Published in final edited form as:

Psychol Methods. 2011 September ; 16(3): 265–284. doi:10.1037/a0024448.

Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates

Kenneth A. Bollen[1] and
University of North Carolina at Chapel Hill

Shawn Bauldry
University of North Carolina at Chapel Hill

Abstract

In the last two decades attention to causal (and formative) indicators has grown. Accompanying this growth has been the belief that we can classify indicators into two categories, effect (reflective) indicators and causal (formative) indicators. This paper argues that the dichotomous view is too simple. Instead, there are effect indicators and *three* types of variables on which a latent variable depends: causal indicators, composite (formative) indicators, and covariates (the "three Cs"). Causal indicators have conceptual unity and their effects on latent variables are structural. Covariates are not concept measures, but are variables to control to avoid bias in estimating the relations between measures and latent variable(s). Composite (formative) indicators form exact linear combinations of variables that need not share a concept. Their coefficients are weights rather than structural effects and composites are a matter of convenience. The failure to distinguish the "three Cs" has led to confusion and questions such as: are causal and formative indicators different names for the same indicator type? Should an equation with causal or formative indicators have an error term? Are the coefficients of causal indicators less stable than effect indicators? Distinguishing between causal and composite indicators and covariates goes a long way toward eliminating this confusion. We emphasize the key role that subject matter expertise plays in making these distinctions. We provide new guidelines for working with these variable types, including identification of models, scaling latent variables, parameter estimation, and validity assessment. A running empirical example on self-perceived health illustrates our major points.

Keywords

causal indicators; composite indicators; measurement; formative indicators; structural equation modeling; factor analysis

Introduction

Over a century ago Spearman (1904) analyzed a group of indicators hypothesized to be functions of individuals' intelligence, the first instance of factor analysis. In his analysis Spearman treated the latent variable as a determinant of the measures (i.e., *effect* or *reflective* indicators). Treating the intelligence indicators as effect indicators made sense because the scores should theoretically respond to changes in intelligence rather than vice versa. As factor analysis spread to diverse subject matters, researchers typically maintained the implicit assumption that their measures were effect indicators. Analyses involving

[1]The authors gratefully acknowledge Eli Lilly and Company for providing funding and we thank Rik Lennox, PhD, Mark E. Boye, PhD, the reviewers, and editor for helpful feedback on this paper.

classical test theory, item response theory, and structural equation models (SEMs) nearly always share the same assumption.

The assumptions underlying effect indicators appear reasonable when measuring most attitudes and beliefs. For some topics, however, such as exposure to stressful life events, socioeconomic status, or the amount of social interaction, the assumptions underlying effect indicators become questionable. For instance, a set of stress indicators including changing jobs, getting married or divorced, and moving to a new home do not conform to the idea that the latent variable (exposure to stress) causes the indicators. One might prefer to think of these life events as causes of stress, in which case the measures are best considered *causal* indicators of exposure to stress.

H. M. Blalock (1964, 1974), a sociologist, appears to be the first to describe the distinction between what he called cause and effect indicators. A few studies followed that examined some of the identification problems associated with causal indicators (Land, 1970), how to estimate the effects of blocks of variables or indicators (Heise, 1972; Marsden, 1982), and paradoxes about the behavior of multiple causal indicators verses multiple effect indicators (Bollen, 1984). In addition, the idea of causal indicators spread from sociology to marketing where the similar but *not* identical idea of *formative* indicators was adopted (Fornell & Bookstein, 1982). Formative indicators, or *composite* indicators (Grace & Bollen, 2008), operate as contributors to a composite variable that is a linear combination of the composite indicators.[2] Sometimes the composite variable is referred to as a "latent" variable, but the researcher assumes the composite indicators perfectly determine the composite (i.e., the latent variable's disturbance is zero).

Following the work of Bollen and Lennox (1991) an awareness of causal indicators as an alternative to effect indicators has rapidly expanded. This has led to a number of empirical studies that include causal indicators as components of the estimated models (e.g., Atkinson & Lennox, 2006; Fayers & Hand, 1997, 2002; Perreria, Deeb-Sossa, Harris, & Bollen, 2005; Glanville & Paxton, 2007). In addition, our understanding of many of the methodological aspects of working with causal indicators has increased.

The ability to distinguish between causal and effect indicators is one area that has seen some progress. A number of researchers have outlined and refined conceptual checks that differentiate causal and effect indicators (Bollen, 1989; Cohen, Cohen, Teresi, Marchi, & Velez, 1990; Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; Diamontopoulos & Winklhofer, 2001; Fayers & Hand, 2002; Jarvis, MacKenzie, & Podsakoff, 2003; Williams, Edwards, & Vandenberg, 2003). In addition, researchers have developed empirical tests involving vanishing tetrads to distinguish between causal and effect indicators (Bollen & Ting, 2000). Establishing the identification of models involving causal indicators has been another active area of research (Bollen, 1989; MacCallum & Browne, 1993; Bollen & Davis, [1994] 2009a, 2009b). Finally, researchers have explored how to assess the validity of causal indicators (Bollen, 1989; Diamantopoulos & Winklhofer, 2001; Fayers & Hand, 2002).

In addition, to causal and composite indicators, *covariates* are another type of variable that can enter these nontraditional measurement models. These are virtually absent from discussions of causal or composite indicators. Of course, covariates are present in models that exclusively consist of observed variables, but our emphasis is on covariates in models with at least some latent variables. In this context, covariates are variables that are not

[2]We use the term "composite indicators" to refer to the variables that are the components of a composite variable. This terminology departs from the usual notion that an "indicator" is in reference to a latent variable as the composite variable is observed.

measures of a latent variable, but they are nonetheless important to include to avoid bias estimates of the relations between latent variable(s) and indicators. One might think of the impact of some covariates as being less immediate, or more distal, than causal indicators, but their associations with the latent variable and its causal indicators necessitates their inclusion to avoid potential omitted variable bias. For instance, we might hypothesize that certain demographic groups are more likely to be exposed to stress so that variables representing these groups have direct effects on stress exposure even after controlling for causal indicators such as job loss, divorce, moving, etc. Failing to take account of these demographic groups could bias our estimates of the impact of job loss, divorce, moving, etc. on stress exposure.

Confusion remains about the similarities and differences between causal indicators, composite (formative) indicators, and covariates. For instance, in a recent exchange in *Psychological Methods* concerning causal indicators the authors base some of their arguments about causal indicators on properties of composite indicators (see Howell, Breivik, & Wilcox, 2007; Bollen, 2007). As another example, a search for formative indicators on the Mplus forum produces numerous questions indicating uncertainty about the properties of causal and formative (composite) indicators. SEMNET, the primary listserv for SEM work, regularly has questions on how to identify and estimate models with causal or formative indicators. Though some authors have discussed composite indicators (e.g., Grace & Bollen, 2008), they do not distinguish these from causal indicators, formative indicators, and covariates.

Past literature has concentrated on heightening awareness among researchers that not all indicators are effect (reflective) indicators. This has supported a dichotomous view of indicators with the traditional effect indicator and the causal indicator as the two categories. Our paper takes the presence of indicators that influence latent variables for granted. But we argue that such variables are not all the same. In other words we seek to move beyond the idea that there are variables that are not effect indicators and to concentrate on better understanding the types of observed variables or indicators that affect latent variables. Distinguishing these different types of variables has two benefits for applied research. It allows a more complete framework for considering respecifications in models that do not fit the data and it provides a more precise link with theory. As we will show, the different types of variables have distinct theoretical interpretations and careful consideration of the differences among them may inform the substantive understanding of the object under study.

We have two purposes with this paper. First, we propose a distinction between causal indicators, composite (formative) indicators, and covariates, and explore the ramifications of the differences in the types of indicators. We also discuss indexes and scales in light of these distinctions between indicators. Second, we provide an overview of the advances in our understanding of working with causal and composite indicators versus covariates and, in the process, provide readers with a new guide for using such variables in their research. As we repeatedly emphasize subject matter expertise is at least as important as empirical techniques in determining the type of indicator. We limit the amount of technical details in our discussion and where possible refer readers to the relevant literature. We develop an empirical example throughout the paper to provide concrete illustrations of working with covariates, and causal, composite, and effect indicators.[3]

We divide the paper into six sections. We first provide an overview of our empirical example, before turning to a discussion of how to distinguish between covariates and effect,

[3]Following the conventions of prior literature, we assume that the latent variables and the endogenous observed variables are (or approximate) continuous variables. In our discussion exogenous observed variables could be dichotomous or continuous.

causal, and composite indicators. In the third section we review approaches to deciding whether variables are best treated as covariates, composite, causal, or effect indicators. In the fourth section we discuss the identification, options for scaling latent variables, and estimation of models involving causal or composite indicators and covariates. In the fifth section we consider methods for evaluating the quality of causal and composite indicators. Finally, in the sixth section we explore a few additional complications that can arise when working with these variables.

