In this instance, the identification problem could be solved by adding two reflective indicators with uncorrelated error terms to the measurement model: "Overall, I am very satisfied with my job" (Y1), and "Generally speaking, I am satisfied with all facets of my job" (Y2). The addition of these two reflective indicators produces what Jöreskog and Golberger (1975) have called a MIMIC (multiple indicators, multiple causes) model structure.

If the conceptual nature of the indicators is ignored, there are several ways this structure might be interpreted. One way is as a composite latent construct (with formative indicators) that causes two other conceptually distinct constructs (Figure 2, Panel A). Another way this structure can be interpreted is as a reflectively measured latent construct that is caused by several conceptually distinct antecedent constructs (Figure 2, Panel B). Finally, this structure can be interpreted as a single latent construct with a mixture of formative and reflective indicators (Figure 2, Panel C). It is important to note that all three of these interpretations are empirically indistinguishable because they produce identical predicted covariance matrices.

However, if the conceptual nature of the indicators is taken into account, not all of these interpretations are equally plausible. For example, it does not make sense to interpret the MIMIC structure as it is shown in Panel A because in this panel the two reflective indicators (Y1 and Y2) are treated as if they are indicators of two different constructs. In the case at hand, this doesn't make sense because both indicators were selected to reflect the conceptual definition of job satisfaction, and it is hard to imagine that these indicators ("Overall, I am very satisfied with my job" and "Generally speaking, I am satisfied with all facets of my job") reflect different constructs. Similarly, if all of the measures are content-valid operationalizations of the same focal construct, we do not think that it is desirable to interpret the MIMIC structure as it is shown in Panel B. In this panel, the model is interpreted as if there are four conceptually distinct constructs represented: each of the antecedent constructs is viewed as having a single reflective indicator, the consequence construct has two reflective indicators, and the antecedent constructs cause the consequence construct. This interpretation of the MIMIC structure is the one preferred by Wilcox et al. (2008, p. 1226). However, from our perspective, this interpretation is undesirable because it (1) ignores the multidimensional nature of the superordinate construct and requires a change in the construct's conceptual definition, and (2) treats the superordinate construct's sub-dimensions as distinct causes that are no different conceptually than any other causes of the superordinate construct (e.g., from a conceptual perspective, pay satisfaction is not viewed as being any more intimately related to, or a part of, job satisfaction than role ambiguity, role conflict, etc.). Instead, we believe it makes the most sense to interpret this entire structure as a measurement model for a single latent construct as shown in Panel C, because each of the measures, whether formative or reflective, is a content-valid operationalization of the same multidimensional focal construct.

As noted by Jarvis, et al. (2003), there are several important advantages to solving the identification problem by adding at least two reflective indicators of the composite latent construct (as shown in Figure 2, Panel C). First, it can be used regardless of whether the focal construct is in an endogenous or exogenous position, or even all by itself. The other methods of achieving identification (e.g., emitting paths to at least two other latent constructs with reflective indicators, or emitting paths to one reflective indicator and at least one other latent construct with reflective indicators) require the focal construct to cause at least one other latent construct in the model. That may not be conceptually appropriate or desirable in some instances. Second, unlike the other two methods, adding two reflective indicators of the focal construct permits it to be included along with other constructs in a confirmatory factor model which could be used to evaluate its measurement properties and discriminant validity (see Anderson and Gerbing 1988).

Third, Jarvis et al. (2003, p. 213) have noted that this procedure diminishes the likelihood of interpretational confounding

[5]Although Bollen and Davis (2009) have noted that "no encompassing necessary and sufficient condition of identification exists for structural equation models with latent variables" (p. 501), their "Exogenous X Rule" provides a useful set of sufficient (but not necessary) identification conditions for formative indicator models with MIMIC-like structures.

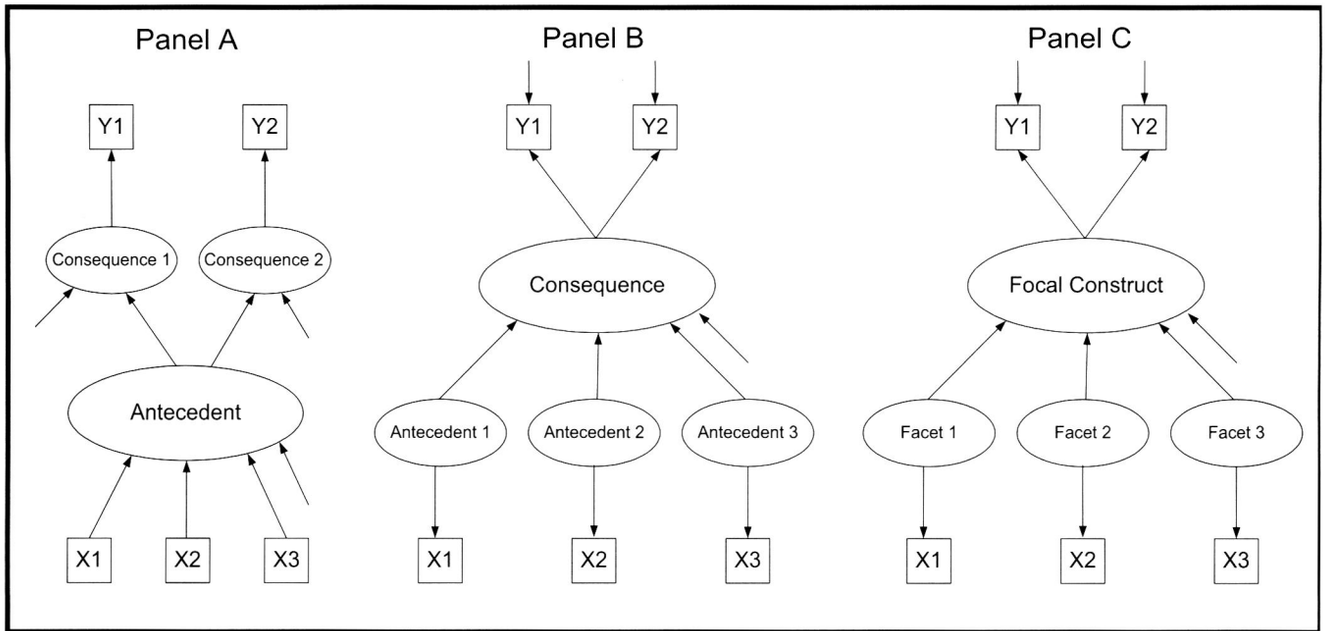


Figure 2. Alternative Interpretations of a MIMIC Structure

because, "the measurement parameters should be more stable and less sensitive to changes in the structural relationships emanating from the formative construct." Interpretational confounding (Burt 1976) is a serious problem arising from a discrepancy between the nominal meaning of a construct (based on its conceptualization) and the empirical meaning of the construct (based on its operationalization) that can affect models with formative or reflective indicators (Anderson and Gerbing 1988; Bollen 2007; Burt 1976). Howell et al. (2007b, p. 207) describe this problem as follows:

> In the context of reflective measurement, Burt (1976), following Hempel (1970, pp. 654-666), distinguished between the nominal meaning and the empirical meaning of a construct. A construct's nominal meaning is that meaning assigned without reference to empirical information. That is, it is the inherent definitional nature of the construct that forms the basis for hypothesizing linkages with other constructs, developing observable indicators, and so forth. A construct's empirical meaning derives from its relations to one or more observed variables. These may be measures of the construct itself (epistemic) or relationships to observable measures of other constructs in a model (structural)....to the extent that the nominal and empirical meanings of a construct diverge, there is an issue of interpretational confounding.

In general, interpretational confounding is present to the extent that the coefficients linking formative or reflective indicators with a focal construct significantly change depending on the other endogenous variables in the model (i.e., those caused by the focal construct). Several researchers (Howell et al. 2007b; Kim et al. 2010; Wilcox et al. 2008) have recently demonstrated that interpretational confounding can be a major problem in models with formative indicators, when the identification of the construct level error term is achieved through structural relationships with "other latent constructs." We agree (Jarvis et al. 2003), and recommend that if identification is achieved through structural relationships with other constructs, an attempt should be made to assess the degree of interpretational confounding present. As noted by Bollen (2007), this can be done in a manner that is similar to that used for models with reflective indicators (i.e., by examining the effect on the measurement parameter estimates of adding other latent constructs to the model that are affected by the focal construct); the main difference being that, "we check for changes in factor loadings (λs) with effect (reflective) indicators, and changes in γs for causal (formative) indicators" (p. 223).

However, we disagree with Wilcox et al.'s (2008, p. 1227) unqualified generalization that

> The empirical meaning of a formatively measured construct depends on the outcome variables in the

> model, such that while the name of a formatively measured construct may remain the same, the construct's empirical realization will vary from model to model and study to study.

We also disagree with Kim et al.'s (2010, p. 363) general conclusion that

> The usage of formative measurement, therefore, may have negative repercussions on the quality of IS research....Construct reusability could especially be jeopardized, leading to inconsistency in theory testing and barriers to building cumulative knowledge.

When the identification of the construct-level error term is achieved through the use of two content valid reflective indicators as we recommend above, Bollen (2007, pp. 223-224) has convincingly demonstrated that interpretational confounding is unlikely to be a problem. Indeed, Howell et al. (2007a, p. 243) acknowledge that "the use of reflective items does indeed go a long way toward fixing the problem of interpretational confounding, and we strongly agree with this approach." The reason interpretational confounding is not likely to be a problem in this instance is that interpretational confounding arises primarily from the combination of model misspecification and underidentification, and our recommendation solves the underidentification problem in a manner that is likely to result in a correctly specified model from a theoretical perspective. This is because, if this method is used, the empirical meaning of the latent construct will be based solely on epistemic relationships. That is, the empirical meaning will be based solely on relationships between the latent construct and a set of observed variables which are all content valid indicators of the construct domain (see Figure 2, Panel C).

The next decision that needs to be made when specifying a construct with formative indicators is whether to freely estimate the covariances among the formative indicators and between these indicators and other exogenous variables. Although it is possible to constrain all of these covariances to be equal to zero, MacCallum and Browne (1993) have noted that freeing them is more consistent with traditional modeling practice. We recommend estimating the covariances among the formative indicators of a construct. However, before freeing the covariances between the formative indicators of one construct and other exogenous constructs or measures, the theoretical meaning and empirical impact of estimating these covariances should be considered (Jarvis et al. 2003; Petter et al. 2007).

One final point worth noting relates to the use of parcels that combine several items into a single index as indicators. Generally speaking, parcels are sometimes used in reflective indicator models to (1) improve the reliability or distributional properties of the indicators, (2) simplify the interpretation of the measurement model parameters, or (3) enhance the goodness of fit of the measurement model (Bandalos and Finney 2001). However, Bandalos (2002) has argued that parcels are not recommended when the goal of the study is scale development, refinement, or testing. She provides several reasons for this. First, using a parcel to represent a set of items is only appropriate when the items are known to be unidimensional. When items that are not unidimensional are combined into parcels, it can bias measurement parameter estimates and obscure the true factor structure (Hall et al. 1999; West et al. 1995), and bias structural parameter estimates in some cases. Second, when parcels are used, the overall test of the fit of the measurement model is not as stringent as when the complete set of individual items is used, because fewer covariances must be fit. This, of course, makes it more likely that the measurement model will appear to fit the data. Third, parceling can mask sources of model misfit. Therefore, she recommends that parcels only be used if the items used to create each parcel already have been shown to be content valid and unidimensional.

Step 5: Collect Data to Conduct Pretest

Once the measurement model has been formally specified, data need to be obtained from a sample of respondents in order to examine the psychometric properties of the scale, and to evaluate its convergent, discriminant, and nomological validity. In choosing this sample, the key issue is how well the sample represents the population for which the measures are designed. This is important, because measurement properties may differ across sub-populations. For example, scale properties may differ cross-culturally, across demographic groups, by experience, by education level, by position within an organization, etc. Another factor that is traditionally considered at this stage is the size of the sample. In exploratory factor analysis (EFA), recommendations of the minimum sample size range from 100 to 500 (Comrey and Lee 1992; Gorsuch 1983), and recommendations of the minimum ratio of the number of respondents to the number of items in the scale range from 3:1 to 10:1 (Cattell 1978; Everitt 1975). However, MacCallum et al. (1999) point out that

> A fundamental misconception about this issue is that the minimum sample size or the minimum ratio of sample size to the number of variables is invariant across studies. In fact, necessary sample size is de-

> pendent upon several aspects of any given study including the level of communality of the variables and the level of overdetermination of the factors (p. 84).

More specifically, MacCallum et al.'s findings indicate that good recovery of population parameters is possible with even small sample sizes (e.g., 60 to 100) when communalities are high and the factors are strongly determined, but that larger sample sizes (e.g., 300 to 500) are necessary when communalities are low and the factors are weakly determined.

To assess convergent validity, alternative measures of the same construct should be included as part of this data gathering effort. To assess nomological validity, measures of constructs that are theoretically related to the focal construct should be obtained. Finally, to assess discriminant validity, measures of similar constructs that are potentially confounded with the focal construct should be obtained.

A final consideration in data gathering relates to whether the construct is expected to be stable over time or not. For example, some constructs like personality traits are expected to be relatively stable over time and/or across situations, while other constructs capture states that are expected to vary depending on the time or place in which they are measured. As noted by Baumgartner and Steenkamp (2006), this issue can be taken into account in the data collection and model specification stage. This requires multiple waves of data in which the same constructs are measured at multiple points in time. Although this obviously requires a great deal of additional effort on the part of the researcher, it may be the best way to determine whether measures that are expected to remain stable (or vary) over time and across situations for theoretical reasons actually do so.

