MIS Quarterly

RESEARCH ESSAY

INTERPRETATION OF FORMATIVE MEASUREMENT IN INFORMATION SYSTEMS RESEARCH[1]

By: **Ronald T. Cenfetelli**
Sauder School of Business
The University of British Columbia
Vancouver, BC V6T 1Z2
CANADA
cenfetelli@sauder.ubc.ca

Geneviève Bassellier
Desautels Faculty of Management
McGill University
Montreal, QC H3A 1G5
CANADA
genevieve.bassellier@mcgill.ca

Abstract

Within the Information Systems literature, there has been an emerging interest in the use of formative measurement in structural equation modeling (SEM). This interest is exemplified by descriptions of the nature of formative measurement (e.g., Chin 1998a), and more recently the proper specification of formatively measured constructs (Petter et al. 2007) as well as application of such constructs (e.g., Barki et al. 2007). Formative measurement is a useful alternative to reflective measurement. However, there has been little guidance on interpreting the results when formative measures are employed. Our goal is to provide guidance relevant to the interpretation of formative measurement results through the examination of the following six issues: multicollinearity; the number of indicators specified for a formatively measured construct; the possible co-occurrence of negative and positive indicator weights; the absolute versus relative contributions made by a formative indicator; nomological network effects; and the possible effects of using partial least squares (PLS) versus covariance-based SEM techniques. We provide prescriptions for researchers to consider when interpreting the results of formative measures as well as an example to illustrate these prescriptions.

Keywords: Structural equation modeling, formative measurement, formative indicators, measurement theory

Introduction

Structural equation modeling (SEM) is of tremendous benefit in social and psychological research. Researchers in Information Systems have used this technique to great advantage as well. This is largely due to the solid guidance on the best practices of employing SEM in IS research whether explicitly stated (e.g., Chin 1998a; Gefen et al. 2000; Gefen and Straub 2005), exemplified in research methods (e.g., Gefen et al. 2003), or a combination thereof (e.g., Salisbury et al. 2002). Conventional uses of SEM rely on the application of latent constructs: hypothetical, conceptual variables that represent some true phenomena that is not directly observable (Bollen 2002).

Reflective measures are by far the most common type of measurement used in SEM. A less common but important alternative is *formative* measurement. Latent constructs are not inherently formative or reflective and the choice of mea-

[1]Wynne Chin was the accepting senior editor for this paper. Tracy Carte served as the associate editor.

surement rests on theoretical considerations[2] (Bollen 2007; Howell et al. 2007b). The primary difference between reflective and formative measurement is that while the construct *causes* variance in its reflective indicators, the direction of causality is reversed such that the formative indicators *cause* variance in the construct. This alternative measurement has been increasingly introduced to the management literature through articles in marketing (Diamantopoulos and Winklhofer 2001), organizational behavior (Edwards and Bagozzi 2000), as well as information systems (Chin 1998a).

One example of a construct with formative indicators found in the IS literature is the perceived user resources construct (Mathieson et al. 2001). The observable indicators for this construct are Likert scale items measuring perceived time, documentation, and access to data, among others, and these indicators are proposed to cause the perceived user resources latent construct. As these indicators increase in magnitude, the perceived user resources construct also increases in magnitude. However, an increase in the construct of perceived user resources does not necessarily result in an increase in perceived time, documentation, or access to data, as would be the case if the measures were reflective.

Formative and reflective measurements are both useful ways to model constructs of interest to IS researchers. Formative measurement is an alternative to reflective measurement that differs at two levels: theoretical and nomological. On a theoretical level, formative measurement provides a means of modeling a construct from a diverse and potentially disparate set of observable phenomena. Consider again the perceived user resources construct. Its formative measurement provides a measurement scheme that is grounded to the context under investigation by bringing together different specific facets of user resources, such as time and access to data. This provides a molar level view of a construct that is potentially distinct from its measured components (Chin and Gopal 1995; Gefen et al. 2000). Formative measurement provides an option for researchers to use measures that tend to be specific, actionable attributes of a phenomenon (Barki et al. 2007; Mathieson et al. 2001).

At a nomological level, formative measurement facilitates the study of the causes and effects of a construct by bringing the analysis of potentially disparate indicators to the level of a holistic, single construct. Formative measurement simplifies what might otherwise be multiple paths in a structural model into a more concise single path emanating from (to) an exogenous (endogenous) formatively measured construct. Such bundling of indicators enhances parsimony through the substitution of a single construct in place of multiple indicators within a theoretical model. Depending on the scope of the research, this coalescence of effects allows the researcher to focus attention on a single structural effect as opposed to multiple observable indicator effects. Of course, the indicators of the construct can still be individually evaluated based on their specific contributions to the construct by evaluating their path weights.

Chin (1995, 1998a) introduced formative measurement to the IS literature by describing its nature and purpose. Since that time, IS researchers have made use of formative measurement to describe and explain important phenomena. For example, Barki et al. (2007) take full advantage of formative measurement to devise a sophisticated and nuanced alternative to the traditionally reflectively measured systems usage construct.

Further substantiating the benefit and importance of formative measurement within IS research, the recent work of Petter et al. (2007) provides important guidance on the proper specification of constructs as either formative or reflective. They identify the statistical conclusion errors that can occur from misspecifying the measurement as reflective when it should be formative, and vice versa. They assert that approximately 30 percent of the constructs previously employed in IS research might best be modeled as formative rather than as reflective.

Interestingly, IS researchers have relied almost exclusively on component-based techniques such as partial least squares (PLS) when testing formatively measured models. Despite the clear availability of covariance-based SEM (CB) techniques, such as LISREL, as an appropriate and valuable alternative, IS researchers default to the use of PLS when they employ formatively measured constructs. Yet this choice of technique may impact the analysis and subsequent interpretation of results (Cenfetelli and Bassellier 2009).

Formative measurement is certainly feasible with a CB technique and is often the technique relied upon in fields outside of Information Systems for formative measurement estimation, but identification of estimates can be more difficult than under PLS, which only requires structurally linking to one other construct (Chin 2009). Instead, one must be aware of somewhat more complex rules such as the 2+ emitted paths rule and the exogenous X rule (Bollen and Davis 2009).

With the nascent use of formatively measured constructs in IS research, there is a need for more guidance in the field for

[2]It is the nature of the observable indicators that are either reflective or formative. A construct represents a true underlying concept that exists independent of its measurement.

researchers as well as for reviewers on the proper *interpretation* of formative measurement results. As we will argue, this is a key issue that can have important consequences on the conclusions researchers come to regarding both theoretical and empirical results. Toward providing this guidance, we address six issues when interpreting formative measurement results accompanied by prescriptions for dealing with these issues. We then illustrate our prescriptions on formative measurement results interpretation through the use of an example model relevant to the IS domain.

Formatively Measured Constructs: Interpreting Results

It is our view that the interpretation of the results of formative measurement is a critical part of empirically substantiating a theory. Validity is also an important part of using formative measurement and Petter et al. (2007) provide some initial guidance for validation both prior to and after data collection. But there is still a lack of guidance for the interpretation of formative measurement results. A key challenge comes from the causal nature of the indicators: as a set, the indicators delineate the coverage of the construct; taken individually, each indicator's contribution toward that coverage of the construct can also be evaluated.

Properly interpreting the results of formative models is essential as both measurement and structural results are used in support of theory. In terms of measurement, formative indicators are used to establish the existence of the latent construct in supporting the theoretical/structural model and require an explicit theory on the formation of the underlying construct. Therefore, formative measures must be carefully evaluated given the influence of formative measures on the content coverage of the latent construct. Addition or removal of a formative indicator could influence whether the construct is adequately described by its indicators (MacKenzie et al. 2005; Petter et al. 2007). Content coverage, in turn, implicates the validity of results obtained in the structural model. In that sense the interpretation of the results of formative measurement bears directly on structural model results and so on the interpretation of results in support of theory.

An additional challenge to the interpretation of formative measurement results is that it takes place within the context of possible structural misspecification—such as a model with omitted paths or variables. There is debate as to whether formatively measured constructs are more susceptible to the effects of structural misspecification than reflectively measured constructs, such as pursued in a recent issue of *Psychological Methods* (Bollen 2007; Howell et al. 2007a, 2007b). Where there is general agreement is that formative indicator results are likely to change when structural misspecification occurs, particularly when there is an incomplete census of indicators—making the interpretation of the formative measures results less straightforward.