A Model for Latent Perceived Health

For our empirical example we create a model of perceived health. The data come from the General Social Survey (GSS). The GSS is a nationally representative repeated cross-sectional survey conducted roughly every year between 1972 and 1994 and then every other year since. Each cross-section consists of a multi-stage stratified random sample of non-institutionalized adults age 18 and older involving approximately 1,500 to 3,000 respondents.

In 1991 the GSS asked a series of questions to a subset of respondents concerning their health. We selected the following five questions to use as illustrations in various models: (1) ill enough to see the doctor in the last year, (2) hospitalized or disabled in the last year, (3) hospitalized or disabled in the last five years (not including last year), (4) self-reported health, and (5) satisfaction with health (see Table 1 for descriptive statistics). In addition, we selected a few demographic variables (age, sex, and race) that are likely to have an impact on perceived health and are therefore important to include as covariates or control variables.

For illustrative purposes we consider a set of three models. One model treats all five measures as effect indicators (see Figure 1, Panel A). This is perhaps the default model most researchers would begin with. The second model treats the first three indicators as causal indicators and the final two indicators as effect indicators (see Figure 1, Panel B). This model, commonly referred to as a multiple indicator-multiple cause or MIMIC model (Jöreskog & Goldberger, 1975), is most consistent with our theoretical understanding of the relationships among the indicators. The third model treats the first three variables as composite indicators and the last two indicators as effect indicators (see Figure 1, Panel C). In Figure 1, Panel C we use a hexagon to represent a composite variable, an exact linear function of the three composite indicators (Grace and Bollen (2008) introduced this notation). It could just as well be represented as a variable in a rectangular box since a composite is simply a linear combination of its component observed variables. However, to emphasize its special role as a weighted combination of other observed variables, we use the hexagon.[4]

Researchers often adjust for potential confounding effects by including covariates that act as control variables in their models. We illustrate this in some of our models by allowing the demographic covariates to have a direct effect on latent health. In the MIMIC models and the models with a composite variable we also allow the demographic covariates to correlate with the causal or composite indicators.

[4]If we were to remove the composite variables from this model and had the *x*s directly influencing the *y*s, the path diagram could represent a multivariate regression with two dependent variables and uncorrelated errors. The presence of the composite makes for a difference in that the weights for the *x*s determine the composites and those weights need not come from regression estimates. For example, a researcher might use equal weights for the composite indicators to form the composite scale.

Types of Indicators and Covariates

In this section we distinguish between effect, causal, and composite indicators, and covariates. Researchers tend to treat causal, composite, and formative indicators as essentially the same. As our later discussion explains, this has led to considerable confusion. The distinction between effect indicators and the latter three variables is sharp whereas the difference between causal indicators, composite indicators, and covariates is less clean. In the next series of subsections we describe each type of variable and then discuss how to distinguish between them.

Effect Indicators

Effect indicators are the most common type of indicators in social and behavioral science research. An effect indicator model is illustrated in Figure 2, Panel A. The dashed lines represent additional observed variables. The model equations are

$$\begin{aligned} y_{1i} &= \alpha_1 + \lambda_{11}\eta_{1i} + \varepsilon_{1i} \\ y_{2i} &= \alpha_2 + \lambda_{21}\eta_{1i} + \varepsilon_{2i} \\ &\vdots \\ y_{Pi} &= \alpha_P + \lambda_{P1}\eta_{1i} + \varepsilon_{Pi} \end{aligned}, \quad (1)$$

where y_{pi} is the *p*th observed variable or indicator that depends on the latent variable $\eta_{1\mathrm{i}}$, and α_{p} are intercepts. The *structural* coefficients λ_{p1} give the expected impact of a one-unit difference in $\eta_{1\mathrm{i}}$ on y_{pi}. We assume $\mathrm{E}[\varepsilon_{\mathrm{pi}}] = 0$ for all *i* and *p*, and $\mathrm{COV}(\eta_{1\mathrm{i}}, \varepsilon_{\mathrm{pj}}) = 0$ for all *i*, *p*, and *j*.

Effect indicators are chosen or created to correspond to the theoretical definition of the concept that the latent variable represents, as such effect indicators may be considered “manifestations” or “demonstrations” of the latent variable The effect indicators of the same latent variable have conceptual unity in that they all correspond to the same dimension of a concept. This common dependence on a single latent variable generally creates an association among the effect indicators (Bollen, 1984). Factor analysis, item response theory, reliability, and validity work in the social sciences all implicitly assume effect indicators.

Causal Indicators

Causal indicators are another general class of indicators. An equation to represent them is

$$\eta_{1i} = \alpha_\eta + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \cdots + \gamma_{1Q}x_{Qi} + \zeta_{1i}, \quad (2)$$

where $\eta_{1\mathrm{i}}$ is a latent variable for the *i*th case, x_{qi} is the *q*th observed variable or indicator, and α_η is the intercept. The coefficients, $\gamma_{1\mathrm{q}}$, are *structural* coefficients that give the expected impact of a one-unit difference in x_{qi} on $\eta_{1\mathrm{i}}$ while holding constant all other *x*s. The term “structural” is important in that it signifies an effect that the researcher assumes to be an invariant causal effect (Goldberger, 1973, pp. 2–6; Bollen, 1989, p. 4). Whether coefficients can reasonably be interpreted as causal must be justified separately from the SEM in which they are embedded. For further discussion see Wright (1934, p. 193) and Pearl (2000, pp. 135–138).

The disturbance, $\zeta_{1\mathrm{i}}$, includes all other determinants of $\eta_{1\mathrm{i}}$ not included in the equation. We assume $\mathrm{E}[\zeta_{1\mathrm{i}}] = 0$ for all *i*, and $\mathrm{COV}(x_{\mathrm{qi}}, \zeta_{1\mathrm{j}}) = 0$ for all *q*, *i*, and *j*. (e.g., Bollen, 1984, 1989). This model assumes that the observed variables influence a single latent variable (see Figure

2, Panel B). Like the effect indicators, the causal indicators have conceptual unity in that they should correspond to the definition of the dimension of the concept that the latent variable represents. The inclusion of the disturbance ζ_{1i} means that the causal indicators do not completely determine the latent variable.

Composite Indicators

Composite indicators are weighted elements that form a composite variable for which there is no disturbance term. That is, the composite variable is an exact linear combination of the composite indicator variables. But beyond having no disturbance, the composite indicator coefficients are not structural or causal coefficients. Rather their coefficients are weights to apply to form the composite variable that is made up of them. To clearly separate these composite indicators from the causal indicators, we represent their coefficients with *w*, a symbol to stand for a weight that is applied to the variable that follows it, or in the case of the intercept it stands for a constant. Figure 2, Panel C pictures a model with composite indicators. The equation for this path diagram is

$$C_{1i}=w_{10}+w_{11}x_{1i}+w_{12}x_{2i}+\cdots+w_{1Q}x_{Qi}, \quad (3)$$

where C_{1i} is a composite variable for the *i*th case. *Formative* indicators as originally defined by Fornell and Bookstein (1982) are the same as composite indicators: "…when constructs are conceived as explanatory combinations of indicators (such as 'population change' or 'marketing mix') which are determined by a combination of variables, their indicators should be formative" (442). Current usage of formative indicators is ambiguous in that sometimes it refers to causal indicators as defined above and other times as its original meaning of composite indicators as defined in this subsection. We stay with the terms of causal indicators and composite indicators to avoid this confusion.

Composite indicators do not necessarily have conceptual unity, but can be an arbitrary combination of variables. Heise's (1972) sheaf coefficient is an example of the use of composite indicators. He proposed the sheaf coefficient as a convenient way of combining the effects of several variables into a single composite whose standardized coefficient can then be compared to the standardized coefficient of other single or composite variables in the model. Combining the effects of dummy variables indicating ethnic group into a sheaf coefficient is one example from Heise (1972), but a researcher can form any set of variables into a composite. Furthermore, the sheaf coefficient is not necessary in an analysis with composite indicators in that the researcher can interpret the variables' influences in their original metric.

Covariates (Controls)

Sometimes there are variables that are *not* measures of a latent variable, but they are associated both with the latent variable and its indicators in such a way that their omission would bias estimates of the relations between latent variables and indicators. We refer to these as covariates. Covariates regularly appear in models that include only observed variables, so their use and meaning extends beyond the measurement context of concern to us. We remind the reader that our focus on covariates is in the context of measurement models.

Some but not all covariates might cause the latent variable, even though they are not causal indicators of the latent variable. For instance, researchers often include demographic variables in a model as covariates rather than as causal or composite indicators. Age, sex, and race may all directly influence a latent variable even though we would not consider

them indicators of the latent variable. Nonetheless, it is important to include covariates, which may have less proximate impacts than the causal indicators, to avoid the potential for omitted variable bias. If a model includes both covariates and causal indicators, then one can interpret some of the *x*s in equation (2),

$$\eta_{1i}=\alpha_{\eta}+\gamma_{11}x_{1i}+\gamma_{12}x_{2i}+\cdots+\gamma_{1Q}x_{Qi}+\zeta_{1i}, \quad (4)$$

as covariates rather than causal indicators. For example, in our perceived health model we sometimes include the covariates of age, sex, and race along with the causal indicators. Sometimes the covariates' coefficients are structural in that they represent separate (typically more distal) causes of the latent variable. Other times the covariates are entered to control for omitted variable bias and their coefficients might not have a clear causal interpretation.

