Psychological Bulletin
1991, Vol. 110, No. 2, 305-314

0033-2909/91/$3.00

Conventional Wisdom on Measurement: A Structural Equation Perspective

Kenneth Bollen
Sociology Department
University of North Carolina at Chapel Hill

Richard Lennox
Institute for Research in Social Science
University of North Carolina at Chapel Hill

The applicability of 5 conventional guidelines for construct measurement is critically examined: (a) Construct indicators should be internally consistent for valid measures, (b) there are optimal magnitudes of correlations between items, (c) the validity of measures depends on the adequacy with which a specified domain is sampled, (d) within-construct correlations must be greater than between-construct correlations, and (e) linear composites of indicators can replace latent variables. A structural equation perspective is used, showing that without an explicit measurement model relating indicators to latent variables and measurement errors, none of these conventional beliefs hold without qualifications. Moreover, a "causal" indicator model is presented that sometimes better corresponds to the relation of indicators to a construct than does the classical test theory "effect" indicator model.

Factor analysis (Spearman, 1904) and classical test theory (Lord & Novick, 1968; Spearman, 1910) have influenced perspectives on measurement not only in psychology but in most of the social sciences. These traditions have given rise to criteria to select "good" measures and to a number of beliefs about the way valid and reliable indicators[1] should behave. For instance, Nunnally (1978, p. 102) warned that if correlations among measures are near zero, they measure different things. Some have argued that high correlations are better than low ones (e.g., Horst, 1966, p. 147), whereas others have claimed that moderate correlations are best (Cattell, 1965, p. 88). As the preceding example illustrates, the guidelines to indicator selection are sometimes contradictory. The result is that one can justify keeping or discarding an indicator depending on whose advice is followed. Obviously, this is an undesirable state of affairs that suggests that the conventional beliefs on measurement and indicator selection require clarification.

We contend that these contradictions can largely be traced to two sources. One is that some items do not conform to the classical test theory or factor analysis models that treat indicators as effects of a construct. We present an alternative model in which indicators influence a construct. Second is the failure to present a measurement model that explicitly shows the assumed relation between constructs, measures, and errors of measurement.

This article explores the applicability of five conventional guidelines for construct measurement: (a) Construct indicators should be internally consistent for valid measures, (b) there are optimal magnitudes of correlations between items, (c) the validity of measures depends on the adequacy with which one samples a specified domain, (d) within-construct correlations must be greater than between-construct correlations, and (e) linear composites of indicators can replace latent variables. Using structural equation models, we show that none of these conventional beliefs hold without qualifications.

As a framework for our discussion, we begin by differentiating between indicators that influence, and those influenced by, latent variables. Building on this distinction, the following sections address the five guidelines and their applicability to the two types of indicators. A conclusion follows where we advocate a structural equation approach to analyzing measures.

Indicators as "Causes" and "Effects" of Latent Variables

Classical test theory and factor analysis in psychology view items or indicators as dependent on a latent variable:[2]

$$y_i = \lambda_{i1}\eta_1 + \epsilon_i, \qquad (1)$$

where y_i is the ith indicator, η_1 is the latent or true variable that affects it, ϵ_i is the measurement error for the ith indicator, and λ_{i1} is the coefficient giving the expected effect of η_1 on y_i.[3] We assume that y_i and η_1 are deviation scores around their means, ϵ_i

We are grateful to Jane Scott-Lennox for her suggestions on several versions of the manuscript and for her help in creating Figures 1 and 2. We also thank Lewis Goldberg, Rick Hoyle, Jeff Tanaka, and Raymond Wolfe for their many helpful suggestions on drafts of this article.

Correspondence concerning this article should be addressed to Kenneth Bollen, Sociology Department, CB#3210, Hamilton Hall, University of North Carolina, Chapel Hill, North Carolina 27599, or to Richard Lennox, who is now at the Pacific Institute for Research and Evaluation, 121 West Rosemary Street, 2nd Floor, Chapel Hill, North Carolina 27516.

[1] We use indicator, item, observed measure, or observed variable interchangeably.

[2] This section and the next one rely heavily on previously published work (Bollen, 1984).

[3] To simplify the notation we do not use an additional subscript to index the individuals in a sample. The i subscript that we use indexes the indicators, so if we had, say, four indicators we would have four equations like Equation 1, one for each indicator.

is uncorrelated with η_1, $\text{COV}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, and $E(\epsilon_i) = 0$. We could elaborate this model by allowing multiple latent variables to influence y_i, but to simplify the discussion we maintain the typical assumption of one underlying factor (i.e., η_1).

Indicators that depend on the latent variable, such as the preceding, are effect indicators. Figure 1a is a path diagram that represents four effect indicators influenced by η_1. The lower arrows in Figure 1a indicate the effects of measurement errors on the observed variables. We use this example to illustrate our points throughout the article. The latent variable might be verbal intelligence and the indicators four measures of it. Or, η_1 could be self-esteem and the y_is, self-esteem measures. What is critical for the effect indicator model and what is implied in the assumptions of classical test theory or factor analysis is that the latent variable determines its indicators.

Another possible measurement model has indicators that cause the latent variable:

$$\eta_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \cdots + \gamma_{1q}x_q + \zeta_1, \qquad (2)$$

where η_1 and all xs are deviation scores, $\text{COV}(x_i, \zeta_1) = 0$ for all i, and $E(\zeta_1) = 0$. Equation 2 differs from Equation 1 in that the indicators determine the latent variable rather than the reverse. We can consider Equations 1 and 2 to be regression equations. In Equation 1, the explanatory variable is latent and the dependent variables (the ys) are observed. In Equation 2, the explanatory variables (the xs) are observed and the dependent variable is latent.