The primary statistic for assessing a formative indicator is its *weight*, the partialized effect of the indicator on its intended construct controlling for the effects of all other indicators of that construct (Cohen and Cohen 1983). To date, IS studies using formative measurement has focused almost exclusively on the assessment of the statistical significance of formative indicator weights. However, the sole analysis of the significance of these weights is not a sufficient interpretation of formative measurement results. To that effect, we suggest that formative measurement be interpreted through (1) the examination of multicollinearity, (2) the number of indicators, (3) the possible co-occurrence of negative and positive indicator weights, (4) the absolute versus relative contributions made by a formative indicator, (5) the nomological network effects, and (6) the possible effects of using PLS versus CB SEM techniques. Examining formative indicators against these criteria allows the researcher to make a more complete interpretation of the importance of the formative indicators[3] and provide stronger support for the decision to cull an indicator. These points do not need be addressed in a specific order, but should all be evaluated before finalizing the iterative cycle of analysis, interpretation, validation, and possible change of formative measurement. Some symptoms of problems in the results of formative measurement, such as nonsignificant weight or negative weight when bivariate correlations are otherwise positive, can be caused by different problems, so it is important to assess all potential causes before deciding on an action to take. These points and prescriptions are discussed in the next section and summarized in Table 1. In the following section, we will discuss these points in the context of our own example derived from an IS empirical study.

Multicollinearity

One issue that can arise in the interpretation of formative measurement results is when indicator weights have very different results in terms of magnitude, sign, or significance as compared to the bivariate correlation between the indicator

[3]For an example of the reporting of weights and interpreting relative importance, see Chwelos et al. (2001, p. 314).

Table 1. Prescriptions for Interpreting Formatively Measured Construct Results

Point	Summary	Prescription
1. Multicollinearity among the indicators	Multicollinearity may create the potential for unstable indicator weights. Test to perform: Assess VIF or eigenvalues Additional information to report: bivariate correlation between indicators and construct	When collinearity between indicators exists, the researcher must: (1) Evaluate the array of formative indicators employed to measure the construct to determine if there is any conceptual overlap among the chosen indicators. (2) If there is conceptual overlap, remove one of the collinear indicators and retest for collinearity, always ensuring that the conceptual meaning of the construct is not affected. (3) If removal would alter the meaning of the construct, guidance and discussion of the conceptual overlap and on how to improve measurement should be provided, knowing that despite the presence of multicollinearity, researchers can still proceed with the evaluation of the structural model.
2. Number of indicators and nonsignificant weights	As the number of indicators determining a formatively measured construct is increased, the more likely it is that there will be indica-tors with low or even nonsignificant weights. The number of indicators has important implications for the statistical significance and, of course, the magnitude of the *weights* for those indicators. Test to perform: Model testing with different groupings of indicators	When a large number of indicators are used and some have nonsignificant weights, researchers have three options: (1) Group the indicators into two or more distinct constructs with separate effects on theoretically relevant outcomes; this results in smaller sets of indicators for each construct and so the likelihood of any given indicator as being significant increases (2) Create separate constructs as above, but also include a second-order aggregate construct that mediates the effects of the separate first order constructs (Jarvis et al. 2003). (3) Keep all indicators forming a single construct and include a discussion of the absolute contribution of the indicators. If the indicator remains nonsignificant across multiple studies, researchers should interpret this as evidence against the conceptual foundations for its inclusion. The choice of the option is to be based on the theoretical relevance for the model being tested.
3. Co-occurrence of negative and positive indicator weights	Negative weights are the result of the pattern of correlations among the formatively measured construct indicators. Suppression occurs when an indicator shares more variance with another indicator than with the formatively measured construct. Test to perform: Suppressor effect Information to report: Bivariate correlation between the suppressor and the indicators and between the indicators and the construct.	When the indicators of a formatively measured construct have both positive and negative indicators weights, the researcher should: (1) Investigate the possibility of suppressor effects. (2) If suppressors are found, and are collinear with other indicators, these indicators may be removed (as per Prescription 1). (3) If negatively weighted items are (a) not suppressors or (b) not collinear, they should be included in the analysis and potentially culled over time if they repeatedly behave differently than other indicators (suggesting the negative items are potentially measuring something else). (4) An indicator with a statistically significant negative weight but otherwise having a positive bivariate correlation with the formatively measured construct should be interpreted as an indicator having a negative effect when controlling for the effects of other indicators.

Table 1. Prescriptions for Interpreting Formatively Measured Construct Results (Continued)

Point	Summary	Prescription
4. Absolute versus relative indicator contributions	Indicators that have a relatively small contribution to a formatively measured construct in comparison to other indicators may still have an important absolute contribution if that indicator is independently assessed from the other indicators. Information to report: Bivariate correlations (loadings) between the indicators and the construct	Researchers should interpret and report both the partialized (relative) indicator weights as well as the zero-order (absolute) bivariate loadings/correlations between the indicators and their associated formatively measured construct. (1) When all weights are significant, there is empirical support to keep all indicators. (2) When an indicator weight is low, and the bivariate correlation is high the indicator should be interpreted as absolutely important but not relatively so and its theoretical relevance and potential overlap should be questioned. (3) If there is no theoretical overlap, the indicator should be kept in the remaining analyses and in subsequent studies. (4) If theoretical overlap is identified, then there is ground to remove the item. (5) When both the weight and loading are nonsignificant, there is no empirical support to keep the indicator and its theoretical relevance should be questioned.
5. Nomological network effects and construct portability	A formatively measured construct and its constituent indicators are inherently subservient to the nomological network in which the construct resides. As a result, indicator weights will change as the nomological network changes. Test to perform: MIMIC/redundancy analysis Information to report: 1. Construct error term if using CB techniques, 2. Indicators weights found in prior studies, 3. Discussion of the structural misspecification and the relevance of the choice of outcomes.	Researcher must test for construct portability: (1) Assess construct error if using a CB technique. (2) When the same construct has been used in prior studies, compare the weights of the indicators across those studies. (3) If the weights across studies differ substantially or if the construct is new (a) perform a MIMIC/redundancy analysis to assess the likelihood of interpretational confounding, and (b) evaluate the structural misspecification and the relevance of the choice of outcomes.
6. The choice of technique	Information to report: 1. Justification for the inclusion/exclusion of construct error term in CB techniques. 2. If construct error term is excluded or if using PLS technique, considerations of the potential inflation in weights.	Researcher must: (1) When there is a theoretical justification for mutual exclusivity between formative indicators of a construct or with other constructs, use a CB technique. (2) When using PLS or when deciding to exclude construct error in CB techniques, interpret the results taking into account that weights may be inflated. (3) When using CB techniques, use the error term as guidance in the assessment of content validity.

and its construct. One potential cause of this issue is multicollinearity. The greater the level of multicollinearity among the indicators, the more likely many of the indicators will have low or nonsignificant path weights, or possibly even an opposite sign from the indicator's bivariate correlation with its construct. This would lead to interpreting these indicators as being unimportant and/or invalid facets of the construct's domain. Of course, while collinearity is a threat to the interpretation of individual formative indicators, it is not a threat to the structural effects within the model (Chin 1998b). As we noted earlier, however, the employment of formative measures increases the relevance that the measurement results play in support of theory. Collinearity, therefore, does present a threat to those results and associated interpretations because certain indicators that were theorized to be relevant to the domain of the construct may be found to be nonsignificant. It is, therefore, essential to always evaluate collinearity whenever formative measures are employed.

To understand how to interpret and respond to collinearity issues, it is useful to review how and why collinearity occurs. As formative indicators are expected to explain unique variance in the construct and not common variance as for reflective indicators, high correlations among indicators is not expected, if not undesirable. In contrast to reflective indicators, where each indicator is by design collinear with other indicators, multicollinearity in formatively measured constructs can potentially lead to unstable indicator weights (Mathieson et al. 2001; Tabachnick and Fidell 2001) and the influence of each indicator on the latent construct cannot be distinctly determined (Bollen 1989). Since formative measures follow similar principles as in multiple regression, we can demonstrate this instability through the equation for the calculation of the standard error (SE) for a given predictor's multiple regression weight β_j:

$$SE_j = \sqrt{\frac{1 - R_Y^2}{(N - k - 1)(1 - R_{(j)}^2)}}$$

where R^2_Y is the coefficient of determination using all of the k predictors, and $R^2_{(j)}$ is the coefficient of determination in predicting variable X_j using all of the remaining (k-1) predictor variables. $R^2_{(j)}$ is the degree to which the predictors are intercorrelated. As this statistic increases, the SE becomes larger, thus making it more likely that symptoms of problems will appear.