Step 6: Scale Purification and Refinement

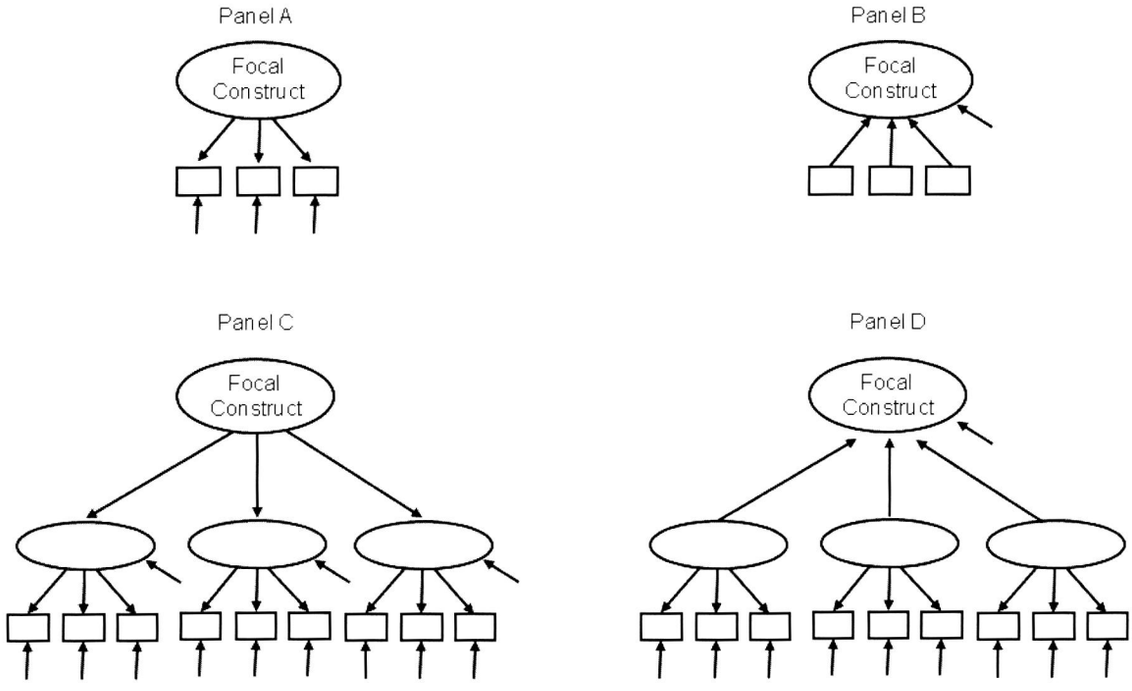
The methods for evaluating a newly developed scale have been widely discussed in the literature (Anderson and Gerbing 1988; Bagozzi 1980; Baggozi et al. 1991; Bollen 1989; Fornell and Larcker 1981; Gefen 2003; MacCallum and Austin 2000; Segars 1997; Spector 1992). However, there are two gaps in this literature. First, this research has focused much more on the criteria for evaluating reflective indicator measurement models than on formative indicator measurement models. A second gap is that there is little discussion of how to apply these criteria to make decisions about which items to omit in order to purify the scale (see the article by Bollen in this issue for an exception). Therefore, in this section we will summarize procedures for scale evaluation that have been previously recommended in the research literature and, in so doing, we will pay particular attention to (1) methods for evaluating constructs with formative indicators, and (2) criteria for eliminating problematic indicators.

Appendix A provides a detailed summary of the steps in the scale purification and refinement process, and Figure 3 provides examples of the types of first- and second-order measurement models to which these steps apply. Panels A and B depict first-order constructs with reflective and formative indicators (respectively). The second-order construct in Panel C has been called a totally disaggregated construct by Bagozzi and Heatherton (1994), a superordinate construct by Edwards (2001), and a Type I construct by Jarvis et al. (2003). The second-order construct in Panel D has multiple first-order sub-dimensions as formative indicators and the sub-dimensions have multiple reflective indicators. Second-order constructs with this model form have been called a spurious model with multiple common causes by Edwards and Bagozzi (2000), an aggregate construct by Edwards (2001), and a Type II construct by Jarvis et al. (2003). Although there are other types of second-order constructs that have different combinations of formative and reflective indicators (Edwards and Bagozzi 2000; Jarvis et al. 2003), we have focused on these two measurement models because there is some evidence that they are the most common types. Finally, it is important to note that the models in Panels B and D are not identified as shown, but can be estimated after resolving the identification problem as recommended in Step 4 (i.e., by adding 2+ reflective indicators to the focal construct).

Evaluating the Goodness of Fit of the Measurement Model

Regardless of the type of measurement model estimated (e.g., Figure 3, Panels A–D), the first step is to determine whether (1) the solution is proper, (2) the individual relationships hypothesized are statistically significant, and (3) the relationships (as a group) are consistent with the sample data. A solution is proper if the estimation procedure converges and none of the variance estimates are negative, both of which can be verified by examining the output. The critical ratios for the estimates of the individual relationships can be evaluated with z-tests at the appropriate level of significance (e.g., $p < .05$ or .01). The chi-square statistic can be used to test whether the model adequately accounts for the sample data. However, in this instance, a nonsignificant ($p > .10$) chi-square statistic is indicative of a good fit because it means that the covariances predicted by the model are not significantly different than the sample covariances.[6] Overall, a proper solution, significant

[6]Note that if a mean structure is also included in the model, a nonsignificant chi-square statistic would indicate that the covariances and means predicted by the model do not differ from the sample covariances and means.



Note: The models in Panels B and D are not identified as shown, but can be estimated after resolving the identification problem as recommended in Step 4.

Figure 3. Examples of First-Order and Second-Order Measurement Models

critical ratios for the individual relationships hypothesized, and a nonsignificant chi-square statistic are evidence of the validity of the hypothesized measurement model.

An alternative means of testing the overall fit of a measurement model, called confirmatory tetrad analysis (CTA), has been proposed by Bollen and Ting (2000). A *tetrad* is the difference between the product of one pair of covariances and the product of another pair of covariances among four random variables. CTA is based on the idea that the structure of a model often implies that certain specific population tetrads should "vanish" (i.e., be equal to zero) and, because of this, a simultaneous test of the vanishing tetrads implied by a model can provide a useful test of the model's fit. If this chi-square test indicates that the tetrads which are supposed to be equal to zero are significantly different from zero, it is evidence that the model is inconsistent with the data. A nonsignificant result indicates that the vanishing tetrad predictions implied by the model are consistent with the data, and can be interpreted as evidence of the validity of the hypothesized model. Although a complete description of CTA cannot be provided here, interested readers are encouraged to consult Bollen and Ting's paper for detailed instructions on how to identify and test the vanishing tetrads implied by many first-order measurement models that include reflective indicators.

Unfortunately, although the chi-square statistic provides the best inferential test of overall model fit, its usefulness is greatly undermined by the fact that it has been found to be related to sample size, model complexity, and nonnormality (see Hu and Bentler 1999). Consequently, it is necessary to rely on other goodness of fit indices to evaluate the extent to which the relationships hypothesized in the measurement model are consistent with the sample data. Some of these alternative goodness of fit indices assess the absolute fit of a model (e.g., GFI, RMSEA, SRMR) and some assess its fit relative to a suitably framed comparison model (e.g., CFI, NFI, TLI). Hu and Bentler (1999) conducted a series of Monte Carlo simulations that showed that in order to balance Type I and Type II error rates it is best to rely on multiple goodness of fit measures from different families of fit indices. Although they note that it is difficult to designate a specific cutoff value for each fit index because the best value depends upon the combination of fit indices used, generally speaking

a cutoff value close to .95 for CFI, .08 for SRMR, and .06 for RMSEA is indicative of a good fitting model, and can be interpreted as evidence in favor of the validity of the hypothesized model.

Assessing the Validity of the Set of Indicators at the Construct Level

For first-order latent constructs with reflective indicators (see Figure 3, Panel A), *convergent validity* can be assessed by calculating the average variance in the indicators that is accounted for by the focal construct. In the typical case where each indicator is related to only one construct, the average variance extracted (AVE) by the focal construct can be calculated by averaging the squared completely standardized factor loadings (λ^2) for the indicators, or by averaging the squared multiple correlations for the indicators (see Fornell and Larcker 1981). An AVE greater than .50 is desirable because it suggests that the latent construct accounts for a majority of the variance in its indicators on average.

For second-order latent constructs with reflective indicators (see Figure 3, Panel C), it is also important to examine the convergent validity of the set of first-order sub-dimensions as reflective indicators of the second-order construct. As noted by Edwards (2001, p. 163), this can be assessed using the multivariate coefficient of determination (R^2_m), which is

> calculated by taking the determinant of the covariance matrix of the multidimensional construct and its sub-dimensions, dividing this quantity by the variance of the multidimensional construct times the determinant of the covariance matrix of the sub-dimensions, and subtracting the resulting quantity from unity

(For an equivalent formula for R^2_m, see Edwards 2003.) Alternatively, AVE could be calculated for the second-order construct by averaging the squared multiple correlations for the first-order sub-dimensions (or averaging the square of each sub-dimension's completely standardized loading on the second-order construct). In either case, values greater than .50 mean that, on average, a majority of the variance in the first-order sub-dimensions is shared with the second-order latent construct.

For first-order latent constructs with formative indicators (see Figure 3, Panel B), convergent validity is not relevant because the composite latent construct model does not imply that the indicators should necessarily be correlated. Instead, Diamantopoulos et al. (2008, p. 1216) have proposed using the magnitude of the construct level error term to assess construct validity based on the rationale that, "if the set of indicators is comprehensive in the sense that it includes all important construct facets, the construct meaning is validly captured; accordingly, the residual variance is likely to be small." This is consistent with the view of Williams et al. (2003, p. 908) that, "as the variance of the residual increases, the meaning of the construct becomes progressively ambiguous....Certainly, the meaning of a construct is elusive when most of its variance is attributable to unknown factors." This reasoning suggests that the construct level error variance should be small and constitute no more than half of the total variance of the construct.[7] Alternatively (but based on similar reasoning), Edwards (2001, p. 163) proposes that the strength of the relationship between a set of formative indicators and a composite latent construct

> can be assessed with the adequacy coefficient (R^2_a), which is used in canonical correlation analysis to assess the relationship between a set of variables and their associated canonical variate... R^2_a is calculated by summing the squared correlations between the construct and its dimensions [i.e., indicators] and dividing by the number of dimensions [i.e., indicators].

R^2_a values greater than .50 would mean that, on average, a majority of the variance in the indicators is shared with the construct. However, it should be recognized that there is little research using either of these methods and a consensus has not been reached regarding their appropriateness. Therefore, it may be prudent to use them in conjunction with other methods of assessing construct validity (e.g., nomological validity, criterion-related validity, etc.) until more evidence from these techniques has accumulated.

For second-order latent constructs with first-order sub-dimensions as formative indicators (see Figure 3, Panel D), start by evaluating whether the average variance extracted (AVE) for the reflective indicators of each individual first-order sub-dimension is greater than .50 (as previously described for first-order constructs with reflective indicators). Then, assess the validity of the set of sub-dimensions serving as formative indicators of the second-order latent construct using Edwards' (2001) adequacy coefficient (R^2_a), which can be calculated by summing the squared correlations between each sub-dimension and the focal construct, and then dividing by the number of sub-dimensions. The information needed

[7]Note that this method is intended to be used in models where the only antecedents of the composite latent construct are its own formative indicators.

for this calculation can be obtained from the completely standardized output reported by most SEM programs (e.g., LISREL).

Assessing Reliability of the Set of Indicators at the Construct Level

For first-order constructs with reflective indicators (Figure 3, Panel A), Cronbach's alpha has traditionally been used to estimate the internal consistency reliability of the measures. This is appropriate, because this type of measurement model implies internal consistency among the measures (see Bollen and Lennox 1991). In addition, Fornell and Larcker (1981) provide a somewhat different index of construct reliability based on the ratio of the variance accounted for by the latent construct to the total variance in the measures.[8] Generally speaking, the accepted standard for both of these indices is .70 or above for newly developed measures (Nunnally and Bernstein 1994). However, it is widely known that alpha increases as the number of measures increases, so a higher value may be appropriate in cases where the number of measures is large (Cortina 1993).

For second-order constructs with reflective indicators (Figure 3, Panel C), start by examining the reliability of the indicators for each individual first-order sub-dimension using one or more of the methods just described. Then, if desired, the reliability of the first-order sub-dimensions as indicators of the second-order construct can be examined by calculating Fornell and Larcker's index of construct reliability for the second-order construct. This can be done by using the completely standardized estimates of the second-order factor loadings and error variances associated with the first-order sub-dimensions. A similar procedure has been recommended by Baumgartner and Steenkamp (2006), and we believe this technique makes sense conceptually in some instances (e.g., when the second-order construct is a stable individual difference that has been measured with the same set of items at several different points in time).

For first-order constructs with formative indicators (see Figure 3, Panel B), the traditional notion of internal consistency reliability does not apply because the model does not predict that the indicators will be correlated (Bollen 1989; Bollen and Lennox 1991). Formative indicators may be negatively correlated, positively correlated, or completely uncorrelated with each other. Consequently, measures of reliability based on the idea of internal consistency like Cronbach's alpha and Fornell and Larcker's construct reliability index are not appropriate, and if applied may result in the removal of indicators that are essential to the domain of the construct (Diamantopolous and Winklhofer 2001; MacKenzie 2003).

For second-order latent constructs with first-order sub-dimensions as formative indicators (see Figure 3, Panel D), the reliability of the indicators for each individual first-order sub-dimension can be evaluated using the procedure described for first-order constructs having reflective indicators. However, traditional notions of internal consistency reliability do not apply to the set of sub-dimensions serving as formative indicators of a second-order construct because the second-order measurement model does not predict that the sub-dimensions will be correlated (Bollen and Lennox 1991; Edwards 2003). Indeed, Edwards (2001, p. 160) argues that, "Reliability is not an issue of debate when a multidimensional construct and its dimensions are treated as latent variables that contain no measurement error."

Evaluating Individual Indicator Validity and Reliability

For first-order constructs with reflective indicators (Figure 3, Panel A), the validity of the individual indicators can be assessed from the measurement model by determining whether the relationship between each indicator and its hypothesized latent construct is large and statistically significant. The significance of the estimate (λ) of a relationship between an indicator and the latent construct can be tested with a z-test of the estimate's critical ratio. The degree of validity of each indicator can be assessed by examining the unique proportion of variance in the indicator accounted for by the latent construct (Bollen 1989). In the typical case where each indicator is hypothesized to load on only one construct, this will be equal to the square of the indicator's completely standardized loading (λ^2). A value greater than .50 would suggest an adequate level of validity (see Fornell and Larcker 1981). The reliability of each indicator can be assessed by examining the squared multiple correlation for the indicator (Bollen 1989). Typically a value greater than .50 is desired because it suggests that the majority of the variance in

[8] An alternative reliability index (coefficient H) that has been proposed by Hancock and Mueller (2001) is the squared correlation between the latent construct and the optimum linear composite formed from the measured indicators. This index has several desirable properties: (1) the sign of a loading cannot impact the assessment of construct reliability, (2) the minimum is 0 and the maximum value is 1 when a single standardized loading is 1 (or -1), and (3) adding an additional indicator will never decrease the value of the coefficient. Unlike Fornell and Larcker's index of reliability, the value of coefficient H can never be less than the best indicator's reliability.

the indicator is due to the latent construct. Of course, in models where each indicator loads on only one construct, the squared multiple correlation and the square of the completely standardized loading will be equal.