Blalock (1964, pp. 162–169) referred to the indicators in Equation 2 as "cause indicators" but other names include *formative* or *composite* indicators. We use the term *causal indicators* borrowing from Blalock but do not attribute any special significance to the term *cause* other than the fact that the indicators determine the latent variable. Note also that this differs from a principal-components model in that a disturbance term is present, and that Equation 2 has just one latent variable for q indicators rather than q latent variables.

In contrast to the effect indicators model in Figure 1a, Figure 1b depicts four x_i indicators that influence the latent variable, η_1. If for example, η_1 represents socioeconomic status (SES), the xs might be education, occupational prestige, income, and neighborhood. Notice that these indicators determine a person's SES rather than the reverse. For instance, according to this model, if income increases, SES increases even if education, job, and neighborhood stay the same. On the other hand, we do not expect an increase in SES to require a simultaneous increase in all four indicators (Hauser, 1973). Thus, modeling these four indicators as dependent on the latent variable (see Equation 1) would not make substantive sense. Other examples of causal indicators include exposure to discrimination, which is indicated by race, sex, age, and disabilities. Similarly, life stress could be the latent variable and job loss, divorce, recent bodily injury, and death in family could be four causal indicators of it. A construct of objective social interaction might be measured with time spent with spouse, with co-workers, with friends, and with others. Accuracy of eye-witness identification could be measured by recall of characteristics of the accused such as eye color, hair color, height, weight, manner of dress, or the presence of a beard. All of these are causal rather than effect indicators (see Bollen, 1984) because in each case the indicators determine the latent construct.

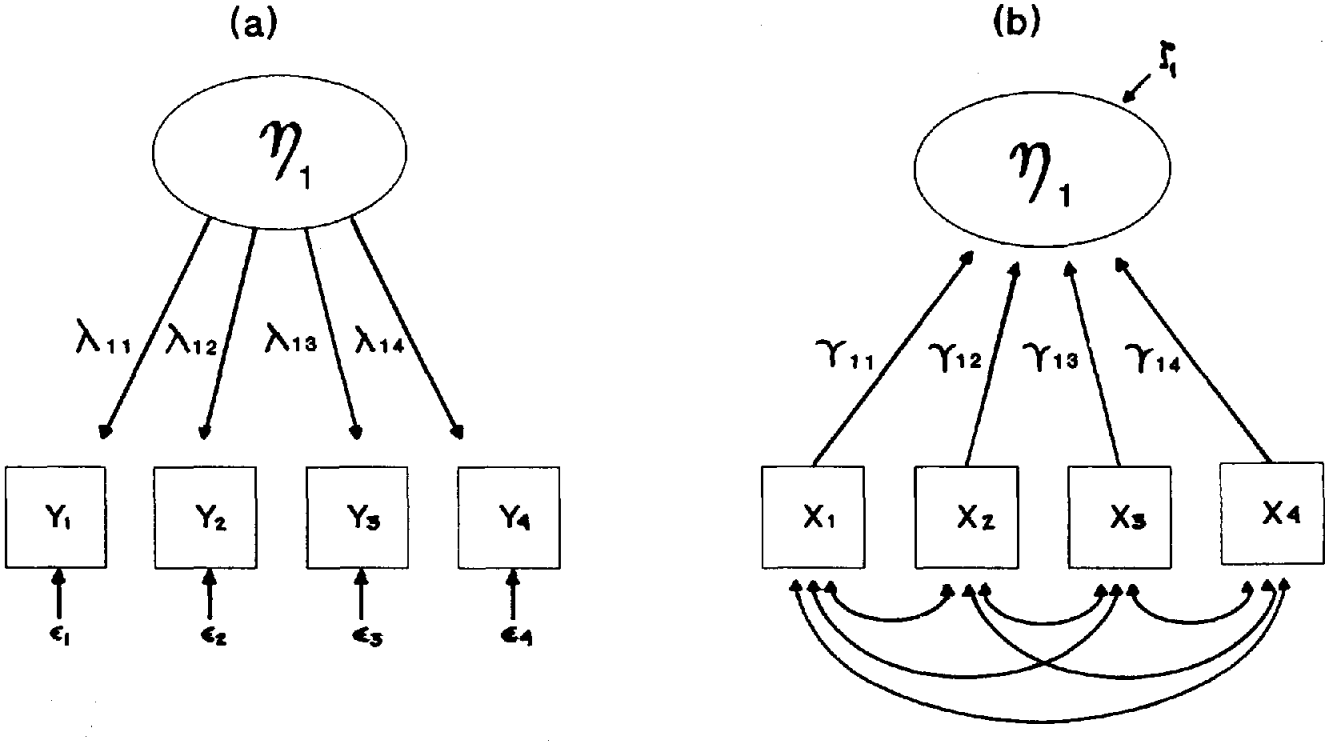

Figure 1. Path diagrams of effect (a) and causal (b) indicator measurement models.

This distinction between indicators as causes and indicators as effects of latent variables has fundamental implications for the conventional ideas about indicators.

Internal Consistency

Perhaps the most widely accepted premise in classical measurement theory is that indicators positively associated with the same concept should be positively correlated with one another.[4] This internal consistency perspective is dominant in psychology, sociology, and the other social sciences. It forms the basis of many reliability estimates and of factor analysis. This belief of the need for positive correlations among indicators of the same concept explains the common practice of screening correlation matrices for items that cluster together and discarding items that have near zero or negative correlations with other measures of the same construct.

We can ascertain the appropriateness of internal consistency by examining the effect and causal indicators represented in Figure 1a and 1b. We begin with effect indicators. To simplify our presentation we standardize η_1 and the ys. Consider the correlation of y_1 with y_2:

$$\text{CORR}(y_1, y_2) = \lambda_{11}\lambda_{21}. \quad (3)$$

With the observed and latent variables standardized, λ_{11} and λ_{21} give the correlation of y_1 with η_1 and y_2 with η_1, respectively. Given that each indicator has a positive association with η_1 (i.e., $\lambda_{11}, \lambda_{21} > 0$), the correlation between y_1 and y_2 should be positive as required by the internal consistency criterion. Of course, the same results follow for any other pair of the y variables: They should be positively correlated. Thus, these results support the internal consistency perspective.