In other models a researcher might specify covariates as causes of the causal indicators such that the covariates would have only indirect or both direct and indirect effects on the latent variables. However, other times researchers can treat the covariates and the causal indicators as correlated exogenous variables without specifying direct effects among them. It is unlikely that covariates would be included among a set of composite indicators in that composite variables are intended to form the composite whereas covariates are intended to control for omitted factors that might bias structural coefficients.

Distinguishing the Three Cs and Effect Indicators

The last section made preliminary distinctions between the different types of variables. This section brings into sharper focus the differences and similarities of covariates and of effect, causal, and composite indicators. As we make clear, theory and substantive knowledge are critical to separating the indicators from each other. We begin by discussing the Three Cs: covariates, and causal and composite indicators. Then we contrast these with effect indicators.

Contrasting the Three Cs

A formal distinction between the causal indicators, composite indicators, and covariates is not always easy. Instead, we can provide informal guidelines to distinguish between causal indicators, composite indicators, and covariates while recognizing that there are situations in which the lines between these blur. The theoretical framing of the analysis is the primary factor in determining the variable type.

Consider causal indicators first. Causal indicators are variables that correspond to the conceptual meaning underlying a latent variable. The latent variable represents a dimension of a concept that has been defined. The causal indicators are variables that measure that dimension of the concept as so defined. If we were to define general socioeconomic status (SES) as the overall standing of a person in the stratification system of a society, then education, income, and occupational prestige would be causal indicators of SES. Or in our running example if we define perceived overall health as an individual's overall perception of his or her well-being, we have indicators of whether the person has seen a doctor in the last year, has been hospitalized or disabled in the last year or in the four years prior to last year. These three variables correspond to the latent variable of perceived health and hence are plausible as causal indicators. We also see these variables as causes of perceived health such that a difference in any one of these variables plausibly causes a change in perceived health. The issue of whether indicators are causal or effect indicators is a separate issue that we turn to later. But for the moment assuming that these variables influence the latent

variable, we would consider them as causal indicators since they correspond to the theoretical definition of the latent variable and are hypothesized to cause changes in perceived health.

Another aspect of causal indicators is that their impacts on a latent variable should be relatively stable across different outcomes of the latent variable. This invariance is an essential part of a structural effect. For instance, suppose that we have three causal indicators of η_{1i} as in

$$\eta_{1i}=\alpha_1+\gamma_{11}x_{1i}+\gamma_{12}x_{2i}+\gamma_{13}x_{3i}+\zeta_{1i}, \quad (5)$$

and that η_{1i} affects η_{2i} and η_{3i}

$$\begin{aligned}\eta_{2i}&=\alpha_2+\beta_{21}\eta_{1i}+\zeta_{2i}\\ \eta_{3i}&=\alpha_3+\beta_{31}\eta_{1i}+\zeta_{3i}\end{aligned}, \quad (6)$$

where we assume that the model is identified by having additional effect indicators of η_{2i} and η_{3i} that are not shown. By setting $\beta_{21} = 1$ and $\alpha_2 = 0$ we can provide a scale for η_{1i}.[5] If x_{1i}, x_{2i}, and x_{3i} are causal indicators, then their coefficient estimates should not vary beyond sampling error in the model shown compared to a model where only η_{2i} is included or only η_{3i} is included. Of course, some differences would be imposed through differences in the scaling of η_{1i}, but the only other source of difference should be sampling fluctuations. In other words, the outcomes of η_{1i} should not have much influence on the coefficients (i.e., the γs) of the causal indicators. If they do fluctuate beyond the amount expected by chance, then this is evidence that the model structure is misspecified or that these are not causal indicators of a single latent variable.

Covariates are a second type of variable. Covariates are variables which are associated with the latent variable and the indicators in such a way that their omission would bias estimates of the relations between the indicators and latent variables. Covariates are not linked to the latent variable through its theoretical definition. Typically, each covariate is a “stand alone” variable that needs to be controlled to prevent omitted variable bias in the structural coefficients of the causal indicators. In those situations where a covariate is a cause of the latent variable, the effects of covariates are often less proximate, or more distal, than causal indicators.

Returning to our example of overall perceived health, we hypothesize that whether the person has seen a doctor in the last year (x_{1i}), has been hospitalized or disabled in the last year (x_{2i}) or in the four years prior to last year (x_{3i}) are three causal indicators of overall perceived health. But in addition, we include age (x_{4i}), sex (x_{5i}), and race (x_{6i}) as covariates that might affect overall perceived health. Age, sex, and race are potential causes of perceived health, but we do not treat them as measures or causal indicators of it. Formally, our statistical model with the three causal indicators (x_{1i}, x_{2i}, x_{3i}) and three covariates (x_{4i}, x_{5i}, x_{6i}) is

$$\eta_{1i}=\alpha_\eta+\gamma_{11}x_{1i}+\gamma_{12}x_{2i}+\gamma_{13}x_{3i}+\gamma_{14}x_{4i}+\gamma_{15}x_{5i}+\gamma_{16}x_{6i}+\zeta_{1i}, \quad (7)$$

[5]Later in the paper we discuss scaling the latent variable by setting a coefficient for the causal indicator to one.

and the causal indicators appear in the equation the same as the covariates. Each coefficient is a structural coefficient. It is the theoretical context that permits us to describe one group of *x*s as causal indicators and another group of *x*s as covariates.

If we switch to a different latent dependent variable, we generally would expect to use a different set of causal indicators. A new latent dependent variable originates from a different concept. The indicators that correspond to the new latent variable are likely to differ from those of the previous latent variable, though there are instances where the same causal indicator influences more than one latent variable. Starting a new job, for example, is a plausible causal indicator of exposure to stress but also a possible causal indicator of a separate latent variable of work boredom. But more frequently a new latent variable leads to a new set of indicators.

In contrast, it is more likely that there is overlap in covariates for two different latent dependent variables. Covariates such as age, race, or gender have effects on many outcomes, which is why they might be part of an equation even with a new latent variable. Since these covariates are predicting a different outcome, we would not expect their coefficients to be the same as they were for the first dependent variable.

Composite indicators are a third type of variable. The primary question is whether a researcher hypothesizes that these indicators have direct causal or structural effects on a latent variable. If the answer is yes, then these are not composite indicators but are causal indicators or possibly covariates. Another question is whether the researcher wants to form a composite (scale or index) that is an exact linear combination of the indicators. If yes, then composite indicators are most likely. This latter condition implies that the error variance of the composite is set to zero. If a model and the error variance are identified, then the null hypothesis that the error variance is zero could be tested.[6] A zero error variance is more plausible for composite indicators than for causal indicators. In addition, being causal indicators assumes that there is conceptual unity among the measures. No conceptual unity implies no causal indicators.

But there are more subtleties to the problem. For instance, suppose that a researcher has a model with several covariates with direct structural effects on a dependent variable. For instance, suppose that η_{2i} is the outcome variable that has age, sex, race, and education as covariates with direct effects on it as in

$$\eta_{2i}=\alpha_{\eta_2}+\gamma_{21}Age+\gamma_{22}Male+\gamma_{23}Black+\gamma_{24}Education+\zeta_{2i}, \quad (8)$$

When these explanatory variables are in the model with direct effects on η_{2i}, they are covariates as per the preceding definition. In a follow-up analysis the researcher might be curious about the combined effect of age, sex, and race in contrast to education. To examine this, the analyst combines age, sex, and race into a composite. The group of variables need not have conceptual unity except in the loosest sense of the word, but the researcher wants to assess their joint impact using a composite as in

$$C_{Di}=w_C+w_{11}Age+w_{12}Male+w_{13}Black, \quad (9)$$

[6]Though theoretically possible it is highly unlikely that causal indicators would completely explain the variance of the latent variable. There has been some confusion over the testing of zero variances in SEMs concerning hypothesis tests on boundary points. If a model is estimated without inequality constraints forcing the variances to be positive, the tests are relatively straightforward. See Savalei and Kolenikov (2008) for clarification of this issue.

and to compare the coefficients of the composite to the coefficient of education as in

$$\eta_{2i}=\alpha_{\eta_2}+C_{Di}+\gamma_{24}Education+\zeta_{2i}. \quad (10)$$

Setting the coefficient of the composite to one helps to identify the model. If one were to standardize the coefficients in this model, then they would be closely related to the "sheaf coefficients" discussed by Heise (1972). In this example the composite is a convenient way of combining variables that have only a vague unity in being demographic variables. As such we would not expect the coefficients from the composite equation (9) to stay the same or even to have the same relative magnitude of effects if we were to change the outcome variable from depression (η_{2i}) to something else. This is true even taking account of any changes in the scaling of the outcome variable.