For second-order constructs with first-order sub-dimensions as reflective indicators (Figure 3, Panel C), the validity of the indicators of the first-order sub-dimensions can be assessed as previously described. Then, the validity of each first-order sub-dimension can be tested by examining whether it is significantly related to the second-order latent construct; and the degree of validity of each sub-dimension can be assessed by examining the unique proportion of variance in the sub-dimension accounted for by the second-order construct. The latter will typically be equal to the square of the sub-dimension's completely standardized loading on the second-order construct (i.e., when each sub-dimension is hypothesized to load on only one second-order construct), and its value should be greater than .50 (see Fornell and Larcker 1981). The reliability of each sub-dimension can be evaluated by examining the squared multiple correlation for the sub-dimension. Values greater than .50 are desired, because they indicate that the second-order latent construct accounts for the majority of the variance in the sub-dimension (Fornell and Larcker 1981).

For first-order constructs with formative indicators (Figure 3, Panel B), indicator validity is captured by the significance and strength of the path from the indicator to the composite latent construct (Bollen 1989; Bollen and Lennox 1991).[9] The significance of the estimate of a relationship between a formative indicator and a latent construct can be tested with a z-test of the estimate's critical ratio. The degree of validity of each formative indicator can be assessed by examining the unique proportion of variance in the construct accounted for by the indicator. This is calculated by subtracting the proportion of variance in the construct accounted for by all of the indicators except for the one of interest from the proportion of variance in the construct accounted for by all of the indicators including the one of interest (see Bollen 1989, pp. 200 and 222).[10] There is no commonly accepted minimum standard for how much of the variance in the latent construct should be uniquely accounted for by each formative indicator. However, it is important to recognize that multicollinearity among the indicators poses the same problems here as it does in any other multiple regression model, because the latent construct is regressed on the indicators (rather than vice versa). This implies that the standard adopted should somehow take into account the number of formative indicators related to the latent construct and the strength of their intercorrelations. The reliability of each individual indicator can be assessed using (1) test–retest reliability (if the indicator is expected to be stable over time), and/or (2) inter-rater reliability (if different raters are expected to agree).

A final method for assessing the validity of a formative indicator can be applied to models that also include some global reflective indicators that capture the meaning of the "overall" latent construct (see Diamantopoulos and Winklhofer 2001). In such models, an estimate of the *indirect* effect of each formative indicator on the global reflective indicator can be obtained by multiplying (1) the estimate of the impact of the formative indicator on the composite latent construct and (2) the estimate of the impact of the composite latent construct on the global reflective indicator (see Bollen 1989). The statistical significance of this indirect effect can then be evaluated using a Sobel (1982) test, or by creating a confidence interval using a bootstrap standard error estimate (MacKinnon 2008). If the estimate of this indirect effect is large and statistically significant, it suggests that the formative indicator is valid. If two or more global reflective indicators are available for the latent construct, the completely standardized indirect effects of a given formative indicator on each of the reflective indicators might be averaged to provide a more robust estimate of the indicator's validity. Two things are worth noting about this procedure. First, estimates of this type of indirect effect can be requested as part of the output of most structural equation modeling programs. Second, this technique puts a premium on the content validity of the global reflective indicators because they are essentially being used as criterion measures to establish the criterion-related validity of the formative indicators. Consequently, this test of validity will only provide useful information to the extent that the global reflective indicators faithfully capture the conceptual domain of the focal construct.

[9]An alternative view advocated by Borsboom (2005, p. 169) is that the concept of validity does not apply to the relationship between formative indicators and a composite latent construct:

> Because validity explicitly concerns the relation of measurement, one wonders whether it is appropriate to view formative models as measurement models in the first place. They might be better conceptualized as models for indexing or summarizing the indicators, or as causal models that do not involve a measurement structure for which one can ask the question of validity.

[10]Note that this will only be equal to the square of the correlation between the formative indicator and the latent construct when the indicator of interest is uncorrelated with the other formative indicators (Bollen 1989, p. 202; Edwards and Bagozzi 2000, p. 165).

For second-order latent constructs with first-order sub-dimensions as formative indicators (Figure 3, Panel D), the validity and reliability of the individual indicators of the first-order sub-dimensions can be evaluated as described for first-order constructs having reflective indicators. Following this, the validity of each individual sub-dimension as an indicator of the second-order latent construct can be tested by examining whether each sub-dimension is significantly related to the second-order latent construct (Bollen 1989; Bollen and Lennox 1991). The degree of validity of each sub-dimension can be assessed using the unique proportion of variance in the construct accounted for by the sub-dimension. This can be calculated by subtracting the proportion of variance in the second-order construct accounted for by all of the sub-dimensions except for the one of interest from the proportion of variance in the second-order construct accounted for by all of the sub-dimensions including the one of interest (see Bollen 1989, pp. 200 and 222). The reliability of each individual sub-dimension can be assessed using Fornell and Larcker's construct reliability index.

Eliminate Problematic Indicators

The preceding analyses can be used to begin to identify problematic indicators. Problematic indicators are ones that have low validity, low reliability, strong and significant measurement error covariances, and/or non-hypothesized cross-loadings that are strong and significant. Reflective indicators of a first-order construct (Figure 3, Panel A) that have nonsignificant ($z < 1.96$) or weak (completely standardized λ^2 less than .50) relationships with the latent construct, or strong and significant measurement error covariances (modification index greater than 3.84 and large expected change estimates) are candidates for elimination, provided that all of the essential aspects of the construct domain are captured by the remaining items. Nonsignificant or weak loadings are an indication of a lack of validity, and measurement error covariances may be a sign of multi-dimensionality (Gerbing and Anderson 1984). Significant measurement error covariances can be identified by looking at the modification indices and their magnitude can be assessed by examining the completely standardized expected change estimates.

In addition, if the items are reflective indicators of a first-order sub-dimension of a second-order construct (Figure 3, Panel C), another consideration is whether the indicators have significant ($z > 1.96$) cross-loadings on a non-hypothesized sub-dimension that are similar in magnitude to their loading on the hypothesized sub-dimension. Significant cross-loadings can be identified by looking at the modification indices and their magnitude can be assessed by comparing the completely standardized loading (λ) of the indicator on the hypothesized sub-dimension to the completely standardized expected change estimates for the loadings on the non-hypothesized sub-dimensions. Large and significant cross-loadings are problematic because they suggest that the item is conceptually confounded due to the fact that it reflects more than one sub-dimension. Finally, one should also consider eliminating first-order sub-dimensions that fail to have significant loadings on the second-order construct, or first-order sub-dimensions with weak (but significant) loadings if the average variance extracted by the second-order construct is less than .50 (Fornell and Larcker 1981). Once again, it is important to remember that sub-dimensions should not be eliminated unless all of the essential aspects of the focal construct domain are captured by the remaining sub-dimensions. Instances where an entire sub-dimension can be dropped without eliminating an essential aspect of the construct domain will probably be rare.

In summary, we have argued that reflective indicators with low validity, low reliability, strong and significant measurement error covariances, and/or non-hypothesized cross-loadings that are strong and significant are candidates for elimination. However, an alternative approach for selecting the final set of reflective indicators has been advocated by Little et al. (1999). They argue that, from a domain sampling viewpoint, reflective indicators are randomly sampled from the universe of possible indicators that represent the construct's domain. Some of these indicators are closer to the core meaning of the construct (i.e., its centroid) than others. Their key idea is that information from the nomological network can be used to identify the best subset of indicators. More specifically, they recommend that (1) on the basis of theory and prior research, the researcher should specify how the focal construct should relate to a set of key marker constructs in its nomological network; (2) using a broad set of potential indicators of the focal construct, estimate the correlations between the focal construct and the key marker constructs with a confirmatory factor analysis; (3) drop one of the indicators of the focal construct and reestimate the construct intercorrelations using a confirmatory factory analysis; and (4) compare the new intercorrelations to the ones obtained from the confirmatory factor analysis of the complete set of indicators and if they match then the indicator can be safely deleted. The ultimate goal of this procedure is to identify a parsimonious subset of indicators that is able to reproduce the construct intercorrelations obtained from the confirmatory factor analysis of the complete set of indicators. Note that this subset "may or may not be the same subset of indicators that would yield the highest internal consistencies" (Little et al. 1999, p. 208). Interested readers are encouraged

to consult the Little et al. paper for additional details because, in our view, this alternative procedure has some merit.[11]

The procedures for eliminating formative indicators are somewhat different. For formative indicators of a first-order construct (Figure 3, Panel B), we recommend that indicators that fail to have strong and significant loadings on the construct be considered for elimination. However, because multicollinearity may make it difficult to separate the distinct influence of the individual indicators on the construct, the redundancy in the indicators should be examined using the variance inflation factor (VIF). Indicators with a nonsignificant loading and a VIF greater than 10 are redundant and could be considered for elimination (Diamantopoulos et al. 2008; Diamantopoulos and Winklhofer 2001; Neter et al. 1996); although some (e.g., Petter et al. 2007) have recommended a lower VIF cutoff level of 3. However, it is particularly important for constructs with formative indicators to ensure that all of the essential aspects of the construct domain are captured by the remaining indicators (Bollen and Lennox 1991; Diamantopoulos and Winklhofer 2001; MacKenzie 2003).

Finally, for second-order constructs with first-order sub-dimensions as formative indicators (Figure 3, Panel D), eliminate indicators that have (1) nonsignificant loadings on the hypothesized sub-dimension, (2) large and significant cross-loadings on non-hypothesized sub-dimensions, and/or (3) large and significant measurement error covariances with indicators of other sub-dimensions. Significant measurement error covariances and cross-loadings can be identified by looking at the modification indices (MI > 3.84) and their magnitude can be assessed by examining the completely standardized expected change estimates. If some of the first-order sub-dimensions fail to have significant loadings on the second-order construct, calculate the VIF to examine multicollinearity among sub-dimensions. Again, the VIF should be less than 10.0; although Petter et al. (2007) have recommended a lower VIF cutoff level of 3. However, we cannot emphasize enough that for constructs with formative indicators, it is important to ensure that all of the essential aspects of the construct domain are captured by the remaining indicators and sub-dimensions (Bollen and Lennox 1991; Diamantopoulos et al. 2008; MacKenzie 2003). Instances where an entire sub-dimension can be dropped without eliminating an essential aspect of the construct domain will probably be rare.

[11]Indeed, we believe that Little et al.'s procedure might also be adapted to select the subset of indicators that maximizes the correlation between the focal construct and a direct manipulation of it.

Step 7: Gather Data from New Sample and Reexamine Scale Properties

Because items are often added, dropped, or reworded in the scale purification process, the next step is to reestimate the measurement model using a new sample of data. This is important in order to assess the extent to which the psychometric properties of the scale may have been based on idiosyncrasies in the developmental sample of data and to permit a valid statistical test of the fit of the measurement model. Because of the limited objectives of this step in the scale validation process, this can be done using a sample that is similar to the one used to purify the items. In instances where items have been dropped, but not added or reworded, this step can use data from a holdout sample of the original data. Using this new sample, the measurement model would be reestimated, its fit reexamined, and the psychometric properties reevaluated.

Step 8: Assess Scale Validity

Assuming the psychometric properties of the purified scale are acceptable, the next step in the construct validation process is to evaluate whether responses to the scale behave as one would expect if they were valid indicators of the focal construct. Generally speaking, the goal is to evaluate whether the indicators of the focal construct (1) are accurate representations of the underlying construct (through experimental manipulation or comparing groups known to differ on the construct), (2) adequately capture the multidimensional nature of the construct, (3) are distinguishable from the indicators of other constructs (discriminant validity), and (4) are related to the measures of other constructs specified in the construct's theoretical network (nomological validity). In the section that follows, we will discuss how each of these types of validity may be examined within the context of constructs having formative and reflective indicators. A summary of these types of validity are provided in Appendix B.

Experimental Manipulation of the Construct

Generally speaking, for constructs with reflective indicators (Panels A and C of Figure 3) it is important to establish whether the measures actually represent the phenomenon that they are intended to represent. For example, if a scale is intended to represent helping behavior, then it is important to determine whether scores on this scale are correlated with actual instances of helping behavior. This is consistent with Stogdill's (1969) observation that, "Validity implies that a

given subscale measures the pattern of behavior that it is intended to measure" (p. 153). More specifically, as noted by Borsboom (2005, p. 150), "a test is valid for measuring an attribute [construct] if and only if a) the attribute exists, and b) variations in the attribute [construct] causally produce variations in the outcomes of the measurement procedure." In this sense, validity refers to the extent to which the scores on a scale are correlated with direct manipulations or measures of the phenomenon the scale purports to measure, and when a scale is highly correlated with the phenomenon it is intended to measure, it can be said to possess veridicality (i.e., that the scores on the scale are in agreement with the object's real properties). This is different from content validity assessments, which focus on the extent to which the indicators represent the conceptual domain of the construct. In the case of content validity, the issue is whether the indicators adequately capture the whole domain of the construct. Here, the concern is whether the indicators accurately capture the real-world phenomena to which they refer. For this reason, Borsboom has argued that this type of evidence is the most important kind for establishing the validity of the measures of a construct.

Although establishing the validity of a measure by directly manipulating the focal construct is a frequently neglected aspect of the scale development process, one exception is the study conducted by Stogdill. In this study, Stogdill was interested in determining whether the subscales of the Leadership Behavior Description Questionnaire (LBDQ) accurately (validly) measured the pattern of behaviors that they were intended to measure. To examine this, he (1) prepared a scenario that depicted a leader acting out the pattern of behavior described in each of the subscales (i.e., consideration, structure, production emphasis, tolerance of freedom, influence with superiors, and representation), (2) made a motion picture of a leader (and followers) acting out the role described in the scenario, (3) showed the movie to observers who were asked to rate the leader's behavior on the LBDQ, and (4) tested whether the leader in the film was rated significantly higher on the focal LBDQ behavior depicted in the film than on other subscales of the LBDQ. In order to insure that it was the leader behaviors depicted in the films that caused the ratings, rather than the actor who portrayed them, Stogdill trained two different actors to play the roles. Consistent with his expectations, he generally found that the same actors playing different roles were described as significantly different on the subscales of the two roles, and that the two actors playing the same role were not described as different on the subscale for that role. Based on these results, Stogdill argued that his findings "constitute evidence that the subscales of the Leader Behavior Description Questionnaire measure what they are purported to measure" (p. 157).

The design of Stogdill's study is instructive because it incorporates many of the steps that are necessary for researchers to use to establish the validity of a scale using experimental procedures, in this specific case with videotape stimuli. Generally speaking, the first steps in the process are to decide on the script, the actors, and the context of the video. Following this, the researcher needs to test the script using subject matter experts to ensure that the scripted events accurately portray the focal construct. The next step is to film and edit the videotape. Finally, the completed videotape should be tested with subjects who are similar to the intended target audience.