In Figure 1b and Equation 2, a different story emerges. Consider η_1 and the xs standardized as before. The correlation of x_1 with x_2 is

$$\text{CORR}(x_1, x_2) = ?. \quad (4)$$

As Equation 4 shows, one cannot know the correlation between x_1 and x_2, nor that between any pair of the xs that is based on the causal indicator model. This means that causal indicators of the same concept can have positive, negative, or no correlation. Researchers relying on factor analysis or the examination of correlation matrices for selecting indicators may be overlooking valid measures of a construct if the indicators determine the latent variable. Consequently, always using internal consistency as a criterion can have dire consequences.

Optimal Correlations of Indicators

Another measurement prescription is that researchers should seek an optimal level of correlation of indicators. One school of thought proposes that high correlations are desirable. For example, Selltiz, Wrightsman, and Cook (1976) argued that "if all variables measure the same general characteristic of an attitude... then we should be able to show that the variables are all highly correlated" (p. 402). Cattell (1965), on the other hand, contended that interitem rs ought not to be too large: "Homogeneity should not exceed a certain point in a good test, else transferability and validity suffer" (p. 88). Finally, still others find that moderate correlations are optimal:

> The optimal level of homogeneity occurs when the mean interitem correlation is in the .2 to .4 range. Lower than .1 and it is likely that a single total score could not adequately represent the complexity of the items; higher than .5 and the items on a scale tend to be overly redundant and the construct measured too specific. The .2 to .4 range of intercorrelations would seem to offer an acceptable balance between bandwidth on the one hand and fidelity on the other. (Briggs & Cheek, 1986, p. 114)

Clearly, the practioner is faced with conflicting advice on the appropriate level of correlations of indicators of the same latent variable. Again, we turn to Figure 1 to gain insight into this problem. For effect indicators (see Figure 1a and Equation 1), Equation 3 shows that for standardized variables the correlation between any y_i and y_j ($i \neq j$) is $\lambda_{i1}\lambda_{j1}$, where λ_{i1} is the correlation of y_i and η_1 and λ_{j1} is the correlation of y_j and η_1. The CORR(y_i, y_j) depends on the magnitude of corelations of y_i and y_j with the latent variable they measure. Does a high correlation indicate a problem? If the indicators are effect indicators that obey Figure 1a, a high correlation of the indicators suggests that one or both measures are highly correlated with η_1, a situation that is desirable. A low CORR (y_i, y_j) does suggest poor reliability (i.e., low squared correlations of the latent with the observed variables, Lord & Novick, 1968, p. 61) for one or both measures. Though we would prefer a moderate correlation over a low correlation, we see no reason to choose indicators with moderate correlations over those with high correlations for effect indicators. Thus, for effect indicators of a single latent variable that conform to Equation 1, we find high correlations superior to moderate or low ones.[5]

But this is only part of the story. Consider Figure 1b with causal indicators. For indicators that determine the latent variable, the magnitude of the indicator correlations is not explained by the model, therefore we cannot say much about the validity of x_i as a measure of η_1. We can say that high correlations between the xs make it difficult to separate the distinct impact of individual xs on η_1. To the extent that one or more of the xs approaches being a perfect linear combination of the others xs, this creates a multicollinearity problem. Low correlations lessen this difficulty.

In brief, we recommend high correlations of indicators for effect indicators. However, we have no recommendations for the magnitude of correlations for causal indicators, because these correlations are explained by factors outside of the model.

Sampling Facets of a Construct

When selecting indicators of a unidimensional construct, researchers are sometimes advised to select ones that represent

[4] This, of course, assumes that negatively worded items are recoded so that all items reflect the same direction.

[5] Remember that this recommendation is conditional on the model that underlies the data. Had we suspected correlated errors of measurement between indicators, then the correlation of indicators could be inflated or deflated by this factor. Using covariance algebra, the composition of the correlation between indicators with correlated errors can be easily derived and the appropriate qualifications placed on the above statements.

all facets of it. Implicit in this argument is the notion that too narrow a set of items undermines construct validity. For example Epstein (1983) suggested that an item should be included in a scale only if it contributes unique variance to the total scale score:

> What is often not realized is that the average interitem correlation for most intelligence tests is between .20 and .30. If items in a scale were more highly correlated with each other, they would be too redundant to sample efficiently the breadth of broad personality variables such as intelligence, honesty, or extraversion. An ideal item in a test that measures a broad trait is one that has a relatively high correlation with the sum of all items in the test (minus itself) and a relatively low average correlation with the other items. (p. 366)

Cattell (1965) had a similar view:

> Usually, in a well functioning machine or living organism, we judge its efficiency by how well it functions as a whole. We do not expect the parts all to be exactly the same, and indeed, such homogeneity is the mark of a lower organism. Similarly, any complex and ingenious test may need to be put together like a watch, with all its parts properly balanced for some final result. If it is chopped up the scores on its parts need not necessarily correlate highly with one another. (p. 162)

One way to assess this advice is to evaluate the consequences of removing indicators. If each of our original indicators are "representative" of distinct facets of a construct, then we should face dire consequences by removing any one of them. Starting with Figure 1a, suppose that we remove y_4. Doing so has no impact on the correlations of the remaining ys with η_1 or their correlations with each other. That is, λ_{i1} is the same correlation of y_i with η_1, and the CORR(y_i, y_j) remains the same. Although the composite formed by all indicators may be less reliable if we use three indicators instead of four, as long as we have sufficient effect indicators to estimate a model and these are at least moderately correlated with η_1, the issue of adequately representing all facets of a concept is specious.

