MIS Quarterly

RESEARCH COMMENTARY

MEASUREMENT AND MEANING IN INFORMATION SYSTEMS AND ORGANIZATIONAL RESEARCH: METHODOLOGICAL AND PHILOSOPHICAL FOUNDATIONS[1]

Richard P. Bagozzi
Ross School of Business, University of Michigan, 701 Tappan Street,
Ann Arbor, MI 48109-1234 U.S.A. {bagozzi@umich.edu}

Despite renewed interest and many advances in methodology in recent years, information systems and organizational researchers face confusing and inconsistent guidance on how to choose amongst, implement, and interpret findings from the use of different measurement procedures. In this article, the related topics of measurement and construct validity are summarized and discussed, with particular focus on formative and reflective indicators and common method bias, and, where relevant, a number of allied issues are considered. The perspective taken is an eclectic and holistic one and attempts to address conceptual and philosophical essentials, raise salient questions, and pose plausible solutions to critical measurement dilemmas occurring in the managerial, behavioral, and social sciences.

Keywords: Construct validity, common method bias, reflective indicators, formative indicators, measurement, structural equation models

Introduction

> …science is not common sense, and its most basic ideas and frames of reference require development through complex intellectual processes which involve not only interpretations of observation but also theoretical and partly philosophical conceptualization.
>
> Talcott Parsons (1968, p. 429)

I have two goals in this paper. The first is to discuss the meaning of formative and reflective indicators. The second is to interpret different senses of measurement error, especially as manifest in common method bias. These topics have received considerable attention recently in a number of thoughtful and important articles (e.g., Le et al. 2007; MacKenzie et al. 2005; Petter et al. 2007; Podsakoff et al. 2003a; Richardson et al. 2009; Sharma et al. 2009). Nevertheless the literature contains conflicting conclusions and recommendations, certain issues remain unexamined, and a need exists for considering the relationships among formative and reflective indicators, measurement error, and specification of scientific theories.

Before turning to these topics, I wish to make two disclosures. Any author proceeds from a number of recognized and hidden assumptions and world views which color one's perspective and approach to research. Like many researchers of my generation, I was indoctrinated and influenced by logical positivism. But as I looked deeper into the philosophy of science, I came to be shaped more by post-positivist outlooks. If I had to pick a label, something I normally try to avoid, I would say that metaphysical and epistemological realism best characterizes my thinking. However, there is more to my outlook than realism, as I have benefitted greatly from study

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

of approaches from social, contextual, neo-Kantian, and postmodern traditions. The personal tension that I have felt between realism and the latter approaches prevents me from embracing realism with the religious-like fervor of realists that I have known or read.[2] Another aspect of my thinking that I wish to disclose is that I attempt to avoid being pulled too far in any direction marked by strictly statistical or methodological standards, philosophical criteria, or substantive concerns when relating observational evidence to proposed theoretical frameworks. Rather I try to follow the spirit of Talcott Parson's assertion quoted above where creative tensions push and pull one in one direction or another and a (temporary) balance must be achieved. This inevitably means that efforts must be made to reconcile often seemingly incompatible and incommensurable policies and dictates at the boundaries of statistical, methodological, social, behavioral, and managerial disciplines and the practices in these disciplines. My orientation is both eclectic and holistic, which might create antagonism in the minds of scientists more circumscribed in perspective than I. But I hope not, for I firmly believe that a dialectic between persons of different points of view is essential for pursuing truth, and that any outcome reached as a consequence is relatively temporary in the scheme of things and likely not to be situated at one extreme or another shaping the debate (at least for long). For me, there are many paths to knowledge, some scientific, some not.

A Framework for Thinking about Theoretical, Empirical, and Spurious Meaning

Some first principles by way of premises undergirding the two main goals of this article (readers most interested in the practical aspects of formative and reflective measurement and in construct validity and common method bias could skip this section and jump to the second and third main sections of the article). A major aim of information systems and organizational research is to formulate theories and hypotheses and test these against observations or experimentation. A superordinate objective, often left unstated, is to uncover truths about the world of experience of information systems and organizations. By experience is meant events, happenings, actions, or behaviors of people, groups, institutions, collectivities, or systems, as well as outcomes influenced by these events, happenings, actions, or behaviors. Explanation, prediction, and understanding are guiding principles here, too, as well as control on occasion.

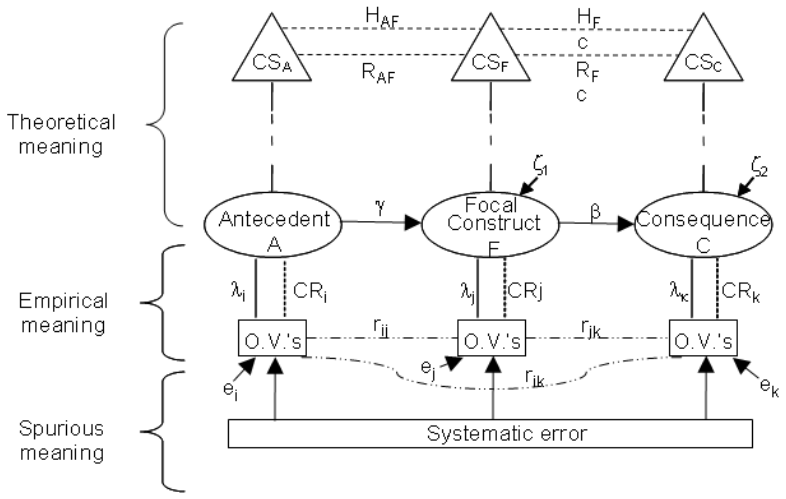
To accomplish such multifaceted aims, researchers draw upon and create conceptual and theoretical ideas, derive hypotheses, design appropriate methods to test hypotheses, measure variables designed to link in some way to the concepts, theories, and hypotheses, implement studies, and interpret findings in the light of the theories and hypotheses.[3] Putting this enterprise into practice is fraught invariably with errors at every point in the process. The stages in the research process typically have been segregated and accomplished in piecemeal ways, with researchers often working in teams, yet functioning as specialists and focusing on only subparts of the whole process. By applying certain standards and policies at each stage or in each piece of the research process in isolation from the other stages or pieces, yet in accordance with the received views therein, errors can be exacerbated and even artificially created because of inconsistencies, trade-offs, and other issues occurring across stages or subparts.

Criteria are needed to better integrate the various parts of the research enterprise and bridge the human, logical, and procedural gaps constituting it. A small beginning is to consider how meaning or sense-making exists and arises in doing research. Sense-making is neither a strictly empirical nor intellectual endeavor but requires the integration of conceptual, theoretical, methodological, and statistical matters of importance. Figure 1 presents a framework for thinking about facets of meaning expressed in a way underlying structural equation modeling or other approaches that might be shown to be special cases of the general structural equation modeling approach.[4] Sense-making entails scrutinizing the specification of theories, estimation of parameters, hypothesis testing, and interpretation of findings in a holistic way. To structure the sense-making process, it is useful to think of three interconnected senses of meaning: theoretical, empirical, and spurious. The framework sketched in Figure 1 and accompanying discussion to follow can be thought of as a premise upon which the treatment of formative and reflective indicators and construct validity and common method bias rest.

[2]A useful history and analysis of logical positivism and the on-going transition to post-positivist philosophies of science can be found in Suppe (1977). Brief commentaries on realism can be found in the following entries on the web from the Stanford Encyclopedia of Philosophy: Realism; Scientific Realism; Semantic Challenges to Realism.

[3]The view of science taken here best fits a positivist or realist, quantitative researcher, and is not meant to apply to or criticize research done in design science and qualitative research.

[4]Knapp (1978) showed that virtually all parametric statistics might be construed to be special cases of the canonical correlation model. Bagozzi et al. (1981) demonstrated that the canonical correlation model is a special case of structural equation models.



Key:
= theoretical or latent variable; = observed or manifest variables; = conceptual specification and/or theoretical definition; ⟶ = inferred (e.g., estimated) causal path; ——— = inferred (e.g., estimated) relation (i.e., factor loading) between latent and manifest variable; - - - - - - = correspondence rule; -------▶ = causal specification; — - - — = relationship between latent variable and its conceptual specification/theoretical definition; – ·· – ·· – = indicates which observed variables are involved in the correlations; CS = conceptual specification and/or theoretical definition; CR = correspondence rule for antecedent, focal construct, and consequence, r's = observed correlations among manifest variables; γ, β, λ's = parameter estimates; H = theoretical hypothesis; R = rationale behind theoretical hypothesis; ζ = theoretical error; o.v.'s = observed variables; e = random error.

Figure 1. A Framework for Thinking about Theoretical, Empirical, and Spurious Meaning

Theoretical Meaning

To introduce key ideas connected to Figure 1, imagine that one desires to consider the meaning of a focal latent construct, *F*. Consider first the *theoretical meaning* of *F*. Theoretical meaning resides in specification of the conceptualization of the focal construct and the theoretical relationships, if any, that the construct has to other constructs in a theoretical network. The conceptualization of a specific construct can be accomplished through a theoretical definition (see triangle CS_F in Figure 1) wherein a focal term referring to the concept that the focal construct is intended to represent is related to one or more other terms in a sentence(s) designating its content or essence. The sentence might specify theoretical attributes or characteristics of the concept, a structure or form in which the attributes relate to each other and to the concept, and/or dispositions (e.g., powers and liabilities) of the concept as a whole or of its attributes. The conceptualization can be evaluated in terms of such ideas as well-formedness, specificity, scope, ambiguity, vagueness, transcendence verus immanence, or other semantic and syntactic criteria. One claim then is that part of the nominal meaning of a construct is captured by the content of its theoretical definition.

In addition to the content of a theoretical definition of a focal construct, *F*, theoretical meaning resides in (1) the antecedents, determinants, or causes of *F*, (2) the consequences, implications, or results of *F*, and (3) the associative (i.e., nonfunctional, noncausal) links to *F* (the latter are not shown in Figure 1).[5] Whereas a theoretical definition specifies what a focal concept is and what it is capable of becoming or doing (through either abstract law-like relations or delineation of its powers and liabilities), its antecedents supply theoretical information as to where it has been (that is, its history and development) and/or how it is influenced or produced. The theoretical meaning here is provided by the content of the hypothesis (H_{AF}) linking antecedent *A* to *F*, and its rationale or reasoning (R_{AF}). A hypothesis might entail a relatively nonspecific statement such as "the greater the magnitude of *A*, the less the level of *F*," or it might contain a more specific statement as to the functional form of the relationship or even

[5]It might be pointed out that the focal concept may be of little interest in itself for some researchers and that the conceptualization becomes most useful when the focal construct is part of a nomological network or theory in which it is embedded and in which it is formally connected to other constructs.

the amount of change expected in F as a function of A. The rationale for the hypothesis is needed to complete the meaning of F provided by A. In general, a rationale for a hypothesis can be achieved through specification of the mechanism or process whereby A influences F, and is typically expressed through theoretical laws and an explication of how A produces change in F (e.g., a causal explanation). Hypotheses and rationales are often expressed contingently, explicating conditions under which, say, A affects F (or F influences C). Note too that any theoretical law will generally correspond, in part, to an empirical law (e.g., a hypothesis of regular succession).

In a parallel fashion, the theoretical meaning of F is also ascertained through its relationship to consequences. That is, the implications of F supply information about where a phenomenon is going, what it can lead to, and/or what influence it has. Again, the meaning here arises through delineation of the form and content of the hypothesis linking F to C (H_{FC}) and its rationale (R_{FC}). Again, certain associations between F and other theoretical variables can provide meaning concerning F (analogous to criterion-related validity, for example).

In sum, the theoretical meaning of a construct inheres in what it is and to what it relates conceptually. A construct standing alone is less rich in meaning than one that is explained by something else or one that also explains or predicts something else. To take an example, consider a construct intended to capture an emotion. The theoretical meaning of an emotion might be specified in terms of happenings experienced by a person and the primary appraisals he/she makes in this regard (e.g., Lazarus 1991). But a fuller theoretical meaning of the emotion would be provided by also considering secondary appraisals and coping responses to the experienced emotion. Indeed, some theorists even go further and propose that integral parts of the meaning of an emotion are the action tendencies seemingly following the experience of the emotion (Frijda 1986). For example, anger, sadness, and fear are often connected intimately to such action tendencies as striking out, seeking comfort, or running away, respectively. An emotion then is a complex representation of the occurrence of events happening to a person, primary appraisals thereof, plus action tendencies peculiar to the emotion and its experience. Secondary appraisals and coping responses are still further consequences of emotions (Lazarus 1991).

It should be pointed out in Figure 1 that the brackets to the left of the figure for theoretical meaning and empirical meaning overlap at the level of the latent variables to suggest that both capture aspects of meaning of the latent variables. More specifically, the nature of any research construct is both conceptual and empirical. In one sense, observed variables are "mapped into" theoretical constructs (or vice versa in some philosophical traditions), and this implies that the essence of construct validity inheres, in part, in the validity of this operationalization. Here issues need to be considered concerning the relationship of a theoretical concept and its attributes, either with the attribute conceived as a component or an instantiation of the concept, and the relationship implying that the attribute can be conceived as ontologically dependent on its concept or vice versa.

What is the relationship of theoretical meaning to formative and reflective approaches to measurement? Here, unlike under empirical and spurious meaning discussed below, it can be seen that the two approaches are similar in terms of theoretical meaning. Most researchers would agree that whether formative or reflective measurement is employed, it is important to provide strong conceptual specifications of the constructs for which the indicators are proposed to measure. Thus, well-formed theoretical definitions are required for constructs, whether one uses formative or reflective indicators. Second, in models where formative or reflective indicators are employed, the theoretical meaning of constructs resides, in part, through connections any focal construct has to other constructs. The specifications of these connections—whether causal, functional, predictive, or associative—can be the same for models with formative and reflective indicators of constructs. That is, theoretical meaning of a focal construct accrues, in part, through specification of hypothesized relations of the focal construct to other constructs and the rationales for these relationships, and this is required for constructs measured with formative and reflective indicators. Criticisms of formative measurement that some researchers have made—to the effect that theoretical meaning changes as a function of the number of formative indicators, the number of constructs measured reflectively, the number of indicators of the reflective constructs, and the relationships between the respective formative construct and reflective constructs—seem misplaced or at least in need of further nuanced interpretation. The criticisms raised do not so much implicate theoretical meaning but rather arise because of certain indeterminancies that inhere in empirical meaning and spurious meaning for models with formative indicators, as discussed below. Yet even here, it is important to recognize that the differences between models with formative and reflective indicators rest on different premises or assumptions, and not on the different empirical implications, *per se*. That is, models using formative indicators and models using reflective indicators can both be meaningful, given their premises, despite yielding different empirical outcomes. More on this below.

Empirical Meaning

Empirical meaning refers to the observational content associated with theoretical constructs after spurious meaning, if any, has been removed. This is accomplished formally through correspondence rules which link theoretical constructs to observed variables. There are at least three kinds of correspondence rules: the operational definition, partial interpretation, and causal indicator models. The operational definition model can be written as $P(t) \equiv (E(t) \rightarrow (R(t))$, which in words reads, "*t* has theoretical property *P* by definition, if and only if, when t is subjected to operation (e.g., experimental test) *E*, it yields result *R*." This correspondence rule can be traced back at least as far as Bridgman (1927, p. 5) who said, "we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations." The implied lack of differentiation between a theoretical construct and its operationalization under the operational definition model means that every construct has one and only one measure at any point in time, wherein the construct and measure are equated in meaning, which not only makes it impossible to speak about internal consistency reliability and construct validity, but also leads to a proliferation of theories and findings with little coherence (because empirical tests with different measures imply that different theories are being tested). Research conducted exclusively with observed variables risks criticism on these grounds.[6] Some cases of formative measurement also rest on the sense of operationalism discussed above.