The use of videotapes may be more appropriate for validating behavioral/performance measures (e.g., leadership behaviors, task performance, customer service, etc.), scales designed to measure interpersonal interactions (e.g., conflict, attraction, etc.), group processes (e.g., communication among team members, encouragement by other group members, etc.), or emotional displays (e.g., anger, disgust, frustration, etc.) than it is for validating internal states (e.g., attitudes, beliefs, values, anxiety, etc.) and personality traits. However, it is important to note that personality traits may be conceptualized in two different ways (Barrick et al. 2001; Hogan et al. 1996; Johnson 1997). On the one hand, one may define a personality trait genotypically as something that exists inside of a person that causes a distinctive pattern of behavior (over time and across situations). On the other hand, one could define a personality trait phenotypically as the tendency to exhibit a particular distinctive pattern of behavior (over time and across situations). Thus, regardless of which conceptualization is adopted, people who are high in the trait of "extroversion" would be expected to exhibit a distinctively extroverted pattern of behavior (over time and across situations). However, the two different conceptualizations have different methodological implications. If personality is defined genotypically, then the person him/herself is the best source of information about the extent to which the person possesses the trait producing the distinctive pattern of behavior. But if personality is defined phenotypically, then others who know the person well may also be good sources of information about the extent to which the person exhibits the distinctive pattern of behavior. This suggests that if one takes the phenotypic perspective, then videotape scenarios can be used to depict the pattern of behavior the personality scale is intended to measure.

Of course, videos are not the only medium that can be used to experimentally manipulate a construct to validate its measures. Indeed, any written, audio, or visual depiction of the focal construct could also be used for this purpose. This is important, because some constructs may be amenable to

manipulation through written scenarios or case histories, but not through video presentations. For example, measures of strategy constructs like the bargaining power of buyers (Porter 1998) may conceivably be validated by presenting subjects with a fictitious case history in which key events and situational factors related to the construct are manipulated. However, the real key is that the depiction must faithfully represent the construct of interest.

The above discussion relates primarily to constructs having reflective indicators like Panels A and C of Figure 3. However, for constructs with formative indicators, as in Panels B and D in Figure 3, a different procedure for assessing experimental validity is needed. In the case of Panel B, we recommend that validity be assessed at the level of the formative indicators. This would involve an experimental manipulation of the attribute that is measured by the formative indicator, and a test of whether this manipulation influences the scores on the measure of the attribute. Note that this implies that when the measurement model is like the one in Panel D of Figure 3, it is the sub-dimensions that serve as the formative indicators of the second-order construct that should be individually manipulated. This approach is consistent with Borsboom (2005), who argues that we

> Consider, as an example of a construct typically addressed with formative models, Socio-Economic Status (SES). A formative model conceptualizes SES as a latent variable that is regressed on indicators such as annual income, educational level, etc. Now, it would be odd to ask whether the question "what is your annual income?" is a valid measure of SES, because, according to the theory proposed here, this question does not measure SES; rather, it measures one of the determinants of SES, namely annual income. And at this level, one can consistently ask the question of validity, namely when one asks whether variation in annual income has a causal effect on variation in the responses to the question (p. 169).

Known-Groups Comparisons

Another, somewhat weaker, method of assessing whether a set of reflective indicators accurately measures the phenomenon of interest was described by Cronbach and Meehl (1955, p. 287),

> If our understanding of a construct leads us to expect two groups to differ on the test [scale], this expectation may be tested directly. Thus, Thurstone and Chave validated the Scale for Measuring Attitude Toward the Church by showing score differences between church members and nonchurchgoers. Churchgoing is not *the* criterion of attitude, for the purpose of the test [scale] is to measure something other than the crude sociological fact of church attendance; on the other hand, failure to find a difference would have seriously challenged the test.

As indicated in Cronbach and Meehl's example, this technique requires using groups with recognized differences on the construct of interest, and testing whether the mean levels of the measures differ across these groups in the hypothesized direction. Although this method of using known group differences to assess the validity of a scale has not been widely used in management and/or MIS research, there are several examples that have been reported in related disciplines such as marketing (see Lastovicka et al. 1999; Netemeyer et al. 1995; Tian et al. 2001) and personality and social psychology (Greenwald and Farnham 2000; Heimberg and Holaway 2007; Webster and Kruglanski 1994).

For example, Webster and Kruglanski (1994) tested the validity of their need for closure scale (NFCS) by comparing groups of individuals who had chosen careers in Holland's (1985) conventional or artistic career domains. Based on Holland's career typology, Webster and Kruglanski reasoned that conventional type people, who had chosen careers characterized by explicit, ordered, and structured tasks, would score higher on the NFCS than artistic type people, who chose careers characterized by ambiguous and unstructured tasks. This expectation was confirmed, providing support for the validity of the NFCS. Other examples of the use of this technique include the study by Lastovicka et al. (1999), who validated their consumer frugality scale by comparing a random sample of adults to a sample of adults subscribing to a publication called the *Tightwad Gazette*; Netemeyer et al. (1995), who validated the physical concern component of their vanity scale by comparing a sample of people who had cosmetic surgery to a sample who had not; and Tian et al. (2001), who validated the various subcomponents of the need for uniqueness scale by comparing a random sample of adults to groups of tattoo and body-piercing artists, owners of customized low-rider automobiles, members of the Society for Creative Anachronism, undergraduate art majors, and recent purchasers of nouveau art posters.

As indicated by the examples above, the known-groups comparison technique has a long history of use in behavioral research in a variety of different disciplines, and it provides a means of assessing the extent to which the scale accurately captures the phenomenon of interest. Although this method

of assessing validity is weaker than the experimental approach because it provides only correlational (rather than causal) evidence, it might be particularly useful in those situations where it is difficult to experimentally manipulate the attribute, property, etc. of interest. Thus, it is surprising that this method has rarely been used in management and MIS research. Indeed, we did not find any examples in our review of the literature. Consequently, this is a method that management and MIS scholars might want to add to their methodological toolkit. For example, Allen and Meyer (1990) identified three aspects of organizational commitment (affective, continuance, and normative), but argued that fundamental to all of them "is a link between the employee and the organization that decreases the likelihood of turnover" (p. 3). This might suggest that the following group differences would be observed: (1) full-time employees should express higher levels of organizational commitment than temporary employees; (2) executives who receive stock options as part of their compensation should be more committed to the organization than executives who do not receive such compensation; (3) people who volunteer for or contribute to political organizations should be more committed to these organizations than people who are merely registered as members of these organizations; and (4) union employees might be expected to be less committed to the organization for which they work than managers of this organization.

Once again, it is important to note that this procedure differs for constructs having formative indicators. In this case, the first step is to identify groups of individuals, organizations, etc. that are high or low on the attribute, property, etc. measured by the formative indicator. Following this, a test should be conducted of whether a dummy variable representing group membership is significantly related to scores on the indicator. Note that this implies that, when the measurement model is like the one in Panel D of Figure 3, it is the individual sub-dimensions that serve as the formative indicators of the second-order construct on which the groups should differ. This means that different groups may be required for establishing the known-groups validity of different sub-dimensions.

Assess the Nomological and/or Criterion-Related Validity

In addition to establishing the validity of the indicators of the focal construct using the experiment-based or known-groups validity methods, it is also important to (1) specify the nature of the lawful relationships between the focal construct and other constructs, and (2) test whether the indicators of the focal construct relate to measures of other constructs in the manner expected. According to Cronbach and Meehl (1955, p. 290), "to 'make clear what something *is*' means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*." As noted by Bagozzi (1980, p. 129), it is important to specify a nomological network because

> it is not sufficient for determining construct validity to focus solely on semantic criteria of the language used to represent concepts and the relationship among concepts and operationalizations. Nor is it sufficient to examine only the empirical criteria of internal consistency of operationalizations or even convergent and discrimiminant validity. Rather, one must also consider the relationship of the concept under investigation to other concepts in an overall context of a theoretical structure. This will involve... the use of syntactic criteria in combination with the modeling of theoretical and empirical relationships....Nomological validity can be interpreted as an extension of the earlier criteria for construct validity in that—given the satisfaction of theoretical and observational meaningfulness, the internal consistency of operationalizations, and convergent and discriminant validity—the meaning of a theoretical concept rests, in part, on its role in the sentences comprising the... [theoretical] propositions.

For example, confidence in the validity of one's measures of "organizational commitment" should increase if they are related to measures of other constructs in a manner that is consistent with prior theory. The other constructs may be antecedents, correlates, or consequences of the construct of interest identified in previous research. This is depicted in Figure 4. Antecedents are constructs that are hypothesized to cause the focal construct. Consequences are constructs that are hypothesized to be caused by the focal construct. Correlates are constructs whose conceptual definitions overlap with the focal construct. In the early stages of the development of a construct, the specification of the nomological network may include only a few other constructs that would be expected to serve as antecedents, consequences, and/or correlates of the focal construct because there has been little or no previous research involving the construct, and little theoretical discussion about why it might be related to other constructs. This is illustrated by the solid lines in Figure 4. However, as indicated by the dashed lines in this figure, additional research on the construct often results in an expansion of the nomological network to include additional antecedents, correlates, and consequences of the focal construct, as well as potential mediators and moderators of its relationships with other variables. Indeed, Cronbach and Meehl (1955, pp. 290-291) argue that

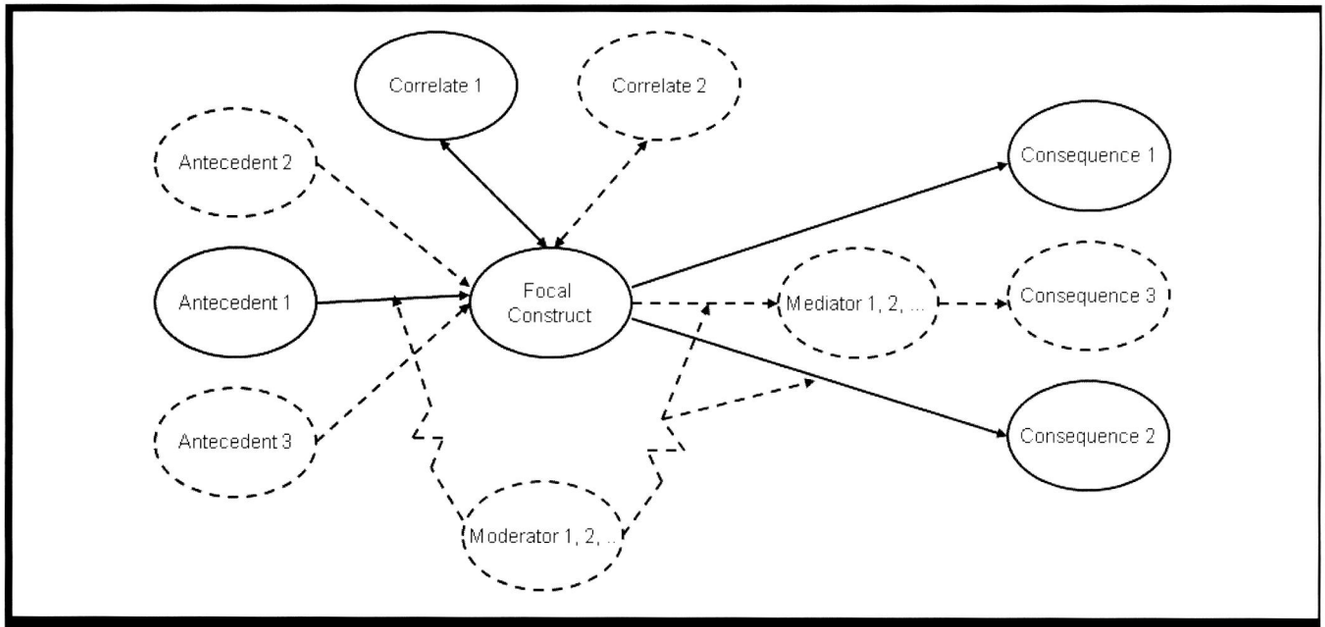


Figure 4. Example of a Nomological Network of a Focal Construct in the Early and More Advanced Stages of Construct Development

> "Learning more about" a theoretical construct is a matter of elaborating the nomological network in which it occurs, or of increasing the definiteness of the components. At least in the early history of a construct the network will be limited, and the construct will as yet have few connections....When a construct is fairly new, there may be few specifiable associations by which to pin down the concept. As research proceeds, the construct sends out roots in many directions, which attach it to more and more facts or other constructs.

Additional research may also lead to a refinement of the focal construct either by expanding its dimensionality or by narrowing its definition over time, because of new empirical findings and/or a more precise theoretical understanding of how the focal construct relates to other constructs.

For example, Mathieu and Zajac (1990) articulate a fairly well developed nomological network for the construct of organizational commitment that includes antecedents (e.g., personal characteristics, role states, job characteristics, group/leader relations, and organizational characteristics), correlates (e.g., job involvement), or consequences (e.g., job performance, turnover, and attendance). However, in the early stages of research on organizational commitment (Mowday et al. 1979), the nomological network was much simpler and only included one of the 26 antecedents, six of the 14 correlates, and three of the 8 consequences identified by Mathieu and Zajac. Indeed, as this example illustrates, it is natural for nomological networks to expand as research on the focal construct increases.