Our point is that, for all practical purposes, equally reliable effect indicators of a *unidimensional* concept are interchangeable. However, heterogeneous facets preclude unidimensionality. If Cattell (1965) and Epstein (1983) are talking about multidimensional constructs, then each dimension should be measured with several indicators. Forcing effect indicators of distinct dimensions into a unidimensional model such as Figure 1a is not an adequate solution.

Causal indicators operate differently (see Figure 1b), for when we remove one indicator, say x_1, the repercussions are more serious. With causal indicators η_1 is composed from all indicators (and the disturbance). Omitting an indicator is omitting a part of the construct. For instance, suppose that η_1 is objective social interaction, x_1 to x_4 are time with friends, with family, with work colleagues, and with others, respectively. Excluding x_1 gives an incomplete picture of objective social interaction. Or, if η_1 is exposure to discrimination, not measuring race will lead us to miss a crucial part of exposure. Initially this may seem to suggest that the idea of "sampling all facets" of a construct is appropriate for causal indicators. However, we take issue with the word sampling. With causal indicators we need a census of indicators, not a sample. That is, all indicators that form η_1 should be included.[6]

In summary, the recommendation that we sample the facets of a construct can be misleading. With effect indicators of a unidimensional construct (see Figure 1a), equally reliable indicators are essentially interchangeable. If many facets mean many dimensions, then each dimension should be treated separately with its own set of effect indicators. For causal indicators (see Figure 1b), excluding an indicator changes the composition of the latent variable and with few exceptions each causal indicator is important to include.

Within-Construct Correlations Versus Between-Constructs Correlations

When selecting indicators, researchers look for items which tend to "cluster together." Specifically, some claim that the correlations of indicators of the same construct should exceed the correlations between indicators from different constructs. A more formal examination determines the suitability of this strategy. Figure 2 is a path diagram of two constructs (η_1 and η_2) each with three effect indicators. In equation form the model is

$$y_i = \lambda_{i1}\eta_1 + \epsilon_i \quad \text{for} \quad i = 1, 2, 3 \tag{5}$$

and

$$y_i = \lambda_{i2}\eta_2 + \epsilon_i \quad \text{for} \quad i = 4, 5, 6, \tag{6}$$

where y_i, η_1, and η_2 are standardized, COV(ϵ_i, η_j) = 0 for i = items 1 to 6 and j = latent variables 1, 2, COV(ϵ_i,ϵ_j) = 0 for $i \neq j$, and $E(\epsilon_i) = 0$ for all i. Consider first the correlation between two indicators of the same construct, say y_1 and y_2:

$$\text{CORR}(y_1, y_2) = \lambda_{11}\lambda_{21}. \tag{7}$$

The correlation between two indicators each from a different construct (e.g., y_1 and y_4) is

$$\text{CORR}(y_1, y_4) = \lambda_{11}\lambda_{42}\rho_{12}, \tag{8}$$

where ρ_{12} is the correlation of η_1 with η_2. According to conventional wisdom, the CORR(y_1, y_2) should exceed CORR(y_1, y_4) whenever y_1 and y_2 are items measuring the same construct and y_1 and y_4 are items that measure different constructs. Of course, if the two latent variables were orthogonal ($\rho_{12} = 0$), this criterion would be satisfied. But typically we would expect two constructs to have a nonzero correlation. Even then, the within-construct correlations often exceed the between-constructs correlations, but a comparison of Equations 8 and 7 reveals that there can be exceptions. Whenever λ_{21} is less than $\lambda_{42}\rho_{12}$, the between-constructs correlation exceeds the within-construct correlation. For instance, suppose that the correlation of y_2 with η_1 is .5 (= λ_{21}), the correlation of y_4 with η_2 is .8 (= λ_{42}), and the correlation between the constructs is .8 (= ρ_{12}). Here, the within-construct correlation would be less than the between-

[6] If an x has no association with the other causal indicators, then it exclusion might have less impact than if it is correlated. Or, if an x is an exact linear combination of the other xs, it provides redundant information and is therefore less critical.

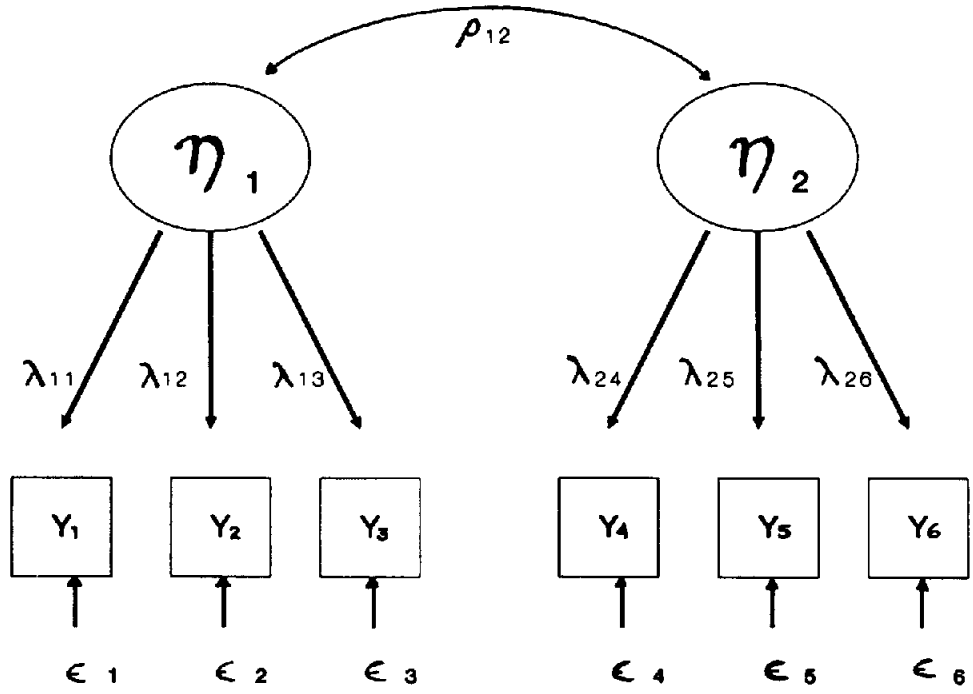

Figure 2. Path diagram of two correlated latent variables with effect indicators.