To overcome problems with such "definitional" operationalism, the logic of multiple operationalism has been promoted (Campbell 1969). The version of multiple operationalism advocated by Carnap (1956), for example, can be written as $E(t) \rightarrow (P(t) \equiv R(x))$, which in words reads, "If t is subject to operation (experimental test procedure) *E*, it will exhibit theoretical property *P*, if and only if it yields result *R*." This correspondence rule gives a partial and empirical interpretation of a theoretical construct because observational meaning is only specified under particular test (i.e., measurement) conditions. Although this correspondence rule permits multiple operationalizations, and hence allows consideration of internal consistent reliability and construct validity, it suffers from the limitation that theoretical constructs have no conceptual meaning independent from the procedures used to obtain observations, and changes in measurement procedures change the meaning of a theoretical construct (Petrie 1971; Suppe 1977, pp. 102-104). We might think of this correspondence rule as analogous to a kind of "supervenience" (Kim 1993) in the sense that a theoretical construct is said to depend on or be determined by an observational procedure. Some models using formative indicators follow, implicitly at least, a kind of partial interpretation logic. I elaborate on this issue below under "On the Meaning of Formative and Reflective Measurement."

A correspondence rule with more desirable properties for psychological, social, and management science constructs has been termed the causal indicator model (Keat and Urry 1975, p. 38), and can be expressed as $(P(t) \rightarrow (E(t) \rightarrow R(t))$, which reads, "If *t* has theoretical property *P*, then if operation (experimental test procedure) *E* is applied, it will yield result *R*." Actually, calling this a causal indicator model is a misnomer. Causality is generally regarded to occur between two observable events or to be described as an inferred law-like relationship between two observable events described abstractly. The "causal" indicator correspondence rule, and indeed the two others described above, relate a theoretical construct to an observation(s). This correspondence rule functions as a scientific law of sorts linking theoretical construct to operational procedure to observation(s) (Schaffner 1969; Sellars 1961).

Notice that this correspondence rule is not part of either the theoretical meaning of a focal construct or the observations, *per se*. Rather it is an auxiliary hypothesis concerning theoretical mechanisms, empirical criteria, and a rule connecting the mechanisms and criteria. This point of view has some affinity with Suppes' (1962) "hierarchical theory" model, where the connection between theoretical construct and observed variable entails a physical theory (e.g., of instrumentation), a theory of operations or experimentation, a theory of data, and *ceteris paribus* conditions. Yet some surplus meaning is allowed for a theoretical construct that cannot be captured fully by observed variables. Notice that the causal indicator correspondence rule is a complex conceptualization consisting of a logical expression, some theoretical meaning, and some empirical meaning. Frequent reference to causality in the literature, when reflective and formative measurement have been discussed, including my own writings at times, have misleadingly characterized measurement in causal terms in the sense of declaring that a latent variable either causes or is caused by an observed variable.

Causality does come into play when specific operations are performed (e.g., an experimental manipulation) and responses are observed or recorded (e.g., in self-report manipulation

[6]This conclusion is implied by the epistemology of realism mentioned earlier. Research limited to observable variables makes it difficult to defend the general existence of concepts to which the observables are presumed to measure, and it makes consideration of theoretical meaning, empirical meaning, and spurious meaning difficult to consider and discriminate. For philosophical discussions touching upon these issues, see Bhaskar (1997).

checks or dependent variable measures), but it is important to recognize that the relationship between a latent variable and an observed variable is not strictly speaking a causal one. Causality occurs in part of the meaning of a correspondence rule, but a correspondence rule has additional logical, theoretical, and empirical meaning; moreover, a factor loading is an inferred parameter derived from empirical associations among observed variables, and therefore constitutes limited empirical meaning (i.e., it reflects only part of empirical meaning, which itself is only part of the meaning of a correspondence rule). With this as background, one can appreciate that a factor loading (λ in Figure 1), as an inferred parameter from associations among observed variables, is distinct from a correspondence rule. Even with no error in estimation, a factor loading, while capturing much of the empirical meaning entailed by a theoretical construct, still fails to represent the full meaning of the construct, which is also contained in the correspondence rule, conceptual specification of the construct, and the theoretical relation(s) of the construct to other constructs, if any.

Nevertheless, when empirical meaning changes (for example, when a purported unidimensional construct is multidimensional or fails tests of unidimensionality, when discriminant validity is lacking, or when systematic biases exist for measures of different constructs), the model that we think we are dealing with (i.e., a specific elaboration of the model in Figure 1) no longer applies, and a lacunae exists between theoretical and empirical meaning. Notice further that adequate empirical meaning depends, in part, on the proper choice of operational procedures and observed variables. For example, such conceptual criteria as logical deducibility of observations from the conceptual definition of a theoretical construct, and consistency and comparable levels of abstraction among multiple measures of a construct, should be met, which are standards going beyond the meaning of factor loadings derived in an empirical analysis (see Bagozzi and Edwards 1998, pp. 79-82).

To the extent that such conceptual criteria are poorly met, we would expect factor loadings to be adversely affected. However, other things affect factor loadings as well, such as spurious meaning. What is the relationship of empirical meaning to formative and reflective approaches to measurement? Here it can be seen that the approaches differ fundamentally.

Under reflective measurement, where indicators are functions of a hypothesized factor and error terms, empirical meaning can be said to be *local* in the sense that the inferred parameters linking each indicator to the construct are in principle particular to the nature of the relationships amongst all indicators of the construct alone, and the residual for each indicator reflects error. Such measurement models can stand on their own so to speak, and the factor loadings and error variances are in general not dependent on indicators of other constructs and the relationship between the focal construct and the other constructs, if the model in which the constructs are embedded is specified correctly and common method bias or systematic error does not occur.

Under formative measurement, by contrast, indicators have no error directly associated with them, and in the most interesting cases (see the next major section in the article), estimates of loadings require that the focal construct be linked to reflective indicators or other constructs that have reflective indicators.

The loadings of the formative indicators on the focal construct depend on information contained in the constructs and indicators of constructs to which the focal construct is connected. In this sense, empirical meaning is *global*. That is, empirical meaning and the estimates of formative loadings are in a sense spread out across the model. Holding constant the number of formative indicators, the estimates of loadings on the focal construct can change if the number of reflective constructs and the number of indicators of the reflective constructs change. Of course, as the number of formative indicators change (e.g., if one or more indicators are deleted or added to a particular specification), the formative loadings can change too. Adding or deleting proper indicators to a reflectively defined construct will not result in significant changes in loadings. As a consequence, empirical meaning differs fundamentally between formative and reflective approaches to measurement. Notice that such differences in empirical meaning could yield differences in inferred linkages between constructs (γ and β in Figure 1) for models with formative indicators versus reflective indicators.

Finally, it should be mentioned that for formative measurement, the number of measures in one sense defines the construct, which is not true for reflective measurement. Thus, for instance, Bollen and Lennox (1991, p. 308) note that "omitting an indicator is omitting a part of the construct" under formative measurement.

Are the above mentioned differences in empirical meaning and inferred linkages between constructs necessarily bad? I would argue that the answer to this question depends on the ontology one entertains with regard to latent variables and with respect to latent variables and the relationships between latent variables. For reflective measurement, it is presumed that the phenomenon that the latent variable is intended to represent exists, and therefore indicators vary, in a sense, when the underlying phenomenon varies. But for formative

measurement, it is presumed that the phenomenon represented by the latent variable does not exist until the indicators are chosen to represent it, where the formative construct then can be said to summarize the indicators. Although the ontologies of the two approaches to measurement differ, either can be appropriate, depending on the researcher's purposes and the ontology one assumes.

Likewise, two different ontologies underlie models with constructs measured with reflective indicators versus models measured with formative indicators. The former presumes that measurement in the local sense mentioned above applies and that relationships between constructs are not dependent on within-construct empirical meaning but rather on across construct empirical meaning. The latter presumes that measurement in the global sense considered above applies and therefore that formative loadings and relationships between constructs measured formatively and constructs measured reflectively will depend on the nature of reflective indicators and their relationship to formative indicators. For technical details, see "On the Meaning of Formative and Reflective Measurement" below.

It can thus be seen that neither formative nor reflective measurement is inherently wrong or right. Each has a different ontology for its latent constructs and a different ontology for the larger models containing latent constructs linked according to a theory or hypotheses of interest, where in the latter case the inferred structural regression parameters potentially differ as a consequence of the estimation procedures applied to the data and models. The choice of ontologies, and hence models, depends on the tastes and needs of researchers and the phenomena under study, and entails a philosophical commitment. Given the assumptions that are associated with such a choice, which differ between the ontologies, formative and reflective measurement models and larger predictive and causal models based on them are both legitimate ways of doing research, as long as they are derived and specified consistently in relation to their respective assumptions. Of course, not only do the assumptions differ, which can be compared, contrasted, and debated, but the empirical findings implied by the different approaches can, and often do, differ and can be compared, contrasted, and debated. The empirical differences should nevertheless be consistent with their respective assumptions. From this perspective, the debates recently aired in the literature between formative and reflective measurement advocates have been too strident in my opinion and have not recognized the ontological premises of favored positions but instead have applied criteria and standards from one perspective to judge the other. The approaches are different and their tests of hypotheses can differ too, but any statement as to which one of the approaches, if any, is better, may have to wait for crucial experiments and differences in predictions of *new* phenomena not contained in explanations of *existing* phenomena common to the different approaches in any particular application.

Some researchers are willing to accept the global dependence of formative loadings on a particular specification, and on estimates related to other constructs and other parameters in a model, and live with the absence of internal consistency reliability and classic construct validity criteria and the indeterminacy of generalizability associated with the ontology of formative measurement. Other researchers who hold to an ontology of reflective measurement and its implications for theory and theory testing will be uncomfortable with the trade-offs. And vice versa perhaps. It is thus important to realize that the choice of formative or reflective measurement entails different ontological assumptions and cannot be resolved by embracing one approach and using it uncritically to criticize the other.

Spurious Meaning

Spurious meaning refers to contamination of empirical meaning and resides in one or more of three sources: random error, systematic error, and measure specificity. It is best to perform measurement procedures so as to eliminate or at least reduce spurious meaning, but when this is not possible to the extent desired, spurious meaning can be controlled for statistically, under certain conditions. For discussion and examples where all three types of spurious meaning have been modeled simultaneously, individually, or in pairs, see Bagozzi, Yi, and Phillips (1991, pp. 438-443; see also Bagozzi et al. 1999). Spurious meaning is especially a concern when method or systematic biases occur, which I discuss more fully below under "Construct Validity and Common Method Variance."

The framework in Figure 1 implies that the meaning of latent variables in structural equation models is complex and goes beyond that found in the mathematical representation of a model and empirical tests of it. Moreover, the three criteria of meaning are interconnected and necessary to consider for a full interpretation of any piece of research. Importantly, the framework and criteria sketched in this regard herein apply well to structural equation models that use formative and reflective indicators.

Another point to consider. Why is it important to consider theoretical meaning and differentiate it from other kinds of meaning? One reason is that it places emphasis on that which

is to be explained and ultimately measured and tested; in particular it puts focus on the content of a conceptualization and its theoretical integrity, and it does so in a way avoiding confounding with empirical issues and contamination in the measurement or testing process. The notion of interpretational confounding is a case in point (Burt 1976). It is important to realize that the parameters linking a focal construct to other constructs (γ and β in Figure 1) are not theoretical hypotheses and their rationales but are empirical manifestations or implications of (H_{AF}), (R_{AF}) and (H_{FC}), (R_{FC}). That is, they are derived from inferential statistics. In this sense, they are imperfect reflections of the theoretical meaning of the relationships that *F* has with other constructs in a theory (and of course they do not address the theoretical definition of a focal concept or of its antecedents and consequences, although they are linked to these through implication or deduction according to a proposed rationale) (Cook and Campbell 1979).[7] Indeed the meaning of *F* has theoretical content that goes beyond the inferred empirical relationships it has with other latent variables or the relationships it has with observed variables represented by factor loadings, which are also inferred from data. Discussions of interpretational confounding have focused on these linkages (i.e., γ and β and λ's) but have done so in a potentially misleading way. The claim that the meaning of a focal construct "can be as much (or more) a function of its relationship(s) to other constructs" (Howell et al. 2007b, p. 208) can be deceptive because under conditions when γ, β, or λ's change, as a consequence of relationships to other constructs, this happens under classically defined interpretational confounding because either poor convergent validity, poor discriminant validity, and/or systematic error (e.g., method effects) occur with measures employed to operationalize or test a theory. But when these outcomes occur, there is a disjunction between the proposed theoretical meaningfulness of either the focal construct, the antecedent and consequent constructs, or the hypotheses and rationales linking them as proposed in the original specification and the empirical content designed to measure or test the theory under consideration. In other words, the model we thought we had in mind (i.e., a specific operationalization of a model based on Figure 1, say) has changed in meaning due to spurious contamination. Interpretational confounding is in essence an instance of misspecification biases at the operational level, not a statement about the theoretical meaningfulness of a focal construct or its linkages to other constructs. But to understand more fully why and how this happens, we need to also consider empirical and spurious meaning for structural equation models. Hence our two main goals, which follow in the next major section of the article: discussions of (1) formative and reflective measurement and (2) construct validity and common method variance.

What is the relationship of spurious meaning to formative and reflective approaches to measurement? Spurious meaning undermines the interpretation of a proposed theoretical specification either because systematic bias prevents or (artificially) creates findings consistent with hypotheses. Ideally, researchers would like to detect and control for such biases. Traditional procedures for detecting and controlling for random and systematic error rely on internal consistency measures of reliability and classic ideas of construct validity. Although reflective approaches to measurement lend themselves to such procedures as Cronbach Alpha and multitrait–multimethod matrices, similar procedures do not exist for formative approaches to measurement at this time. For the case of reflective measurement, present technologies permit one to detect and take into account spurious meaning by use of at least five procedures (see the final major section of this article). For the case of formative measurement, no general procedures exist for detecting and taking into account spurious meaning, at present.

A final related issue of difference concerns establishment of generalizability. While models containing formatively measured constructs and models containing only reflectively measured constructs can both, in principle, be subjected to cross-validation, tests of generalizability are more indicator-dependent for the formative versus reflective case. That is, because formative loadings change as (1) the number of formative indicators changes and (2) the number of reflective indicators of other constructs and the relationships of these indicators to the formative indicators, in a model containing a formatively specified construct, change (e.g., Bagozzi 2007), models containing formative indicators are likely to lack generalizability when the two aforementioned changes in specification occur. With models limited entirely to reflectively measured constructs, adding or deleting items to the constructs in the models will not affect generalizability (as long as systematic error does not occur). Of course, because formatively measured constructs rest on an ontology where the meaning of such constructs and their relationship to other constructs depends on the property of "globalness" described

[7]Note that *A*, *F*, and *C*, which correspond to factors or latent variables in structural equation models, have what MacCorguodale and Meehl (1948) called surplus meaning. This meaning derives from the conceptual specification/theoretical definition of each latent variable and its explicit relationship(s) with other conceptual specification(s)/theoretical definition(s) of other latent variables connected to a latent variable under scrutiny. Surplus meaning does not refer to the estimate of error variance for ζ_1 or ζ_2 shown in Figure 1. The latter error captures unexplained variance in a latent variable, which is inferred from data, and in this sense is similar to inferred random or spurious meaning. Note finally that another kind of surplus meaning can be found in correspondence rules, as discussed under "Empirical Meaning."

above, researchers accepting this ontology would be comfortable with the changing measurement and structural parameters that occur in such instances. For this ontology, lack of generalizability in the classic sense is not a problem because empirical meaning and inferred structural parameters are presumed to be dependent on a specific model and its particular indicators. Empirical meaning and inferred structural parameters are not dependent on particular indicators for models operationalized only with reflective indicators, when these indicators achieve comparable empirical meaning and systematic error is absent.