The researcher is not assuming a causal influence from the indicator to the composite, but rather the composite is a convenient weighted sum of the indicators. It is even possible that the weights for the composite indicators are set equal to the structural coefficients of the covariates from equation (8) where they were covariates. But this does not change their status as composite indicators. As long as there is a composite variable that is an exact linear combination of observed variables as in equation (9), these observed variables are composite indicators. If the composite is not in the model and the observed variables have direct effects on the outcome variable as in equation (8), then they are covariates.

Figure 3 is a path diagram to illustrate these points more fully. Figure 3a has three composite indicators (x_1, x_2, x_3), two composites (C_1, C_2) and two outcome variables (y_1, y_2). The same three composite indicators form the two different composites. Each composite is associated with a different outcome (either y_1 or y_2). The weights (coefficients) for the composite indicators differ depending on the composite (either C_1 or C_2). The weights (coefficients) could be optimized to predict a particular outcome As the outcome changes so do the weights. The exception would be if the weights were fixed in advance, such as is implicit when indexes or scales are used and the composite indicators are equally weighted or weighted in a predetermined way. In these situations, the weights are not free to vary.

If we modified this diagram to have a single composite with y_1 and y_2 as two effects of it, this would represent strong assumptions about the composite. First, it would suggest that all of the impact of the composite indicators (x_1, x_2, x_3) on y_1 and y_2 are mediated through the single composite. However, as discussed above the composite is a convenient summary of the effects of several variables. It is not intended as a mediating latent variable (with disturbance) that in turn influences several outcomes. Second, it would assume that a single set of weights for the composite indicators works well for both y_1 and y_2. In addition, it would assume that such a model would fit the data even with the implicit restriction that the composite has zero error variance.

In contrast, Figure 3b has three causal indicators (x_1, x_2, x_3), one latent variable (η_1), and two effect indicators (y_1, y_2). Here the causal indicators (x_1, x_2, x_3) and the two effect indicators (y_1, y_2) correspond to the same concept that is represented by the latent variable. The disturbance variable for the latent variable is not zero since there are other influences that affect the latent variable. In addition, in this causal indicator model we would expect the coefficients of the causal indicators to be stable regardless of whether only y_1 or only y_2 were used (other than differences due to the change of scaling of the latent variable). If changes were observed, it would be a symptom of a structurally misspecified model (Bollen, 2007).[7]

This difference between composite and causal indicators provides insight into a recent article (Howell, et al, 2007) and comment (Bollen, 2007) in this journal. Howell, et al. (2007) suggested that formative indicators were more susceptible to instability of coefficients depending on the outcome. Bollen (2007) argued that these coefficients would only change beyond sampling and scaling effects if there were structural misspecifications in the model. We could explain these different perspectives as the differences we would expect for composite (formative) versus causal indicators. The coefficients of composite indicators can change by outcome variable. The coefficients of causal indicators should be stable and if they are not it is a symptom of something wrong in the model.

Another contrast between causal and composite indicators is that there is some arbitrariness in what variables go into a composite. Unlike causal indicators where their inclusion is based on whether a variable measures the concept represented by the latent variable, the inclusion of variables in a composite is more arbitrary than being closely guided by a theoretical definition. In the previous example, a researcher could have included education as part of the demographic composite since there is not an explicit standard of what should be in or out of such a composite.

Returning once again to our example of overall perceived health, we hypothesize that whether the person has seen a doctor in the last year (x_{1i}), has been hospitalized or disabled in the last year (x_{2i}) or in the four years prior to last year (x_{3i}) are three causal indicators of overall perceived health. But in addition, we include age (x_{4i}), sex (x_{5i}), and race (x_{6i}) as covariates that might affect overall perceived health. Age, sex, and race are potential causes of perceived health, but we do not treat them as measures or causal indicators of it. Formally, our statistical model with the three causal indicators (x_{1i}, x_{2i}, x_{3i}) and three covariates (x_{4i}, x_{5i}, x_{6i}) is

$$\eta_{1i}=\alpha_{\eta}+\gamma_{11}x_{1i}+\gamma_{12}x_{2i}+\gamma_{13}x_{3i}+\gamma_{14}x_{4i}+\gamma_{15}x_{5i}+\gamma_{16}x_{6i}+\zeta_{1i}, \quad (11)$$

and the causal indicators appear in the equation the same as the covariates. Each coefficient is a structural coefficient. It is the theoretical context that permits us to describe one group of *x*s as causal indicators and another group of *x*s as covariates. At other times, the covariates are not directly influencing the latent variable, but are such that they must be taken into account to avoid biased estimates of the causal indicators' effects.

It is possible that the relation between causal indicators and their latent variable differs across categories or values of covariates. The causal indicators might have a different impact on the latent variable for men versus women or for young versus old. This would be analogous to Differential Item Functioning (DIF) in the IRT literature except here we are dealing with causal indicators rather than effect indicators. One way to test for this would be to form interactions between the suspected covariate and the causal indicator to test if the causal indicator has significantly different effects across values of the covariate. Multiple group analysis where the researcher forms separate groups (e.g., male vs. female) and tests the equivalency of the effects of the causal indicators would be another way to test the invariance of causal indicators' effects on the latent variable.

The distinction between composite indicators and covariates is different. As we described above, composites are linear combinations of their composite indicators where there is no disturbance for the composite. The coefficients associated with composite indicators are best

[7]It is theoretically possible, but extremely unlikely that the disturbance variance in Figure 3b is zero, which would then make the η_1 variable completely determined by its causal indicators. In this hypothetical case the coefficients of the causal indicators would remain structural even though the disturbance is absent.

understood as weights rather than the structural coefficients. Since the composite acts as convenient summary device, it does not make sense to consider including covariates among a set of composite indicators. Researchers select variables to construct a composite with the intent to estimate the joint (usually standardized) coefficient of the group of variables on an outcome rather than to estimate the net effect of the composite indicators on the composite controlling for other variables.[8] When we have a composite, then all of the indicators going into it are composite indicators. A covariate would then be a variable that does not enter the composite, but that does influence other endogenous variables in the model.

In brief, composite (formative) indicators are the indicators that form the composites and composites are an exact linear function of the composite indicators so that the implicit error variance is zero. Causal indicators are structural determinants of a latent variable that correspond to the definition of the latent variable and that do not completely determine it. In the context of our measurement models, covariates do not correspond to the definition of the concept that the latent variable represents, but they are variables that need to be controlled to prevent bias estimates of the causal indicators' effects on the latent variable.

Distinguishing Effect Indicators

The preceding contrasts causal indicators, composite indicators, and covariates. We have not yet compared these to effect indicators. The most basic difference between the three Cs and effect indicators is that the latent variable is a function of causal indicators, composite indicators, and covariates while the reverse is true for effect indicators.[9] Covariates and composite indicators are the easiest to distinguish from effect indicators and we treat them first.

Covariates influence a latent variable and must be controlled when assessing the impact of other variables (e.g., causal indicators) on the latent variable. Given this role as determinants of the latent variable and not representatives of the concept of a latent variable, most of the time there is little danger of mixing up covariates with effect indicators.

Composite indicators may consist of quite dissimilar variables with little conceptual unity even though the researcher is interested in their aggregated effect. Because of this, composite indicators might be positively, negatively, or even unrelated to each other. In addition, the coefficients or weights of the composite indicators are not structural coefficients of the effects of the composite indicators on the composite. Effect indicators should follow the theoretical definition of the concept that the latent variable represents and the latent variable determines the values of the effect indicators. The factor loadings from the latent variable to the effect indicators are structural coefficients in that the latent variable causes a change in the effect indicators. Generally effect indicators will be associated due to this common dependency on the same latent variable.[10]

Causal indicators often are the most difficult to separate from effect indicators. Most importantly, causal and effect indicators are measures that correspond to a theoretical definition of the concept represented by a latent variable. Hence, they are indicators that

[8]It is possible that a researcher might claim that some composite indicators are really covariate controls, but it is difficult to imagine when this would occur.

[9]In addition, the three Cs involve allowing many more covariances to be estimated than do effect indicators. This is because the three Cs are typically exogenous observed variables that are allowed to correlate whereas effect indicators are endogenous variables with a common dependence on a latent variable. These additional covariance parameters are a small cost to pay for a properly specified model.

[10]In special cases of correlated errors or factor complexity greater than one, it is possible for effect indicators not to be associated. But it is highly unusual for such complications to destroy the usual associations found when effect indicators measure the same latent variable.

have conceptual unity, something not necessary for covariates or composite indicators. To distinguish causal and effect indicators the key question is whether the indicators influence the latent variable or vice versa. Unfortunately no definitive statistical tests exist to adjudicate between the two types of indicators. There are, however, a number of conceptual and empirical checks that can help determine how best to model a set of indicators.

The first conceptual check involves conducting a *mental experiment* in which one imagines changes in the latent variable and subsequent changes in the indicators, and vice versa (Bollen, 1989, pp. 65–67; Edwards & Bagozzi, 2000, pp. 157–160; Diamontopoulos & Winklhofer, 2001, p. 271; Fayers & Hand, 2002, p. 238; Jarvis et al., 2003, p. 203; Williams et al., 2003, p. 906). If one imagines that a change in the latent variable should lead to a change in all of the observed indicators, then it would be best to treat the indicators as effect indicators. Alternatively, if one imagines a change in any given observed variable should result in a change in the latent variable, then it would be best to treat that indicator as causal.