constructs correlation even though y_1 and y_2 are measures of the same construct and y_1 and y_4 belong to different constructs. Clearly, the same results hold for other pairs of indicators and for other values of the correlations.[7] Thus, if we rely on inspection of the correlation matrix alone, we may incorrectly identify which variables gauge the same construct and which measure different ones.[8]

It is easy to demonstrate that this guideline has even less basis with causal indicators of different latent variables. With causal indicators such as those displayed in Figure 1b, the model places no restriction on the magnitude of the correlations between indicators. Therefore, we can make no definitive statements about between- versus within-construct correlations.

In summary, the guideline that within-construct indicator correlations should exceed between-construct correlations can lead to incorrect indicator selection for effect and causal indicators.

Linear Composites as Substitutes for Latent Variables

Another common practice is to calculate the sum of equally weighted items to form a composite scale to measure the latent variable. Often these composites stand in for the latent variable in a regression, analysis of variance (ANOVA) or similar types of analysis. The appropriateness of the linear composite, like other conventions, depends on whether the latent variable is measured with effect or causal indicators.

Consider Figure 1a with effect indicators. The equally weighted linear composite, say c_1, is

$$\begin{aligned} c_1 &= y_1 + y_2 + y_3 + y_4 \\ &= (\lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41})\eta_1 + (\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4), \\ &= \lambda\eta_1 + \epsilon, \end{aligned} \tag{9}$$

where $\lambda = \lambda_{11} + \lambda_{21} + \lambda_{31} + \lambda_{41}$ and $\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$. (The second line follows from substitution of Equation 1 for each indicator and simplifying.) The last line of Equation 9 shows that the equally weighted linear composite has a relation to η_1 that is similar in form to that of each of the individual indicators: λ is the expected change in c_1 for a one unit change in η_1 and ϵ is the error term in c_1. Thus, the linear composite's scale

[7] Orthogonal factor analysis would appear to be a possible exception. However, exploratory factor analysis with orthogonal factors differs from the situation described here in several ways. First, and most important, the advice on within-construct correlations being greater than between-constructs correlations is based on the examination of the correlation matrix of indicators, not on a factor analysis of these measures. Second, exploratory factor analysis virtually always leads to each indicator having nonzero loadings on more than one factor. This differs from the assumption in the above guidelines, that each measure is only influenced by a single construct. If the factor complexity of the variable is greater than one, *within* and *between* constructs loses its meaning.

[8] Campbell and Fiske's (1959) multitrait-multimethod (MTMM) approach to measurement validity relied heavily on comparing correlations for measures of traits found by different methods. Here too, the comparisons of correlations can be misleading. See Althauser and Heberlein (1970) and Werts and Linn (1970) for an alternative structural equations approach to the MTMM model.

differs from η_1 (except in the unlikely case that $\lambda = 1$) and the composite contains measurement error.

One can gain further insight into the nature of the composite (c_1) by examining its squared correlation with η_1. Using the definition of the squared correlation and Equation 9, we see that

$$\rho^2_{c_1\eta_1} = [\mathrm{COV}(c_1, \eta_1)]^2/[\mathrm{VAR}(c_1)\mathrm{VAR}(\eta_1)]$$
$$= (\lambda^2\mathrm{VAR}(\eta_1))/[\lambda^2\mathrm{VAR}(\eta_1) + \mathrm{VAR}(\epsilon)]. \quad (10)$$

The numerator is the variance in the composite associated with the true score or latent variable. The denominator is the total variance of the composite. As such, Equation 10 gives the reliability of the linear composite. The linear composite and η_1 are perfectly correlated (or c_1s reliability is one) only in the unlikely case that VAR(ϵ) is zero. As is well-known, the reliability of the composite typically is greater than that of its components (Lord & Novick, 1968), but this is not the same as saying that it has perfect reliability. An implication of these results is that using the composite as an explanatory variable in a regression analysis still leads to an inconsistent ordinary least squares (OLS) estimator of the coefficient for the latent variable that the composite reflects. If the composite is the only variable measured with error, then the coefficient estimated for that variable will tend to be too low. In the more realistic situations of more than one explanatory variable containing error, the coefficient estimates can tend to be downwardly or upwardly "biased" (Bollen, 1989, chap. 5).

What if we use a weighted, rather than unweighted, composite? How does this alter the above conclusions? If we knew λ and weighted each y_i by $1/\lambda$ we would have

$$c_2 = \eta_1 + (1/\lambda)\epsilon, \quad (11)$$

where $c_2 = (1/\lambda)c_1$. Now c_2 and η_1 have the "same scale" in the sense that we expect a one unit shift in c_2 for a one unit shift in η_1. However, this does not eliminate the random measurement error in c_1 or c_2, as is clear from the last term in Equation 11. Other weighting schemes for the ys could increase the squared correlation between c_1 (or c_2) and η_1, but except for some unlikely cases, this correlation cannot be raised to one.