One means to check for indicator collinearity, primarily centered on the correlations among the predictors, is from the size of the $(1 - R^2_{(j)})$ term seen in the denominator of the equation above. This term is the tolerance statistic and its inverse is the variance inflation factor (VIF). Different standards of acceptable values for VIF exist, such as 3.33 (Diamantopoulos and Siguaw 2006) or 10.00 (Hair et al. 1998; Mathieson et al. 2001) with lower values being better. Eigenvalues of the correlation matrix of predictors that significantly depart from 1.00 may also be indicative of collinearity. Mathieson et al. (2001) provide a good example of reporting collinearity statistics for formative indicators.

Multicollinearity can be seen as a useful signal that there is conceptual redundancy among the chosen indicators, particularly for very closely worded perceptual measurement items. Such redundancy needs to be identified at the time of interpreting results and guidance on how to improve the construct's coverage should be provided for possible uses of the construct in future research. If there is redundancy, it may be appropriate to remove the overlapping indicator (Diamantopoulos and Winklhofer 2001). The tradeoff with formative measures showing multicollinearity is that the removal of problematic indicators can decrease the content coverage and, therefore, have an effect on the construct's definition.

Thus, collinearity is a risk to the stability of formative indicator weights and so challenges the interpretation of formative measures. But culling or failing to include conceptually valid indicators risks changing the nature of the formatively measured construct whereas structural predictive capability of the formatively measured construct is not threatened by collinearity (although structural misspecification may also affect structural results). Removal of an indicator is appropriate where there is clear conceptual overlap and a high degree of correlation (e.g., 0.90) between that indicator and another indicator(s). However, one should consider whether the indicator to be removed provides at least some degree of additional predictive power, which would be the case for more moderate intercorrelations (e.g., 0.80). This points to the importance of assessing formatively measured constructs across different nomological networks, as we will discuss later.

Prescription 1: Researchers must test for multicollinearity. When collinearity between indicators exists, the researcher must:

(1) Evaluate the array of formative indicators employed to measure the construct to determine if there is any conceptual overlap among the chosen indicators.

(2) If there is conceptual overlap, remove one of the collinear indicators and retest for collinearity, always ensuring that the conceptual meaning of the construct is not affected.

(3) If removal would alter the meaning of the construct, guidance and discussion of the conceptual overlap and on how to improve measurement should be provided, knowing that despite the presence of multicollinearity, researchers can still proceed with the evaluation of the structural model.

The Number of Indicators and Nonsignificant Measures

The number of indicators used for formative measures has important implications for the statistical significance and the magnitude of each indicator's weight. A greater number of indicators will result in a greater likelihood that many of the indicator weights will be low in magnitude as well as statistically nonsignificant. For example, Mathieson et al. identify seven formative indicators for their construct *perceived resources*, a relatively large number of indicators. As their results show, four of the indicators, such as *financial resources*, are nonsignificant. This nonsignificance is present even though they explicitly test for and exclude the possibility of multicollinearity.

The potential occurrence of formative indicator nonsignificance lies in contrast to reflective indicators where the number of reflective indicators used to measure a construct has little bearing on the measurement results. Reflective measurement is analogous to simple regression where an unlimited number of simple regression statements can be specified without effect on any other regression statement. With formative measurement, however, there is an inherent limit to the number of indicators that can retain a statistically significant weight. Formative measurement is analogous to multiple regression; a construct cannot have more than 100 percent of its variance explained by its formative indicators. If two formative indicators are specified, and we make the assumption that these indicators are orthogonal so as to maximize their potential for individual contribution, then the maximum possible *average* standardized indicator weight will be 0.707. If the number of specified indicators, again uncorrelated, is increased to 10, then the maximum possible average standardized indicator weight will decrease to 0.316.[4]

However, it is highly unlikely that all or even most of the indicators will be uncorrelated, and for CB techniques, it is also unlikely that the indicators will jointly explain all of a formative measured construct's variance. As a result, formative measures with a relatively large number of indicators will typically have indicator weights that are substantially lower than those presented in these hypothetical examples. Formative indicators essentially "compete" with one another to be explanatory of their targeted construct. In this competition to explain variance, only a limited number of indicators will likely be significant while the others will be nonsignificant. As we will discuss in greater detail later, if the *loading* (bivariate correlation) between an indicator and its construct is significant, retaining the indicator should be considered.

To deal with the potential impact of the number of indicators, we prescribe three alternative approaches. One approach is to consider whether the facets described by the formative indicators tap into a single construct or possibly multiple constructs, or sub-dimensions of the same construct, dependent upon the desired level of abstraction. As one employs a greater number of formative indicators, it may be more difficult to conceptually align all indicators under the umbrella of a single construct. However, the identification of a wide array of formative indicators may be relevant to the research study.

[4]Orthogonality among indicators maximizes the set of possible beta weights as compared to when there are correlations among those indicators. When the indicators are orthogonal, the maximum average beta weight is $n^{-1/2}$ where n is the number of indicators. This is the maximum average weight under the constraint of unit variance in the formatively measured construct as found from the solution to the associated Lagrangian multiplier.

To effectively model all indicators within a single model, distinct sets of indicators should be identified that are conceptually aligned and create separate formatively measured constructs with independent effects on theoretically relevant outcomes. For example, if we were to consider project performance as a construct of interest, we could identify a number of formative indicators designed to tap this construct. Indicators might include adherence to project budget, project schedule, quality of output, and technical innovation. On closer inspection, however, the first two may be more representative of project efficiency while the latter two may be more representative of effectiveness. Thus it may be useful to model two separate formatively measured constructs for each of these two aspects of performance with separate effects on the relevant outcome variables in the model.

A second related approach is to create a second order construct that does provide an overall conceptual relation among the identified array of formative indicators. That second order construct is itself formed by first order formatively measured constructs. To continue the performance example above, we could use a second order construct "performance" that is formed by efficiency and effectiveness as previous described. Jarvis et al. (2003) provide an example of such an operationalization (p. 205, Figure 2, Panel IV).

Finally, one can consider removing indicators to increase the likelihood that the remaining indicators are statistically significantly in explaining variance in the construct. Given current guidance in the literature, this is the likely option researchers would take. But as with removing items as a result of multicollinearity, there should be caution to avoid changing the conceptual meaning of the construct. This issue favors the first two alternative approaches involving separate formatively measured constructs.

An illustrative example from the IS literature is the study of business and IT alignment by Chan et al. (1997). Their model includes seven dimensions of effectiveness that influence four dimensions of business performance. Three of the effectiveness dimensions concern satisfaction. The average effect of the satisfaction dimensions on performance was 0.15. Modeling the three satisfaction dimensions under a second order "satisfaction" construct would likely lead to an increase in the average effect of satisfaction on performance.

Prescription 2: When a large number of indicators are used and some have nonsignificant weights, researchers have three options:

(1) Group the indicators into two or more distinct constructs with separate effects on theoretically relevant outcomes; this results in smaller sets of indicators for each con-

struct and so the likelihood of any given indicator as being significant increases.

(2) Create separate constructs as above, but also include a second-order aggregate construct that mediates the effects of the separate first order constructs.

(3) Keep all indicators forming a single construct and include a discussion of the absolute contribution of the indicators. If the indicator remains nonsignificant across multiple studies, researchers should interpret this as evidence against the conceptual foundations for its inclusion.

The choice of the option is to be based on the theoretical relevance—content coverage and level of abstraction—for the constructs and model being tested.

Co-occurrence of Negative and Positive Indicator Weights

When interpreting formative indicator weights, there may be a possible co-occurrence of both positive and negative indicator weights.[5] Negative formative indicator weights are counterintuitive in the sense that the immediate interpretation would be that an *increase* in the indicator would *diminish* the formatively measured construct to which it is a part. Such an occurrence is particularly difficult to interpret when the indicator otherwise has a positive zero-order bivariate correlation with the other indicators of the construct and with the construct itself.

This co-occurrence of positive and negative indicator weights can potentially lead to a serious misinterpretation of the results. For example, one might conclude that the negatively weighted indicator has an overall negative effect on its associated construct and any downstream constructs. But this is likely not the case. The co-occurrence of negative with positive indicator weights when bivariate correlations are otherwise positive is directly the result of the pattern of correlations among the indicators. This can be seen in multiple regression when expressed in matrix form

$$\mathbf{Y} = \mathbf{XB}$$

where $\mathbf{Y}$ is the n × 1 vector of criteria cases, $\mathbf{X}$ is the matrix of n × k predictors, and $\mathbf{B}$ is the k × 1 vector of multiple regression (beta) coefficients (we will assume standardized variables and zero errors for simplicity). Sample size is represented by n and k is the number of predictor variables. Solving for $\mathbf{B}$,

$$\mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Thus the beta weights are a product of $\mathbf{X'X}$, the correlation matrix among the predictor variables, and $\mathbf{X'Y}$, which is a vector of bivariate correlations between the predictor variables and the criterion variable. That a predictor variable has an oppositely signed beta weight as compared to its bivariate correlation with the criterion would be the direct result of negative values in the inverse of $\mathbf{X'X}$, the pattern of predictor correlations. Again, we are describing a case where correlations among indicators and bivariate correlations between indicator and formatively measured construct are positive.