Consider the five measures of perceived health we introduced above. Each measure seems to tap the concept of a person's subjective assessment of their well-being. That is, the five measures have conceptual unity. It seems reasonable to imagine that an increase in perceived health is likely to lead to an increase in a respondent's self-rated health and satisfaction with health. But, this is less clear with the other three measures. If a respondent visits a doctor, becomes hospitalized or disabled, then one might imagine a change in their perceived health, but a change in perceived health will not necessarily lead to all three outcomes. As a result, we might consider treating self-rated health and satisfaction with health as effect indicators and the three measures related to visiting a doctor or the hospital as causal indicators.

Another conceptual check involves evaluating how *interchangeable* or *essential* the indicators are in measuring the latent variable (Bollen & Lennox, 1991, p. 308; Diamontopoulos & Winklhofer, 2001, p. 271; Fayers & Hand, 2002, p. 245; Jarvis et al., 2003, p. 203). Given a set of effect indicators of a unidimensional concept represented by one latent variable, dropping a single indicator does not affect the relationships between the remaining indicators or their relationships with the latent variable. In this sense, one might consider effect indicators with roughly the *same* reliability and validity as interchangeable or non-essential, though we could distinguish which indicator is most closely related to the latent variable. This is not the case with causal indicators. Consider, for example, a latent variable for social interaction measured by time spent with friends, family, strangers, and coworkers. Leaving out one of the indicators alters the latent social interaction variable, potentially in a significant way. In addition, the omission of one or more of these causal indicators might bias the coefficients of the remaining causal indicators, assuming that the omitted ones correlate with the included. In contrast to effect indicators, causal indicators are not interchangeable and are usually essential.

A final conceptual check involves considering the *expected level of association* among the indicators (Bollen, 1984; Cohen et al., 1990; Bollen & Lennox, 1991; Diamontopoulos & Winklhofer, 2001; Jarvis et al., 2003). A set of effect indicators positively associated with a latent variable should all have positive correlations with each other. Causal indicators do not require such an association. Returning to our empirical example involving perceived health, one would expect a positive correlation among self-rated health and health satisfaction, but it's not clear whether one would expect a strong relationship between going to the doctor and being hospitalized.

Summary

Table 2 is a summary of the contrasts between covariates and effect, causal, and composite indicators. The columns of the table list the type of variable. The rows list possible characteristics of the variable types. A check means that the characteristic is necessary for a type of variable. No mark means that it is not needed. To illustrate, the first row is "Conceptual Unity" and only effect indicators and causal indicators are checked. This means that to be one of these types of indicators, the variable must correspond to the definition of the concept that is represented by the latent variable. Composite indicators and covariates need not correspond to the concept. Similarly, effect and causal indicators should have "stable coefficients across outcomes" of the latent variable. Sampling fluctuations or scaling differences should account for any differences. All but effect indicators have the latent variable or composite as a function of the indicator. A composite indicator is the only type where the error variance for the receiving variable should be zero. Effect indicators are the only ones predicted to necessarily correlate with each other. Finally, the next to last row lists whether omitting the indicator type leads to bias. This refers to whether the parameter estimates of the remaining variables change if an indicator type is left out. This is most clearly seen with causal indicators and covariates which are essential to maintain unbiased estimation.[11] Composite indicators' weights or coefficients are somewhat arbitrary and sometimes might even be assigned a value (e.g., equal weights) in advance. With fixed weights, the weights of the other composite indicators do not change if one or more composite indicators are absent. However, if there are other variables in the model with estimated coefficients, then the values of these coefficient estimates are likely to shift depending on what composite indicators are included. Or if the researcher estimates the weights of the composite indicators, then these estimates are likely to shift depending on which composite indicators are included. It is for these reasons that we check the row for possible bias if a composite indicator is in or out of a model.

Empirical Checks of Causal versus Effect Indicators

If the conceptual checks prove inconclusive or the researcher requires empirical evidence to support treating variables as causal or effect indicators, then there are two empirical checks to help distinguish these types of indicators. The first check involves examining the patterns of association among the indicators (Bollen, 1984; Cohen et al., 1990, pp. 187–188; Bollen & Lennox, 1991, p. 307; Diamontopoulos & Winklhofer, 2001, p. 271; Jarvis et al., 2003, p. 203). As described in the conceptual check, one would expect positive correlations among effect indicators positively associated with a latent variable. In contrast, with a set of causal indicators any type of correlation (including no correlation) is possible. Therefore, if one finds low or negative correlations among a set of indicators with positive relations to the latent variable, then this is evidence of either poor effect indicators or of causal indicators. Checking the associations among the indicators, however, may not be useful if the researcher suspects a mixture of causal and effect indicators among a set of variables, correlated errors among the indicators, or that the indicators may measure more than one concept (multidimensionality).

To illustrate these points we return to our indicators of perceived health. We find a moderately strong positive correlation (r = 0.56) between self-rated health and health satisfaction (see Table 3). We do not, however, find a clear pattern of correlations among the other variables. We see that visiting a doctor has a moderate correlation with being hospitalized or disabled last year (r = 0.34) and essentially no correlation with being

[11]If a causal indicator or covariate is uncorrelated with the other exogenous variables in the model, then its omission would not bias results. Or if the direct effect of one of these variables on the latent variable were zero, then its omission would not bias the other coefficients.

hospitalized in the four years prior to the last year (r = 0.01). Furthermore, being hospitalized last year has a negative correlation with being hospitalized in the previous four years (r = −0.23).

An alternative empirical check involves testing for vanishing (zero) tetrads among a set of indicators using Bollen and Ting's (2000) confirmatory tetrad analysis. A tetrad refers to the difference between the product of a pair of covariances and the product of another pair among four random variables. With four variables we can derive three tetrads

$$\begin{aligned}\tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24} \\ \tau_{1342} &= \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32} \\ \tau_{1423} &= \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43}\end{aligned} \quad , \quad (12)$$

where τ_{ghij} refers to the tetrad defined by $\sigma_{gh}\sigma_{ij} - \sigma_{gi}\sigma_{hj}$ and σ_{ij} refers to the population covariance between the *i*th and *j*th indicators (notation based on Kelley (1928)). A vanishing tetrad refers to those tetrads that equal zero. Different model structures often imply different sets of vanishing tetrads. This fact allows one to test the overall fit of a model by testing the implied vanishing tetrads. If an implied vanishing tetrad is significantly different than zero, then this constitutes evidence against the hypothesized model.[12]

Since testing for vanishing tetrads only involves the covariances among the indicators, it is particularly useful in helping to determine how to treat a set of indicators. A model consisting of four effect indicators implies three vanishing tetrads, while a model consisting of four causal indicators implies no vanishing tetrads [see Bollen & Ting (2000, p. 7) for a derivation]. Models with a mixture of causal and effect indicators as well as models allowing for correlated disturbances also often imply different sets of vanishing tetrads and can be tested. In addition, Bollen and Ting (2000) demonstrate how to work with models including less or more than four indicators.

Bollen (1990) provides a simultaneous vanishing tetrad test statistic that enables a researcher to test whether two or more tetrads are zero. It is the basis for confirmatory tetrad analysis (Bollen & Ting, 1993) in general, and can be used for the vanishing tetrad test for causal indicators. The test is constructed as

$$T = N\mathbf{t}^T \sum\nolimits_{tt}^{-1} \mathbf{t}, \quad (13)$$

where *N* is the sample size, **t** is a vector of the independent sample tetrad differences, and $\sum_{tt}^{-1}$ is the inverse of the covariance matrix of the limiting distribution of **t** as *N* goes to infinity. The *T* statistic asymptotically approximates a chi-square statistic with *df* equal to the number of independent vanishing tetrads considered in the test. A non-significant result suggests that the observed tetrad differences are not significantly different than zero, indicating that the data are consistent with the vanishing tetrads implied by the hypothesized model.

Bollen (1990) and Bollen and Ting (1993) explain how to estimate the covariance matrix, Σ_{tt}, in equation (13). Bollen and Ting (1998) and Johnson and Bodner (2007) propose bootstrap methods to which the statistic can be compared. Hipp and Bollen (2003) extend

[12]Bollen and Ting (2000) note that testing for vanishing tetrads for causal versus effect indicators is less useful as an exploratory tool when there are a large number of indicators. The sheer number of possible combinations of causal and effect indicators would lead to a massive number of tests. Often conceptual formulations restrict the plausible possibilities to a more reasonable number.

the analysis of vanishing tetrads to models involving dichotomous or ordinal observed variables. SAS macros are available to implement these tests (Hipp, Bauer, & Bollen, 2005; Ting, 1995).

For our example involving five variables related to perceived health we performed vanishing tetrad tests using the SAS macro for three candidate models (see Figure 1). We consider two models: (1) a model treating all of the variables as effect indicators and (2) a MIMIC model with three causal indicators and two effect indicators. Given the conceptual unity of the five indicators and that we expect structural relations to the latent overall perceived health variable, we rule out composite indicators. We also run a set of vanishing tetrad tests that include our demographic covariates.