On the Meaning of Formative and Reflective Measurement

Technical discussions of formative and reflective measurement go back to earlier treatments of structural equation models (e.g., Blalock 1964; Jöreskog 1969; Jöreskog and Goldberger 1975). Recent discourses elaborate on many conceptual, operational, and interpretive issues arising over the years (e.g., Bagozzi 2007; Bollen 2007; Diamantopoulos and Winklhofer 2001; Howell et al. 2007a, 2007b; MacKenzie et al. 2005; Petter et al. 2007). Readers of the literature are apt to come away with considerable confusion and uncertainty about the meaning and viability of different formulas of measurement, for opinions of authors span the spectrum from seemingly concluding that formative measurement should never be used and only reflective measurement is meaningful (e.g., Howell et al. 2007a, 2007b) to asserting that the nature of theoretical variables alone *could* dictate whether reflective or formative measurement should be used (e.g., Diamantopoulos and Siguaw 2006; Podsakoff et al. 2003b).[8] For example, Howell et al. (2007b, p. 216) "strongly suggest that when designing a study, researchers should attempt to measure their constructs reflectively" because the classic conceptualization of validity does not apply to formative measurement, whereas Diamantopoulos and Siguaw (2006, p. 265) allow that "constructs such as socio-economic status are typically conceived as combinations of education, income and occupation" and therefore should be represented formatively, and Podsakoff et al. (2003b, p. 650) stipulate that "some constructs (e.g., leadership performance/effectiveness, articulating a vision, charismatic leadership) are fundamentally formative in nature and should not be modeled reflectively."

Howell et al. (2007b) seem to proceed from the ontology of reflective measurement and use the criteria and meaning therein to criticize formative measurement. Another perspective is provided by Petter et al. (2007) who appear to recognize that formative measurement is based on a different ontology than reflective measurement and recommend that we keep each approach straight. Petter et al. further show that a significant number of articles have misspecified formative constructs by mistakenly taking them for reflective constructs (see also Diamantopoulos and Winklhofer 2001). My position is philosophically closer to that espoused by Petter et al. and Diamantopoulos and Winklhofer who recognize that a (philosophical) choice is required in measurement. I also believe that it is important to recognize the differences pointed out by Howell et al. concerning what I termed above local and global measurement consequences and the implications of the different approaches for structural parameters, and hence the confirmation or interpretation of theoretical meaningfulness in any test of theory. While recognizing that one's choice of ontology supports the use of either formative or reflective approaches to measurement, I wish to consider the implications of both approaches by examining a number of basic and generic cases of each.

Consider first the measurement model for reflective indicators of factors:

$$\mathbf{y} = \mathbf{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y}$ is a $p \times 1$ vector of p indicators (i.e., observed scores or measures), $\mathbf{\Lambda}_y$ is a $p \times k$ matrix of regression weights (i.e., factor loadings), $\boldsymbol{\eta}$ is a $k \times 1$ vector of latent variables (i.e., factors) underlying the p indicators, and $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of disturbances (i.e., error terms or uniquenesses).[9] The error term, $\boldsymbol{\varepsilon}$, is sometimes taken as random or pure measurement error, but it is important to realize that it might be written as

$$\boldsymbol{\varepsilon} = \mathbf{e} + \mathbf{s} + \mathbf{m}$$

where $\mathbf{e}$ is a random component, $\mathbf{s}$ is a component specific to each measure (hence, termed measure specificity), and $\mathbf{m}$ is a component specific to systematic error (e.g., method bias). Researchers often assume that $\mathbf{s}$ and $\mathbf{m}$ are small in comparison to $\mathbf{e}$ and therefore can be ignored, but some researchers are increasingly discovering that such assumptions are unwarranted and failure to take $\mathbf{s}$ and $\mathbf{m}$ into account may lead to Type I and Type II errors. Importantly, structural equation models within the context of multimethod research

[8] I wish to express my gratitude to the editor for pointing out errors in my initial presentation of formative and reflective measurement and for helping me realize the choice that must be made philosophically with respect to ontological considerations in measurement and modeling (Detmar Straub, personal communication, December 30, 2009).

[9] An equivalent specification for the reflective measurement model is $\mathbf{x} = \mathbf{\Lambda}_x\xi + \delta$, where the terms have analogous meaning and dimensionality as presented for equation (1).

Figure 2. Reflective and Formative Measurement: Single Factor Case

designs can be used to represent different sources of variance and scrutinize specific and method biases, as well as trait variance and random error.

The left-hand panel of Figure 2 shows a graphical representation of a reflective measurement model for the case where four measures indicate a single factor. Under unrestricted exploratory factor analysis with maximum likelihood estimation, it is possible to test whether k factors account for a set of measures. Confirmatory factor analysis, where restrictions are placed on $\mathbf{\Lambda}_y$ (and on the variance–covariance matrix of factors, $\boldsymbol{\psi}$, and the variance–covariance matrix of disturbances, $\boldsymbol{\theta}_e$, if warranted), goes farther than exploratory factor analysis by permitting one to test the viability of different models, not merely the number of factors. For example, one might have reason to test for the significance of correlated factors, particular factor loadings, or specific correlated disturbances, as well as an overall model.

There really is not a measurement model, *per se*, for formative measurement, such as shown in the right-hand panel of Figure 2, and no test of the entire model can be done, as is possible for reflective measurement. Nevertheless a set of weights can be determined corresponding to $\gamma_1 - \gamma_4$. However, no measure error is designated for the formative measurement model. One way that weights can be ascertained is by use of principal components analysis where the measures, $x_1 - x_4$, are transformed into a component (or multiple orthogonal components, if applicable) that retains the original amount of variance in a data set under study (Chin 1995). Sometimes researchers add an error term to η under the formative model shown in Figure 2, but it is important to recognize that neither the error variance, factor loadings, nor the model as a whole can be tested in the way that the reflective model implied by equation (1) can (e.g., Bollen and Lennox 1991, p. 312).

To move formative measurement into a specification permitting testing of hypotheses and interpreting the meaning of formative "indicators" and a latent formative construct, it is necessary to add either two or more reflective measures to η in the model in the right-hand panel of Figure 2 or a latent variable(s) with its own reflective measures dependent on η. Before we discuss these and other possibilities, consider the formative model shown in Figure 3. Here we have added a reflective measure, y, to the formative model in Figure 2. This model (Case I) is estimable and testable. However, one should not interpret η as a formative construct and the x's as formative indicators of η. The model is, in fact, a multiple regression model with one dependent variable and four independent variables, as can be seen when the equation is written out, where we have made use of the facts that ε = 0 and λ =1:

$$\eta(=y) = \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \zeta \qquad (2)$$

Although the Case I model can be used to predict η, it is not actually a formative model, and one should not interpret η as a latent variable. Rather all variables are measured variables and are fully interpretable under classic conventions for multiple regression. With two or more y's, the Case I model would be a MIMIC model (see Case II below).

A somewhat analogous issue of interpretation can be seen when we add a single predicted latent variable, η_2, measured with a single indicator, to the reflective model in Figure 2. Figure 4 presents this model. There are two ways to interpret this model. If y_5 is highly and proportionally correlated with

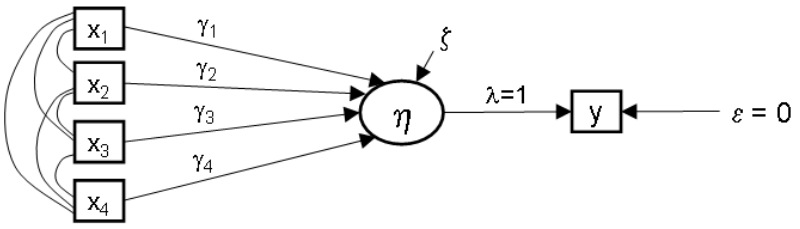


Figure 3. Formative Measurement and the Simplest Predictive Model (Case I)

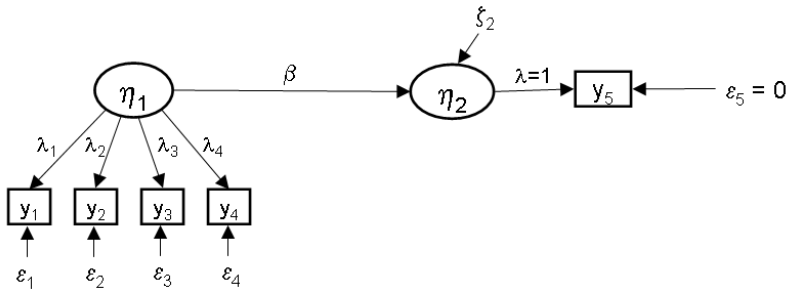


Figure 4. Reflective Measurement and the Simplest Predictive Model

$y_1 - y_4$, and $y_1 - y_5$ are conceptually and operationally similar (i.e., they have similar empirical meaning), then one interpretation is that one factor exists and $y_1 - y_5$ measure this factor (assuming the model fits well overall). Alternatively, if the model fits well overall, but (1) β is significant yet sufficiently different from $\lambda_1 - \lambda_4$, (2) $y_1 - y_4$ are highly correlated and comparably defined, and (3) y_5 is conceived differently than $y_1 - y_4$ and correlated at a different level with them and is subsumable under a different theoretical specification than $y_1 - y_4$ (i.e., η_1 and η_2 differ), then it may be possible to interpret β as a parameter relating η_1 to η_2. However, with but a single measure of η_2, there will normally be some ambiguity whether η_2 is different from η_1, or whether y_5 can be conceived as a poor measure of η_1. It would be better to have at least two measures of η_2 instead of one, if one desires to claim that η_1 influences η_2 and the measures of the two constructs (1) achieve convergent and discriminant validity and (2) support a predictive link between the two latent variables. A meaningful predictive model under reflective measurement that does not harbor the above mentioned ambiguities can be seen in Figure 5.

To recap up to this point, let us summarize what we have learned about reflective and formative measurement, for these principles put measurement into perspective, reveal differences, and suggest building blocks for bare-bones baseline models discussed below, while pointing to extensions of the baseline models: The simplest stand-alone reflective measurement model has one factor, is meaningful, and in the most interesting case has at least four measures; moreover, parameters are estimable and hypotheses can be tested. With two or more reflective factors, as few as two measures of each factor are required to avoid ambiguity, although three or more measures per factor would be better (because a tougher test of hypothesized factors is provided, the greater the number of measures per factor).

The simplest, pure formative measurement model assumes that measures have no error and is not estimable and testable in the way that reflective measurement models are. However, weights can be estimated that relate measures to component(s). But it should be mentioned that a gap exists between the pure formative measurement model and formative measurement where a formative construct predicts two or more measures or latent variables. This occurs because, for principal components analysis, each component typically shows high weights based on moderate to high sharing of variance for the measures corresponding to a component, whereas for formative measurement models, where the formative construct predicts other variables, which is the most com-

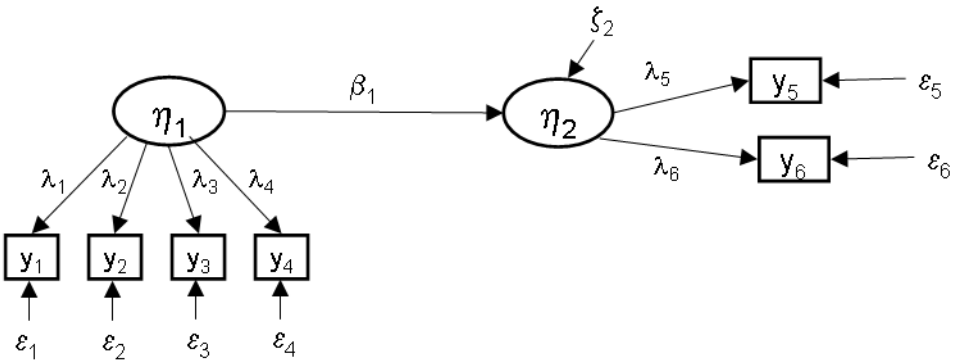


Figure 5. Reflective Measurement and a Meaningful Predictive Model

mon case, it is "not necessary for indicators to covary with each other" (Jarvis et al. 2003) and indeed researchers recommend that formative indicators be dropped when VIF $\geq$ 3.3 (e.g., Diamantopoulos and Siguaw 2006). *Thus a trade-off might exist between what is required for a meaningful formative measurement model under principal components analysis and what is required for a meaningful formative model when formative measurement is combined with prediction*.

Finally, with regard to prediction, a single predicted measure cannot be a basis for concluding that a formative construct exists (see Figure 3 and the discussion in text where it is claimed this model is really a multiple regression equation), nor can it be concluded unambiguously that such a predicted measure is explained, as an indicator of a second latent variable, by another reflective latent variable antecedent to it (see Figure 4 versus Figure 5 and the discussion in the text).

What, then, is the simplest, meaningful formative model? A candidate is the MIMIC (multiple indicator, multiple causal) model where two or more independent formative measures impinge upon a single latent variable and the latent variable influences two or more reflective measures. Figure 6 (Case II) presents an example for the particular case with four x's, three y's, and a single η.[10] This model is fully estimable and testable (see Bagozzi et al. 1981). But how should the Case II formative model be interpreted? Can the x's be construed as formative measures of η? Because MIMIC models are actually close versions of the canonical correlation model, a special interpretation is in order. I submit that η under a MIMIC model can be interpreted, figuratively, as an operation mathematically transforming or linking information in the x's to information contained in the y's. More formally, the MIMIC model finds a linear combination of the x's that maximally correlate with a linear combination of the y's (e.g., Anderson 1958; Stewart and Love 1968). Under this interpretation, we do not have a formative construct measured by formative measures. What we have is a model focused on prediction: prediction of y's by x's. The MIMIC model is valuable in forecasting the effects of a group of measures of independent variables on a group of measures of dependent variables.

For instance, researchers might be interested in predicting effort on the job (y_1), bad mouthing (y_2), and withdrawal intentions (y_3), as a function of satisfaction with pay (x_1), supervision (x_2), coworker relations (x_3), and opportunity for advancement (x_4). However, such a MIMIC model would not permit the interpretation of η as, say, latent job satisfaction. Rather, the latent variable η should be interpreted figuratively as a transformational mechanism between the x's and the y's, and might have additional practical utility, for instance, in forecasting y's for a new sample of employees who have scores on the x's.

[10]The Case II model differs from the Case I model, which is not a MIMIC model. The Case I model is a multiple regression model with one dependent variable indicator, and the operation entailed is not one of a linear transformation of a set of predictors into a set of predicted variables. As a MIMIC or canonical correlation model, the Case II model requires at least two predictors and two predicted variables for such a transformation as entailed by canonical correlation to take place; with only one independent variable and two or more dependent variable indicators, the model is a reflective model with one latent variable measured by two indicators and the latent variable is predicted by a single independent variable measured with one indicator. Only when we have two or more independent variable indicators and two or more dependent variable indicators in the Case II model will we have a transformation of a set of independent variable predictors into a set of dependent predicted variables.

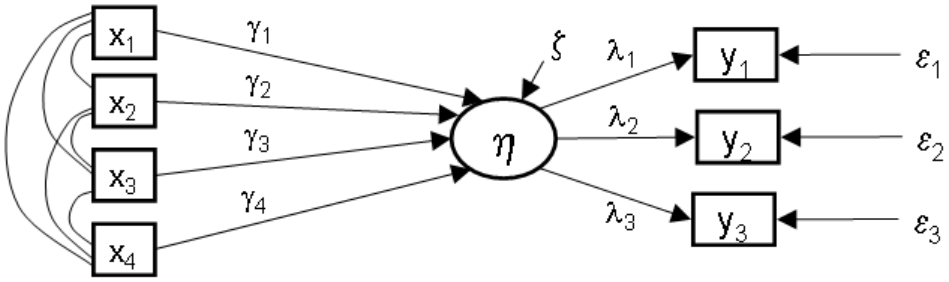


Figure 6. Formative Measurement and a Meaningful Model: The MIMIC Model (Case II)

The interpretation of latent formative constructs changes when these predict latent constructs that are measured by two or more reflective indicators. For example, returning to the MIMIC model described above, for measures of satisfaction predicting job outcomes, an interesting expansion consists for the case where job effort, bad-mouthing, and job withdrawal intentions are each latent variables ($\eta_2 - \eta_4$), with two or more reflective measures each, and are predicted by formative measures through an η. We will return to this distinct formative construct model when we consider Case IV below. But first an ambiguous special case.