Negative values result directly from the *magnitude* of correlations among indicators. As the correlations of these indicators increase, it is more likely that negative indicator weights will occur in the presence of otherwise positive correlations among indicators and between indicators and the formatively measured construct. This effect may occur even if collinearity is not a threat. The application to formative measurement is clear. Researchers typically develop a set of formative indicators that have positive correlations among themselves (although this is not required). A resulting negative indicator weight presents an interpretation conundrum. How can adding what is otherwise a positively oriented dimension result in *diminishing* the construct of which it is a component? Again, such a "flip" in sign is directly the result of the magnitude of correlations among the formative indicators as noted above. Although this provides a mathematical derivation for this effect, the question remains as to how to ascribe interpretive meaning.

The clearest interpretation is that suppression effects are involved (Cohen and Cohen 1983). Suppression effects occur when one of the predictor variables explain significant variance in other predictor variables not otherwise associated with the criterion. Suppressor variables have the effect of reducing the common variance between one or more of the predictor variables and the criterion variable and so reduce or even reverse the magnitude of the partial correlation coefficient of those predictor variables with the criterion variable. One way of interpreting such suppressor effects is by considering the suppressor variable as controlling for the variance

[5]It is important to clarify that by negative weights we are not referring to situations where PLS may model a construct with *all* negative weights in a "reverse coded" manner. Chin explains this phenomenon and how to deal with it (http://disc-nt.cba.uh.edu/chin/plsfaq/negative_weights_and_loadings.htm). Our discussion above assumes that there is a co-occurrence of *both* positive *and* negative weights in a single formatively measured construct and those indicators otherwise have positive intercorrelations. For those cases where both weight and loading are negative, this is also a case of reverse coding but for a single indicator and thus not relevant to the issue of a weight being of *opposite valence* to its loading.

not directly associated with the formatively measured construct. Another way of understanding suppressor effects is that the primary effect of the suppressor variable is to suppress and explain the *error variance* in one or more of the other indicators, rather than explaining the dependent formatively measured construct.

An example of such a shift in sign from bivariate to partial correlation is in the model of *supply chain flexibility* investigated by Gosain et al. (2004). The formatively measured construct *breadth of information sharing* is posited as a predictor of flexibility and has a positive correlation of 0.245 with *flexibility*. But when placed as a predictor alongside other predictors of *flexibility*, such as *interconnected processes* and *coordination-related knowledge*, the partial correlation (path weight) between *breadth of information sharing* and *flexibility* becomes a statistically significant -0.523, a negative value. Gosain et al. comment that this result was "surprising" given the expectation that the result would be significantly positive, not negative. Indeed, given our description of this type of suppression effect above, the relationship is a generally positive one. The negative value indicates that *breadth of information* has a negative influence when all other factors, such as *interconnected processes*, are otherwise equal between firms.

Prescription 3: When the indicators of a formatively measured construct have both positive and negative indicators weights, the researcher should:

(1) Investigate the possibility of suppressor effects.
(2) If suppressors are found and are collinear with other indicators, these indicators may be removed (as per Prescription 1).
(3) If negatively weighted items are (a) not suppressors or (b) not collinear, they should be included in the remaining analysis and potentially culled over time if they repeatedly behave differently than other indicators (suggesting the negative items are potentially measuring something else).
(4) An indicator with a statistically significant negative weight that otherwise has a positive bivariate correlation with the formatively measured construct should be interpreted as an indicator having a negative effect when controlling for the effects of other indicators.

Absolute Versus Relative Indicator Contributions

Another point to consider in the interpretation of formative measurement results is the difference between the absolute and relative importance of an indicator to its construct. Given that indicators are analogous to predictors in a multiple regression, it is useful to consider guidance on the two types of importance researchers can ascribe to a predictor in relation to a criterion. Nunnally and Bernstein (1994) refer to one type of importance as the relation between a predictor and a criterion holding constant all other predictors. This is the importance of the predictor *relative* to the other predictors. Analogous to beta weights in a multiple regression, the weights of formative indicators provide the *unique* importance of each indicator by partializing the variance of criterion that is predicted by the other indicators. As we have noted, the weight of an indicator is the prevailing and focal statistic for evaluating the importance of a formatively measured construct indicator.

As important as formative indicator weights are for determining their *relative* contribution to their assigned construct, it is also possible to evaluate the *absolute* importance of an indicator to its construct. This is provided by the *loading* of the indicator and so its bivariate correlation with the formatively measured construct. This is identical to what Nunnally and Bernstein refer to as "validity," the *zero-order correlation* between a predictor and a criterion; this statistic "describes the information a predictor provides about the criterion, *ignoring all other predictors*" (p. 192, emphasis added). Just as formative indicator weights are analogous to the beta weights of a multiple regression, formative indicator loadings are analogous to this zero-order correlation. In some cases, indicators may have a low or even nonsignificant weight, and therefore a low or nonsignificant *relative* contribution to the construct. However, an indicator with a low or nonsignificant weight may still have an important *absolute* contribution if the indicator is assessed independently from the other indicators. As a result, there is the potential for misinterpretation of formative indicator results. Quite simply, a researcher may conclude from a low or nonsignificant indicator weight that the indicator is *unimportant* despite what may be a significant zero-order correlation, thus supporting that the indicator is, indeed, important.

A caveat when interpreting formative indicator loadings is that one should not compare formative indicator loadings with one another as the intraset correlations for each block was never taken into account in the estimation process (Chin 1998b). One exception to this rule is if a formative indicator needs to be used as a surrogate for a reflective indicator—with theoretical justification to do so—then the indicator with the highest loading should be used.[6] In general, loadings

[6]We thank Wynne Chin for pointing this out.

should be interpreted in isolation from other indicators to assess the one-to-one relationship between an indicator and its construct. The interpretation of formative indicator loadings is an often-overlooked but important statistic. Again, a nonsignificant weight may lead someone to conclude that an indicator has *no relationship* with the construct it measures when a more valid interpretation may be that the indicator does not contribute beyond the influence of the other indicators in the set. If conceptual overlap is identified, then removal of the indicator can be considered. Of course, if both the weight and loading of an indicator are nonsignificant, then there is some question as to whether that indicator is a valid component of the construct and removal of the indicator should be considered.

A good example of the issue of absolute versus relative contribution comes from Wixom and Todd (2005). They propose various antecedents of *system quality* including *timeliness*. *Timeliness* has a large and significant bivariate correlation with system quality of 0.67. This absolute correlation would lead one to believe that *timeliness* is highly related to *system quality*. However, when placed as an antecedent alongside other predictors, the relative effect of *timeliness* becomes nonsignificant (weight = 0.04). One should, therefore, interpret the result as follows: while significantly related to *system quality*, *timeliness* does not provide additional explanatory power once other antecedents have been taken into account, but *timeliness* is still an important aspect of system quality of its own accord.[7]

Prescription 4: Researchers should interpret and report both the partialized (relative) indicator weights as well as the zero-order (absolute) bivariate loadings/correlations between the indicators and their associated formatively measured construct.

(1) When all weights are significant, there is empirical support to keep all indicators.

(2) When an indicator weight is low, and the bivariate correlation between the indicator and its construct is high, the indicator should be interpreted as absolutely important but not relatively so and its theoretical relevance and potential overlap should be questioned.

(3) If there is no theoretical overlap, the indicator should be kept in the remaining analyses and in subsequent studies;.

(4) If theoretical overlap is identified, then there is ground to remove the item.

(5) When both the weight and loading are nonsignificant, there is no empirical support to keep the indicator and its theoretical relevance should be questioned.

Nomological Network Effects

Another important issue of formative measurement interpretation is construct portability: the relative invariance of a construct's indicator weights as the construct is used in different nomological networks. Some degree of change in indicator weights should always be expected as the estimation of a formatively measured construct depends on other constructs in the model (Diamantopoulos 2006; Howell et al. 2007b). But large changes imply a lack of portability and so threaten the generalizability of the interpretation of a given indicator's contribution and so also the interpretation of the results of the model. A formative indicator weight that changes, for example, from being large in one nomological network to small in another, would make interpretation of its importance difficult.