With causal indicators the meaning of the linear composite shifts somewhat. Call the equally weighted linear composite for the four xs in Figure 1b c_3:

$$c_3 = x_1 + x_2 + x_3 + x_4 = \mathbf{1}'x, \quad (12)$$

where $\mathbf{1}' = [1\ 1\ 1\ 1]$ and $x' = [x_1\ x_2\ x_3\ x_4]$. Recall that η_1 is

$$\eta_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3 + \gamma_{14}x_4 + \zeta_1 = \Gamma'x + \zeta_1, \quad (13)$$

where $\Gamma' = [\gamma_{11}\ \gamma_{12}\ \gamma_{13}\ \gamma_{14}]$. Comparing Equation 12 with Equation 13, one can see that c_3 only equals η_1 in the unlikely case that all the γs are one, and ζ_1 is zero. More generally c_3 does not equal η_1.

We just examined the squared correlation of the equally weighted composite of effect indicators and the latent variable that they measure. We can do the same for the causal indicators. Using the definition of the squared correlation, we get

$$\rho^2_{c_3\eta_1} = [\mathrm{COV}(c_3, \eta_1)]^2/[\mathrm{VAR}(c_3)\mathrm{VAR}(\eta_1)]. \quad (14)$$

Substituting Equations 12 and 13 into Equation 14 and simplifying we get

$$\rho^2_{c_3\eta_1} = [\mathbf{1}'\Sigma_{xx}\Gamma]^2/[\mathbf{1}'\Sigma_{xx}\mathbf{1}(\Gamma'\Sigma_{xx}\Gamma' + \Psi_{11})], \quad (15)$$

where Σ_{xx} is the population covariance matrix of the xs and Ψ_{11} is the variance of ζ_1. As long as Γ does not equal 1 and $\Psi_{11} \neq 0$, this squared correlation is less than one. Introducing c_3 as a proxy for an explanatory latent variable in an OLS regression analysis leads to an inconsistent estimator of the coefficient for η_1.

The similarities between linear composites of effect indicators or causal indicators are intriguing. In both cases the linear composite has less than perfect correlation with the latent variable η_1. In both cases this correlation can be altered by weighting the indicators. Even with optimal weighting, the correlation of η_1 with the composites is reduced from one due to error variance. But, the source of error variance differs by type of indicator. For effect indicators, the origin of the error in the composite is the measurement error in the indicators. For the causal indicators, error variance is due to the equation disturbance ζ_1, which affects η_1, but which is uncorrelated with the xs. Regardless of these differences, the linear composite is not equivalent to the latent variable.

Conventional Wisdom in Practice

Until now we have nearly exclusively dealt with analytic results to demonstrate the flaws in several conventional beliefs about measurement. In this section we illustrate our analytic findings with several psychological examples to clarify the implications of our work for applied research. We chose these examples because they are representative of much of psychological measurement efforts. In some cases we find that the procedures followed by researchers are consistent with the analytic results. In other cases they are not. But in all cases the measurement model is not formally specified.

Example 1

Watson and Friend's (1969) Fear of Negative Evaluation Scale (FNE) has 30 items that are listed in Example 1 in the Appendix. Although the authors did not present a formal measurement model in their original article, they did make it clear that they relied heavily on internal consistency for item selection. Our earlier analytic results showed that using internal consistency as a standard implies effect indicators, an unidimensional construct, and uncorrelated measurement errors. The items seem to reflect the latent variable. For instance, we expect that as a person's fear of negative evaluation increases, so will his or her agreement with the item "I am afraid that people will find fault with me." Thus, effect indicators rather than causal indicators seem reasonable. The unidimensionality of the construct measured by the indicators seems plausible. All items address the fear of being evaluated by others and use the words "worry" and "afraid" frequently. Though we could conceive of arguments suggesting more than one dimension underlying these items, to simplify matters we accept the unidimensionality of the construct. Finally, although correlated errors are possible among items using similar wordings or appearing near to each other on the questionnaire, for purposes of discussion we assume that errors are generally independent of one another.

How internally consistent should the items be? The homogeneity of the items translates directly into extremely high inter-

nal consistency with coefficient alphas routinely ranging between .92 and .96 (Watson & Friend, 1969; Wolfe, Lennox, Welch, & Cutler, 1984). In Wolfe et al.'s analysis, the scale also produced an average interitem correlation of .39, which according to Briggs and Cheek's (1986) prescription for optimal correlations, is only marginally in the zone of tolerance and therefore risks narrowing the construct given that by their definitions they may correlate too highly. In contrast our analytic results showed that for effect indicators of unidimensional constructs, the higher the correlations are, the better. So we see no problems with moderately large correlations among these items.

The large number of items in the FNE scale has led some to suggest that the scale be shortened (Leary, 1983: Wolfe et al., 1984). Some might think that removing indicators destroys the breadth of the construct being measured and can introduce inaccuracies. But, reference to our section on removing indicators reveals that no effect indicator is indispensable in measuring a unidimensional construct. And removal may have only minor effects on the overall reliability of the scale when there are many highly correlated items. The coefficient alpha for Leary's 12-item FNE and Wolfe et al.'s 14-item version were both .90.

Example 2

Berscheid, Snyder, and Omoto (1989) recently devised a scale to assess the impact of diversity of activities in relationships. They did so by counting the number of different specific activities that partners in a relationship performed on a regular basis. Their measure, a 38-item checklist, contains a broad range of activities such as "attended class," "went to a movie," and "went to a party." The items are listed in Example 2 in the Appendix. The authors do not define a measurement model that specifically states the hypothesized relations between the latent variable and items. But, it makes most sense to view these items as causal indicators of interaction diversity rather than effect indicators of this construct. We expect the diversity to increase as more of these activities are engaged in rather than the reverse. For instance, if a couple "engaged in sexual relations" and "went to a grocery store," these increase the diversity of the activities. But, we do not expect that an increase in diversity causes sexual relations and trips to the grocery store.