The top panel of Figure 7 (Case III) presents a formative model with one formative latent variable, η_1, predicting one reflective latent variable, η_2. Ostensibly, researchers might intend this specification to represent the effects of a latent variable measured formatively by $x_1 - x_i$ on a latent variable measured reflectively by $y_1 - y_j$. To take an example, imagine that a researcher proposes that η_1 is job satisfaction, with measures $x_1 - x_4$ as described above, and η_2 is performance on the job with two measures, y_1 and y_2 (e.g., a supervisor rating and a self-rating). Can we interpret η_1 as job satisfaction and β its effect on performance? Not unambiguously. It turns out that the model in the top panel of Figure 7 is actually equivalent to the model in the bottom panel of the figure (see MacCallum and Browne 1993; Rindskopf 1984). That is, the former model is indistinguishable from the latter, which is a MIMIC model. This conclusion, in fact, generalizes: all formative constructs predicting a single latent variable that is measured reflectively with two or more indicators can be transformed into a MIMIC model. Substantively, therefore, we have two seemingly distinct models that are statistically equivalent. Hence the aforementioned ambiguity. To the extent that the rule of parsimony is valid, this would seem to suggest that the MIMIC model should be chosen, if one has to make a choice. But this means that we lose the interpretation of a formative latent construct influencing a reflective latent construct. Instead, we are left with the less conceptually precise predictive interpretation mentioned above for MIMIC models. The model in the top panel of Figure 7 provides more information than the MIMIC model conceptually but statistically is indistinguishable from it. There is yet another ambiguity with formative construct models that unfortunately applies in nearly any conceivable configuration going beyond the models described up to this point.

Figure 8 (Case IV) can be used to demonstrate the issue. Here we have three formative measures, $x_1 - x_3$, for a single formative construct, predicting three latent reflective constructs, where each is shown with two measures for simplicity.[11] For a similar model, but with only two latent reflective variables instead of three, Howell et al. (2007b) showed that the formative parameters ($\gamma_1 - \gamma_3$) are dependent on the association between the x's and y's and amongst the y's. For the model shown in Figure 8, I further showed that the γ's are dependent on the relationships between x's and y's, even holding the associations amongst the y's constant (Bagozzi 2007). This means that formative measurement parameters, which are claimed to relate x's to a latent variable and somehow measure it, can change (1) when measures are added to or subtracted from existing latent reflective variables that are dependent on the formative latent construct, or (2) when additional latent reflective variables and new measures are added as variables predicted by the formative construct, or (3) when one or more latent reflective variables that are dependent on the formative latent construct or measures of existing latent reflective constructs are removed.

Why is this an issue of concern? The dependence of γ's on relationships between x's and y's (and possibly amongst y's) means that formative constructs potentially have ambiguous meanings that shift from analysis to analysis on the same set

[11]The principles developed here apply also to cases with two or more formative measures predicting four or more latent reflective constructs. We will consider the case with two latent reflective constructs shortly.

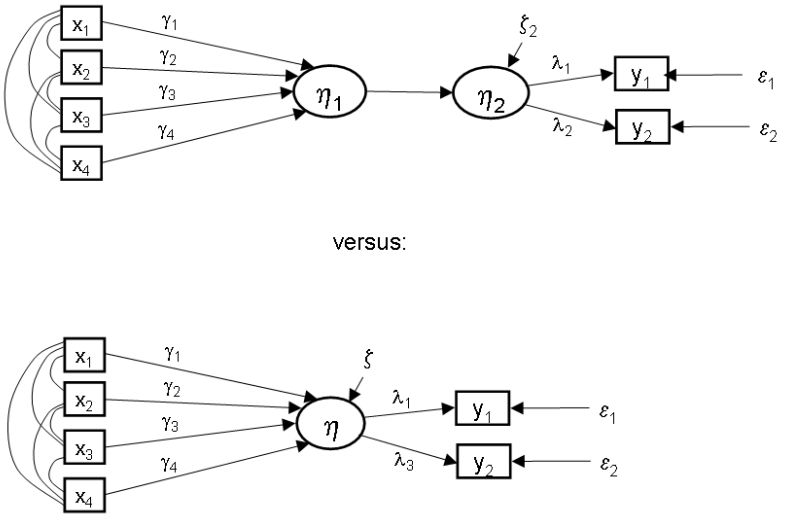


Figure 7. Two Equivalent Formative Models (Case III)

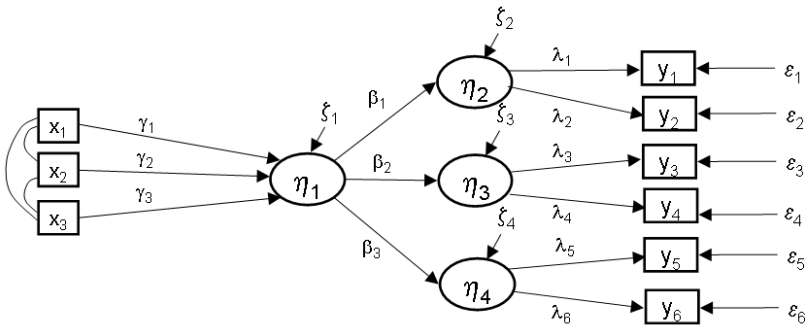


Figure 8. A Formative Measurement Predictive Model Exhibiting a General, Fundamental Ambiguity (Case IV)

of data when the inclusion of y's change, within a study, and from analysis to analysis across studies that propose similar theoretical variables and hypotheses and use similar or different measures. One implication is that the empirical meaning of formative constructs is unstable and the real possibility exists for a lack of comparability across studies as to the meaning of theoretical constructs and the generalizability of findings in this regard as noted above. The problem is analogous to classic drawbacks with definitional operationalism which stipulates that to understand the meaning of a concept an operational procedure is needed and every concept is nothing more than its specific operationalization. Every operation, then, implies a different concept. Compared to reflective measurement practices, formative measurement tends to proliferate concepts and findings of relationships between concepts when different formative indicators or different numbers of formative indicators and different reflectively measured latent variables are used across studies. Note, however, that this consequence is not a problem under the ontology presumed in formative measurement, and researchers following this approach maintain that the empirical meaning of concepts can be spread throughout a model, and the structural parameters derived therein are dependent on the particular specification and measures used, such that formative loadings and structural parameters may well change from context to context and study to study, depending on the formative indicators and reflective indicators used in a particular application.(e.g., Campbell 1969). Nevertheless, this property makes it difficult to make comparisons with and across data sets and to accumulate knowledge.

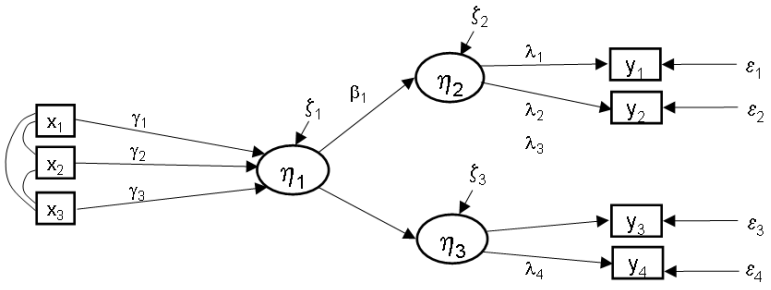


Figure 9. A Formative Measurement Predictive Model Exhibiting Additional Limitations (Case V)

A related implication is that it is difficult to know what it is that is being measured if purported measures of a proposed formative construct are dependent on the measures of hypothesized effects of the formative construct. It is also difficult in such cases to make strong claims about the distinctiveness of measures of formative constructs from measures of the reflective constructs it predicts. To the degree that the γ's are ambiguous and measures of formative and reflective constructs are confounded in the same model, it is problematic to make strong interpretations and claims about prediction and causality between formative and reflective constructs. This is especially the case when one desires to interpret the formative construct as a meaningful latent variable. Of course, if one is more concerned about the predictions in the entire model and accepts the ontological assumptions with formative indicators and their associated constructs, then these concerns are not an issue, given the assumptions.

A further limitation of formative measurement can be seen in Figure 9 (Case V). Here we have a special case of Case IV in that a single formative construct predicts two latent reflective constructs. The model has two potentially restrictive assumptions that are untestable: namely, ζ_2 and ζ_3 must be presumed uncorrelated and no path(s) can be estimated between η_2 and η_3. The first assumption would hold only if all variance in η_2 and η_3 were fully explained by η_1, except for random error, and no omitted variables explaining both η_2 and η_3 existed. This is unlikely to occur in many substantive tests of the model, and in any case cannot be ascertained because the hypothesis of uncorrelated errors is not testable.[12] The assumption of no causal paths between η_2 and η_3 for the model in Figure 9 also substantively limits its applicability and testability.[13]

Clearly the bottom line is that researchers contemplating the use of formative measures and constructs should make explicit their ontological assumptions and carefully assess the theoretical and empirical meaningfulness of any model in this regard. Meaningfulness cannot be established fully by scrutinizing the nature or conceptual meaning of a construct in isolation from empirical meaning or spurious meaning. To reiterate points from the presentation so far, formative measurement and formative constructs have a place in research, but it is crucial to recognize that their applicability is restricted to a few narrowly defined models, unless one is willing to make a commitment to the ontology behind the approach and forgo the ontology and implications of the use of reflective measures exclusively. The main conclusions are presented in Table 1.

Construct Validity and Common Method Variance

Common method variance has received detailed discussion in recent years in the information systems and organization research literatures (e.g., Le et al. 2007; Malhotra et al. 2006; Podsakoff et al. 2003a; Richardson et al. 2009; Sharma et al. 2009). Although some debate exists concerning how prevalent and significant common method variance is, with some claiming that such bias is often low (e.g., Malhotra et al. 2006; Spector 1987) and others concluding that the bias may be substantial (e.g., Doty and Glick 1998; Podsakoff et al. 2003a; Sharma et al. 2009; Williams et al. 1989), there is rea-

[12]For the Case IV model (Figure 8) and for models with four or more latent reflective constructs dependent on at least one formative construct, correlated errors amongst the reflective constructs can be estimated and tested.

[13]Causal paths amongst $\eta_2 - \eta_4$ for the model in Figure 8 can be estimated and tested, as can correlated errors for $\zeta_2 - \zeta_4$.

Table 1. Summary of Conclusions Comparing Formative and Reflective Measurement

1.	*The formative measurement model* (e.g., $\eta = \gamma_1 x_1 + \gamma_2 x_2 + \ldots \gamma_n x_n$) *assumes that no measurement error exists* (see right-hand panel of Figure 2) *and does not provide a model to test* of the sort provided for the analogous reflective measurement case where, further, error terms are included in the specification (see left-hand panel of Figure 2). Nevertheless, it is possible to derive weights for measures corresponding to formative components (e.g., by principal components analysis; Chin 1995). Sometimes researchers characterize the formative measurement model as follows: $\eta = \gamma_1 x_1 + \gamma_2 x_2 + \ldots \gamma_n x_n + \zeta$. But here *it is important to realize that this model is not identified and to achieve identification one must add either reflective indicators to η or one or more latent variables that η predicts* (see Figures 6–9). As Bollen and Lennox (1991, p. 312) stress for formatively measured latent constructs, Because the latent construct is a linear combination of its causes (and a disturbance), its validity, and indeed its psychological meanings cannot be judged from its item covariances. Without external criteria, a cause induced latent trait is psychologically uninterpretable. Also, the causal indicator model in isolation is statistically underidentified. Only when imbedded in a causal model that includes consequences of the latent construct can the causal indicator model be estimated.
2.	A seemingly formative model, wherein the formative construct predicts a single observed variable, is not really a formative model (see Figure 3). Indeed, the model is simply a multiple regression model. *The formative construct is illusory and should not be interpreted as a latent variable in this case*. Unlike under the MIMIC model (see below), the formative measurement model predicting a single indicator does not provide for a linear combination of a set of independent variables predicting a linear combination of a set of dependent variables.
3.	*A formative construct predicting two or more observed variables is called a MIMIC model and is estimable, testable, and meaningful* (see Figure 6). *However, the latent variable* is similar to a phantom or imaginary latent variable and *should not be interpreted substantively*. Rather it functions figuratively in a transformative sense similar to that found in canonical correlation analysis. That is, a linear combination of independent variables predicts a linear combination of dependent variables. Focus of the MIMIC model is more on prediction than explanation, *per se*, because no identifiable latent independent or dependent variable exists. Instead, one gets an optimum prediction (in the sense of maximum correlation) from independent variable indicators to dependent variable indicators; the indicators on either side of the transformation may arise from, or represent, one or multiple distinct constructs; it is not possible to interpret the empirical meaning of constructs in the MIMIC model.
4.	The case of a formative construct predicting a single latent variable measured reflectively (see Figure 7) *is ambiguous* because it is indistinguishable mathematically from a MIMIC model (e.g., MacCallum and Browne 1993, p. 538). Here two apparently different substantive interpretations cannot be adjudicated by findings. Thus, *it is unclear that the formative construct one might think is present is, in fact, valid and meaningful, and the predictive meaning implied by the MIMIC model may be the most justifiable interpretation here* (see point 3 above).
5.	In formative models, where two or more reflective latent variables are predicted by a formative construct (see Figures 8 and 9), a fundamental dependency exists between the relationships of x and y measures, which makes parameter estimates for γ's potentially unstable, as one adds or subtracts latent reflective endogenous variables and their measures or adds to or subtracts from measures of existing latent reflective variables (Bagozzi 2007). This makes interpretations of the meaning of formative measurement potentially indeterminate and generalizations across studies or even interpretations across analyses within a study potentially problematic, unless one is willing to commit to the ontology of formative measurement. Note also that adding to or subtracting from formative indicators of a construct will also in general change loadings of existing indicators or change the meaning of the formative construct (Bollen and Lennox 1991).
6.	Most of the above conclusions apply whether the formative construct occurs as an exogenous or endogenous variable embedded in a larger model. Moreover because the models shown in Figures 8 and 9 are often parts of larger models, formative measurement within such models will frequently be difficult to interpret in such cases. The interpretation of any such formative construct in terms of its relationship with formative indicators will likely be ambiguous. Mixing formatively and reflectively measured constructs in the same model entails accepting two different ontologies and should be justified by the researcher doing so. The two ontologies seem to be incommensurable because the interpretation of indicators and the constructs they are purported to measure, including the meaning of correspondences between indicators and latent variable, are based on different theoretical assumptions. Likewise, the meanings of inferred structural parameters between latent variables might also differ between the two ontologies because of the differences in meanings of constructs and the empirical differences in dependencies of structural parameters across models based on formative versus reflective measurement.
7.	There is a place for formative measurement in information systems and organization research. For example, the MIMIC model might be useful when prediction is a central concern. Also some theoretical variables might seemingly fit formative conceptualizations better than reflective ones (e.g., Diamantopoulos and Winklhofer 2001; Petter et al. 2007; Podsakoff et al. 2003b). However, the caveats mentioned in points 1–6 above and in the text should be kept in mind when considering formative constructs. Further, given our present technologies, when issues of internal consistency reliability, construct validity, and generalizability are of specific interest, it is best to consider reflective measurement whenever feasible. Also, common method biases are more straightforwardly handled by reflective measures than by formative measurement.

son to believe that the bases for making assertions as to the magnitude and prevalence of common method biases may be, in part, due to different procedures used to assess such bias (e.g., Bagozzi and Yi 1990; Bagozzi et al. 1991). In any case, editors are calling increasingly for researchers to ascertain the validity of their findings: "authors need at a minimum to address potential threats to validity occasioned by common methods…[so] method issues…cannot be ignored" (Ashkanasy 2008, quoted in Richardson et al. 2009, p. 36).