Regardless of whether one is assessing the nomological validity of constructs with formative or reflective indicators, the procedure is the same. It entails estimating the latent constructs (so that measurement error is controlled) and testing whether estimates of their relationships with hypothesized antecedents, correlates, and consequences are significantly different from zero (with the anticipated sign). For example, the models shown in Figure 5 all depict relationships between the focal construct (represented as a second-order latent construct) and its antecedents or consequences. In each of these figures, the relationship between the focal construct and one of its antecedents or consequences is marked with an asterisk (*). The statistical significance of the coefficients for these paths provides the key test of nomological validity of the focal construct's indicators. If these paths are significant, it suggests that the focal construct relates to other constructs as specified in the nomological network, thus increasing confidence in the validity of the indicators. In addition, the magnitude of the completely standardized coefficient associated with these paths provides an indication of the strength of the relationship between the focal construct and its antecedents or consequences, which can be evaluated for consistency with theoretical expectations. For example, if the focal construct

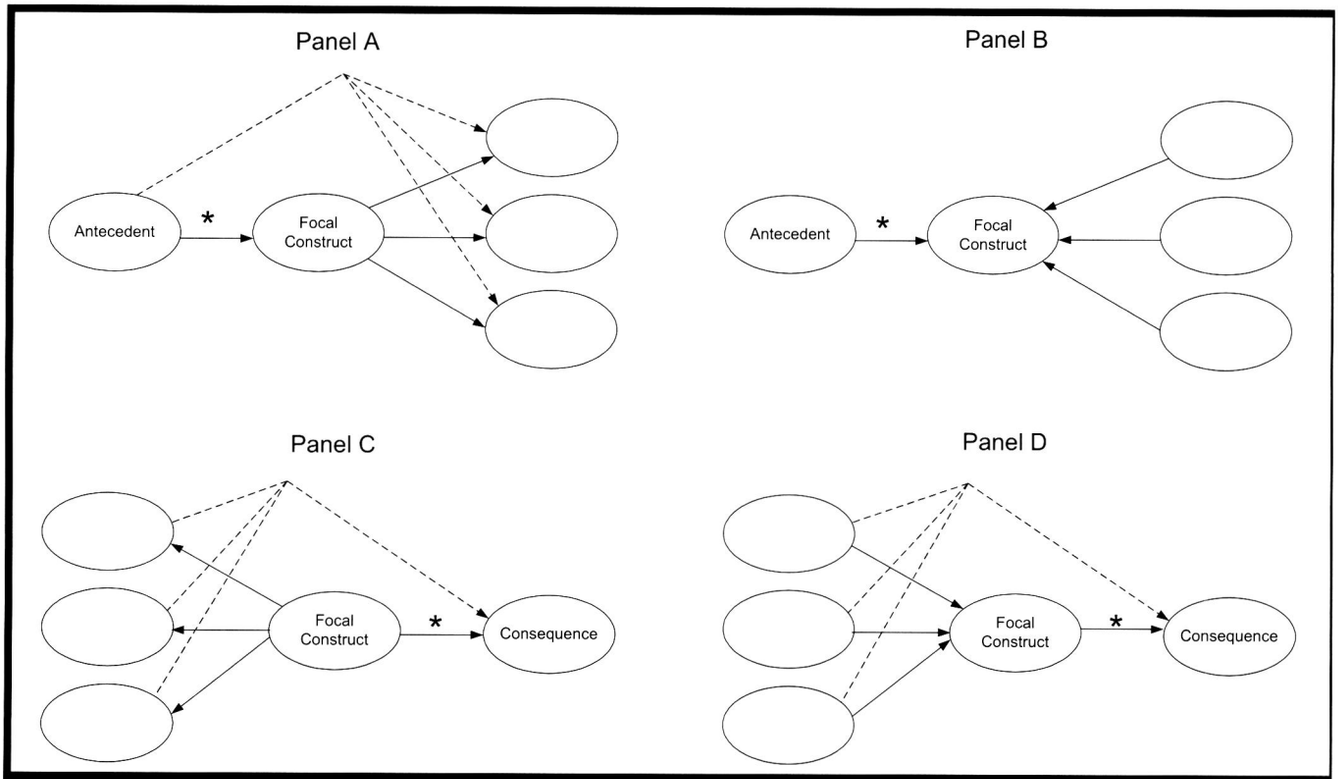


Figure 5. Illustration of Relationships Between a Multidimensional Focal Construct and its Antecedents or Consequences

is expected to be strongly related to one of the other constructs in the nomological network due to conceptual similarity, then the coefficients should not only be significantly different from zero, but also large in magnitude.

Finally, it is important to control for common method biases when conducting these tests (DeVellis 1991). Otherwise, the relationships observed in support of the nomological validity of the indicators of the focal construct with other constructs may be spurious. As discussed by Podsakoff et al. (2003a), method biases can be at least partially controlled either procedurally or statistically using a variety of techniques. For constructs with reflective indicators, either means of control can be used. For constructs with formative indicators, the procedural means of control can be readily applied, but not all of the statistical control methods can be used in every model structure due to identification problems.

Using the Nomological Network to Assess the Validity of the Multidimensional Structure

Edwards (2001) has noted that the relationships in the nomological network can also be quite helpful in further evaluating the adequacy of the multidimensional structure of the focal construct. In the case of an endogenous multidimensional focal construct with reflective indicators (Figure 5, Panel A), this can be done by assessing whether an antecedent construct has a direct effect on the sub-dimensions of the focal construct over and above the indirect effects that this antecedent construct has on the sub-dimensions through the focal construct itself. (These direct effects are represented in Panel A of Figure 5 by dashed lines.) If the direct paths are significant, it means that the antecedent construct accounts for additional variance in one or more of the sub-dimensions over and above that accounted for by the focal construct, and it is likely that the sub-dimension has some unique variance that is not captured by the focal construct. This can be interpreted as evidence that the hypothesized multidimensional structure is inconsistent with the data. However, if the magnitudes of the indirect effects of the antecedent on the sub-dimensions of the focal construct are substantially greater than the magnitudes of the direct effects of the antecedent on the sub-dimensions, then it still may be reasonable to conclude that the hypothesized measurement model is supported.

Similarly, in the case of an exogenous multidimensional focal construct having reflective indicators (Figure 5, Panel C), this

can be done by assessing whether the sub-dimensions of the multidimensional construct have significant direct effects on a consequence construct over and above the direct effect that the focal construct has on the consequence (see dashed lines Panel C of Figure 5). In both cases, the significance of the direct paths can be tested with a chi-square difference test of the models with and without these paths, or by examining the modification indices, which show the expected improvement in fit if a constrained parameter is freed. However, it is important to note that, for the models to be identified, all of the dashed paths cannot be added at the same time (Edwards 2001, p. 186). If the direct paths are significant, it means that the sub-dimensions of the construct account for additional variance in the consequence construct over and above that accounted for by the focal construct, and this suggests that the focal construct does not adequately capture all of the important aspects of its sub-dimensions (Edwards 2001). However, if the magnitude of the effect of the focal construct on the consequence is substantially larger than the combined magnitudes of the direct effects of the sub-dimensions on the consequence construct, then it may still be reasonable to conclude that the measurement model is supported.

In the case of an endogenous multidimensional focal construct with formative indicators (Figure 5, Panel B), the testing of the multidimensional structure is complicated by the fact that this model implicitly includes relationships among all exogenous constructs, thus suggesting that the direct effects of the antecedent on the sub-dimensions cannot be used to test the adequacy of the hypothesized measurement model. Instead, Edwards (2001) argues that the adequacy of the hypothesized multidimensional structure can be assessed by using a confidence interval (based on standard errors obtained from bootstrap procedures) to evaluate whether the R^2 for the effect of the antecedent on the focal construct is equal to the R^2_m for the focal construct's sub-dimensions. If the two are equivalent, it can be interpreted as support for the hypothesized multidimensional structure of the focal construct. Alternatively, he suggests this multidimensional structure can be tested by examining whether the direct effects of the antecedent on each sub-dimension (without the focal construct in the model) are equal. If they are, he reasons that no information is lost by using the focal construct to represent the sub-dimensions in the model. This test is based on the assumption that a single coefficient (the coefficient of the effect of the antecedent on the focal construct), must *completely* represent the effect of the antecedent on all sub-dimensions, or the focal construct is concealing potentially useful information.[12]

In the case of an exogenous multidimensional focal construct having formative indicators (Figure 5, Panel D), the adequacy of the hypothesized multidimensional structure can be assessed by testing whether the sub-dimensions of the multidimensional focal construct have significant direct effects on a consequence construct, over and above the direct effect that the focal construct has on the consequence (see dashed lines in Panel D of Figure 5).[13] As noted above, the significance of these direct paths can be tested with a chi-square difference test of the models with and without these paths, or by examining the modification indices, which show the expected improvement in fit if a constrained parameter is freed. Once again, it is important to note that (1) for the model in Panel D to be identified, all of the dashed paths cannot be added at the same time, and additional consequences (or overall reflective measures) would have to be added (Edwards 2001, p. 186), and (2) if the magnitude of the effect of the focal construct on the consequence construct is substantially larger than the combined magnitudes of the direct effects of the sub-dimensions on the consequence, then it may still be reasonable to conclude that the measurement model is supported.

Assess Discriminant Validity

In addition to showing that the indicators provide an accurate representation of the focal construct, and that they behave in

[12]However it is important to note that, although this test is reasonable, it is not necessarily implied by the structure of the model so its appropriateness depends upon whether it makes sense conceptually. One could argue that this requirement is too stringent, particularly if the single coefficient captures the vast majority of the effect of the antecedent on all of the sub-dimensions of the construct. For example, if 80 to 90 percent of the effect of the antecedent on all of the sub-dimensions is adequately captured by the focal construct that would seem to suggest that the focal construct is useful. Of course, it is true that you could account for 100 percent of the effect of the antecedent on the sub-dimensions by treating each sub-dimension as a separate construct, but that additional explanatory power comes at a price. Theoretically, one would need to develop unique theoretical reasons why each antecedent influences each sub-dimension without reference to the higher-order focal construct previously thought to link the sub-dimensions together. Similarly, unique theoretical rationales for why each sub-dimension influences other consequence constructs would be needed. In addition, treating the sub-dimensions as unrelated constructs would eliminate any sense of the higher-order focal construct. This is problematic because, as we have noted earlier in this paper, many of the constructs in management and MIS are conceptualized at a higher-order level.

[13]For an alternative perspective, see Franke et al. (2008). They argue that a formative measurement model specification implies that the latent construct should *completely* mediate the effect of its indicators on all other outcome variables (e.g., on reflective indicators, or other consequence constructs with reflective indicators), and that if this is not the case, then the measurement model should be rejected. More specifically, based on Bollen and Davis (2009), they reason that if the latent construct completely mediates the effects of the indicators on all outcome variables, then certain proportionality constraints on the model coefficients must hold. If these proportionality constraints do not hold for a particular indicator, then one has reason to doubt its validity.

a manner that is consistent with the nomological network, it is also important to show that these indicators are distinguishable from the indicators of other constructs. As noted by Campbell and Fiske (1959, p. 84), "one cannot define without implying distinctions, and the verification of these distinctions is an important part of the validation process." This is called discriminant validity. After setting the scale of measurement for each construct by fixing its variance at 1.0, discriminant validity can be assessed for any pair of constructs (1) by constraining the estimated correlation between the constructs to 1.0 and then performing a chi-square difference test on the values obtained for the constrained and unconstrained models, or (2) by constructing a confidence interval around the correlation estimate and seeing whether it includes the value of 1.0. Note that although a correlation of less than 1.0 is a necessary condition for demonstrating discriminant validity, it is best not to regard it as a sufficient condition. A more stringent method of assessing discriminant validity might be to test whether the construct intercorrelation is less than .71. Although we are not aware of any uses of this procedure in the literature, it may make sense because it would test whether the constructs have significantly less than half of their variance in common.

Alternatively, for constructs with reflective indicators, Fornell and Larcker (1981) have suggested examining whether the average variance extracted (AVE) for each construct is greater than the square of the correlation between the constructs. Conceptually, this requires that each latent construct account for more of the variance in its own indicators than it shares with another construct. This test should be performed for one pair of constructs at a time, by averaging the squared multiple correlations (or averaging its squared completely standardized item loadings) for each of the two construct's indicators (separately), and then comparing the AVE values to the square of the intercorrelation between the two constructs of interest. If the two constructs are distinct, then the average squared multiple correlation for each of them should be higher than the square of their intercorrelation.

For constructs with formative indicators, discriminant validity can be assessed using the first procedures described above for constructs with reflective indicators (i.e., testing whether the constructs are less than perfectly correlated or have less than half of their variance in common) provided that the composite latent construct has been identified using two reflective indicators and the scale of measurement for the construct has been set by fixing the construct variance at 1.00. An alternative that might be used when this condition is not met (i.e., when the latent construct is not identified) is to calculate a scale score for the focal construct and test whether it is less than perfectly correlated with a measure of another distinct construct. However, the disadvantage of using this procedure is that it ignores measurement error, which can attenuate the relationship between the scale score representing the focal construct and other constructs. This is obviously problematic, since the attenuation makes it more likely to conclude that the constructs are distinct.

Step 9: Cross Validate the Scale

The next step in the scale development process is to cross validate the psychometric properties using new samples. This is particularly important if model modifications were made in the scale development and refinement process. The new samples should be another population to which the construct would be expected to apply. For constructs with reflective indicators, the measurement estimates obtained from the developmental sample could be compared to the estimates obtained from the validation samples using the procedures recommended by Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000). These authors recommend using multigroup analysis to compare a series of nested models with systematically increasing equality constraints across groups to test (1) the equivalence of the covariance matrices, (2) the configural equivalence of the factor structure, (3) the metric equivalence of the factor loadings, and (4) the scalar equivalence of the item intercepts. Following this, these authors present several options for how to proceed in testing for the equivalence of the factor variance and mean across the developmental and validation samples. Steenkamp and Baumgartner note that often full metric equivalence of the factor loadings will not be supported, and in those instances, researchers should test whether partial metric equivalence is supported.

Note that this procedure applies to a first-order factor structure. However, its logic could be applied to a second-order factor structure like the one depicted in Figure 3 Panel C using a variation of this general procedure described by Byrne and Stewart (2006). Interested readers are encouraged to consult their article.

Although the tests of equivalence discussed above were originally developed for constructs having reflective indicators, analogous procedures might be used for constructs with formative indicators like the one shown in Figure 2, Panel C. With the exception of the recent paper by Diamantopoulos and Papadopoulos (2010), this is an issue that has not received much attention in the literature. They recommend that researchers should test (1) the configural equivalence of the pattern of relationships between the formative and reflective indicators and the composite latent construct, (2) the metric equivalence of the estimates of the relationships

between the composite latent construct and its reflective indicators, (3) the equivalence of the estimates of the effects of the formative indicators on the composite latent construct, and (4) the equivalence of the residual error term. One limitation of this procedure (as we have described it) is that it does not permit a test of whether the relationship between the latent construct and the variable used to set the scale of measurement is equivalent across groups, because this parameter is fixed at 1. Diamantopoulos and Papadopoulos recommend a procedure for determining whether this parameter is equivalent across groups *prior to* making the comparisons described above. However, its details are not described here because others have argued that partial metric equivalence across groups is sufficient (Byrne et al. 1989; Reise et al. 1993; Steenkamp and Baumgartner 1998). More specifically, these authors have suggested that full metric invariance is not necessary for further tests of invariance and substantive analyses to be meaningful, provided that at least one item (other than the one fixed at unity to define the scale of each latent construct) is metrically invariant.

Once again, it is important to note that this procedure applies to a first-order factor structure. However, its logic might be adapted to cross-validate a second-order factor structure similar to the one depicted in Figure 3, Panel D but with two reflective indicators of the second-order construct added for identification purposes. For this type of model, we speculate that researchers might test (1) the configural equivalence of the complete structure, (2) the metric equivalence of the estimates of the relationships between the first-order factors and their reflective indicators, (3) the metric equivalence of the estimates of the relationships between the second-order composite latent construct and its reflective indicators, (4) the equivalence of the estimates of the effects of the first-order formative indicators on the composite latent construct, and (5) the equivalence of the residual error term. If tests for differences in the intercepts of the reflective indicators and/or the means of the first-order factors in this model are also desired, the logic of Byrne and Stewart's (2006) procedure might be used to modify the above steps.