There are different reasons for changes in indicator weights across networks. One cause is interpretational confounding: the divergence in the nominal and empirical meaning of a construct. As noted by Burt (1976), interpretational confounding occurs when the empirical meaning assigned to an unobserved variable differs from the meaning assigned to that variable by the researcher prior to estimating the unknown parameters. As a result, there is a mismatch between the true model of the latent construct and the model being estimated empirically. This can be related to the content validity points raised by Petter et al. (2007). Any resulting inferences made from the estimated construct are, therefore, ambiguous and will not likely be consistent across different networks (Burt 1976; Howell et al. 2007b).

Another possible cause related to the issue of interpretational confounding is the misspecification of the structural model (Bollen 2007)[8]. Structural misspecification can lead to interpretational confounding as a result of estimating a model with missing variables or missing structural paths as well as variables/paths that should not be included. As Bollen notes in his rejoinder to Howell et al., it is misspecification that is the underlying cause of indicator weight instability and not

[7]The antecedents of system quality in the Wixom and Todd model are themselves constructs, not indicators, but the same principles apply directly to indicator weights.

[8]Bollen (2007) identifies structurally misspecified models as being a model with omitted paths, omitted variables, incorrect dimensionality, or omitted correlated disturbances or correlated exogenous variables that are present in the true model but absent in the estimated model, as well as unneeded paths, variables, or correlations that are absent in the true model.

the nature of formative measures themselves (e.g., reflectively measured constructs can also be subject to misspecification and interpretational confounding).

One additional reason for indicator weights to change in different networks is whether the outcome predicted is within a reasonable scope of possible outcomes. One cannot expect that all outcomes can be explained by a specific construct. The set of items forming the construct has to converge with the predicted outcome. In that sense, if more than one outcome can be explained by the same variable, they have to be part of the nomological scope of that construct. Outside of that scope a different cause should be considered, including a different set of indicators and/or a new construct. For example, a formatively measured construct of perceived effectiveness of IT-enabled institutional structures would be expected to explain outcomes within a scope of relevant behavior, such as the trust in the community of online sellers (Pavlou and Gefen 2005). Other outcomes, such as trust in brick-and-mortar sellers, would be outside of the reasonable scope of outcomes and thus there would not be an expectation of weight invariance in such models.

Both reflectively and formatively measured constructs are susceptible to structural misspecification and interpretational confounding. Chin and Marcolin (1995) provide a useful illustration of a lack of reliability of *reflective* indicator loadings in differing networks. What distinguishes formative from reflective measurement is that it is more straightforward to distinguish the source of instability as either misspecification or interpretational confounding when measuring a construct reflectively. This is accomplished by analyzing the reflective measurement model for item reliability, high loadings, convergent validity, and so forth, so as to identify problematic indicators that may contribute to misspecification. There are no clear measurement model metrics for formatively measured constructs and this was the primary basis for Howell et al. (2007a) suggesting abandonment of formative measurement in favor of reflective.

One means of helping to assess the correct measurement specification of formatively measured constructs is to evaluate the construct disturbance term. This term is the variance in the construct not explained by its formative indicators. However, this disturbance term can only be estimated using CB techniques. The meaning and interpretation of this error term is the topic of some debate. One the one hand, some researchers suggest using it as an indication of the construct reliability (Jarvis et al. 2003; Petter et al. 2007) or of the extent to which the indicators represent the construct (Diamantopoulos 2006; Diamantopoulos et al. 2008), while others claim that even that term depends on other constructs in the model and, therefore, are not an accurate assessment of the quality of the construct. (Howell et al. 2007b; Wilcox et al. 2008). Even if the error term is prescriptive of measurement validation, it is only so in terms of the indicators in aggregate, not for individual indicators (Jarvis et al. 2003). But at a minimum, it can be used as a diagnostic of the construct in that larger error terms make the meaning of the construct increasingly vague.

A better means of assessing the specification of the formative measure may be through the use of MIMIC or redundancy analyses. Both analyses involve two sets of indicators to estimate a given latent construct; one set is reflective of the construct, while the other set is formative. For a MIMIC analysis, both sets of indicators are tied to the exact same construct. Redundancy analysis involves separate constructs, one measured formatively and the other reflectively. CB techniques can perform a MIMIC analysis but are less capable of doing a redundancy analysis. Conversely, in current implementations of PLS software tools, MIMIC cannot be directly implemented as all indicators for a given construct in PLS must be specified as either formative or reflective; mixed modeling is not currently available. Instead, a redundancy analysis can be implemented in at least two ways. First, each of the formative indicators can be replaced by a corresponding formatively measured, single-item construct. These constructs mediate the effects of each individual formative indicator on to the reflectively measured construct and the structural path weights are identical to formative indicator path weights. A second means to perform a redundancy analysis through PLS is to create two latent constructs: one construct is measured with formative indicators and that construct is causal to a second, conceptually equivalent construct measured with reflective indicators. This last implementation is demonstrated by Chin (1998b) and by Mathieson et al. (2001). The strength of the structural path in this implementation can be used to assess the validity of the designated set of formative indicators in tapping the construct of interest: a magnitude of ideally 0.90 or at least 0.80 and above is desired (Chin 1998b).

In CB techniques, MIMIC can be implemented by modeling the formative and the reflective indicators tapping the same latent construct. A redundancy analysis can also be implemented in CB techniques by separating the formative indicators from the reflective ones with each set tapping their own separate but causally related constructs. However, a minimum of three reflective indicators is required in order to achieve model identification. Both the MIMIC and redundancy models in CB techniques are equivalent (Bollen 1989).

The benefit of MIMIC/redundancy analyses is that they serve to validate the proposed set of formative indicators (Chin

1998a). Since the reflective and formative indicators are tapping the same latent construct, MIMIC and redundancy analysis help to assure a content valid set of formative indicators as well as help to support any claims that the construct is portable to different nomological contexts. If formative indicator weights do not remain stable in predicting outcomes within a reasonable nomological scope despite tests that support MIMIC/redundancy analyses, then it is likely that there is structural misspecification. There is the possibility of confounds beyond structural misspecification, such as sampling issues. In the next section, we offer an illustrative example of a changing nomological network and its effect on a set of formative indicators that control for sampling biases by using the same dataset.

Prescription 5: Researcher must test for construct portability:
(1) Assess construct error if using a CB technique.
(2) When the same construct has been used in prior studies, compare the weights of the indicators across those studies.
(3) If the weights across studies differ substantially or if the construct is new, (a) perform a MIMIC/redundancy analysis to assess the likelihood of interpretational confounding, and (b) evaluate the structural misspecification and the relevance of the choice of outcomes.

Other Considerations

There are two other considerations that may impact the results and so the interpretation of those results and both of these considerations directly bear upon the choice of estimation technique. As we noted in the introduction, both PLS and CB techniques are capable of estimating models with formative measures. One consideration is whether and how construct error is modeled for the formative construct. When using PLS, the researcher is making the assumption that there is no error at the construct level as all constructs in PLS are modeled without error. However, CB techniques allow for the estimation of a construct error term that contributes to the construct's value and the researcher must specify whether the construct error term should be constrained to zero or estimated. The modeling of a construct disturbance (error) leads to a construct that is not equivalent to an exact linear composite, while a PLS estimation results in the construct being an exact weighted sum of its indicators.

As a result, including or excluding a construct error term changes the results of the model to be estimated. More specifically, a lack of error term will tend to increase the weights, therefore inflating their importance (Bollen and Lennox 1991). Given the criticality of interpreting formative weights in support of theory, if the weights of formative indicators are overestimated when a construct error term is not estimated, the contribution and validity of specific indicators may potentially be misinterpreted as being more significant than they truly are. In essence, this may lead to an overestimation of the specific effects of the indicator. The precise conditions under which this bias in weights can occur require further research. In addition to providing more accurate estimates for the weights, the error term also provides guidance in the assessment of the content validity (as discussed earlier).

A second consideration that also involves the choice of estimation technique is that when using PLS, the researcher is making the assumption that the formative indicators of a construct freely covary with one another, as well as with other constructs in the model. The exogeneity of formative indicators when using CB techniques allows for the specification of a hypothesized covariance both among the formative indicators as well as between the indicators and any exogenous constructs in the model (Jarvis et al. 2003 provide some guidance on how to do so). Fixing the covariance matrix may be appropriate if there is a theoretical justification for mutual exclusivity among the formative indicators or between the indicators and other constructs. For example, the 43-item social readjustment rating (Holmes and Rahe 1967) is a formative measure of life stress. Some of the most significant indicators of stress include marriage, divorce, and death of a spouse. These are arguably mutually exclusive indicators and thus expected to have no covariance with one another, in theory.