If we accept that these are causal indicators, our earlier discussion establishes that the items in the diversity of activities in relationships scale need not be internally consistent or have a high reliability coefficient. As the following quote indicates, Berscheid et al. (1989) seem at least partially aware that internal consistency might not be necessary:

> Because of the heterogeneity of the behavioral domains sampled, high internal consistency would not necessarily be expected. We did, however, compute ... a measure of internal consistency for this 38-item scale of dichotomous responses and found that the scale is internally consistent. (Berscheid et al., 1989, p. 795)

Perhaps under the pressure of current conventions Berscheid et al. (1989) report internal consistency. Reliability of causal indicators is not demonstrated by internal consistency (Bollen, 1989, chap. 6). Thus, the justification for reporting such measures is unclear.

Example 3

In the previous two examples the utility of item correlations in evaluating reliability and validity are relatively straightforward. In the FNE Scale, item correlations are expected to be high, and indicative of good reliability and low measurement error. On the other hand, the relationship diversity measures are best conceived of as causal indicators and therefore need not correlate. Indeed the correlations of the items offer no information on reliability. These expections follow from a measurement model that has a single latent variable with all items being its effects, as in the FNE, or all items representing its causes, as in the relationship diversity case. When we have a construct that has a mix of effect and causal indicators, the situation is further complicated. The Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977) is a case in point.

The CES-D is a 20-item self-report measure constructed by collecting "good" items from several previously validated depression scales including the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961); Minnesota Multiphasic Personality Inventory; the Self-Rating Depression Scale (Zung, 1965); and other unpublished measures. Example 3 in the Appendix lists the measures. Items were selected to operationalize major components of depression identified in the literature and include depressed mood, feelings of guilt and worthlessness, feelings of helplessness and hopelessness, psychomotor retardation, loss of appetite, and sleep disturbance. These descriptors suggest that CES-D items measure a multidimensional and not unidimensional construct of depression. Furthermore, the items appear to be a mixture of effect and causal indicators. For instance, "I felt depressed" and "I felt sad" appear to be effect indicators of depressed mood. That is, we expect that a change in the latent depression variable leads to a change in responses to these items. However, "I felt lonely" could be a causal indicator of the same construct in that loneliness may cause depression rather than vice versa. Alternatively, loneliness could be a separate dimension that requires several indicators to measure it. To further complicate things, one could argue that some items are reciprocally related to depression. For example, individuals may become depressed because they think people dislike them, which makes them appear unattractive, thus other people may avoid them and actualy dislike them (i.e., the low affect may be unattractive and offputting).

Radloff (1977) used Cronbach's alpha and assessed the scale's reliability by examining the internal consistency of items. Considering the multidimensionality of the construct, the presence of causal indicators, and the possibility of reciprocally related indicators, inspecting item correlations, internal consistency, and reliability estimates are difficult to rationalize. If depression was hypothesized to be a single unidimensional construct that had several effect indicators with uncorrelated errors, then our previous analytic results would justify a check of internal consistency for this subset. But, certainly we cannot expect the same for the mixture of items that are used in the CES-D scale.

Conclusions

We examined five conventional beliefs about measurement, none of which held without qualification. We found the following: (a) The claim that valid measures of a unidimensional

construct must be internally consistent is true for effect indicators but not for causal indicators; (b) the prescription of optimal levels of correlations for indicators often is misguided; (c) the idea that indicators must tap all facets of an unidimensional concept makes sense for causal indicators but not for effect ones; (d) comparing correlations of indicators within and between constructs can be misleading regardless of the indicator type; and (e) linear composites of indicators are not the same as the latent variable with which they are associated.

Conventional wisdom on item selection and scale evaluation is thus shown to be qualified by consideration of the specific directional relationship between the indicators and the latent construct. Traditional measures of reliability and the examination of the correlation matrix of indicators are so ingrained that researchers have failed to realize that these are not appropriate under all situations. We presented several examples to illustrate these points.

Some of our results hold the potential for abuse. For instance, the causal indicators model seems to offer a handy excuse for low internal consistency, but the model is not without limitations. 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 (Bollen, 1989, chap. 6). For example, we suggested that the diversity of social interactions measures proposed by Berschied et al. (1989) are causal indicators. To estimate a model with these causal indicators would necessitate the inclusion of consequences of the social interactions diversity construct such as exposure to socially transmited illnesses or ideas. Similarly, depression as measured by the CES-D scale might be linked to suicidal ideation or alcohol consumption. The point is that causal indicators are not invalidated by low internal consistency so to assess validity we need to examine other variables that are effects of the latent construct.

We also should emphasize that our models relating latent and observed variables are simplified. Many indicators are probably influenced by more than one latent variable, whereas we kept one latent variable per indicator. In addition, correlated measurement errors are likely, but our models assume uncorrelated errors. Adding these characteristics would complicate our demonstrations but the basic message would be the same: The applicability of the conventional practices of indicator selection are potentially misleading.

We do not recommend new guidelines based on indicator or factor correlations. Rather we advocate that researchers specify a model to relate their indicators to latent variables. In developing a model researchers should not automatically confine themselves to the unidimensional classical test model. Causal indicator models, multidimensional models, and other alternative specifications are in some cases more suitable. Once specified and identified, structural equation procedures can test whether the proposed model is consistent with the data. Structural equation models alone cannot lead to correct indicator selection. However, formal model specification relating observed and latent variables determines which pieces of conventional wisdom are appropriate.

References

Althauser, R. P., & Heberlein, T. A. (1970). A causal assessment of validity and the multitrait-multimethod matrix. In E. Borgatta (Ed.), *Sociological methods* (pp. 151–169). San Francisco: Jossey-Bass.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Berscheid, E., Snyder, M., & Omoto, A. M. (1989). The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology, 57,* 792–807.

Blalock, H. M. (1964). *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press.

Bollen, K. A. (1984). Multiple indicators: Internal consistency or no necessary relationship? *Quality and Quantity, 18,* 377–385.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the evaluation of personality scales. *Journal of Personality, 54,* 106–148.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cattell, R. B. (1965). *The scientific analysis of personality.* New York: Penguin Books.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360–392.