We begin with the model treating all of the variables as effect indicators. Following Bollen and Ting (2000), one can determine algebraically that a model with five effect indicators has 15 tetrads, all of which vanish (are zero). Only five of the fifteen tetrads, however, are independent, which gives us 5 implied vanishing tetrads to test. For the MIMIC model we can also determine algebraically that there are 2 implied vanishing tetrads that are a subset of the vanishing tetrads for the all effect indicator model. This indicates that we can perform nested tetrad tests for the MIMIC model nested within the model with all effect indicators.

Determining the implied vanishing tetrads algebraically becomes more difficult when we add covariates to the model. Fortunately, the SAS macro automates this procedure and determines the number of implied vanishing tetrads based on the implied covariance matrix. In practice, for this to work correctly it is important to include a number of significant digits when entering implied covariance matrices into the SAS macro (for our tests we used 8 significant digits). Bollen, Lennox, and Dahly (2009) provide a step-by-step guide to using the SAS macro for conducting tetrad tests.

We find a significant chi-square for the model treating all of the variables as effect indicators, but a non-significant chi-square for the MIMIC model (see Table 4). The significant chi-square for the effect indicators model is evidence that the model does not fit the data well. By contrast the non-significant chi-square for the MIMIC model suggest that this model is consistent with the data. The nested test between the MIMIC model and the effect indicators model is also significant, which suggests the MIMIC model is preferred. We see the same pattern of results when we add the demographic covariates to the model.

It is important to keep in mind that empirical checks can provide support for the presence of one type indicator over another, but they do not provide conclusive evidence. In our example, the MIMIC model has a better fit with the data, but it does not preclude the possibility that other models treating the indicators differently could fit the data equally as well (or even better). One must rely on substantive theory in conjunction with empirical tests to select the most appropriate form for the indicators.

Model Identification and Estimation

In our discussion of identification we focus on causal indicator models as they create more problems than either effect indicators models or models with composite variables. The main challenge to estimating models with causal indicators lies in identifying all of the model parameters. In this section, we briefly review what happens with respect to model identification when a latent variable measured by causal indicators emits 1 or 2 or more paths. When no paths are emitted, the model is virtually never identified.

The parameters of a SEM include regression coefficients, factor loadings, and the variances and covariances of latent variables and error variables. Identification concerns whether

unique values of the model parameters exist that we can determine using information provided by the distribution of the observed variables. Typically we use the population means (μ) and the population covariance matrix (Σ) to identify the model parameters.

In general, for any SEM one can form an implied mean vector ($\mu(\theta)$) and an implied covariance matrix ($\Sigma(\theta)$) where θ contains all of the parameters in the model. The moment structure hypotheses are given by

$$\begin{array}{l} \boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\theta}) \\ \sum=\sum(\boldsymbol{\theta}) \end{array} . \quad (14)$$

If the model is correctly specified, then these equalities should hold exactly. Bollen (1989) provides general expressions of $\mu(\theta)$ and $\Sigma(\theta)$ that hold for any SEM. In terms of (14), the question of identification is whether it is possible to uniquely solve for each parameter in θ in terms of μ and Σ, the means, variances, and covariances of the observed variables.

If, as shown in Figure 4, Panel A, a latent variable measured by causal indicators affects one other variable it is possible to identify some of the parameters by removing the latent variable and allowing the indicators to have direct effects on the outcome (Bollen, 1989, pp. 312–313; MacCallum & Browne, 1993; Bollen & Davis, [1994] 2009a). The equations for this model are

$$\begin{array}{l} \eta_{1i}=\alpha_{\eta}+\gamma_{11}x_{1i}+\gamma_{12}x_{2i}+\gamma_{13}x_{3i}+\zeta_{1i} \\ y_{1i}=\eta_{1i}+\varepsilon_{1i} \end{array} , \quad (15)$$

where we have set $\alpha_1 = 0$ and $\lambda_{11}=1$. These restrictions provide the latent variable η_{1i} with a metric and anchor its mean, a requirement for all latent variables.

By solving for η_{1i} in the second equation of (15), we can combine the two equations into a single equation

$$y_{1i}=\alpha_{\eta}+\gamma_{11}x_{1i}+\gamma_{12}x_{2i}+\gamma_{13}x_{3i}+\zeta_{1i}+\varepsilon_{1i}. \quad (16)$$

We now have an equation with a composite disturbance ($\zeta_{1i} + \varepsilon_{1i}$) that is uncorrelated with the covariates (x_{qi}). This enables us to estimate (17) as a standard regression model using an Ordinary Least Squares (OLS) estimator available in any statistical or SEM software package. This demonstrates that the intercept and regression coefficients are identified, and therefore we can estimate the effects of the causal indicators on the latent variable. We cannot, however, separately identify the variances of ζ_{1i} and ε_{1i}, though we can identify the variance of the sum ($\zeta_{1i} + \varepsilon_{1i}$). In practice, this means that it is impossible to know how much of the combined error variance is due to measurement error and how much is due to the variance in the latent variable not explained by the causal indicators. We also cannot estimate the variance of η_{1i}. Finally, we have no degrees of freedom for the chi-square test statistic for the overall fit of the model or for the various indices of fit. Nonetheless, this model represents some progress over a model with no emitted paths in that we can identify the effects of the causal indicators on the latent variable.

In equation (15) we assume that y_{1i} is a suitable scaling indicator for η_{1i}. What happens if y_{1i} is unrelated to η_{1i} ($\lambda_{11}=0$) contrary to our scaling assumption? How would we know this since we have implicitly set $\lambda_{11}=1$? Fortunately, we can find indirect evidence of $\lambda_{11}=0$. Consider the original equations for the model prior to imposing any scaling restrictions

$$\begin{aligned} \eta_{1i} &= \alpha_\eta + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \gamma_{13}x_{3i} + \zeta_{1i} \\ y_{1i} &= \alpha_y + \lambda_{11}\eta_{1i} + \varepsilon_{1i}. \end{aligned} \quad (17)$$

Substituting the equation for η_{1i} into the equation for y_{1i} leads to the reduced form equation of

$$y_{1i} = (\alpha_y + \lambda_{11}\alpha_\eta) + \lambda_{11}\gamma_{11}x_{1i} + \lambda_{11}\gamma_{12}x_{2i} + \lambda_{11}\gamma_{13}x_{3i} + \lambda_{11}\zeta_{1i} + \varepsilon_{1i}. \quad (18)$$

The compound disturbance of this equation is uncorrelated with the *x*s so that the OLS estimator is a consistent estimator of the coefficients. If $\lambda_{11}=0$, then all of the coefficients of the covariates should be within sampling fluctuations of zero. If only a subset of the coefficients are significantly different from zero, then $\lambda_{11}\neq 0$.

An alternative way to scale the latent variable would be to set one of the γs to one, say $\gamma_{11}=1$. If this were done, then the coefficient of x_{1i} would equal λ_{11}. Suppose that x_{1i} was a bad choice of scaling indicator since $\gamma_{11}=0$. If this was true but the other *x*s had effects and $\lambda_{11}\neq 0$, then the coefficient of x_{1i} in the reduced form equation would not be significantly different from 0 and the other coefficients would be. So here too we have evidence that the scaling indicator is bad. Analogous arguments are valid if the single outcome variable is another latent variable rather than a single effect indicator of a latent variable.

The situation improves significantly if one has a second outcome variable for the latent variable (see Figure 4, Panel B). We can write the equations for this model as

$$\begin{aligned} \eta_{1i} &= \alpha_\eta + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \gamma_{13}x_{3i} + \zeta_{1i} \\ y_{1i} &= \eta_{1i} + \varepsilon_{1i} \\ y_{2i} &= \alpha_{y_2} + \lambda_{21}\eta_{1i} + \varepsilon_{2i} \end{aligned} \quad . \quad (19)$$

This is a special case of the well-known MIMIC model in which all of the parameters are identified (Hauser & Goldberger, 1971; Jöreskog & Goldberger, 1975). More generally, a model with a single latent variable that has both causal and effect indicators is identified if there are at least two effect indicators with uncorrelated errors[13] and at least one causal indicator (Jöreskog & Goldberger, 1975). This rule establishes that the MIMIC model for perceived health is identified. In addition, Bollen and Davis ([1994] 2009a) generalize this rule to cover models with multiple latent variables measured by causal indicators, which they term the exogenous X rule. Heuristically, identification by the exogenous X rule is established if: (1) “Each latent variable has at least one effect indicator that loads only on it (a ‘unique indicator’) and their errors are uncorrelated,” (2) “Each latent variable directly affects at least one other effect indicator and errors for these variables are uncorrelated with the errors of the unique indicators,” (3) there are least *m* causal indicators and the matrix of coefficients for the effects of the causal indicators on the latent variables has rank(*m*), and (4) “the structural model relating the causal indicators to the latent variables and the latent variables among themselves has an identified structure” (Bollen & Davis, [1994] 2009a, pp. 503–505).