If there is an assumption of mutual exclusivity among formative indicators, or between indicators and constructs, and the indicators are allowed to freely covary, this may result in lower fit when using a CB technique, therefore decreasing the overall results of the model. Whereas CB techniques allow for the testing of explicit hypotheses about these relationships, PLS replaces these testable hypotheses with untested assumptions.[9]

Prescription 6:
(1) When there is a theoretical justification for mutual exclusivity between formative indicators of a construct or with other constructs, use a CB technique.
(2) When using PLS or when deciding to exclude construct error in CB techniques, interpret the results taking into account that weights may be inflated.
(3) When using CB techniques, use the error term as guidance in the assessment of content validity.

[9]Our thanks to a reviewer for this observation.

Summary

To summarize, evaluating the results of formative measures requires more than the assessment of the significance indicator weights alone. Multicollinearity may impact the significance, sign, and magnitude of the weights. The number of indicators can impact the magnitude and significance of the weight. The absolute and relative contribution of each indicator will influence the assessment of their importance. Changes in the weights across networks may determine the generalizability of the construct and so the generalizable importance of a given indicator.

Our broader prescription is that when performing the evaluation of formative measures, authors need to perform more tests and report more information than is usually done, such as (1) assess the bivariate correlations among the indicators and with their construct, (2) test for indicator multicollinearity (VIF or bivariate correlation), and (3) evaluate indicator weights, signs, and loadings. The tests and information needed for each prescription are identified in Table 1.

Ultimately, the conclusion of these evaluations may be that some indicators should be removed. In such a case, it is important that the impact on the conceptual definition of the construct first be evaluated. If it changes the definition, it should be made clear whether the amputated construct is different from the one initially conceptualized. These changes should be discussed and substantiated in studies. However, we believe that very few reasons, if any, would lead to the decision to remove an item after a single study showing some concerns in the results, when the theoretical definition of the construct justifies its inclusion. A same formative measure should be tested multiple times before making decisions on the relevance of removing an indicator.

This guidance should also point to the fact that formative measures cannot be used casually, simply because, as we have seen in some studies, the results with reflective indicators were not satisfactory. Proper analysis of formative measures requires the research to go through several steps to assure proper interpretation.

Illustrative Example

To help ground our explanations of formative measurement interpretation, we present a simple illustrative example model as shown in Figure 1. This model consists of three constructs: an exogenous formatively measured construct with five observed indicators, and two endogenous reflectively measured constructs, one with three indicators and the other with four indicators. We should emphasize two caveats about this model. First, although the model is based on an actual set of data collected by one of the authors, the model we present is intended solely for illustrative purposes and is not proposed in support of any particular theory. Second, our example is just one permutation of many possible types of models with alternatives including models with multiple formatively measured constructs or endogenous formatively measured constructs. For simplicity and ease in interpretation, we use standardized variables and correlation matrices in place of original scales and covariance matrices. One last comment about this specific example is that the formatively measured *SERVQUAL* construct is measured with five observable indicators. This differs from what readers may be familiar with where *SERVQUAL* is modeled with five latent constructs. We use PLS-Graph (Chin 2001; Roberts and Thatcher 2009) to perform the estimation and we align our discussion of this example with the prescriptions in the previous section. Overall results are in Figure 2 and descriptive statistics and correlations are reported in Appendix A.

Collinearity Testing

We began our analysis with an assessment of collinearity among the formative indicators. A decomposition of the correlations among the five service quality formative indicators revealed one eigenvalue of 3.93, thus suggesting the possibility of collinearity. Other tests also suggest this possibility: Bivariate correlations were more equivocal in diagnosing collinearity with correlations among the five indicators ranging from 0.64 to 0.82, and the VIF was of 4.16, above the threshold 3.33 (Diamantopoulos and Siguaw 2006). As prescribed, we determine if there is any conceptual overlap among the chosen indicators. A reading of the wording of the indicators suggests no major overlap and therefore no need at this point to decide to remove an indicator.

Where collinearity has been ruled out as a cause of low indicator weights, one should investigate the possibility that low weights may be the result of the number of indicators used to measure the construct. We assess this next.

The Effect of the Number of Indicators for a Formatively Measured Construct

Our example demonstrates the effect the number of indicators has on formative indicator weight significance. Even with just five indicators, two of the indicators, *assurance* and *empathy*, are not statistically different from zero (Table 2). This is despite otherwise high bivariate correlations between

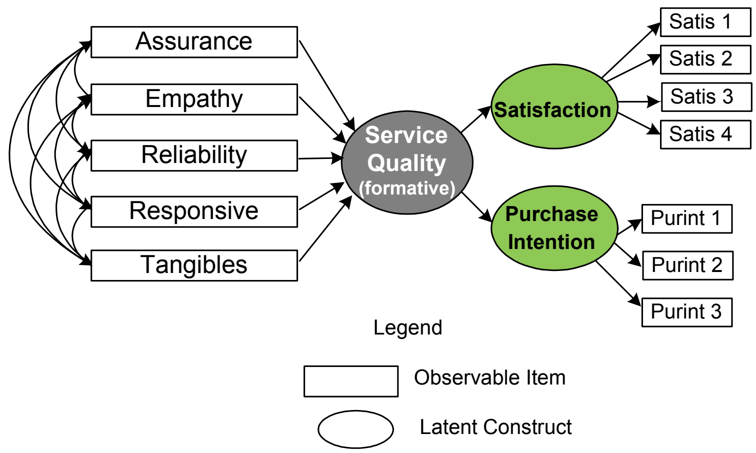

Figure 1. Illustrative Example

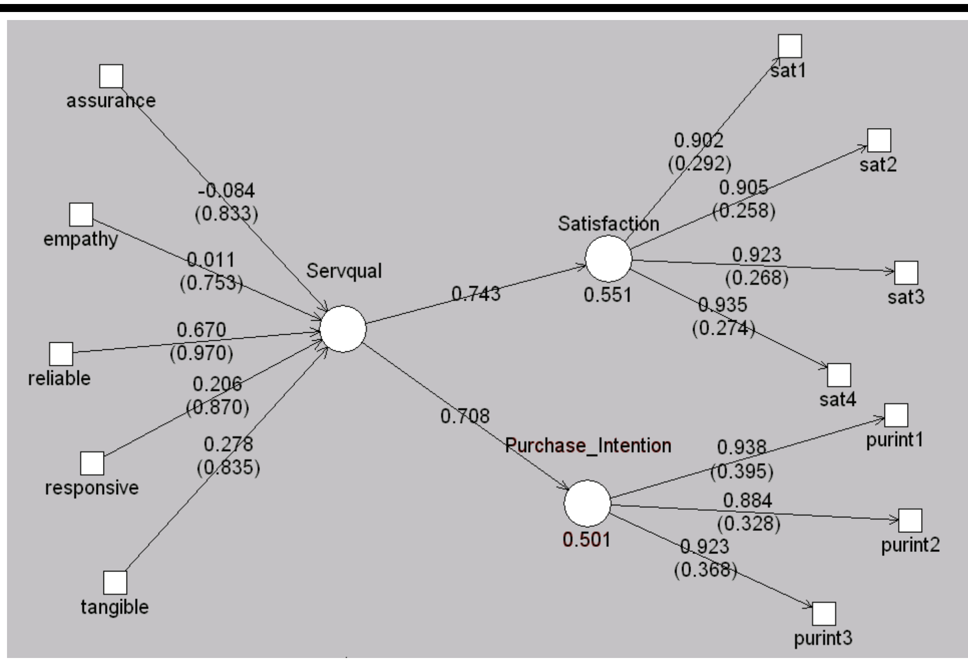

Figure 2. PLS-Graph Output for Illustrative Example

Table 2. Indicator Weights and Significance for Illustrative Example

	Path	t-stat	p value
Assurance	–0.08	–0.98	0.33
Empathy	0.01	0.18	0.86
Reliability	0.67	9.27	0.00
Responsive	0.21	2.38	0.02
Tangibles	0.28	4.80	0.00

these two indicators and the SERVQUAL construct. PLS reports these bivariate correlations as loadings in tandem with the weights as can be seen in Figure 2 (e.g., loading of 0.83 for assurance).

As prescribed, we can evaluate the possibility of conceptualizing the construct differently by grouping indicators into more constructs or dimensions. For example, grouping *assurance* and *empathy* under one construct, and grouping *reliability*, *responsive,* and *tangibles* under another.