Another means of assessing the extent to which the measurement model parameters will cross-validate with other samples is by calculating Browne and Cudeck's (1983) cross validation index (CVI). The process of calculating the CVI is as follows: (1) split the sample into developmental and validation samples; (2) estimate the model using the developmental sample (S_{Dev}) to obtain the predicted covariance matrix ($\hat{\Sigma}_{Dev}$); and (3) calculate the minimum value of the fitting function for the difference between validation sample data (S_{Val}) and the predicted covariance matrix from the developmental sample ($\hat{\Sigma}_{Dev}$). Double cross validation indices can be created by first estimating the model using the validation sample data (S_{Val}) to obtain the predicted covariance matrix ($\hat{\Sigma}_{val}$) and then calculating the minimum value of the fit function for the difference between the developmental sample data (S_{Dev}) and the predicted covariance matrix from the validation sample ($\hat{\Sigma}_{val}$). The CVI can be multiplied by n-1 to produce a chi-square statistic that reflects how well the predicted covariance matrix from the developmental sample ($\hat{\Sigma}_{Dev}$) accounts for validation sample data (S_{Val}). Similarly, the double cross validation index can be multiplied by n-1 to produce a chi-square statistic that reflects how well the predicted covariance matrix from the validation sample ($\hat{\Sigma}_{Val}$) accounts for developmental sample data (S_{Dev}). In principle, the minimum values of the fit functions for the cross validation and double cross validation comparisons can be used to calculate a variety of different goodness-of-fit indices including the chi-square statistic and GFI. Ideally, one would want the chi-squares to be nonsignificant and the GFIs to be above .90. Although to our knowledge this procedure has only been applied to models with constructs having reflective indicators, it may be applicable for constructs with formative indicators as well. Future research should explore this issue.

Step 10: Develop Norms for the Scale

The final step in the scale development process is to develop norms to aid in the interpretation of scores on the scale. This step is important because, as noted by Spector (1992, p. 67),

> In order to interpret the meaning of scores, it is helpful to know something about the distribution of scores in various populations. The scale of measurement for most constructs in the social sciences is arbitrary. The meaning of a score can only be determined in relation to some frame of reference. The normative approach, which forms the basis for much social science measurement, uses the *distribution* of scores as that frame of reference. Hence, the score for an individual is compared to the distribution of scores. The score is considered high if it is greater than most of the distribution, and it is considered low if it is smaller than most of the distribution.

Estimating the population distribution requires administering the scale to a representative sample of members of the population of interest (Urbina 2004). If the population of interest is managers at a particular level of the organizational hierarchy, then a representative sample of that population should be obtained. On the other hand, if a scale is designed to measure attributes or characteristics of IT workers, then a

representative sample of that population should be obtained. Norms for a given population can be obtained by calculating the mean and standard deviation of the scores. In addition, the shape of the distribution (e.g., skewness or kurtosis) should also be examined. Aside from the extra effort associated with collecting this data (which is not trivial), perhaps the biggest barrier to the development of scale norms is the difficulty of obtaining "representative" samples of the population to which one desires to generalize.

Another important consideration in the development of scale norms is the size of the normative samples. As noted by Urbina (2004), the sample needs to be large enough to ensure that the scale scores obtained are stable. However, the required normative sample size varies depending on the size of the population for which the researcher wants to generate norms. For example, generalized ability measures that are used for college admission might require normative samples that number in the tens of thousands, whereas normative samples for measures that are applicable only to members of specialized occupational groups may require only hundreds.

A final consideration in the development and use of scale norms is to recognize that norms may change over time. For example, norms for the Scholastic Aptitude Test (SAT) are known to have changed over time, and this is undoubtedly true for many other types of scales. This suggests that norms need to be periodically updated, and that the time frame during which the norms were established needs to be specified.

For additional information on the development of norms, readers are encouraged to consult the extensive literature on this topic (e.g., Jaeger 1984; Levy and Lemeshow 1999; Sudman 1976). However, because of the difficulty in obtaining representative samples of sufficient size, it is probably fair to say that few scales reported in the field of management information systems have well established norms. Nevertheless, this is useful information and having some data on the distribution of scores in different samples can be helpful, since having some context is probably better than having none at all.

Additional Issues Related to Constructs with Formative Indicators

Our discussion of the steps of the construct validation process above is based on the assumption that conceptual considerations determine the measurement model specification for the focal construct. We have argued that, for conceptual reasons, it may sometimes be desirable to use formative indicators in order to faithfully and fully represent a particular construct. However, others (Howell et al. 2007a, 2007b; Kim et al. 2010; Wilcox 2008) have argued that the use of formative indicators should be avoided whenever possible. For example, Howell et al. (2007b) have argued that

> formative measurement is not an equally attractive alternative to reflective measurement and that whenever possible, in developing new measures or choosing among alternative existing measures, researchers should opt for reflective measurement (p. 205).

They go on to

> strongly suggest that when designing a study, researchers should attempt to measure their constructs reflectively with at least three, and preferably as many as is (practically) possible, strongly correlated indicators that are unidimensional for the same construct (p. 216).

In our view, their recommendation to measure everything exclusively with reflective indicators can sometimes lead to problems. First, as noted by Bollen (2007), appropriate reflective measures may not always be available. This may be the case if a researcher only has access to secondary data, and only formative indicators of the focal construct are available. Second, in order to capture the entire conceptual domain of a complex multidimensional construct, it is usually necessary for the overall reflective measures to be more general than the formative indicators used to represent the sub-dimensions would need to be. Unfortunately, as noted by Converse and Presser (1986),

> The more general the question, the wider the range of interpretations it may be given. By contrast, wording that is specific and concrete is more apt to communicate uniform meaning (p. 31).

Related to this, another disadvantage of relying exclusively on generally worded reflective measures is that research has shown that these types of measures are more subject to context effects than measures worded more specifically (Schuman and Presser 1996). Fourth, Bollen (2007) has noted that relying exclusively on global reflective indicators to represent a construct that has traditionally been measured with formative indicators might change the nature (or the meaning) of the construct:

> socioeconomic status (SES) might be measurable through subjective assessments of a person's

> standing either with self-reports or reports of others. Unfortunately, this confounds subjective SES with objective SES, in which the former taps perceptions whereas the latter gets at a person's standing through variables that are more objective, such as a person's education, income, occupation, and so on. Subjective SES and objective SES are different constructs, possibly with different causes and consequences. I can more easily imagine measuring subjective SES with effect (reflective) indicators than I can imagine measuring objective SES with effect (reflective) indicators. Or consider actual exposure to media violence as a latent variable. We could ask individuals their *perceptions* of the total amount of violence they typically observe in the media, and effect (reflective) indicators could tap this perceived exposure to violence. But these perceptions might differ from the actual total exposure to media violence. This latter latent variable might require causal (formative) indicators that measure amount of time spent watching violent movies, violent television, violent games, violent magazines, and so on. Whether actual or perceived exposure should be used depends on the hypotheses to be tested, but these concepts are not necessarily interchangeable (p. 227).

Fifth, using *only* global reflective indicators in place of the formative indicators of a higher-order construct may diminish the correspondence between the empirical meaning of the construct and its nominal meaning, because we do not know whether the respondent is considering all of the sub-dimensions (facets) of the focal construct that are part of the nominal definition when responding to the global question. This can be illustrated in the case where job satisfaction is nominally defined as a multidimensional construct. For example, according to Smith et al. (1969), job satisfaction can be defined as consisting of five distinct facets (satisfaction with work, pay, promotion, supervision, and coworkers). If we measure this construct exclusively with a global reflective item like "Overall, how satisfied are you with your job," one person may consider how satisfied s/he is with the nature of the work and/or the level of pay s/he receives and report a value of 5 on the seven-point global job satisfaction measure. Another person may consider how satisfied s/he is with his/her coworkers and/or the opportunities for advancement in the organization and report a value of 4 on the global job satisfaction measure. Still another person may evaluate how s/he feels right now as s/he sits at work filling out a rather boring questionnaire from a researcher s/he doesn't know and report a value of 2 on the global job satisfaction measure. Two points are worth noting. First, in none of the cases does the empirical meaning of the latent construct (i.e., the values of 5, 4, and 2) match the nominal meaning of the construct as defined on the basis of theory. Second, across these cases, the differences in the empirical meaning of the construct are due to differing perceptions of the meaning of the question (i.e., differing interpretations of what is being measured), rather than to any differences in the respondents' overall level of job satisfaction. If instead these people had been asked to report how satisfied they were with each of the facets included in the construct definition, and these items were used as formative indicators of the job satisfaction construct, the empirical meaning of the construct would have more closely corresponded to the nominal meaning of construct as defined by Smith et al. Thus, in cases like this, relying exclusively on global reflective indicators of the focal construct may produce a greater discrepancy between the nominal meaning of the construct and its empirical meaning than if formative indicators, or a combination of formative and reflective indicators, are used.

At first glance, one might think that one way to reduce the discrepancy between the empirical and nominal meaning of the focal construct would be to explicitly refer to all of the facets in the "global" reflective indicator. For example, a global measure of job satisfaction may be worded in such a way that it asks respondents to explicitly think about how satisfied they are with various aspects of their job when they are rating their overall satisfaction (e.g., "Generally speaking, how satisfied are you with your work, pay, coworkers, supervisor, and opportunities for promotion?"). However, the disadvantage of this approach is that (1) it makes the question cognitively complex because it is double, triple, or quadruple-barreled (Tourangeau et al. 2000), and (2) the researcher has no way to tell precisely how the facets are being combined by the respondents to generate their response to the question. Although this does not suggest that explicitly mentioning the sub-dimensions is a bad idea, it does suggest that one may not want to rely *exclusively* on this type of question to measure the construct.

Alternatively, one could try to reduce the discrepancy between the empirical and nominal meaning of the construct by measuring each sub-dimension with separate items and then using the set of items as *reflective* indicators of the focal construct. However, in this instance where the sub-dimensions are viewed as defining characteristics of the focal construct, using formative measures *as if* they were reflective measures will (1) result in measurement model misspecification and estimation bias (Jarvis et al. 2003; MacKenzie et al. 2005; Petter et al. 2007), and (2) cause the empirical

meaning of the construct to be defined in terms of common variance, rather than total variance[14] (Law and Wong 1999).

Thus, although relying *exclusively* on reflective indicators may work well for some constructs, for others the representation of the construct would be improved by supplementing the reflective indicators with formative indicators, and for still others formative indicators may be a researcher's only choice. Howell et al. (2007b) have recommended that when reflective measurement of a construct is infeasible or impossible, formative indicators should be modeled as separate unidimensional constructs. The principle advantage of this approach is that it avoids the problems associated with the use of measurement models with formative indicators.

However, this approach would present several challenges. First, treating the sub-dimensions as separate constructs would require researchers to clarify the theoretical rationale for the relationships between each of the individual sub-dimensions and the other constructs they include in their hypothesized model. Although this may lead to better articulated theories, it may present a considerable challenge to researchers particularly in those cases where there are several multidimensional constructs in a model or when the multidimensional constructs are new. Second, it is important to note that if a researcher treats the sub-dimensions as separate constructs, then s/he cannot decompose the total effect of each sub-dimension on an outcome variable into the proportion of the effect that is due to the superordinate construct and the proportion that is due to sub-dimension specific factors, without modeling the superordinate construct explicitly.

Finally, perhaps the most serious problem with this approach is that it ignores the fact that many of the most widely studied constructs in the fields of MIS (e.g., source credibility, perceived user resources, seller's performance, observational learning, firm performance) and management (e.g., job satisfaction, organizational commitment, trust, job performance, transformational leadership) are often conceptualized as being multidimensional or multifaceted. This poses a problem because multidimensional constructs are valued for their generality, breadth, and simplicity (Edwards 2001). When the sub-dimensions are modeled as separate constructs, the empirical model tested does not recognize any conceptual connection between these sub-dimensions and the superordinate construct. One way that this conceptual connection might be taken into account has been discussed by Edwards (2001). He suggests that the total coefficient of determination or multivariate R^2 which represents the proportion of generalized variance in a set of dependent variables explained by one or more independent variables can be used as an estimate of the total effect of an antecedent on a set of separate sub-dimensions of a construct. Similarly, an estimate of the total effect of a set of sub-dimensions on a consequence construct can be assessed by summing the squared correlations between the consequence construct and the sub-dimensions and dividing by the number of sub-dimensions. Note, however, that this procedure assumes that the entire effect of the sub-dimensions on a consequence construct, or the entire effect of an antecedent construct on the set of sub-dimensions, can be attributed to the superordinate construct hypothesized to be associated with the sub-dimensions. The only way to decompose the effect of (on) the superordinate construct on (of) a consequence (an antecedent) construct, from the total effects of (on) the sub-dimensions on (of) the consequence (an antecedent) construct would be to estimate a model that includes a second-order construct.

Thus, although treating the sub-dimensions of a multidimensional construct as separate constructs has some advantages, it has some disadvantages as well. We believe that the decision about whether to explicitly include a superordinate multidimensional construct in a model or not is one that researchers will have to decide based on their conceptual needs. We agree with Edwards (2001, p. 152) that both approaches have merit:

> Advocates of multidimensional constructs endorse generality, breadth, and simplicity, whereas critics promote specificity, precision, and accuracy. Given that both sets of objectives are laudable, researchers would be better served by an integrative approach than by admonitions to adopt one side of the debate.

General Discussion and Conclusion

We believe that this paper makes several contributions to the literature. First, there is little guidance in the literature on how to go about defining theoretical constructs. This is a serious omission, because the failure to adequately define a construct creates a number of problems that undermine not only construct validity, but also internal validity and statistical conclusion validity (MacKenzie 2003). Second, although there are some papers that have been written on scale development processes and construct validation over the past few decades, these papers have tended to focus more on reflective measurement models than on formative measurement models.

[14]Note that empirically defining the latent construct in terms of total variance would be more consistent with the conceptual definition of the construct than defining the latent construct in terms of common variance would be.

Therefore, one of the contributions of this research is that it provides a more comprehensive set of recommendations on how the scale development process should take the directional relationship between the indicators and the latent construct they are intended to measure into account. Third, this paper identifies a key limitation of widely used tests of nomological validity and discusses how the experimental and known-groups comparison techniques can be used to obtain critical information that does not suffer from this limitation. Both of these techniques for establishing construct validity have a long tradition in the social sciences (Cronbach and Meehl 1955), but are rarely used in the fields of management or MIS. Finally, this paper addresses these issues and integrates them into a comprehensive set of steps which researchers can use to guide their scale development and construct validation efforts.