It can be argued that common method bias is but one contributor to variance in any measure and that a full accounting of measure variance requires representation of five sources (e.g., Bagozzi et al. 1999; Le et al. 2007): (1) an underlying concept, construct, or trait, (2) method bias, (3) measure specificity, (4) occasion specific effects when measurement is done over time, and (5) random error. Further, to ascertain construct validity in its fullest sense, one needs to carefully consider all five sources of variance. Construct validity is the extent or degree to which an operationalization measures a concept it is supposed to measure (e.g., Cook and Campbell 1979). The classic procedure for assessing construct validity examines convergent and discriminant validity by use of the multitrait–multimethod (MTMM) matrix, which is a correlation matrix consisting of two or more measures of two or more constructs obtained by two or more methods (Campbell and Fiske 1959). The MTMM matrix approach is a strongly empirical one that attempts to ascertain from the pattern and magnitude of correlations whether substantial trait variance and method variance exist. Because the variances of measures reflected in the correlation coefficients in a MTMM matrix are in a sense the resultants of the five effects mentioned above, it is difficult to make definitive conclusions about the presence and magnitude of any of the sources of variance from inspection of any MTMM matrix. As a consequence, a number of more formal statistical procedures have been developed to look at method bias and other aspects of construct validity.

Below, I discuss five general procedures for examining method variance and construct validity. The approaches are ordered roughly from least to most useful and comprehensive, but it should be acknowledged that all exhibit pros and cons. My aim is to briefly describe each procedure, point out key limitations and advantages, and bring some coherence to the topic, as the literature is fragmented and occasionally contradictory and misleading. Table 2 presents a summary of these five procedures and their pros and cons. A point to keep in mind when thinking about the five procedures is that no single one is applicable in all or even many cases; each rests on strong assumptions and each needs to be reconciled with the nature of the data at hand as well as the meaning of the model and statistical methods needed to implement it.

Unmeasured Latent Method Factor

Perhaps the easiest approach to apply is the addition of a factor to a test of a substantive model, wherein all measures in the substantive model are specified to load on the factor. The substantive model could be a confirmatory factor analysis (CFA) model or causal model, and the extra factor added to the substantive model has been purported in the literature to represent method variance. About 50 documented studies have employed this approach to date (Richardson et al. 2009, p. 9). Figure 10 provides an example.

Two procedures have been followed to implement the unmeasured latent method factor approach. One is to run a model without the method factor and compare this model to the one with the method factor added. If the introduction of the method factor fails to change substantive conclusions (e.g., γ_1, γ_2, and β are significant in both models in Figure 10), then it is concluded that the amount and extent of method variance do not pose a threat to the validity of tests of hypotheses.

The other way to implement the unmeasured latent method factor approach is where we take a CFA as an example (e.g., Williams et al. 1989). First a CFA is run with hypothesized trait factors (the trait-only model). Then a single factor model is run with all measures loading on it (the method-only model). Finally a CFA is run with the focal traits and the single factor added with all measures loading on the latter (the trait–method model). The trait-only model can be compared with a χ^2 difference test to a null model of modified independence (i.e., a model where only error variance is estimated) to ascertain the significance of trait variance; likewise the method-only model can be compared to the null model to determine the significance of method variance. The trait–method model can be compared with a χ^2 difference test to the trait-only model to evaluate the significance of method variance; and the trait–method model can be compared to the method-only model to appraise the significance of trait variance.

The advantage of the unmeasured latent method factor approach for detecting and correcting for method variance is its ease of implementation. It is not necessary to acquire additional data such as required by the other procedures described below.

But two problems with the unmeasured latent method factor approach should be mentioned. First, the basis for claiming that the added factor with all measures loading on it captures method bias has not been convincingly established. Can we interpret the factor and significant loadings on it as method variance? No clear answer to this question has been given. I submit that significant loadings on the added factor represent

Table 2. Summary of Five Procedures Used in the Assessment or Control of Common Method Bias and Their Pros and Cons

Procedure	Description	Pros	Cons
Unmeasured latent method factor (see Figure 10)	All indicators in a CFA or causal model are allowed to load on one common "method" factor, and this model is compared to the CFA or causal model without the method factor. Changes in factor loadings, correlated factors, paths between factors, and model fit are taken to reflect method bias.	Easy to implement. No measures or indicators other than those used in the focal CFA or causal model are needed.	Unclear whether significant loadings on the method factor actually represent or correct for method bias. Significant method factor loadings could reflect some (unknown) combination of measure specificity, method bias, and/or some other source of systematic error. Adding a method factor to a model that already fits satisfactorily could induce such consequences of over-fitting as improper solutions (e.g., out of range factor loadings, negative error variances); failures in the estimation program to converge; incongruous, counterintuitive, or inconsistent causal parameter estimates; and different signs of method factor loadings on the same or different factors or patterns of loadings where some are significant, others nonsignificant, and no convincing rationale can be provided for this.
(Correlated) marker variable approach	A variable and its measures are chosen as surrogates for method variance and then used to partial out method bias. One way that this has been done is to compare regression parameters for a model without taking into account the marker variable to regression parameters for a model where the method bias has been partialled out.	Relatively easy to implement once an appropriate marker variable has been found.	• Difficult to find measures of a marker variable that are unrelated theoretically to measures of the substantive variables already in the model. • For the case where variables have a single measure or indicator, measurement error might be high but unknown. • Not clear whether systematic variance is due to the marker or a combination of method, measure specificity, or other confounds. • If the marker is related theoretically to one or more substantive variables, the approach could remove substantive variance. • Assumes that common method biases have the same effect on all observed variables.
(Dedicated) marker variable approach	Measures of a hypothesized contaminator (e.g., social desirability) are modeled as indicators of a factor, and measures of the remaining variables in the focal model load on this factor.	Controls for explicit biases associated with the dedicated marker.	Method bias may not, and generally would not, be controlled for, beyond the dedicated marker. Method bias, measure specificity, and other systematic biases are not dealt with explicitly and may be confounded.
Method–Method Pair technique	Within the context of a meta-analysis, ANOVA is used to model variation in a correlation of interest across studies, as a function of within-study and between-studies variances.	Suggests the consequences of using different methods to measure variables in a theory.	• Unknown whether method bias identified is confounded with method–method pairs and error and what the extent of such confounding might be. • No explicit representation is provided of the type of error found in each method. • Relies on between-studies variation in methods to assess common method bias.
Confirmatory Factor Analysis Approaches:			
Additive trait–method–error model (see Figure 11)	Each measure is modeled as a function of a specific trait–method combination by use of a CFA in a multitrait–multimethod matrix design. The fullest and most interesting design for this and the other CFA approaches requires the use of as different and as similar methods as possible.	Provides a partitioning of measure variance into trait, method, and error components. Overcomes limitations of the Campbell-Fiske procedure, yet gives greater intuition and stronger statistical criteria to assess achievement of convergent and discriminant validity (as well as reliability).	• Measure specificity and random error are confounded. • Often yields ill-defined solutions: empirically under-identified models, failure of estimation program to converge, parameter estimates outside allowable ranges, excessively large standard errors. • When correlations among traits and/or methods are too high, trait and method component partitions may not yield trait-free or method-free interpretations.

Table 2. Summary of Five Procedures Used in the Assessment or Control of Common Method Bias and Their Pros and Cons (Continued)

Procedure	Description	Pros	Cons
Correlated uniqueness model (see Figure 12)	Under a CFA specification, each measure is modeled as a function of a specific trait–error combination and residuals of measures are correlated, corresponding to methods.	Overcomes certain limitations of the Campbell-Fiske procedure, yet gives greater intuition and stronger statistical criteria to assess achievement of convergent and discriminant validity (as well as reliability). Gives an estimation of trait and error variance. Seldom produces ill-defined solutions. Methods not assumed to be unidimensional.	• Methods and traits assumed independent of each other. • Measure specificity and random error confounded. • Interpretation of correlated uniquenesses can be difficult (see text). • Assumes methods uncorrelated. • Factor loadings may be underestimated to the degree measure specificity occurs. • Effects of specific methods difficult to interpret.
Direct product model	Hypothesizes that traits and methods statistically interact to produce variation in measures,while error is additive. The multiplicative effects occur such that sharing a method across traits exaggerates the correlations between highly correlated traits relative to traits that are relatively independent.	Useful when self-, peer-, and expert-ratings are used in a study. The stronger the true associations are between traits, the more likely they are to be noticed and exaggerated. Useful also when multiple occasions are methods. A high correlation between two traits will be more attenuated over time than will a low correlation.	• The assessment of convergent and discriminant validity is complex. • Convergent validity assessment is rather global and nonspecific. • Trait and method variation confounded.
Additive trait–method–error model with explicit measures of methods	Each measure is modeled as a function of a specific trait–method–error combination. Specific measures of methods are modeled.	Provides a partitioning of measure variance into trait, method, and error components. Overcomes certain limitations of the Campbell-Fiske procedure, yet gives greater intuition and statistical criteria to assess achievement of convergent and discriminant validity (as well as reliability). Gives the most precise interpretation of the meaning of methods of all procedures. Ill-defined solutions less common than with the additive trait–method–error model where no explicit measures of methods are provided.	• Measure specificity and random error are confounded. • May be difficult to identify source of method bias and obtain appropriate measures.
Direct product model with measurement occasion	Traits, methods, and measurement occasions interact statistically to produce variation in measures, while error is additive.	Models both differential augmentation and differential attenuation.	• The assessment of convergent and discriminant validity is complex. • Convergent validity assessment is rather global and nonspecific. • Trait, method, and occasion variation confounded.

Table 2. Summary of Five Procedures Used in the Assessment or Control of Common Method Bias and Their Pros and Cons (Continued)

Procedure	Description	Pros	Cons
Correlated trait–correlated method minus one model (see Figure 13)	Indicators are modeled as functions of traits, methods, and error, except for one method which is omitted to achieve a comparison standard. Works for meaningful methods that are structurally different (as opposed to interchangeable methods).	Measurement error separated from true trait and method effects. Method effects modeled as trait specific rather than assumed uniform across traits. Observed variance partitioned into trait-specific, method-specific, and error components. Trait factors are the true-score variables of the comparison standard, and method-specific factors are functions of the residuals. Method factors specific to a trait can be allowed correlated with trait factors for the other traits, if desired.	Multiple indicators are needed for each trait–method unit to specify trait-specific method effects which may be difficult to obtain in many information systems and organization contexts. May be difficult to specify a meaningful method as the comparison standard.
Multilevel confirmatory factor analysis model (see Figure 14)	A special case of the random effects ANOVA, this model represents deviations in a target trait from the mean across all traits, where multiple trait factors occur for each trait–method unit. The model has one method factor per trait, where method effects reveal the deviation of the error-free or true-score of raters from the trait value on target traits.	Method-specific sources of variance are separated from error-specific sources. Trait, method, and error variance can be partitioned. Useful formulas exist for computing reliability, consistency, and method specificity coefficients.	Method factors corresponding to common trait–method units are assumed to be unidimensional. The fullest implementations require multiple trait–method unit indicators, which may be difficult to obtain in many organizational and information systems contexts.

unknown systematic variance. This variance could be a combination of measure specificity, method bias, and/or some other unknown source of systematic bias not related to the method of data collection (e.g., social desirability). With only a single method employed in the typical study, it is impossible to separate method variance from other sources of systematic bias and from true-score variance. The problem is somewhat analogous to the issue of reliability when only a single item is available. Internal consistence reliability requires multiple items. To validly ascertain the source and amount of method variance may well require multiple methods.

A second problem with the unmeasured latent method factor approach stems from possible ambiguous or invalid empirical outcomes with its application. Researchers sometimes find that a causal model or the trait-only model fits the data satisfactorily, yet they then go ahead and add the method factor to the model. Whenever an additional factor is added to a good fitting model, there is always the possibility of over-fitting the model to the data. A common consequence of over-fitting is the occurrence of improper solutions (e.g., out-of-range factor loadings, negative error variances). In addition, failures of the estimation program to converge may occur. Still further, incongruous, counterintuitive, or inconsistent parameter estimates for causal paths can happen. For example, some method factor loadings might be negative and significant, others positive and significant. Finally, many loadings on a method factor may be nonsignificant. How are we to interpret differential effects of a purported single method factor when in fact all measures were obtained by use of a common method? How are we to interpret items loading on one trait factor that show differential significance or different signs for loadings on the method factor?

In sum, despite its ease of use, the unmeasured latent method factor approach has serious problems in both conceptual and operational senses. It is difficult to be sanguine about its use, except perhaps in a loose, suggestive sense as revealing the possible presence of systematic error.

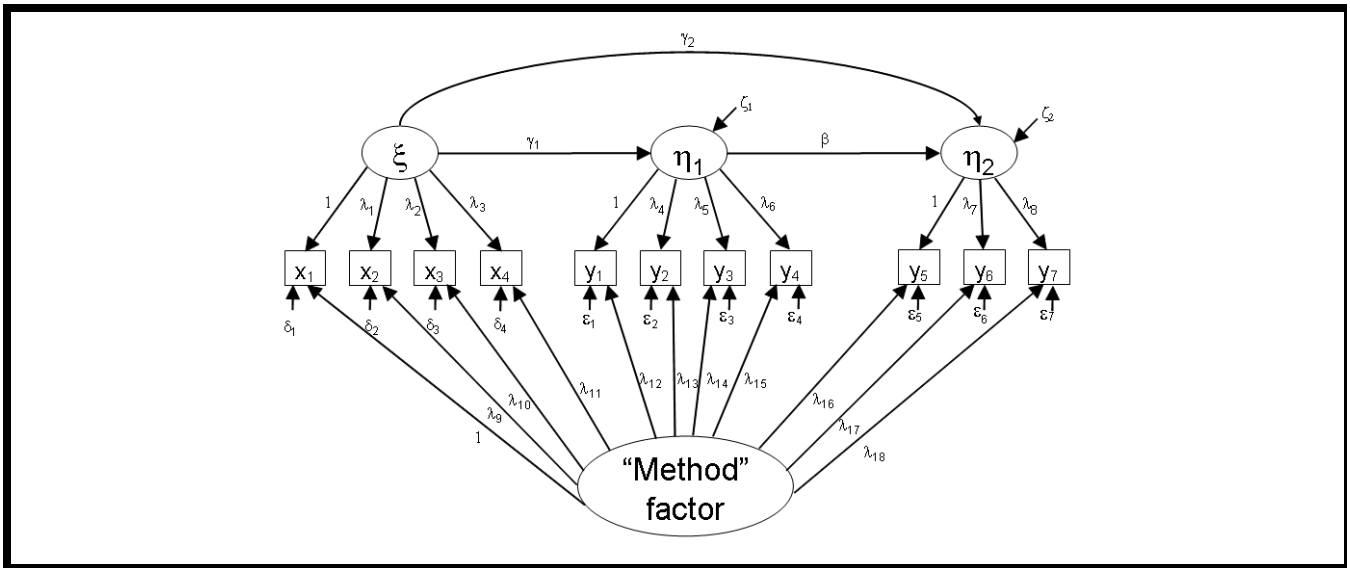


Figure 10. Unmeasured Latent Method Factor Approach to Common Method Variance

The Marker Variable Approach

The marker variable approach attempts to find a variable that might be a surrogate for method variance and then to partial out the shared variance between measures of variables in a model and the measure(s) of the marker variable (e.g., Lindel and Whitney 2001). Common method bias is estimated as a function of the smallest positive correlation between a measure in the substantive model and the marker. Ideally, a marker is chosen and designed into a study *a priori* by selecting a measure(s) of a variable that is expected, theoretically, to be unrelated to measures of the substantive variables already included in the study. Alternatively, if one is unable to provide an ideal marker, it may be possible *ex post facto* to choose the "smallest correlation among the manifest variables…[as] a reasonable proxy" for common method bias (Lindell and Whitney 2001, p. 115), although it has been argued that a more conservative approach is to select the second smallest positive correlation amongst manifest variables as a conservative estimate of common method bias.