Co-occurrence of Negative and Positive Indicator Weights

Our illustrative example also helps to demonstrate the potential occurrence of negative and positive indicator weights, despite otherwise positive bivariate correlations. The *assurance* indicator has a negative beta weight (–0.08). In this particular example, the weight is not significantly different from zero (p = 0.33). An increase in sample size or other minor changes in modeling could, however, make this negative indicator weight statistically significant. In such a case, this negative weight would exist despite the positive correlation of assurance both with the other four indicators of the SERVQUAL construct as well as with the construct itself.

As prescription 3 suggests, the negative weight may be caused by a suppression effect. The negative weight can then be interpreted as follows: when *empathy*, *reliability*, *responsive,* and *tangibles* are otherwise equal, increased amounts of *assurance* will reduce the degree of SERVQUAL. However, as this weight is not significant in our sample, it may be best to consider removing it instead.

Absolute Versus Relative Indicator Contributions

Our example also illustrates our point about comparing absolute to relative indicator contributions. As we just described, *assurance* and *empathy* have indicator weights that are not significantly different from zero. This result may lead a researcher to report that these indicators have no significant effect on their construct and thus infer that these two indicators are not important for other consequential effects, such as satisfaction.

As prescription 4 suggests, both the relative and absolute contribution of the indicators should be assessed and reported. The PLS output (Figure 2) shows loadings (zero-order correlations) with SERVQUAL of 0.83 and 0.75 for *assurance* and *empathy*, respectively. This suggests that although the unique contributions of each of these indicators to SERVQUAL is small in comparison to the other three indicators, there is a still a strong bivariate relationship between these two indicators and SERVQUAL. Contrary to what we observe from the indicator weight results alone, these indicators are important in an absolute sense, if not a relative sense and therefore can potentially be used as a surrogate of the underlying construct if necessary.

Nomological Network Effects

With our example, we can identify a change in weight when we change the nomological network of SERVQUAL. We changed the nomological network by removing one of the endogenous constructs (*purchase intention*) and replaced it with a different reflectively measured construct (*functionality*). Figure 3 shows the changes to the relative magnitudes of the indicator weights when the revised model is analyzed with PLS. In the original model, the *empathy* indicator had a weight that was not significantly different from zero. In the revised model, however, it has a significant weight of 0.24.

As prescribed, we first performed a redundancy analysis. Following both Chin (1998b) and Mathieson et al. (2001), we estimated a model with two different operationalizations of SERVQUAL as depicted in Figure 4. The formative operationalization of SERVQUAL was as before and this construct predicted a second reflectively measured operationalization of SERVQUAL. The reflective operationalization was measured with two items. The path coefficient between the two constructs was 0.94, indicative of an extremely strong degree of formative indicator validity (Mathieson et al. 2001). We do note, however, that this is an assessment of the formative indicators in aggregate, not individually. *Empathy* and *assurance* maintain their low indicator weights even in the redundancy analysis. *Tangibles* also has a low indicator weight.

We then looked for other possible causes of the changes in the weights, such as structural misspecification or choice of outcomes. If we assume that the new outcome *functionality* is a theoretically relevant outcome within the same nomological scope as *purchase intention*, the significant change in this indicator weight may be indicative of an issue with structural misspecification of the SERVQUAL construct.

General Comments

At this point it is relevant to assess the results of the different tests above more globally. Taken together, the redundancy analysis, the prior assessments of collinearity, the number of

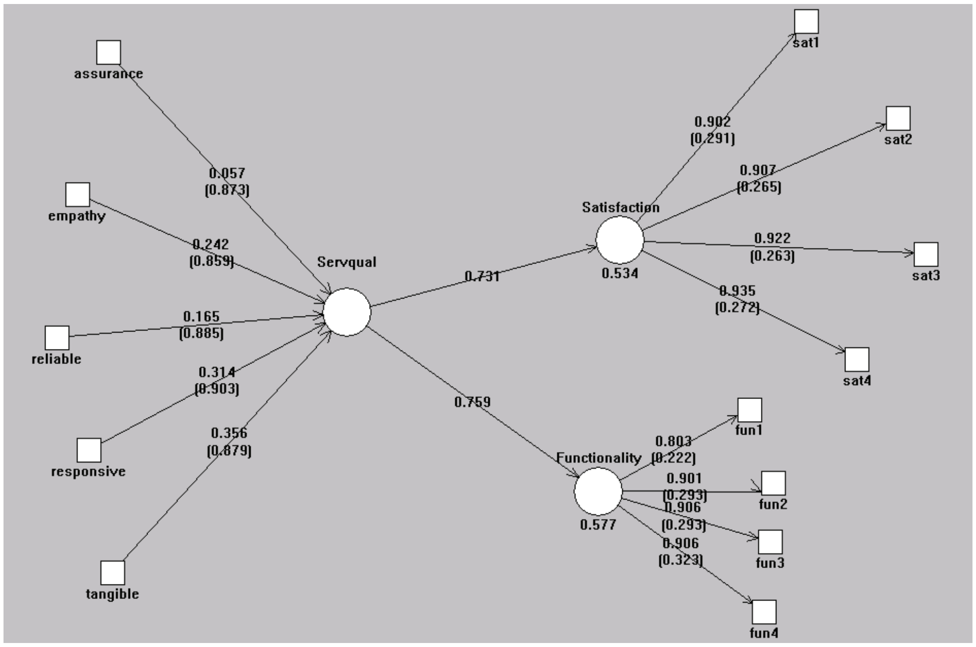

Figure 3. Illustrative Example Results for a Changed Nomological Network (PLS)

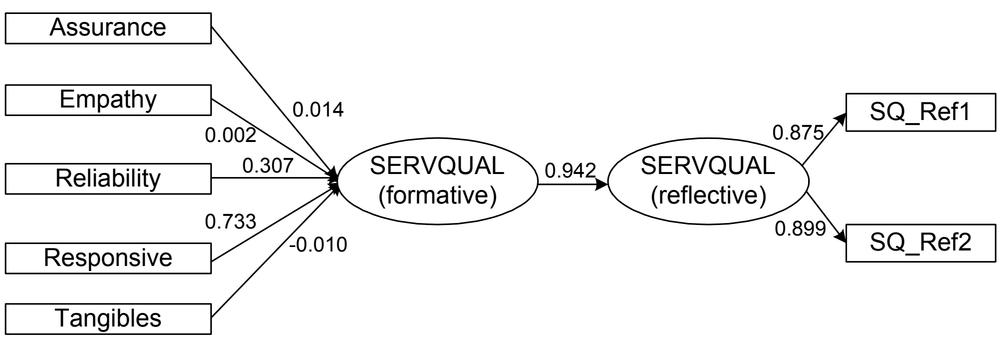

Figure 4. Redundancy Analysis

indicators, absolute versus relative contributions, and nomological network effects might help in determining that SERVQUAL is best measured formatively with only the *reliability* and *responsive* indicators. *Assurance* and *empathy* may fall under a separate construct umbrella, and *tangibles* under a third. Such a determination would have to be defended as consistent with the intended meaning of the SERVQUAL construct, or any other constructs, so as to avoid interpretational confounding. Again, we caution in drawing any generalized conclusions from this particular dataset and offer these suggestions to help guide the reader through the interpretation of his or her own results.

Conclusion

Formative measurement is a useful alternative to reflective measurement but one that continues to lack in the same degree of guidance as compared to reflective measurement. With this in mind, we sought to provide further guidance in terms of the interpretation of formative measurement results. First, we emphasize the increased role formative measurement results can play in empirically supporting theory and the corresponding importance in reporting and interpreting formative measurement results. Second, we identified six issues to be considered when researchers interpret the results of formative measurement, a step often not performed in IS research. These six issues concern (1) multicollinearity, (2) the number of indicators specified for a formatively measured construct, (3) the possible co-occurrence of negative and positive indicator weights, (4) the absolute versus relative contributions made by a formative indicator, (5) nomological network effects, and (6) the possible effects of using partial least squares versus covariance-based SEM techniques. Further, we prescribe six points to consider when reporting and interpreting formative indicator weights. Our prescriptions for interpreting formative measurement results are summarized in Table 1. We hope that these contributions foster a continued and expanded use of formative measurement in IS research.

Acknowledgments

The authors would like to thank Henri Barki, Joshua Slive, and Peter Todd for their very helpful comments on earlier versions of this paper. We would also like to thank Senior Editor Wynne Chin, the associate editor, and the five anonymous reviewers for their very helpful and constructive comments. This research was supported by grants from the Social Sciences and Humanities Research Council of Canada.