Another advantage of having a second effect indicator is that the model is overidentified as long as there is more than one causal indicator and no correlated errors. This allows one to

[13]More accurately, we assume that the effect indicators are conditionally uncorrelated net of the latent variable so that we also exclude direct effects between the effect indicators.

compare the hypothesized model to the saturated model with the chi-square test of model fit and to consider various other indices of model fit.

So far we have restricted our discussion to models with one or two effect indicators for the latent variable. Analogous results hold if the outcome variables of the latent variable measured by causal indicators are themselves latent variables. The 2+ emitted paths rule requires each latent variable to emit at least two paths to other variables, either latent or observed (MacCallum & Browne, 1993; Bollen & Davis [1994] 2009a, 2009b). In this case, the outcome latent variables should be measured by at least two effect indicators.

To summarize this section, if the latent variable measured by causal indicators has no outcome variable, then there is little a researcher can do to estimate the effects of the causal indicators on the latent variable. If there is at least one outcome, such as an effect indicator or another observed variable, directly influenced by the latent variable, then it is possible to identify and estimate the coefficients of the causal indicators on the latent variables. With two or more outcome variables with uncorrelated errors and no direct effect between the outcome variables, the situation is even better and a researcher can proceed as with any other SEM including testing the overall model fit and calculating a variety of fit indices using any of the standard SEM software packages.

Options for Scaling Latent Variables Measured by Causal Indicators

All latent variables require a scale to identify and interpret parameters. The options for scaling latent variables when working with effect indicators are well studied and understood, but this is not the case for latent variables measured by causal indicators. In this section, we briefly review three options and provide recommendations as to when to use them.

First, if a latent variable measured by causal indicators also emits paths to effect indicators, then a natural choice is to scale the latent variable to one of the effect indicators. If there is more than one effect indicator, choose the one that appears most closely related to the latent variable. This is the approach we use in our example involving perceived health. We scale latent perceived health to our first effect indicator, respondent's self-rated health.

In some cases, a latent variable measured by causal indicators emits paths only to anticipated outcomes of the latent variable. One must use theory or substantive knowledge to distinguish outcomes from effect indicators. For instance, one might measure latent socioeconomic status (SES) by a set of causal indicators (e.g., education, income, and occupational status) and then model the effect of SES on political attitudes. In this case, it is possible to scale the latent variable by setting one of the paths to the outcomes to 1, but this may feel awkward as it sets the scale based on an outcome rather than a measure of the latent variable. An alternative is to scale the latent variable to one of the causal indicators. In this situation, this may make more sense as the causal indicator maintains conceptual unity with the latent variable and therefore may provide a more interpretable metric for the latent variable. A one-unit difference in the scaling causal indicator leads to an expected one-unit difference in the latent variable that it scales.

Finally, some researchers scale the latent variable by setting its variance to 1 or the error variance of the latent variable to 1. These options are only available for models in which the latent variable emits more than one path. In models where the latent variable emits only a single path none of the parameters will be identified if one chooses these scaling options (see Appendix 1 for a proof). If there is a single emitted path from a latent variable, then the researcher needs to scale the latent variable using the outcome variable or one of the causal indicators.

Estimation

Once one has established a model is identified, a number of estimators are available for obtaining parameter estimates. The most popular estimator for SEMs is a maximum likelihood estimator (MLE), which one can use for models involving causal, composite, or effect indicators. The fitting function for the MLE is given by

$$F_{ML} = \log|\sum(\boldsymbol{\theta})| + \mathrm{tr}[\mathbf{S}\sum^{-1}(\boldsymbol{\theta})] - \log|\mathbf{S}| - (p+q) + [\overline{z} - \boldsymbol{\mu}(\boldsymbol{\theta})]^T \sum^{-1}(\boldsymbol{\theta})[\overline{z} - \boldsymbol{\mu}(\boldsymbol{\theta})], \quad (20)$$

where |.| is the determinant, tr[.] is the trace, **S** is the sample covariance matrix, *p* is the number of effect indicators, and *q* is the number of causal or composite indicators and covariates. The θ that maximizes this function is the MLE.

For our example involving variables related to health we estimated the models described above (see Figure 1) using AMOS 17 (Arbuckle, 2008). We first consider the overall model fit statistics for each model. The all effect indicator model has a significant chi square test statistic and fit measures that all fall short of conventional standards. The same is true for the composite indicator model. The composite indicator model's one degree of freedom comes from the uncorrelated errors of the y variables and the poor fit of the model is evidence against this restriction. We find clear evidence that the MIMIC model has the best fit with the data of the three we considered (see Table 5). The overall chi-square is non-significant, the CFI and TLI are right around 1.00, the RMSEA is 0.00, and we have a negative BIC. As with the composite indicators model, the MIMIC model has uncorrelated errors of the ys, but it specifies their common dependence on the latent variable and this contributes to a well fitting model. When we add covariates to the model, the MIMIC model still has the best fit of the three, but there are some indications that the model does not fit the data as well. The chi-square is significant, the TLI is low, and the RMSEA is above 0.05. This may be because we constrained all of the demographic covariates to have only direct effects on latent perceived health. If we were to develop this model further, it would be worth considering whether the covariates have direct effects on one of the two effect indicators, self-reported health and health satisfaction.

Turning to the parameter estimates, we see the factor loadings or regression coefficients depending on the model are mostly significant and in the expected direction (see Table 6). The one exception is hospitalized/disabled in the prior 4 years, which is not a significant indicator in any model. We find significant variance in the latent variable for the all effect indicator model. The disturbance variance of our latent variable for health in the MIMIC model is significantly different from zero, which rules out the composite indicator model. In the effect indicators model we see low R-squares for the indicators that we treat as causal or composite indicators in the other two models. These are important clues that the model using effect indicators may not be correct. We find in the MIMIC model that the three causal indicators account for 19 percent of the variance in the latent health variable. We find the same pattern of results when we include covariates in the model (results not reported).

Evaluating Indicators

Once we have determined the type of indicator or covariate in the model and have estimated the model, it is possible to evaluate the performance of the variables. Effect indicators are the predominant implicitly assumed type of indicator and the literature on assessing them is vast. To save space we will not repeat it here. Similarly, covariates and their effects are well studied in regression courses and will not be discussed. The assessment of composite and causal indicators is less developed, so we will concentrate on them.

One simple statistic to assess both composite and causal indicators is to estimate the squared correlation between the indicator and the latent variable or composite. For a composite indicator, this squared correlation is

$$\rho^2_{x_qC}=\frac{[\,COV(x_q,C)]^2}{VAR(x_q)\,VAR(C)}, \quad (21)$$

where $COV(x_q,C)$ is the covariance of the composite indicator and the composite, $VAR(x_q)$ is the variance of the composite indicator, and $VAR(C)$ is the variance of the composite. The higher this squared correlation is, the greater the shared variance of the composite indicator and the composite. In a sense this is a measure of how good a proxy a single composite indicator is for the full composite with high values suggesting better proxies.

Sample estimates of the quantities in equation (21) enable estimation of a squared correlation for each composite indicator. Upon estimation, the composite equals

$$\widehat{C}_{1i}=\widehat{w}_{10}+\widehat{w}_{11}x_{1i}+\widehat{w}_{12}x_{2i}+\cdots+\widehat{w}_{1Q}x_{Qi}. \quad (22)$$

This enables a value of the composite and we can use any statistical software to estimate the correlation between it and each composite indicator. After squaring this, we have the squared correlation measure.

A similar squared correlation measure is available for each causal indicator,

$$\rho^2_{x_q\eta}=\frac{[\,COV(x_q,\eta)]^2}{VAR(x_q)\,VAR(\eta)}, \quad (23)$$

where $COV(x_q,\eta)$ is the covariance of a causal indicator with the latent variable and the $VAR(\eta)$ is the variance of the latent variable. Given the assumption that the error or disturbance is uncorrelated with the causal indicators, then

$$COV(x_q,\eta)=COV(x_q,\widehat{\eta}), \quad (24)$$

with

$$\widehat{\eta}=\widehat{\alpha}_\eta+\widehat{\gamma}_{11}x_{1i}+\widehat{\gamma}_{12}x_{2i}+\cdots+\widehat{\gamma}_{1Q}x_{Qi}. \quad (25)$$

Forming $\widehat{\eta}$, we can estimate COV(x_q, $\widehat{\eta}$) which equals $COV(x_q,\eta)$. An estimate of $VAR(\eta)$ is often available from SEM software output.[14] With these parts, we can estimate the squared correlation of the causal indicator and its latent variable. This quantity tells us the amount of shared variance of the latent variable and the causal indicator.

The question of validity seems out of place for composite indicators since the composite indicator is not intended to measure a latent variable. Rather, the composite is simply a weighted sum of its composite indicators. We can ask, however, about the validity of a causal indicator. This question can be addressed at several stages of the research process. In

[14]If not available, an analyst can estimate it as $VAR(\eta) \mid VAR(\zeta)$.

an early stage causal (and effect) indicators should be selected because they correspond to the theoretical definition of the concept that the latent variable represents. That is, the causal indicator should have face validity in that it is plausible to think that it influences the latent variable. The plausibility derives from substantive expertise of the researcher and comparing the indicator to the theoretical definition of the latent concept.