Under the correlational marker approach, the following equation is used to remove shared variance between a measure of the marker variable and measures of other variables in a study:

$$r_{yi.m} = (r_{yi} - r_s)/(1 - r_s)$$

where $r_{yi.m}$ is the partial correlation between a measure of the marker and another measure purportedly controlling for common method bias, r_{yi} is the observed correlation between the marker measure and measure i, and r_s is the smallest (or second-smallest) correlation found between the measure of the marker variable and a measure of one of the substantive variables. This formula can be used to compute a new matrix of correlations corrected for common method bias, and the adjusted correlation matrix can be used in multiple regression, path analysis, or causal modes to test hypotheses (e.g., Malhotra et al. 2006; Podsakoff et al. 2003a; Richardson et al. 2009). A comparison of regression parameters with and without the correction for the marker provides an indication of the effects of common method bias on hypothesized relationships.

A number of limitations with the correlational marker approach should be mentioned. Especially when all variables are measured with single items, measurement error may not only be high but is not taken into account in the procedure. Likewise, it remains unknown whether all of the remaining systematic variance can be attributed to hypothesized variables or whether it contains some proportion of systematic error due to method, measure specificity, or other confounds. Further, to the extent that a marker is related theoretically to substantive variables, the correlational marker approach might remove some substantive variance (e.g., Richardson et al. 2009, pp. 6-7). Finally, the approach "assumes that common method biases have the same effect on all observed variables" (Podsakoff et al. 2003a, p. 890), which may not be realistic.

A creative extension of sorts to the correlational marker procedure has been termed "controlling for the effects of a directly measured latent method factor" (Podsakoff et al. 2003a, p. 893) or the "congeneric common method variance

model" (Richardson et al. 2009, p. 5). Here measures are obtained for a hypothesized contaminator, and the contaminator is specified as a factor with all measures of the substantive variables loading on it. For example, if one believes that social desirability systematically influences people's responses, over and above the presumed true meaning of measures of substantive variables, items from a social desirability scale could be included in a study, and a social desirability factor created to test for and partial out the effects of social desirability.

The dedicated marker variable approach, as the above extension might be called, may be an effective way to test for explicit biases. Findings with and without the dedicated marker factor can be compared to ascertain such biases. However, it should be emphasized that such an approach focuses on the bias specific to whatever contamination one suspects and measures. The dedicated marker approach does not, in general, control for common method bias, except perhaps to the extent that the hypothesized factor and measures actually measure bias associated with the method itself. This would normally be difficult to accomplish. For instance, social desirability biases may be distinct forms of biases unique to the method(s) used in a study. The dedicated marker variable approach assumes that method bias, measure specificity, and other systematic bias (e.g., that associated with time when multiple occasions are investigated in a study) are small in comparison to random error, and in any case, these are not tested for under the approach. In sum, although the dedicated marker procedure permits the test of particular biases, it does not address common method and other systematic biases, if any, and may even be contaminated with these.

The Method–Method Pair Technique

The approaches to construct validity discussed above and the ones mentioned in the sections following the present one can be applied in principle to a single study and a single sample. The method–method pair technique is used in conjunction with a meta-analysis to explain variance in observed between-studies correlations of measures of substantive variables (e.g., Sharma et al. 2009). For example, the following random effects ANOVA model can be used in this regard:

$$\mathrm{Var}(r_i) = \mathrm{Var}\,(u_i + e_i) = \sigma + \Gamma^2$$

where r_i = observed correlation reported in study i, u_i = effect of between-studies differences on the correlation coefficient of study i, e_i = within-studies error, and σ^2 and Γ^2 are the within-study variance (sampling error) and between-studies variance, respectively.

In an innovative study, Sharma et al. (2009) examined the effects of common method variance on the correlation between perceived usefulness and usage from the technology acceptance model literature (Davis et al. 1989). Common method variance (u_i) was operationalized by method–method pairs across five categories: system-captured (e.g., from historical records or archives), behavioral continuous (e.g., as recorded on open-ended scales), mixed behavioral continuous and behaviorally anchored, behaviorally anchored (e.g., as recorded on close-ended scales), and perceptually anchored (e.g., agree–disagree scales). The aforementioned ordering was hypothesized to reflect common method variance going from very low to very high. All measures of perceived usefulness employed perceptually anchored scales, whereas measures of usage utilized across data sets came from all five categories. Using information from 75 data sets, Sharma et al. found that 56.09 percent of the variance in the correlations between perceived usefulness and usage could be attributed to method variance, with 36.28 percent due to error and the rest partitioned amongst control variables. These findings point to a considerable amount of method bias.

The main advantage of the method–method pair technique is that it can be used to suggest the consequences of employing different methods in the measurement of variables in a theory to test. Of course, conclusions drawn from such a study have to be taken as a matter of faith and incorporated into a subsequent research study someone conducts; the method–method pair technique does not address construct validity in any specific study and cannot be implemented as such.

The primary limitations of the method–method pair technique are the following. Because the explained variable is an observed correlation, and therefore is differentially affected potentially by up to five sources of variation, it is unclear that the effect of method–method pairs can be attributed entirely to method bias. Indeed, it is possible that trait and method variance are confounded. Likewise, error and method bias may be confounded. At least it seems to be unknown whether method bias of different sorts might be confounded with method–method pairs and error and what the extent of such confounding might be. Another problem is the nonspecifity of the nature of error entailed in any method–method pair ranking. No explicit representation is provided of the type of error found in each method. Moreover, in some research contexts system-captured bias might be of an entirely different sort and greater than behaviorally and perceptually anchored biases; yet there is no way to ascertain this with the proposed method–method pair ranking. In fact, the rank ordering of method–method pairs assumes that this variable has no error of its own. Perhaps future applications of the method–method pair technique could be performed on correlations corrected for measurement error or both measurement

and method error. Sharma et al. (2009, pp. 485-486) discuss additional limitations of the method–method pair technique.

The CFA Approach Applied to MTMM Matrix Data

To more definitively ascertain method variance, it is necessary to formally consider multiple methods and multiple traits in an integrated way. A number of models have been proposed to accomplish this, and we will briefly consider seven here.

The approach that seems to be close in spirit to the MTMM matrix perspective proposed by Campbell and Fiske (1959) is the *additive trait–method–error model* shown in Figure 1 for the case of three traits and three methods. Here the variance in each measure, x_{mt}, is partitioned into three parts: that due to method, trait, and error. For example, $x_1 = \lambda_{14}\xi_{m1} + \lambda_{11}\xi_{t1} + \delta_1$, where ξ_{m1} is method factor 1, ξ_{t1} is trait factor 1, λ_{14} is the factor loading relating x_1 to ξ_{m1}, λ_{11} is the factor loading relating x_1 to ξ_{t1}, and δ_1 is a disturbance. Convergent validity is achieved when the overall model fits satisfactorily and factor loadings are significant and high in value. Ideally, standardized trait factor loadings of about .7 or greater are desired. The logic seems to be that one wants at least 50 percent of variance in a measure to be attributable to a trait. This may not be feasible in many practical applications, and a more realistic minimum might be .6 or greater for trait factor loadings, as this still demonstrates a strong relationship between trait factor and measure. Certainly, as trait loadings fall below .5, however, they point to rather low trait variance. Discriminant validity is attained when correlations amongst traits are significantly less than 1.00 and can be tested by inspection of the confidence interval for correlations, or better yet by chi-square difference tests, where chi-squares for a model with and without a φ_{ij} constrained to 1.00 are compared with one degree of freedom. Finally, the trait–method–error model yields a convenient partitioning of measure variance into trait, method, and error components, which is a useful diagnostic not found in many approaches to construct validity.

An important design issue should be mentioned. Campbell and Fiske originally asserted that the MTMM matrix should be formed so as to employ maximally different methods, and researchers using CFA to analyze construct validity often echo this recommendation. The rationale appears to be the belief that, to the extent that two or more very different methods agree, and construct validity is achieved, we should come away with the greatest assurance that this indeed is the case. But I think that the strongest evidence for construct validity will be accomplished when a set of maximally similar *and* a set of maximally dissimilar methods are employed. Why? Well, it should be more difficult to demonstrate discriminant validity when similar methods are used, and likewise more difficult to settle convergent validity when different methods are applied. Using only similar methods makes it too easy perhaps to achieve convergent validity; but using only different methods makes it too easy to attain discriminant validity. Hence the recommendation to use as different and as similar methods as possible.

In sum, the trait–method–error model elegantly operationalizes the intent of appraising construct validity originally proposed under the MTMM matrix procedure and indeed goes beyond this procedure. It is intuitive and easy to apply. It permits methods to correlate freely and allows for differential effects of methods on measures (the MTMM matrix procedure assumes that methods are uncorrelated and methods influence all traits equally). It provides for a statistical test of the model, as well as parameter estimates of key aspects of construct validity (the MTMM matrix procedure does not provide these). It permits the computation of reliability (the MTMM approach assumes measures are equally reliable). Finally, it yields estimates of the proportion of variance due to trait, method, and error.

The trait–method–error model has shortcomings too. Measure specificity and random error are confounded (as they are in most approaches, except in two mentioned below). In practice, the model often yields ill-defined solutions: empirically under-identified models, failure of the estimation program to converge, parameter estimates outside allowable ranges, and excessively large standard errors (Marsh 1989). Of course, such outcomes occur when models are over-fitted to data or when a model is fitted to data not appropriate to the data at hand. Nevertheless, Marsh and Bailey (1991) report that about 75 percent of their attempts to run the trait–method–error model yielded ill-defined solutions in their simulations and analyses covering 435 MTMM matrices. What is one to do when ill-defined solutions occur? In rare cases, it may be possible to provide one's own starting values when failures to converge occur; also, it may be appropriate to fix a negative error variance to zero and rerun the model (when the negative variance is nonsignificant), but this requires some skill and judgment to achieve meaningful results (see Bagozzi 1993). More often than not, ill-defined solutions suggest model misspecification, and the best course of action is to try another model, such as those discussed below. Finally, when correlations amongst traits and/or methods are too high, the trait–method–error model may not yield partitions into trait and method components with "trait-free" and "method-free" interpretations (e.g., Kumar and Dillon 1992), and trait and method variance may become confounded (e.g., Marsh 1989).

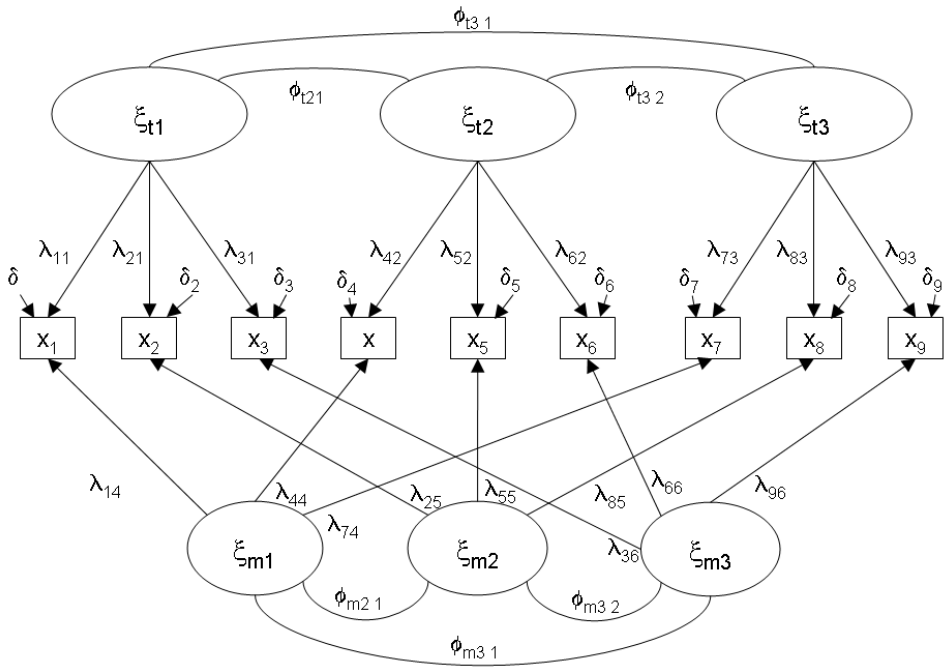


Figure 11. Trait–Method–Error Model

The *correlated uniqueness model*, shown in Figure 12, is similar to the trait–method–error model, but instead of explicit method factors, error terms for measures corresponding to method effects are allowed to be correlated (e.g., Marsh 1989). This specification permits the estimation of differential impacts of each method on the multiple measures corresponding to that method.

Three advantages of the correlated uniqueness model over the trait–method–error model are the following. First, the correlated uniqueness model seldom produces ill-defined solutions; Marsh and Bailey, for example, found that only 2 percent of the MTMM matrixes they examined exhibited improper solutions. Second, methods are not assumed to be unidimensional. The confounding of method variance with trait variance is avoided (when this is due to common trait variation across methods and traits are highly correlated). Third, when four or more traits are measured with at least three methods, one can test the assumption that all correlated uniquenesses associated with one particular method can be explained in terms of a single, unidimensional factor (the test can be conducted by comparing χ^2 tests). It turns out that the trait–method–error model with correlations among methods constrained to be zero is a special case of the correlated uniqueness model. For cases where three traits and three methods are used, the models are identical. But when four or more traits are examined, more parameters are associated with each method under the correlated uniqueness model than the trait–method–error model with orthogonal methods.

Four shortcomings of the correlated uniqueness model should be mentioned. First, the interpretation of correlated uniqueness as method effects is not always clear. Two possible empirical outcomes make the meaning of findings potentially ambiguous: the presence within the same method of (1) significant positive and negative correlations and (2) significant and nonsignificant correlations. The former is incongruous, since it is difficult to conceive of reasons why the same method has opposite effects on measures of different traits when the traits are expected to covary in *either* a positive or negative direction. The latter finding is possible in theory, but in practice is difficult to explain unless one has *a priori* methodological reasons accounting for differences in the significance and nonsignificance of correlated uniquenesses for a common method. In sum, whereas a correlated uniqueness model may fit the data well, the presence of one or both of the above outcomes may be a consequence of model misspecification or capitalization on chance.