All 10 of the steps that we have outlined are important for the development of valid scales. Without a clear definition, it is difficult to avoid contamination and deficiencies in the set of items used to represent the construct or to specify the relationship between the construct and its indicators. If the indicators do not adequately capture the domain of the construct, there may be little value in examining their psychometric properties or the relationships between these indicators and the indicators of other constructs. If the measurement model is improperly specified, it may lead to inappropriately dropping items that are necessary to capture the complete domain of the construct, result in the use of inappropriate scale evaluation indices, and bias estimates of the relationships between the construct and other constructs. If the researcher does not properly test the measurement model and evaluate the scale, it is difficult to determine whether the hypothesized measurement relationships are consistent with the data or to know how to refine the scale to improve its psychometric properties. Unless the validity of the scale is adequately assessed using experimental manipulations, by comparing groups known to differ on the construct, testing the relationships between the focal construct and other constructs in its nomological network, and examining whether the focal construct is distinct from other, similar constructs, the researcher will be uncertain about whether scores on the scale covary with the phenomenon the scale is intended to measure. Unless the scale is cross-validated across subject populations, situations, and time, it will be difficult to evaluate the limits of its generalizability or its usefulness in other contexts. Finally, without adequately established norms for the scale, it is difficult to determine in an *absolute sense* the meaning of a particular score on the scale.

That being said, we recognize that practical limitations may prevent researchers from being able to follow all of the recommendations discussed in this paper in a single study, either because of a lack of time or resources, or both. Indeed, as noted by Nunnally and Bernstein (1994),

> Each scientist can perform only a relatively small number of major studies in a lifetime. This leaves insufficient time to do all that is required to specify the domain of a construct, develop measures of the construct, and relate these measures to other variables of interest (pp. 87-88).

With this in mind, we would like to offer some suggestions for establishing priorities in situations where it is not practical to conduct all of the steps.

First, because so many things depend upon having a clear conceptual definition, this is one step in the process that should never be neglected in a scale validation study. More generally, we recommend focusing more attention on the front-end of the process—on providing a clear conceptual definition and developing indicators that adequately tap the construct domain and properly specifying the measurement model—than on cross-validating the scale and developing norms for it. Another way to economize might be to combine some of the steps in the process. For example, a researcher can gather data from one large sample, and split this sample into developmental and validation subsamples. The developmental sample can then be used to purify and refine the scale, and the validation sample can be used to examine nomological, discriminant, and convergent validity. However, as we noted earlier, this strategy will only work if the researcher is dropping items from the scale, rather than adding or modifying the items. Similarly, researchers can also combine cross-validation efforts with the development of scale norms because these two steps are synergistic in that they both require the collection of data from different samples. Finally, researchers may choose to omit some of the more sophisticated techniques, such as the procedure we described for using the nomological network to assess the validity of a multidimensional structure. That being said, one set of activities that we think should not be omitted is rigorously testing whether scores on the scale covary with the phenomenon the scale is intended to measure, either through the use of experimental manipulations of the focal construct, comparisons of groups known to differ on the focal construct, or tests of the relationships between the focal construct and other constructs.

In summary, the goal of this paper was not to articulate every possible technique that researchers should use to validate newly developed scales, but to follow the lead of Churchill (1979) by providing an updated set of guidelines that

researchers could use to improve the quality of measures used in research in the behavioral sciences. Hopefully, they will prove useful to those conducting research in MIS and the behavioral sciences.

References

Allen, N. J., and Meyer, J. P. 1990. "The Measurement and Antecedents of Affective, Continuance and Normative Commitment to the Organization," *Journal of Occupational Psychology* (63:1), pp. 1-18.

Anastasi, A., and Urbina, S. 1997. *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, J. C., and Gerbing, D. W. 1988. "Structural Equation Modeling in Practice – A Review and Recommended Two-Step Approach," *Psychological Bulletin* (103:3), pp. 411-423.

Anderson, J. C., and Gerbing, D. W. 1991. "Predicting the Performance of Measures in a Confirmatory Factor Analysis with a Pretest Assessment of Their Substantive Validities," *Journal of Applied Psychology* (76:5), pp. 732-740.

Anderson, J. C., Gerbing, D. W., and Hunter, J. E. 1987. "On the Assessment of Unidimensional Measurement—Internal and External Consistency, and Overall Consistency Criteria," *Journal of Marketing Research* (24:4), pp. 432-437.

Avolio, B. J., Bass, B. M., and Jung, D. I. 1999. "Re-examining the Components of the Transformational and Transactional Leadership Using the Multifactor Leadership Questionnaire," *Journal of Occupational and Organizational Psychology* (72:4), pp. 441-462.

Bagozzi, R. P. 1980. *Causal Models in Marketing*, New York: John Wiley.

Bagozzi, R. P. 2007. "On the Meaning of Formative Measurement and How it Differs from Reflective Measurement: Comment on Howell, Breivik, and Wilcox (2007)," *Psychological Methods* (12:2), pp. 229-237.

Bagozzi, R. P. 2010. "Structural Equation Models Are Modeling Tools with Many Ambiguities: Comments Acknowledging the Need for Caution and Humility in Their Use," *Journal of Consumer Psychology* (20:2), pp. 208-214.

Bagozzi, R. P., and Heatherton, T. F. 1994. "A General Approach to Representing Multifaceted Personality Constructs: Application to State Self-Esteem," *Structural Equation Modeling* (1:1), pp. 35-67.

Bagozzi, R. P., and Phillips, L. W. 1982. "Representing and Testing Organizational Theories: A Holistic Construal," *Administrative Science Quarterly* (27:3), pp. 459-489.

Bagozzi, R. P., Yi, Y., and Phillips, L. W. 1991. "Assessing Construct Validity in Organizational Research," *Administrative Science Quarterly* (36:3), pp. 421-458.

Bagozzi, R. P., Tybout, A. M., Craig, C. S., and Sternthal, B. 1979. "The Construct Validity of the Tripartite Classification of Attitudes," *Journal of Marketing Research* (16:1), pp. 88-95.

Bandalos, D. L. 2002. "The Effects of Item Parceling on Goodness-of-Fit and Parameter Estimate Bias in Structural Equation Modeling," *Structural Equation Modeling* (9:1), pp. 78-102.

Bandalos, D. L., and Finney, S. J. 2001. "Item Parceling Issues in Structural Equation Modeling," in *Advanced Structural Equation Modeling: New Developments and Techniques*, G. A. Marcoulides and R. E. Schumacker (eds.), Mahwah, NJ: Lawrence Erlbaum, pp. 269-296.

Barrick, M. R., Mount, M. K., and Judge, T. A. 2001. "Personality and Job Performance at the Beginning of the New Millennium: What Do We Know and Where Do We Go Next?," *International Journal of Selection and Assessment* (9:1-2), pp. 9-30.

Baumgartner, H., and Steenkamp, J-B. E. M. 2006. "An Extended Paradigm for Measurement Analysis of Marketing Constructs Applicable to Panel Data," *Journal of Marketing Research* (43:3), pp. 431-442.

Bharadwaj, A. S. 2000. "A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Study," *MIS Quarterly* (24:1), pp.169-196.

Blalock, H. M., Jr. 1964. *Causal Inferences in Nonexperimental Research*, New York: W. W. Norton.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*, New York: John Wiley.

Bollen, K. A. 2007. "Interpretational Confounding Is Due to Misspecification, Not to Type of Indicator: Comment on Howell, Breivik, and Wilcox (2007)," *Psychological Methods* (12:2), pp. 219-228.

Bollen, K. A. 2011. "Evaluating Effect, Composite, and Causal Indicators in Structural Equation Models," *MIS Quarterly* (35:2), pp. 359-372.

Bollen, K. A., and Davis, W. R. 2009. "Causal Indicator Models: Identification, Estimation, and Testing," *Structural Equation Modeling* (16:3), pp. 498-522.

Bollen, K. A., and Lennox, R. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective," *Psychological Bulletin* (110:2), pp. 305-314.

Bollen, K. A., and Ting, K-F. 2000. "A Tetrad Test for Causal Indicators," *Psychological Methods* (5:1), pp. 3-22.

Borsboom, D. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*, Cambridge, UK: Cambridge University Press.

Boudreau, M-C., Gefen, D., and Straub, D. W. 2001. "Validation in Information Systems Research: A State-of-the-Art Assessment," *MIS Quarterly* (25:1), pp. 1-16.

Browne, M. W., and Cudeck R. 1983. "Cross-Validation of Covariance Structures," *Multivariate Behavioral Research* (18:2), pp. 147-167.

Burt, R. S. 1976. "Interpretational Confounding of Unobserved Variables in Structural Equation Models," *Sociological Methods and Research* (5), pp. 3-52.

Byrne, B. M., Shavelson, R. I., and Muthen, B. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance," *Psychological Bulletin* (105:3), pp. 456- 466.

Byrne, B. M., and Stewart, S. A. 2006. "The MACS Approach to Testing for Multigroup Invariance of a Second-Order Structure:

A Walk Through the Process," *Structural Equation Modeling—A Multidisciplinary Journal* (13:2), pp. 287-321.

Cammann, C., Fichman, M., Jenkins, D., and Klesh, J. 1983. "Assessing the Attitudes and Perceptions of Organizational Members," in *Assessing Organizational Change: A Guide to Methods, Measures and Practices*, S. Seashore, E. E. Lawler, P. Mirvis, and C. Cammann (eds.), New York: John Wiley & Sons, pp. 71-138.

Campbell, D. T., and Fiske, D. W. 1959. "Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix," *Psychological Bulletin* (56:2), pp. 81-105.

Carlson, L., and Grossbart, S. 1988. "Parental Style and Consumer Socialization of Children," *Journal of Consumer Research* (15:1), pp. 77-94.

Cattell, R. B. 1978. *The Scientific Use of Factor Analysis*, New York: Plenum Press.

Chaplin, W. F., John, O. P., and Goldberg, L. R. 1988. "Conceptions of States and Traits: Dimensional Attributes with Ideals as Prototypes," *Journal of Personality and Social Psychology* (54:4), pp. 541-557.

Churchill, G. A., Jr. 1979. "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research* (16:1), pp. 64-73.

Clark, L. A., and Watson, D. 1995. "Constructing Validity: Basic Issues in Objective Scale Development," *Psychological Assessment* (7:3), pp. 309-319.

Comrey, A. L., and Lee, H. B. 2002. *A First Course in Factor Analysis*, Hillsdale, NJ: Lawrence Erlbaum.

Converse, J. M., and Presser, S. 1996. *Survey Questions: Handcrafting the Standardized Questionnaire*, Newbury Park, CA: Sage Publications.

Cortina, J. M. 1993. "What Is Coefficient Alpha? An Examination of Theory and Application," *Journal of Applied Psychology* (78:1), pp. 98-104.

Cronbach, L. J., and Meehl, P. E. 1955. "Construct Validity in Psychological Tests," *Psychological Bulletin* (52:4), pp. 281-302.

Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), pp. 318-346.

DeVellis, R. F. 1991. *Scale Development: Theory and Applications*, Newbury Park, CA: Sage Publications.

Diamantopoulos, A. 2006. "The Error Term in Formative Measurement Models: Interpretation and Modeling Implications," *Journal of Modeling in Management* (1:1), pp. 7-17.

Diamantopoulos, A. 2011. "Incorporating Formative Measures into Covariance-Based Structural Equation Models," *MIS Quarterly* (35:2), pp. 335-358.

Diamantopoulos, A., and Papadopoulos, N. 2010. "Assessing the Cross-National Invariance of Formative Measures: Guidelines for International Business Researchers," *Journal of international Business Studies* (41:2), pp. 360-370.

Diamantopoulos, A., Reifler, P., and Roth, K. P. 2008. "Advancing Formative Measurement Models," *Journal of Business Research* (61:12), pp. 1203-1218.

Diamantopoulos, A., and Siguaw, J. A. 2006. "Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration," *British Journal of Management* (17:4), pp. 263-282.

Diamantopoulos, A., and Winklhofer, H. M. 2001. "Index Construction with Formative Indicators: An Alternative to Scale Development," *Journal of Marketing Research* (38:2), pp. 269-277

Doll, W. J., and Torkzadeh, G. 1988. "The Measurement of End-User Satisfaction," *MIS Quarterly* (12:2), pp. 259-274.

Edwards, J. R. 2001. "Multidimensional Constructs in Organizational Behavior Research: An Integrative Analytical Framework," *Organizational Research Methods* (4:2), pp. 141-192.

Edwards, J. R. 2003. "Construct Validation in Organizational Behavior Research," in *Organizational Behavior: The State of the Science*, J. Greenberg (ed.), Mahwah, NJ: Lawrence Erlbaum Associates, pp. 327-371.

Edwards, J. R., and Bagozzi, R. P. 2000. "On the Nature and Direction of Relationships Between Constructs and Measures," *Psychological Methods* (5:2), pp. 155-174.

Everitt, B. S. 1975. "Multivariate Analysis: The Need for Data, and Other Problems," *British Journal of Psychiatry* (126:2), pp. 237-240.

Fornell, C., and Larcker, D. F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18:1), pp. 39-50.

Franke, G. R., Preacher, K. J., and Rigdon, E. E. 2008. "Proportional Structural Effects of Formative Indicators," *Journal of Business Research* (61:12), pp. 1229-1237.

Gefen, D. 2003. "Unidimensional Validity: An Explanation and an Example," *Communications of the AIS* (12:2), pp. 23-47.

Gerbing, D. W., and Anderson, J. C. 1984. "On the Meaning of Within-Factor Correlated Measurement Errors," *Journal of Consumer Research* (11:1), pp. 572-580.

Gerbing, D. W., and Anderson, J. C. 1988. "An Updated Paradigm for Scale Development Incorporating Unidimensionality and its Assessment," *Journal of Marketing Research* (25:2), pp. 186-192.

Goertz, G. 2006. *Social Science Concepts: A User's Guide*, Princeton, NJ: Princeton University Press.

Gorsuch, R. L. 1983. *Factor Analysis* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum.

Greenwald, A. G., and Farnham, S. D. 2000. "Using the Implicit Association Test to Measure Self-Esteem and Self-Concept," *Journal of Personality and Social Psychology* (79:6), pp. 1022-1038.

Hall, R. J., Snell, A. F., and Singer-Foust, M. 1999. "Item Parceling Strategies in SEM: Investigating the Subtle Effects of Unmodeled Secondary Constructs," *Organizational Research Methods* (2:3), pp. 233-256.

Hancock, G. R., and Mueller, R. O. 2001. "Rethinking Construct Reliability Within Latent Variable Systems," in *Structural Equation Modeling: Present and Future – A Festschrift in Honor of Karl Jöreskog*, R. Cudeck, S. du Toit, and D. Sörbom (eds.), Lincolnwood, IL: Scientific Software International Inc., pp. 195-216.

Haynes, S. N., Richard, D. C. S., and Kubany, E. S. 1995. "Content Validity in Psychological Assessment: A Functional Approach

to Concepts and Methods," *Psychological Assessment* (7:3), pp. 238-247.