References

Barki, H., Titah, R., and Boffo, C. 2007. "Information System Use-Related Activity: An Expanded Behavioral Conceptualization of Individual-Level Information System Use," *Information Systems Research* (18:2), pp. 173-192.

Bollen, K. A. 1989. *Structural Equations with Latent Variables* (1st ed.), Malden, MA: Wiley-Interscience.

Bollen, K. A. 2002. "Latent Variables in Psychology and the Social Sciences," *Annual Review of Psychology* (53:1), pp. 605-634.

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

Bollen, K. A., and Davis, W. R. 2998, "Causal Indicator Models: Identification, Estimation, and Testing," *Structural Equation Modeling: A Multidisciplinary Journal* (16:3), pp. 498-522.

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

Burt, R. S. 1976. "Interpretational Confounding of Unobserved Variables in Structural Equation Models," *Sociological Methods Research* (5:1), pp. 3-52.

Cenfetelli, R. T., and Bassellier, G. 2009. "The Analysis of Formatively Measured Constructs in IS Research: Choosing Between PLS and Covariance-Based Techniques," Working Paper 09-MIS-002, University of British Columbia.

Chan, Y. E., Huff, S. L., Barclay, D. W., and Copeland, D. G. 1997. "Business Strategic Orientation, Information Systems Strategic Orientation, and Strategic Alignment," *Information Systems Research* (8:2), pp. 125-150.

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

Chin, W. W. 1998a. "Issues and Opinion on Structural Equation Modeling," *MIS Quarterly* (22:1), pp. vii-xvi.

Chin, W. W. 1998b. "The Partial Least Squares Approach for Structural Equation Modelling," in *Modern Methods for Business Research,* G. A. Marcoulides (ed.), Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 295-336.

Chin, W. W. 2001. "PLS-Graph User's Guide, Version 3.0" (available online at http://www.pubinfo.vcu.edu/carma/Documents/OCT1405/PLSGRAPH3.0Manual.hubona.pdf).

Chin, W. W. 2009. "How to Write Up and Report PLS Analyses," in *Handbook of Partial Least Squares: Concepts, Methods and Applications,* V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang (eds.), Berlin: Springer, Forthcoming.

Chin, W. W., and Gopal, A. 1995. "Adoption Intention in GSS: Relative Importance of Beliefs," *Data Base for Advances in Information Systems* (26:2-3), pp. 42-64.

Chin, W. W., and Marcolin, B. L. 1995. "The Holistic Approach to Construct Validation in IS Research: Examples of the Interplay Between Theory and Measurement," in *Proceedings of the Annual Conference of the Administrative Sciences Association of Canada, Information Systems Division* (Vol. 16, No. 4), Windsor, Ontario, Canada, pp. 33-43.

Chwelos, P., Benbasat, I., and Dexter, A. S. 2001. "Research Report: Empirical Test of an EDI Adoption Model," *Information Systems Research* (12:3), pp. 304-321.

Cohen, J., and Cohen, P. 1983. *Applied Multiple Regression/Correlation Analysis for Behavioral Sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

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

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

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

Gefen, D., Karahanna, E., and Straub, D. W. 2003. "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly* (27:1), pp. 51-90.

Gefen, D., and Straub, D. W. 2005. "A Practical Guide to Factorial Validity Using PLS-Graph: Tutorial and Annotated Example," *Communications of AIS* (16:5), pp. 1-28.

Gefen, D., Straub, D., and Boudreau, M. 2000. "Structural Equation Modeling and Regression: Guidelines for Research Practice," *Communications of AIS* (4:7), pp. 1-78.

Gosain, S., Malhotra, A., and El Sawy, O. A. 2004. "Coordinating for Flexibility in e-Business Supply Chains," *Journal of Management Information Systems* (21:3), pp. 7-45.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. 1998. *Multivariate Data Analysis* (5th ed.), Englewood Cliffs, NJ: Prentice Hall.

Holmes, T. H., and Rahe, R. H. 1967. "Social Readjustment Rating Scale," *Journal of Psychosomatic Research* (11:2), pp. 213-218.

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

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

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

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.

Mathieson, K., Peacock, E., and Chin, W. W. 2001. "Extending the Technology Acceptance Model: The Influence of Perceived User Resources," *Data Base for Advances in Information Systems* (32:3), pp. 86-112.

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

Pavlou, P. A., and Gefen, D. 2005. "Psychological Contract Violation in Online Marketplaces: Antecedents, Consequences, and Moderating Role," *Information Systems Research* (16:4), pp. 372-399.

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

Roberts, N., and Thatcher, J. 2009. "Conceptualizing and Testing Formative Constructs: Tutorial and Annotated Example," *Database for Advances in Information Systems* (40:3), pp. 9-39.

Salisbury, W. D., Chin, W. W., Gopal, A., and Newsted, P. R. 2002. "Research Report: Better Theory Through Measurement—Developing a Scale to Capture Consensus on Appropriation," *Information Systems Research* (13:1), pp. 91-103.

Tabachnick, B. G., and Fidell, L. S. 2001. *Using Multivariate Statistics* (4th ed.), New York: HarperCollins.

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

Wixom, B. H., and Todd, P. A. 2005. "A Theoretical Integration of User Satisfaction and Technology Acceptance," *Information Systems Research* (16:1), pp. 85-102.

About the Authors

Ronald T. Cenfetelli is an assistant professor of Management Information Systems at the University of British Columbia's Sauder School of Business. He holds a Ph.D. from the University of British Columbia, an M.B.A. from Indiana University, and a B.S. in engineering from Purdue University. Ron's research interests include e-business, the behavioral and emotional aspects of technology usage, the role of technology in influencing immoral behavior, and structural equation modeling. Ron has prior careers as a United States naval officer and as a manager in the pharmaceutical industry.

Geneviève Bassellier is an assistant professor at the Desautels Faculty of Management, McGill University, Montréal, Canada. She received her Ph.D. from the University of British Columbia, and her M.Sc. and B.Com from HEC Montréal. Her current research interests are in the areas of knowledge management and strategic management, in particular focusing on cross-functional knowledge and knowledge sharing between IT and business people, and on the role of top management in IT strategic initiatives. Her work has been published in *Information Systems Research*, *MIS Quarterly*, and *Journal of Management Information Systems*.

Appendix A

Illustrative Model Information

Table A1. Correlation Matrix of Observable Indicators

		1	2	3	4	5	6	7	8	9	10	11	12
1	assurance	1.00	0.75	0.82	0.75	0.62	0.56	0.55	0.58	0.74	0.59	0.45	0.53
2	empathy	0.75	1.00	0.71	0.73	0.56	0.52	0.5	0.54	0.64	0.52	0.38	0.5
3	reliability	0.82	0.71	1.00	0.8	0.69	0.59	0.64	0.65	0.71	0.7	0.57	0.65
4	responsive	0.75	0.73	0.8	1.00	0.64	0.59	0.58	0.6	0.67	0.6	0.5	0.56
5	satis1	0.62	0.56	0.69	0.64	1.00	0.74	0.76	0.79	0.64	0.7	0.55	0.66
6	satis2	0.56	0.52	0.59	0.59	0.74	1.00	0.79	0.8	0.59	0.59	0.47	0.55
7	satis3	0.55	0.5	0.64	0.58	0.76	0.79	1.00	0.84	0.58	0.65	0.5	0.57
8	satis4	0.58	0.54	0.65	0.6	0.79	0.8	0.84	1.00	0.59	0.68	0.53	0.61
9	tangible	0.74	0.64	0.71	0.67	0.64	0.59	0.58	0.59	1.00	0.53	0.46	0.52
10	purint1	0.59	0.52	0.7	0.6	0.7	0.59	0.65	0.68	0.53	1.00	0.74	0.81
11	purint2	0.45	0.38	0.57	0.5	0.55	0.47	0.5	0.53	0.46	0.74	1.00	0.71
12	purint3	0.53	0.5	0.65	0.56	0.66	0.55	0.57	0.61	0.52	0.81	0.71	1.00

Table A2. Means and Standard Deviations of Observable Indicators

	Mean	Std. Dev.
assurance	6.08	0.98
empathy	5.93	1.13
reliability	6.35	0.96
responsive	5.90	1.26
satis1	6.12	1.09
satis2	5.74	1.10
satis3	5.83	1.31
satis4	5.99	1.17
tangible	6.21	0.98
purint1	6.58	0.98
purint2	6.70	0.88
purint3	6.65	0.88