A second, broad limitation of the correlated uniqueness model is that it assumes that methods are uncorrelated. This may be reasonable when highly different methods are purposively

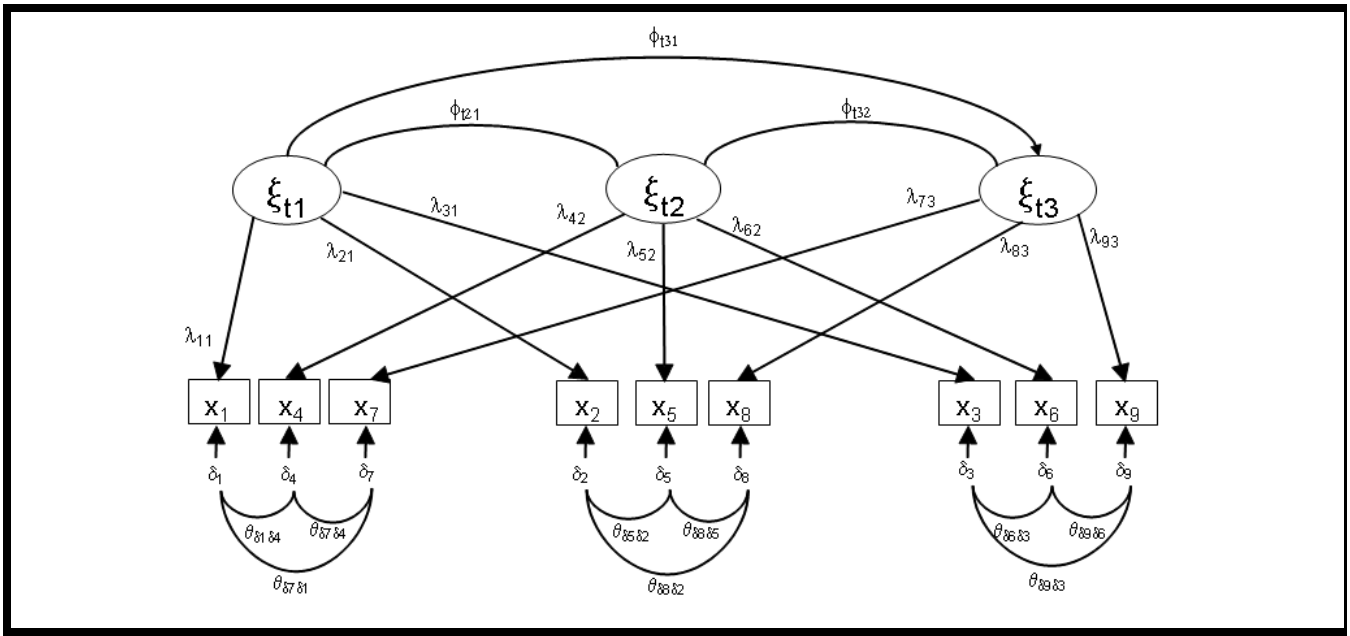


Figure 12. Correlated Uniqueness Model

chosen in a construct validation study. But for the typical study perhaps, where different kinds of self-reports constitute the methods, methods would be expected to be significantly correlated, possibly highly so. Even different methods may be significantly correlated under some conditions. To the extent that the assumption of uncorrelated method effects is violated, parameter estimates for trait variances and covariances will be biased (e.g., Conway et al. 2004). Another limitation to note is that factor loadings will be underestimated to the degree that measure specifity occurs, although measure specifity is often low in practice. Finally the correlated uniqueness model does not permit the flexibility of measuring specific sources of method bias and modeling these (because method factors, *per se*, are not part of its specification).

The trait–method–error model and the correlated uniqueness model both hypothesize that traits and methods supply independent additive effects to variation in a measure. But in some circumstances it may be possible for traits and methods to interact in the sense that "the higher the basic relationship between two traits, the more that relationship is increased when the same method is shared" (Campbell and O'Connell 1982, p. 95). Here the *direct product model* may apply. The general equation for the direct product model can be written as

$$\Sigma - Z(P_m \otimes P_t + E^2)Z$$

where Σ is the variance-covariance matrix of observed measures, Z is a diagonal matrix of scale constants, P_m and P_t are method and trait correlation matrixes, respectively, whose elements are particular multiplicative components of common score correlations (i.e., correlations corrected for attenuation) $\otimes$ is a right-direct (Kronecker) product, and E^2 is a diagonal matrix of unique variances. This model can be implemented in standard structural equation model programs (e.g., Bagozzi and Yi 1990), although Browne's (1990) MUTMUM program may be easier to use. From an intuitive perspective, the direct product model hypothesizes multiplicative effects of methods and traits such that sharing a method across traits exaggerates the correlations between highly correlated traits relative to traits that are relatively independent. That is, the higher the inter-trait correlation, the more the relationship is enhanced when both measures share the same method, whereas the relationship is not affected when inter-trait correlations are zero.

The main advantage of the direct product model over other approaches mentioned so far is that it is the only one to explicitly take into account trait–method interaction effects. Yet, the direct product model has been criticized for not being readily implemented and for not fitting many contexts. By contrast, Campbell and O'Connell (1967, p. 44) imply that trait–method interactions may be the rule rather than the exception. Where might the direct product model fit in research? One case is where self- and peer-ratings or self- and expert-ratings are employed. Each rater might have an

implicit theory and set of expectations about the co-occurrence of certain traits, which lead to rater-specific biases. In such cases, the stronger the "true" associations are between traits, the more likely they are to be noticed and exaggerated, thus producing the multiplicative trait–method pattern. This is called *differential augmentation* in the literature (e.g., Campbell and O'Connell 1967, 1982). Another case that fits the direct product model is termed *differential attenuation*. This occurs, for example, when measurement is done over time and multiple occasions are employed as methods. Here correlations between measures over time are typically lower for longer than shorter lapses in time, demonstrating an auto-regressive or Markov process. Accordingly, a high correlation between two traits will be more attenuated over time than will a low correlation. In contrast, a correlation of zero can erode no further, and thus remains zero when computed across methods (i.e., occasions, here).

The direct product model has a number of drawbacks. One is that it is difficult to assess convergent and discriminant validity. For some guidance here, as well as examples and description of a set of useful hypotheses under the direct product model, see Bagozzi and Yi (1992). Compared to the trait–method–error model, the direct product model applies rather global interpretations of convergent validity and does not supply the degree of convergent validity. Second, trait and method variation are confounded under the direct product model. So one cannot assess variation in a measure due to trait and method separately.

An *extension of the trait–method–error model* should be noted. When one has explicit measures related to the nature of two or more methods, it may be useful to use these measures as indicators of method factors, with the appropriate measures of traits loading on these factors. For example, if under the key informant approach, one gathered data from CEOs and subordinates at each of a number of organizations on properties or processes in the organizations, then separate "method" factors for the CEO and one or more subordinates could be specified. Because CEOs and subordinates have differential knowledge of, investment in, concerns about, etc. organization properties and processes, measures of these (e.g., extent of knowledge) could be obtained and used to operationalize the method factors. Notice that such a specification is different than the dedicated marker variable approach in that specific measures of the multiple methods are acquired, whereas under the dedicated marker variable approach measures for biases other than that specifically reflecting biases of the methods are obtained. Indeed, it may be possible to model the effects of both method biases tied directly to each method in a multimethod study, in addition to a systematic bias such as social desirability modeled as a dedicated marker effect.

Another model we wish to consider under a CFA specification occurs in at least two senses. One is the *additive model* where trait, method, measure specificity, and error are represented explicitly. Bagozzi et al. (1991) consider such a model and provide an example. An alternative model that permits estimation of measure specificity is the *panel model* (e.g., Bagozzi and Yi 1993). The four sources of variance can also be studied by use of a *three-facet design*. For instance, Bagozzi et al. (1999) extended and illustrated the direct product model to incorporate measurement occasion (p_o) as a measure specific-like factor:

$$\Sigma = Z(p_o \otimes p_m \otimes p_t + E)Z$$

The additive and multiplicative models for incorporating the four sources of variance go farther than the ones considered heretofore, and in this sense overcome the limitations therein. However, each shares the other limitations pointed out under the descriptions provided above.

For the case where methods are structurally different (i.e., when they are not randomly selected but rather come from different sources), Eid et al. (2003) propose an approach that explicitly compares and contrasts the different methods. Figure 13 presents an example for the case where each trait–method–error pair has one indicator (the approach is more informative when each trait–method pair has two or more indicators; for an illustration with three indicators for each trait–method pair, see Eid et al. 2008). An example might be the key informant method where three different key informants (e.g., the physician, pharmacist, and nurse on hospital pharmaceutical and therapeutics committees) estimate properties of the committees they sit on (e.g., degree of conflict, information sharing, and trust). The approach has been termed, the *correlated trait–correlated method minus one model*. Notice in Figure 13 that the first indicator of each trait does not load on a method factor, whereas every other indicator loads on either the second or third method factor. The first factor is taken as the comparison standard.

For the generalization of the model shown in Figure 13 where multiple indicators exist for each trait–method pair, a number of benefits can be mentioned for this perhaps seemingly strange specification (Eid et al. 2003, pp. 54-55). One is that measurement error is separated from true trait and method effects. Second, method effects can be modeled as trait-specific rather than assumed uniform across traits. Third, the observed variance in measures can be partitioned into trait-

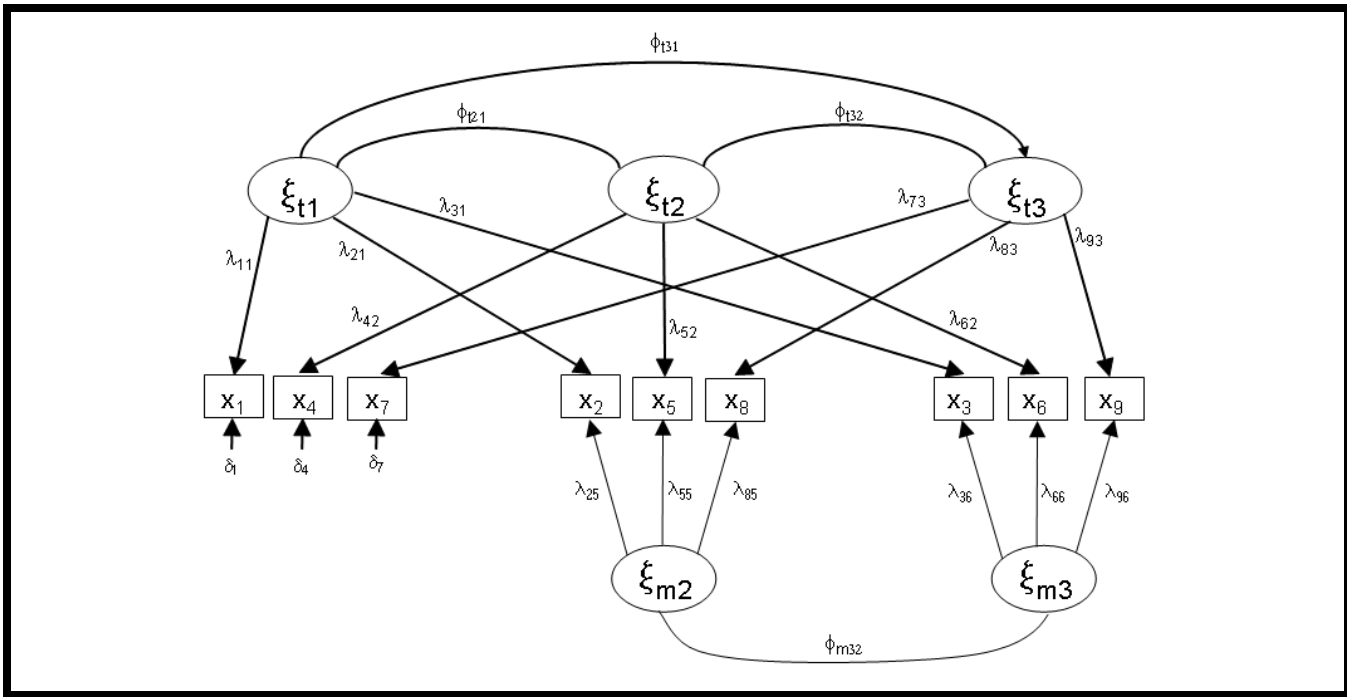


Figure 13. Correlated Trait–Correlated Method Minus One Model for Multitrait–Multimethod Data Where Methods Are Structurally Different

specific, method-specific, and error components. Fourth, trait factors are the true-score variables of the comparison standard, and the method-specific factors are functions of the residuals, which gives trait and method factors clear meanings. Finally, if desired, method factors specific to a trait can be allowed correlated with trait factors for the other traits.

A limitation of the correlated trait–correlated method minus one model is that multiple indicators are needed for each trait–method unit to specify trait-specific method effects. This makes implementation of such models difficult because of the increased data demands over most of the other confirmatory factor analysis approaches discussed herein. Nevertheless the above mentioned benefits make such an approach desirable if one has the resources and opportunity to collect such data. Another drawback is that the method selected as a standard must be meaningful. Eid et al. (2003) provide an example where self-ratings are the standard of comparison for two peer ratings, one by friends and the other by acquaintances. Here the contrast of self with two peers seems to make sense, but other meaningful cases may be difficult to find.

The last confirmatory factor analysis case I wish to consider is the *multilevel confirmatory factor analysis model* (see Eid et al. 2008). Figure 14 presents an illustration, where again the single indicator trait–method pair case is shown for simplicity. The multilevel CFA model applies when methods are interchangeable. For example, a researcher might select three employees at random in a sample of employees across organizations to express their evaluation of the empathy, fairness, and trust of their supervisors. The model shown in Figure 14 is a special case of the random effects analysis of variance model and, continuing our example, values for each trait factor represent the deviations of supervisors from the means across all supervisors for each trait characteristic, where it can be seen that three trait factors occur for each characteristic. Note also that there is one method factor per trait, which means that employee-specific sources of variance are unidimensional for each trait characteristic. As Eid et al. (2008, p. 234) point out, method effects show the deviation of the error-free or true-score of employees as raters from the trait value (mean of all supervisors plus source of variance from trait) on the target characteristics.

A primary advantage of the multilevel confirmatory factor analysis model is that it separates method-specific from error-specific sources of variance. It also permits the partitioning of variance into trait, method, and error components, and yields straightforward ways to compute the proportion of total

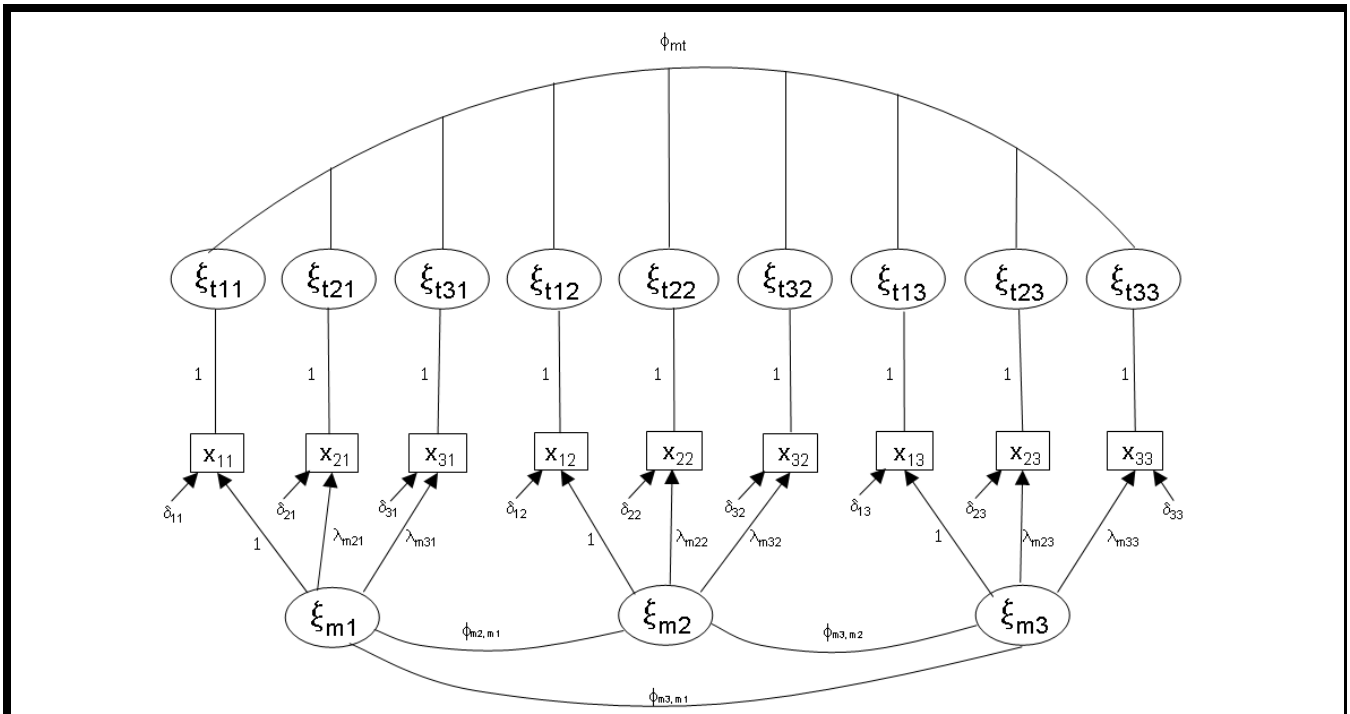


Figure 14. Multilevel Confirmatory Factor Analysis Model for Multitrait–Multimethod Data and Interchangeable Methods

variance that is not due to random error (reliability coefficient), the degree to which true differences between ratings by employees represent differences between target characteristics (consistency coefficient), and the proportion of true variance of ratings that is due to differences between employees (and not due to differences between supervisors) (method specificity coefficients) (Eid et al. 2008, p. 235). A restrictive assumption of this model is that the method factors corresponding to common trait–method units are unidimensional. It should be noted that Eid et al. (2008) consider models where both interchangeable and structurally different methods apply, but of course this requires special data requirements. Again the most useful and powerful analyses occur when one has multiple indicators for each trait–method combination, but this is also the most difficult design to implement and will be difficult to accomplish in many organizational behavior and information systems studies.