Hauser, R. M. (1973). Disaggregating a social-psychological model of educational attainment. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 255–289). San Diego, CA: Academic Press.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont, CA: Wadsworth.

Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika, 57,* 239–251.

Leary, M. (1983) A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin, 9,* 371–375.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401.

Selltiz, C., Wrightsman, L. S., & Cook, S. W. (1976). *Research methods in social relations* (3rd ed.). New York: Holt, Rinehart & Winston.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15,* 663–671.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Watson, D., & Friend, R. (1969). Measurement of social–evaluative anxiety. *Journal of Consulting and Clinical Psychology, 33,* 448–457.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin, 74,* 193–212.

Wolfe, R. N., Lennox, R. D., Welch, L. K., & Cutler, B. L. (1984, April). *A shortened version of the Fear of Negative Evaluation Scale.* Paper presented at the 55th annual meeting of the Eastern Psychological Association, Baltimore.

Zung, W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12,* 63–70.

Appendix

Example 1: The Fear of Negative Evaluation Scale (Watson & Friend, 1969)[a]

1. I rarely worry about seeming foolish to others.[b]
2. I worry about what other people will think of me when I know it doesn't make any difference.
3. I become tense and jittery if I know someone is sizing me up.
4. I am unconcerned even if I know people are forming an unfavorable impression of me.[b]
5. I feel very upset when I commit some social error.
6. The opinions that important people have of me cause me little concern.[b]
7. I am often afraid that I may look ridiculous or make a fool of myself.
8. I react very little when other people disapprove of me.[b]
9. I am frequently afraid of other people noticing my shortcomings.
10. The disapproval of others would have little effect on me.[b]
11. If someone is evaluating me I tend to expect the worst.
12. I rarely worry about what kind of impression I am making on someone.[b]
13. I am afraid that others will not approve of me.
14. I am afraid that people will find fault with me.
15. Other people's opinions of me do not bother me.[b]
16. I am not necessarily upset if I do not please someone.[b]
17. When I am talking to someone, I worry about what they may be thinking of me.
18. I feel that you can't help making social errors sometimes so why worry about them.[b]
19. I am usually worried about the kind of impression I make.
20. I worry a lot about what my superiors think of me.
21. If I know someone is judging me, it has little effect on me.[b]
22. I worry that others will think I am not worthwhile.
23. I worry very little about what others may think of me.[b]
24. Sometimes I think I am too concerned with what other people think of me.
25. I often worry that I will say or do the wrong thing.
26. I am often indifferent to the opinions other have of me.[b]
27. I am usually confident that others will have a favorable impression of me.[b]
28. I often worry that people who are important to me won't think very much of me.
29. I brood about the opinions my friends have about me.
30. I become tense and jittery if I know I am being judged by my superiors.

[a] From "Measurement of social–evaluative anxiety" by D. Watson and R. Friend, 1969, *Journal of Consulting and Clinical Psychology, 33,* p. 450. Copyright 1969 by the American Psychological Association. Reprinted by permission.

[b] Indicated reversed-scoring.

Example 2: The Relationship Closeness Inventory–Diversity Subscale (Berscheid, Snyder, & Omoto, 1989)[a]

1. did laundry
2. prepared a meal
3. watched TV
4. went to an auction/antique show
5. attended a non-class lecture or presentation
6. went to a restaurant
7. went to a grocery store
8. went for a walk/drive
9. discussed things of a personal nature
10. went to a museum/art show
11. planned a party/social event
12. attended class
13. went on a trip (e.g., vacation or weekend)
14. cleaned house/apartment
15. went to church/religious function
16. worked on homework
17. engaged in sexual relations
18. discussed things of a non-personal nature
19. went to a clothing store
20. talked on the phone
21. went to a movie
22. ate a meal
23. participated in a sporting activity
24. outdoor recreation (e.g., sailing)
25. went to a play
26. went to a bar
27. visited family
28. visited friends
29. went to a department, book, hardware store, etc.
30. played cards/board game
31. attended a sporting event
32. exercised (e.g., jogging, aerobics)
33. went to an outing (e.g., picnic, beach, zoo, winter carnival)
34. wilderness activities (e.g., hunting, hiking, fishing)
35. went to a concert
36. went dancing
37. went to a party
38. played music/sang

[a] From "The Relationship Closeness Inventory: Assessing the Closeness of Interpersonal Relationships" by E. Berscheid, M. Snyder, and A. M. Omoto, 1989, *Journal of Personality and Social Psychology, 57,* p. 806. Copyright 1989 by the American Psychological Association. Adapted by permission.

Example 3: The Center for Epidemiological Studies Depression Scale (Radloff, 1977)[a]

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I felt that I could not shake off the blues even with help from my family or friends.
4. I felt that I was just as good as other people.
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt that everything I did was an effort.
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I was happy.
13. I talked less than usual.
14. I felt lonely.
15. People were unfriendly.
16. I enjoyed life.
17. I had crying spells.
18. I felt sad.
19. I felt that people dislike me.
20. I could not get going.

[a] From "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population" by L. S. Radloff, 1977, *Applied Psychological Measurement, 1*, p. 387, Copyright 1977 Applied Psychological Measurement, Inc., under agreement with West Publishing Company. Reproduced by permission.

Received September 20, 1990
Revision received January 11, 1991
Accepted January 19, 1991 ■

1992 APA Convention "Call for Programs"

The "Call for Programs" for the 1992 APA annual convention will be included in the October issue of the *APA Monitor*. The 1992 convention will be held in Washington, DC, from August 14 through August 18. Deadline for submission of program and presentation proposals is December 13, 1991. Additional copies of the "Call" will be available from the APA Convention Office in October.