Heimberg, R. G., and Holaway, R. M. 2007. "Examination of the Known-Groups Validity of the Liebowitz Social Anxiety Scale," *Depression and Anxiety* (24:7), pp. 447-454.

Hempel, C. G. 1970. "Fundamentals of Concept Formation in Empirical Science," in *Foundations of the Unity of Science* (Vol. 2), O. Neurath, R. Carnap, and C. Morris (eds.), Chicago: University of Chicago Press, pp. 653-740.

Hinkin, T. R. 1995. "A Review of Scale Development Practices in the Study of Organizations," *Journal of Management* (21:5), pp. 967-988.

Hinkin, T. R., and Tracey, J. B. 1999. "An Analysis of Variance Approach to Content Validation," *Organizational Research Methods* (2:2), 175-186.

Hogan, R., Hogan, J., and Roberts, B. W. 1996. "Personality Measurement and Employment Decisions," *American Psychologist* (51:5), pp. 469-477.

Holland, J. L. 1985. *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*, Englewood Cliffs, NJ: Prentice Hall.

Hovland, C., Janis, I., and Kelley, H. 1953. *Communication and Persuasion*, New Haven, CT: Yale University Press.

Howell, J. M., and Hall-Merenda, K. E. 1999. "The Ties That Bind: The Impact of Leader–Member Exchange, Transformational and Transactional Leadership, and Distance on Predicting Follower Performance," *Journal of Applied Psychology* (84:5), pp. 680-694.

Howell, R. D., Breivik, E., and Wilcox, J. B. 2007a. "Is Formative Measurement Really Measurement? Reply to Bollen (2007) and Bagozzi (2007)," *Psychological Methods* (12:2), pp. 238-245.

Howell, R. D., Breivik, E., and Wilcox, J. B. 2007b. "Reconsidering Formative Measurement," *Psychological Methods* (12:2), pp. 205-218.

Hu, L. T., and Bentler, P. M. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives," *Structural Equation Modeling* (6:1), pp. 1-55.

Jaeger, R. M. 1984. *Sampling in Education and Social Sciences*, New York: Longman, Green and Co.

Jarvenpaa, S. L., Dickson, G. W., and DeSanctis, G. L. 1984. "Methodological Issues in Experimental IS Research: Experiences and Recommendations," in *Proceedings of the Fifth International Conference on Information Systems*, Tucson, AZ, pp. 1-30.

Jarvis, C. B., MacKenzie, S B., and Podsakoff, P. M. 2003. "A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research," *Journal of Consumer Research* (30:2), pp. 199-218.

Johnson, J. A. 1997. "Units of Analysis for the Description and Explanation of Personality," in *Handbook of Personality Psychology*, R. Hogan, J. Johnson, and S. Briggs (eds.), San Diego: Academic Press, pp. 73-93.

Jöreskog, K. G., and Goldberger, A. S. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable," *Journal of the American Statistical Association* (70:351), pp. 631-639.

Kerlinger, F. N. 1973. *Foundations of Behavioral Research* (2nd ed.), New York: Holt McDougal.

Kim, G., Shin, B., and Grover, V. 2010. "Investigating Two Contradictory Views of Formative Measurement in Information Systems Research," *MIS Quarterly* (34:2), pp. 345-365.

Kozlowski, S. W. J., and Klein, K. J. 2000. "A Multilevel Approach to Theory and Research in Organizations: Contextual, Temporal, and Emergent Processes," in *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Direction*, K. J. Klein and S. W. J. Kozlowski (eds.), San Francisco: Jossey-Bass, pp. 3-90.

Lastovicka, J. L., Betencourt, L. A., Hughner, R. S., and Kuntze, R. J. 1999. "Lifestyle of the Tight and Frugal: Theory and Measurement," *Journal of Consumer Research* (26:1), pp. 85-98.

Law, K. S., and Wong, C. S. 1999. "Multidimensional Constructs in Structural Equation Analysis: An Illustration Using the Job Perception and Job Satisfaction Constructs," *Journal of Management* (25:2), pp. 143-160.

Law, K. S., Wong, C. S., and Mobley, W. H. 1998. "Toward a Taxonomy of Multidimensional Constructs," *Academy of Management Review* (23:4), pp. 741-755.

Lawshe, C. H. 1975. "A Quantitative Approach to Content Validity," *Personnel Psychology* (28:4), pp. 563-575.

Levy, P. S., and Lemeshow, S. 1999. *Sampling of Populations: Methods and Applications* (3rd ed.), New York: John Wiley.

Little, T. D., Lindenberger, U., and Nesselrode, J. R. 1999. "On Selecting Indicators for Multivariate Measurement and Modeling of Latent Variables: When 'Good' Indicators are Bad and 'Bad' Indicators are Good," *Psychological Methods* (4:2), pp. 192-211.

MacCallum, R. C., and Austin, J. T. 2000. "Applications of Structural Equation Modeling in Psychological Research," *Annual Review of Psychology* (51:1), pp. 201-226.

MacCallum, R. C., and Browne, M. W. 1993. "The Use of Causal Indicators in Covariance Structure Models: Some Practical Issues," *Psychological Bulletin* (114:3), pp. 533-541.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. 1999. "Sample Size in Factor Analysis," *Psychological Methods* (4:1), pp. 84-99.

MacKenzie, S. B. 2003. "The Dangers of Poor Construct Conceptualization," *Journal of the Academy of Marketing Science* (31:3), pp. 323-326.

MacKenzie, S. B., Podsakoff, P. M., and Jarvis, C. B. 2005. "The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions," *Journal of Applied Psychology* (90:4), pp. 710-730.

MacKinnon, D. P. 2008. *Introduction to Statistical Mediation Analysis*, New York: Taylor & Francis Group.

Mathieu, J. E., and Zajac, D. M. 1990. "A Review and Meta-Analysis of the Antecedents, Correlates, and Consequences of Organizational Commitment," *Psychological Bulletin* (108:2), pp. 171-194.

Mowday, R. T., Steers, R. M., and Porter, L. W. 1979. "The Measurement of Organizational Commitment," *Journal of Vocational Behavior* (14:2), pp. 224-247.

Nederhof, A. J. 1985. "Methods of Coping with Social Desirability Bias: A Review," *European Journal of Social Psychology* (15:3), pp. 263-280.

Netemeyer, R. G., Burton, S., and Lichtenstein, D. R. 1995. "Trait Aspects of Vanity: Measurement and Relevance to Consumer Behavior," *Journal of Consumer Research* (21:4), pp. 612-626.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. *Applied Linear Statistical Models* (4th ed.), Boston: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. 1994. *Psychometric Theory* (3rd ed.), New York: McGraw Hill.

Peterson, R. A. 2000. *Constructing Effective Questionnaires*, Thousand Oaks, CA: Sage Publications.

Petter, S., Straub, D., and Rai, A. 2007. "Specifying Formative Constructs in Information Systems Research," *MIS Quarterly* (31:4), pp. 623-656.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., and Podsakoff, N. P. 2003a. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *Journal of Applied Psychology* (88:5), pp. 879-903.

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., and Lee, J. Y. 2003b. "The Mismeasure of Man(agement) and its Implications for Leadership Research," *Leadership Quarterly* (14:6), pp. 615-656.

Porter, M. E. 1998. *Competitive Strategy*, New York: Free Press.

Rai, A., Patnayakuni, R., and Seth, N. 2006. "Firm Performance Impacts of Digitally Enabled Supply Chain Integration Capabilities," *MIS Quarterly* (30:2), pp.225-246.

Reise, S. P., Widaman, K. F., and Pugh, R. H. 1993. "Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance," *Psychological Bulletin* (114:3), pp. 552-566.

Rotundo, M., and Sackett, P. R. 2001. "The Relative Importance of Task, Citizenship, and Counterproductive Performance to Global Ratings of Job Performance: A Policy-Capturing Approach," *Journal of Applied Psychology* (87:1), pp. 66-80.

Sartori, G. 1984. "Guidelines for Concept Analysis," in *Social Science Concepts: A Systematic Analysis*, G. Sartori (ed.), Beverly Hills, CA: Sage Publications, pp. 15-85.

Scandura, T. A., and Williams, E. A. 2000. "Research Methodology in Management: Current Practices, Trends, and Implications for Future Research," *Academy of Management Journal* (43:6), pp. 1248-1264.

Schuman, H., and Presser, S. 1996. *Questions and Answers in Attitude Surveys: Experiments on Questions Form, Wording, and Context*, Thousand Oaks, CA: . Sage Publications.

Schwab, D. P. 1980. "Construct Validity in Organizational Behavior," in *Research in Organizational Behavior* (Vol. 2), B. M. Staw and L. L. Cummings (eds.), Greenwich, CT: JAI Press, pp. 3-43.

Schriesheim, C. A., Cogliser, C. C., Scandura, T. A., Lankau, M. J., and Powers, K. J. 1999. "An Empirical Comparison of Approaches for Quantitatively Assessing the Content Adequacy of Paper-and-Pencil Measurement Instruments," *Organizational Research Methods* (2:2), pp. 140-156.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., and Lankau, M. J. 1993. "Improving Construct Measurement in Management Research: Comments and a Quantitative Approach for Assessing the Theoretical Adequacy of Paper-and-Pencil Survey-Type Instruments," *Journal of Management* (19:2), pp. 385-417.

Segars, A. H. 1997. "Assessing the Unidimensionality of Measurement: A Paradigm and Illustration Within the Context of Information Systems Research," *Omega* (25:1), pp 107-121.

Serva, M. A., Fuller, M. A., and Mayer, R. C. 2005. "The Reciprocal Nature of Trust: A Longitudinal Study of Interacting Teams," *Journal of Organizational Behavior* (26:6), pp. 625-648.

Smith, P. C., Kendall, L. M., and Hulin, C. L. 1969. *The Measurement of Satisfaction in Work and Retirement: A Strategy for the Study of Attitudes*, Chicago: Rand McNally.

Sobel, M. E. 1982. "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models," in *Sociological Methodology 1982*, S. Leinhardt (ed.), Washington, DC: American Sociological Association, pp. 290-312.

Spector, P. E. 1992. *Summated Rating Scale Construction: An Introduction*, Newbury Park, CA: Sage Publications.

Spector, P. E. 1997. *Job Satisfaction: Application, Assessment, Causes, and Consequences*, Thousand Oaks, CA: Sage Publications.

Steenkamp, J-B. E. M., and Baumgartner, H. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research* (25:1), pp. 78-90.

Stogdill, R. M. 1969. "Validity of Leader Behavior Descriptions," *Personnel Psychology* (22:2), pp. 153-158.

Straub, D. W. 1989. "Validating Instruments in MIS Research," *MIS Quarterly* (13:2), pp. 147-169.

Straub, D. W., Boudreau, M-C, and Gefen, D. 2004. "Validation Guidelines for IS Positivist Research," *Communications of the AIS* (13), pp. 380-427.

Sudman, S. 1976. *Applied Sampling*, New York: Academic Press.

Tian, K. T., Beardon, W. O., and Hunter, G. L. 2001. "Consumers' Need for Uniqueness: Scale Development and Validation," *Journal of Consumer Research* (28:1), pp. 50-66.

Tourangeau, R., Rips, L. J., and Rasinski, K. 2000. *The Psychology of Survey Response*, Cambridge, UK: Cambridge University Press.

Urbina, S. 2004. *Essentials of Psychological Testing*, Hoboken, NJ: John Wiley & Sons.

Vandenberg, R. J., and Lance, C. E. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research," *Organizational Research Methods* (3:1), pp. 4-70.

Vroom, V. H. 1964. *Work and Motivation*, New York; John Wiley & Sons.

Webster, D. M., and Kruglanski, A. W. 1994. "Individual Differences in Need for Cognitive Closure," *Personality Processes and Individual Differences* (67:6), pp. 1049-1062.

West, S. G., Finch, J. F., and Curran, P. J. 1995. "Structural Equation Modeling with Nonnormal Variables: Problems and Remedies," in *Structural Equation Modeling: Concepts, Issues,*

and Applications, R. H. Hoyle (ed.), Thousand Oaks, CA: Sage Publications, pp. 56-75.

Wilcox, J. B., Howell, R. D., and Breivik, E. 2008. "Questions About Formative Measurement," *Journal of Business Research* (61:12), pp. 1219-1228.

Williams, L. J., Edwards, J. R., and Vandenberg, R. J. 2003. "Recent Advances in Causal Modeling Methods for Organizational and Management Research," *Journal of Management* (29:6), pp. 903-936.

Winer, B. J. 1971. *Statistical Principles in Experimental Design*, New York: McGraw-Hill.

Wong, C. S., Law K. S., and Huang, G. H. 2008. "On the Importance of Conducting Construct-Level Analysis for Multidimensional Constructs in Theory Development and Testing," *Journal of Management* (34:4), pp. 744-764.

Yao, G., Wu, C-H., and Yang, C-T. 2008. "Examining the Content Validity of the WHOQOL-BREF from Respondents' Perspective by Quantitative Methods," *Social Indicators Research* (85:3), pp. 483-498.

Yi, M. Y., and Davis F. D. 2003. "Developing and Validating an Observational Learning Model of Computer Software Training and Skill Acquisition," *Information Systems Research* (14:2), pp. 146-169.

About the Authors

Scott B. MacKenzie (Ph.D., University of California, Los Angeles), is the Neal Gilliatt Chair and Professor of Marketing at the Kelley School of Business at Indiana University. His current research focuses on research methodology, advertising effectiveness, sales team performance, leadership, and organizational citizenship behavior. He currently serves on the editorial review boards of *Journal of Marketing Research*, *Journal of Marketing*, *Journal of Consumer Research*, and several other journals.

Philip M. Podsakoff (DBA, Indiana University) is professor of organizational behavior and human resource management, and holder of the John F. Mee Chair of Management in the Kelley School of Business at Indiana University. His current research focuses on the determinants of leadership effectiveness, the antecedents and consequences of organizational citizenship behaviors, relationships between employee attitudes and behaviors, and methodological issues in the social and behavioral sciences. He currently serves on the editorial review boards of *Journal of Applied Psychology, Organizational Behavior and Human Decision Processes, Personnel Psychology*, and *Journal of Management.*

Nathan Podsakoff (Ph.D., University of Florida) is an assistant professor at the Eller College of Management at the University of Arizona. His research focuses on the effects of work-related stressors, the effects of citizenship behaviors in the workplace, identifying and explaining scholarly impact in the organizational behavior field, and issues related to the measurement and modeling of constructs in the behavioral sciences. He currently serves on the editorial review boards of *Academy of Management Journal*, *Journal of Applied Psychology*, and *Organizational Behavior and Human Decision Processes*.