Comments on Approaches to Common Method Variance

There are so many ways to approach construct validity and analysis of method bias, and so many issues to consider when implementing these and interpreting results, that we might be apt to throw up our hands in frustration and conclude that no approach is viable. Certainly all approaches have many pros and cons. But it would be misleading and self-defeating to conclude that construct validity cannot be assessed and bias corrected for in certain instances. It is important to keep the claims and limitations of the many approaches in perspective.

The method–method pair technique, for example, is not implementable within a single study with one or a few samples, but under this approach, we can learn which methods are better than others and try to incorporate findings from any meta-analysis into our next study. Likewise, we can benefit from looking across studies employing one or another construct validity approach. This might provide guidance on what methods or measures to accentuate or avoid in the future and what is to be learned by abstracting up from individual studies and looking for useful patterns of findings and conclusions, both methodologically and substantively. An argument can be made that it is important to approach construct validity through a program of interconnecting studies over time.

Some drawbacks of approaches should not be taken as absolute stigmas and lead one to categorically avoid them.

For instance, while the trait–method–error approach frequently fails to yield interpretable results, it does on occasion succeed and gives useful information on trait, method, and error variance to help in appraising construct validity and help in the selection of measures and items for further research. Other approaches may confound trait and method variance or neglect measure specificity, say, yet still be useful in a predictive model where an interesting dependent variable is examined and certain systematic and random errors are controlled for, while predicting the dependent variable. Here we at least correct for errors even though we do not know their separate contributions. Yet other research contexts will dictate what can and cannot be done because the context is additive, multiplicative, or in some other way restrictive, requiring that one consider the context–approach fit. And we should recognize that something is to be gained by combining different approaches in a single study, as in the above mentioned example of integrating the dedicated marker variable approach with the trait–method–error approach. Other combinations are possible as well.

The study of construct validity is a time and energy intensive endeavor and, done right, will require the implementation of multiple methods and traits and the application of advanced statistical and methodological procedures. But this does not mean that everyone must apply a MTMM matrix design in any piece of research or that editors should necessarily require that every study should demonstrate lack of contamination due to method or other systematic bias. Sometimes a well-done, thorough multimethod study will establish precedents for future studies where multiple methods are not required and emphasis is placed more on theory development and "adequate" tests of the theory. We should encourage and reward exemplary studies, yet what we learn from them might not need to be repeated in their entirety for researchers building on these studies. Of course, real opportunities exist for doctoral students, faculty members, and other researchers willing to put the time and effort into developing valid scales and constructs and testing substantive hypotheses with them, while examining and controlling for construct validity. The thorough study of construct validity, whether in and of itself or as part of a broader piece of research or program of research, is a high investment, high risk undertaking but offers the possibility of high rewards. Administrators of any portfolio of research, whether by an individual researcher, team of researchers, or journal, should think about including such a study.

A final comment to note is the following. Different types of data require different types of models and statistical procedures. Sometimes the type of data will dictate, or at least narrow, the choices of models appropriate for analysis. For example, if one has random or interchangeable methods, then the multilevel confirmatory factor analysis model might be appropriate, whereas the correlated trait–method minus one model would not be a good fit; conversely, if one has structurally different methods and a meaningful comparison method can be identified, then the correlated trait–method minus one model would be a good choice, while the multilevel model would not. Of course, both of the aforementioned models require multitrait–multimethod data, preferably with multiple trait–method unit indicators. When such data are not available, the only recourse may be to use one or more of the other approaches reviewed herein. The advantages and disadvantages discussed above for each model also provide some guidance narrowing choices, in addition to the type of data at hand, *per se*. So although many models have been proposed for analyzing construct validity and method bias, the problem at hand and the researcher's purposes will shape the choice of approach(es). Generally, it is safe to say that there is no single approach that dominates all others, so the hope for a "gold standard' is, at present at least, unrealistic, and no substitute exists for sound judgment (in the face of imperfect methods and uncertain data).

Conclusion

There is an inherent tension between our desire for precise guidelines and standards for designing, conducting, and interpreting research, on the one hand, and the characteristic complexities, uncertainties, and ambiguities of research problems, on the other hand. Measurement and construct validity are at the messy end of the spectrum of things and defy simple solutions. Yet we do not believe or want to hear this, and we ln search for an elusive research elixir to make the messiness go away. An admittedly exaggerated analogy might help to demonstrate my point. Students and researchers accustomed to the seemingly absoluteness of the meaning of F tests in regression and ANOVA analyses, say, often expect analogous standards of interpretation for structural equation models. But no single test, not even the χ^2 test, can be applied definitively to ascertain the meaningfulness of most models, and instead, one must rely on a holistic interpretation of the χ^2 test along with a set of additional goodness-of-fit tests (some of which sometimes conflict with each other), as well as other diagnostics.

I do not think it is wise to make broad, either–or categorical statements concerning formative and reflective measurement and construct validity and method bias. Rather, I think these

areas fit the classic dictum where scholars are encouraged to embark on ever deeper question posing and to regard answers along the way as temporary guidelines but not necessarily conclusive criteria. My intent in this paper was to provide some language for looking into the issues and encourage the reader to use this language to discover his or her own perspective on the issues. What seems to be undeniable is that if researchers in information systems and the other organizational sciences make careful and concerted attempts to validate their scales and instrumentation, the rigor of our research efforts will gradually improve and the credibility of our scientific results will be enhanced. This is an admirable goal that should be sought out.

Acknowledgments

I am most grateful to the senior editor, associate editor, and two reviewers for comments and suggestions made on an earlier draft of this article. Their feedback caused me to rethink a number of issues and fundamentally shaped the content herein.

References

Anderson, T. W. 1958. *Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Ashkanasy N. M. 2008. "Submitting Your Manuscript," *Journal of Organizational Behavior* (29), pp. 263-264.

Bagozzi, R. P. 1993. "Assessing Construct Validity in Personality Research: Application to Measures of Self-Esteem," *Journal of Research in Personality* (27), pp. 49-87.

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

Bagozzi, R P., and Edwards, J. R. 1998. "A General Approach for Representing Constructs in Organizational Research," *Organizational Research Methods* (1), pp. 45-87.

Bagozzi, R. P., Fornell, C., and Larcker, D. F. 1981. "Canonical Correlation Analysis as a Special Case of a Structural Relations Model," *Multivariate Behavioral Research* (16), pp. 734-454.

Bagozzi, R. P., and Yi, Y. 1990. "Assessing Method Variance in Multitrait–Multimethod Matrices: The Case of Self-Reported Affect and Perceptions at Work," *Journal of Applied Psychology* (75), pp. 547-560.

Bagozzi, R. P., and Yi, Y. 1992. "Testing Hypotheses about Methods, Traits, and Communalities in the Direct-Product Model," *Applied Psychological Measurement* (16), pp. 373-380.

Bagozzi, R. P., and Yi, Y. 1993. "Multitrait–Multimethod Matrices in Consumer Research: Critique and New Developments," *Journal of Consumer Psychology* (2), pp. 143-170.

Bagozzi, R. P., Yi, Y., and Nassen, K. D. 1999. "Representation of Measurement Error in Marketing Variables: Review of Approaches and Extension to Three-Facet Designs," *Journal of Econometrics* (89), pp. 393-421.

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

Bhaskar, R. A. 1997. *A Realist Theory of Science*, London: Version.

Blalock, H. M. 1964. *Causal Inferences in Nonexperimental Research*, Chapel Hill, NC: University of North Carolina Press.

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

Bridgman, P. W. 1927. *The Logic of Modern Physics*, New York: Macmillan.

Browne, M. W. 1990. "MUTMUM PC User's Guide," unpublished manuscript, Department of Psychology, Ohio State University.

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

Campbell, D. T. 1969. "Definitional Versus Multiple Operationalism," *Et Al.* (2:1), pp. 14-17.

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

Campbell, D. T., and O'Connell, E. J. 1967. "Method Factors in Multitrait–Multimethod Matrices: Multiplicative Rather than Additive?," *Multivariate Behavioral Research* (2), pp. 409-426.

Campbell, D. T., and O'Connell, E. J. 1982. "Methods as Diluting Trait Relationships Rather than Adding Irrelevant Systematic Variance," in *Forms of Validity*, D. Brinberg and L. Kidder (Eds.), San Francisco: Jossey-Bass, 1982, pp. 93-111.

Carnap, R. 1956. "The Methodological Character of Theoretical Concepts," in *Minnesota Studies in the Philosophy of Science* (Vol. 1), H. Feigl and M. Scriven (eds.), Minneapolis: University of Minnesota Press, pp. 38-76.

Chin, W. W. 1995. " Partial Least Squares is to LISREL as Principle Components Analysis is to Common Factor Analysis," *Technology Studies* (2), pp. 315-319.

Conway, J. M., Lievens, F., Scullen, S. E., and Lance, C. E. 2004. "Bias in the Correlated Uniqueness Model for MTMM Data," *Structural Equation Modeling* (11), pp. 535-559.

Cook, T. D., and Campbell, D. T. 1979. *Quasi-Experimentation: Design and Analysis Issues in Field Settings*, Boston: Houghton Mifflin.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science* (35:8), pp. 982-1003.

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), 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), pp. 269-277.

Doty, D. H., and Glick, W. H. 1998. "Common Method Bias: Does Common Methods Variance Really Bias Results," *Organization Research Methods* (1:4), pp. 374-406.

Eid., M., Lischetzke, T., Nussbeck, F. W., and Trierweiler, L. I. 2003. "Separating Trait Effects from Trait-Specific Method Effects in Multitrait–Multimethod Models: A Multiple-Indicator CT-C(M-1) Model," *Psychological Methods* (8),pp. 38-60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., and Lischetzke, T. 2008. "Structural Equation Modeling of Multitrait–Multimethod Data: Different Models for Different Types of Methods," *Psychological Methods* (13), pp. 230-253.

Frijda, N. 1986. *The Emotions*, Cambridge, UK: Cambridge University Press.

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), pp. 238-245.

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

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), pp. 199-218.

Jöreskog, K. G. 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis," *Psychometrika* (34), pp. 183-202.

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), pp. 631-639.

Keat, R., and Urry, J. 1975. *Social Theory as Science*, London: Routledge & Kegan Paul.

Kim, J. 1993. *Supervenience and Mind*, Cambridge: Cambridge University Press.

Knapp, T. R. 1978. "Canonical Correlation Analysis: A General Parametric Significance-Testing System," *Psychological Bulletin* (85), pp. 410-416.

Kumar, A., and Dillon, W. R. 1990. "An Integrative Look at the Use of Additive and Multiplicative Covariance Structure Models in the Analysis of MTMM Data," *Journal of Marketing Research* (29), pp. 51-64.

Lazarus, R. S. 1991. *Emotion and Adaptation*, Oxford, UK: Oxford University Press.

Le, H., Schmidt, F. L., and Putka, D. J. 2007. "The Multifaceted Nature of Measurement Artifacts and its Implications for Estimating Construct-Level Relationships," *Organizational Research Methods* (12:1), pp. 165-200.

Lindell, M. K., and Whitney, D. J. 2001. "Accounting for Common Method Variance in Cross-Sectional Research Designs," *Journal of Applied Psychology* (86:1), pp. 14-121.

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.

MacCorguodale, K., and Meehl, P. E. 1948. "On A Distinction Between Hypothetical Constructs and Intervening Variables," *Psychological Review* (55), pp. 95-107.

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.

Malhotra, N. K., Kim, S. S., and Patil, A. 2006. "Common Method Variance in IS Research: A Comparison of Alternative Approaches and a Reanalysis of Past Research," *Management Science* (52:12), pp. 1865-1883.

Marsh, H. W. 1989. "Confirmatory Factor Analyses of Multitrait–Multimethod Data: Many Problems and A Few Solutions," *Applied Psychological Measurement* (13), pp. 335-361.

Marsh, H. W., and Bailey, M. 1991. "Confirmatory Factor Analysis of Multitrait–Multimethod Data: A Comparison of the Behavior of Alternative Methods," *Applied Psychological Measurement* (15), pp. 47-70.

Parsons, T. 1968. "Social Interaction," in *International Encyclopedia of the Social Sciences* (Vol. 7), D. L. Sills (Ed.), New York: Crowell Collier and Macmillan, pp. 429-441.

Petrie, H. G. 1971. "A Dogma of Operationalism in the Social Sciences," *Philosophy of the Social Sciences* (1), pp. 145-160.

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 Bias 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), pp. 615-656.

Richardson, H. A., Simmering, M. J., and Sturman, M. C. 2009. "A Tale of Three Perspectives: Examining Post Hoc Statistical Techniques for Detection and Correction of Common Method Variance," *Organizational Research Methods* (12:1), pp. 1-39.

Rindskopf, D. 1984. "Using Phantom and Imaginary Latent Variables to Parameterize Constraints in Linear Structural Models," *Psychometrika* (44), pp. 157-167.

Schaffner, K. F. 1969. "Correspondence Rules," *Philosophy of Science* (36), pp. 280-290.

Sellars, W. 1961. "The Language of Theories," in *Current Issues in the Philosophy of Science*, H. Feigl and G. Maxwell (Eds.), New York: Holt, Rinehart, and Winston, pp. 57-77.

Sharma, R., Yetton, P., and Crawford, J. 2009. "Estimating the Effect of Common Method Variance: The Method–Method Pair Technique with an Illustration from TAM Research," *MIS Quarterly* (22:2), pp. 473-490.

Spector, P. E. 1987. "Method Variance as an Artifact in Self-Reported Affect and Perceptions at Work: Myth or Significant Problem?," *Journal of Applied Psychology* (72:3), pp. 438-443.

Stewart, D., and Love, W. 1968. "A General Canonical Correlation Index," *Psychological Bulletin* (70), pp. 160-163.

Suppe, F. 1977. *The Structure of Scientific Theories* (2nd ed.), Urbana, IL: University of Illinois Press.

Suppes, P. 1862, "Methods of Data," in *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress,* E. Nagel, P. Suppes, and A. Tarski (Eds.), Stanford, CA: Stanford University Press, pp. 252-261.

Williams, L. J., Cote, J. A., and Buckley, M. R. 1989. "Lack of Method Variance in Self-Report Affect and Perceptions at Work: Reality or Artifact?," *Journal of Applied Psychology* (74:3), pp. 462-468.

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), pp. 903-936.

About the Author

Richard P. Bagozzi is the Dwight F. Benton Professor of Behavioral Science in Management, Ross School of Business, and Professor of Social and Administrative Sciences, College of Pharmacy, at the University of Michigan. A Ph.D. graduate of Northwestern University, he has honorary doctorates from the University of Lausanne, Switzerland, and Antwerp University, Belgium. Professor Bagozzi does basic research in human emotions, social identity, philosophy of action, philosophy of mind, and the interface between philosophy, statistics, and psychology. His applied research occurs in consumer behavior, health behavior, organizational behavior, sales force behavior, ethics, and the role of structural equation models in measurement and construct validity.