Even experts can be wrong, so it makes sense to look for empirical means of assessment. If a researcher has no outcome variables for the latent variable, then the only empirical check is the vanishing tetrad test. This provides information on the appropriateness of considering all or a subset of the indicators causal versus all or a subset as effect indicators. If the vanishing tetrad test supports effect indicators or a mixture of causal and effect indicators, this cast doubts on treating them all as causal indicators. Alternatively, a vanishing tetrad test supportive of causal indicators does not guarantee their quality.

With at least two outcome variables that have no direct effects between them and no correlated errors, we can estimate the coefficients and their standard errors for the causal indicators along with the variance of the latent variable. This information allows one to assess the validity of the indicators. One approach to assessing validity involves examining the magnitude of the direct structural relationships between the indicators and the latent variable (Bollen, 1989). Bollen (1989, pp. 197–205) outlines three ways of considering the magnitude of the direct relationship: (1) the unstandardized validity coefficient, (2) the standardized validity coefficient, and (3) the unique validity variance.

The unstandardized validity coefficient is simply the estimate of the effect of the indicator on the latent variable in their original metrics.[15] The standardized validity coefficient corresponds to the standardized regression coefficient. Finally, the unique validity variance for causal indicators measures the part of the explained variance in $\eta_{1\mathrm{i}}$ that is *uniquely* attributable to x_{qi}. The formula for the unique validity variance $U_{\eta 1\mathrm{xq}}$ is

$$U_{\eta_1 x_q} = R^2_{\eta_1} - R^2_{\eta_1(x_q)}, \quad (26)$$

where $R^2_{\eta_1}$ is the proportion of variance in $\eta_{1\mathrm{i}}$ explained by all variables with a direct effect on the latent variable and $R^2_{\eta_1(x_q)}$ is the proportion of variance in $\eta_{1\mathrm{i}}$ explained by all variables with a direct effect excluding x_{qi}. The unique validity variance ranges from 0 to 1 where higher values suggest greater validity than lower ones.

If the causal indicators have no impact on the latent variable then their coefficients will not be statistically significant and this will be reflected in the measures of validity. Rather than being poor measures, it is possible that the latent variable does not affect the outcome variable as hypothesized. In the absence of prior research or good substantive knowledge, we cannot know whether we have bad causal indicators or a nonexistent relationship between the latent variable and an outcome. If we find that some causal indicators have significant coefficients and others do not, then this provides indirect evidence that the causal indicators with nonsignificant coefficients are more likely the problem. In contrast, if all the causal indicators are not significant, then this suggests that problem lies with the relationship between the latent variable and the outcome variable or that the causal indicators are highly collinear with one another.

[15]This measure of validity is available for models in which the latent variable emits a single path because it only requires the estimate of the effect of the causal indicator on the latent variable.

Using the five variables we find mixed results for the shared variance and the validity of our three hypothesized causal indicators of perceived health. Visiting the doctor and being hospitalized in the last year have shared variances with perceived health of 0.11 and 0.12, unstandardized validity coefficients of −0.28 and −0.45, and standardized coefficients of −0.24 and −0.29 respectively, all of which are statistically significant (see Table 7). The shared variance for hospitalized in the 4 years prior is close to zero, the unstandardized validity coefficient is −0.13, and the standardized coefficient is −0.09, which are only marginally significant. We find a similar pattern of results with the unique validity variance (see Table 7). These results indicate that hospitalized in the 4 years prior to the last year may not be a good causal indicator of perceived health.

One complication that can affect models with causal indicators is collinearity among the indicators (Bollen, 1989, pp. 205–206; Diamantopoulos & Winklhofer, 2001, p. 272). If a causal indicator is perfectly predicted by the other causal indicators, then we cannot estimate its separate effect on the latent variable. More often when there are high correlations among a set of causal indicators, it is difficult to estimate their individual effects. This is the same issue as multicollinearity in multiple regression and analogous diagnostics are appropriate here. The presence of multicollinearity is particularly problematic for assessing the validity of causal indicators. If the researcher is not specifically interested in estimating the effect of the causal indicators on the latent variable, then dropping nonsignificant variables from the model is a possibility. This strategy can result in biased coefficients for the other causal indicators, so it only should be considered if these coefficients are not of primary interest.

In many cases when we have two outcome variables of the latent variable the model will be overidentified. This permits the analyst to test the overall fit of the model and to consider a variety of model fit indices. If these fit statistics suggest a problem with the model, it is possible that problem lies with the causal indicators (though there are many other reasons a model may not fit the data well).

We can use these measures of validity in the selection of causal indicators (Bollen, 1989, pp. 197–205; Diamantopoulos & Winklhofer, 2001, pp. 272–274; Fayers & Hand, 2002, pp. 241–245). All else being equal, we prefer causal indicators that have significant values of the unstandardized and standardized validity coefficients, high values of the unique validity variance, high squared correlation with the latent variable, and little evidence of collinearity. Additionally, one would like evidence that the overall model fits the data well.

A Few Potential Issues

Scales and Indexes

Social and behavioral scientists commonly use scales and indexes as exogenous or endogenous variables in models.[16] Concepts as diverse as socioeconomic status, depression (Radloff, 1977), and activities of daily living (Lawton & Brody, 1969) are routinely measured with scales or indexes. Typically indexes and scales are linear combinations of variables that can be included in a model in lieu of the individual items. Scales and indexes allow researchers to estimate the effect of a single summary variable rather than the individual effects of a set of variables. As such, scales and indexes result in a more parsimonious model. But this advantage comes with potential costs. To discuss these costs we distinguish two situations: (1) when the items are causal indicators, composite indicators, or covariates (the three Cs), (2) when the items are effect indicators.

[16]Some treatments associate indexes with causal indicators and scales with effect indicators (e.g., DeVellis 1991). This distinction does not appear to be universally maintained, so we use the terms indexes and scales interchangeably in our discussion.

The Three Cs

Suppose we have a model with a single outcome variable (y_i) that depends on a latent variable (η_i), which in turn is influenced by a vector of causal indicators ($\mathbf{x}_i$). We can write this model as

$$\begin{aligned} y_i &= \alpha_y + \eta_i + \zeta_y \\ \eta_i &= \alpha_\eta + \boldsymbol{\gamma}\mathbf{x}_i + \zeta_\eta \end{aligned}, \quad (27)$$

where αs are intercepts, ζs are disturbances with mean 0 and uncorrelated with each other and the explanatory variables in their respective equations. To simplify the situation we assume homoscedastic and nonautocorrelated errors. The vector $\boldsymbol{\gamma}$ contains the coefficients for the effect of the vector $\mathbf{x}_i$ on η_i.

If we substitute the second equation in (27) into the first, we get

$$y_i = \alpha_y + \alpha_\eta + \boldsymbol{\gamma}\mathbf{x}_i + \zeta_\eta + \zeta_y, \quad (28)$$

which we can estimate with OLS to obtain consistent and unbiased estimators of the coefficients of the causal indicators.

What happens if an index or scale, S_i, replaces the latent variable (η_i) in the model? Typically, a scale or index is constructed as a linear combination of the indicators and the researcher uses it as a proxy for a latent variable. We represent this linear combination as

$$S_i = \mathbf{w}\mathbf{x}_i, \quad (29)$$

where $\mathbf{w}$ is a vector of weights for the items. Sometimes a scale or index is a simple sum of the items, in which case $\mathbf{w}$ = [1 1 … 1]. For instance, the Activities of Daily Living scale and some exposure to stressful live events scales are often constructed as the sum of the dummy coded items. In other situations, such as with many indexes of socioeconomic status, each item is first standardized to have mean 0 and standard deviation 1 before summing. Standardizing variables and summing them is an implicit weighting scheme.

If the equations in (27) represent the true model, but an analyst replaces η_i with S_i, then in general the coefficients for $\mathbf{x}_i$ will be biased. We can see this by examining the relationship between S_i and η_i,

$$S_i = \eta_i - \alpha_\eta - \zeta_\eta + (\mathbf{w} - \boldsymbol{\gamma})\mathbf{x}_i, \quad (30)$$

If $\mathbf{w} = \boldsymbol{\gamma}$, then the weights represent the true coefficients and we would obtain the correct impact of the causal indicators on η_i since we are assuming that the weights in the index are the same as the coefficients of the indicators. It seems unlikely, however, that the weights used to form an index or scale would coincidentally match the true impact of the causal indicators on the latent variable. Therefore, the use of indexes or scales will generally lead to biased estimates of the effects for $\mathbf{x}_i$.

Though it is unlikely that a researcher would include covariates along with causal indicators to form a scale or index, this does not change the results. In addition, forming an index or scale from composite indicators is no different. If $\mathbf{w} \neq \boldsymbol{\gamma}$, then one will obtain biased and inconsistent coefficients. The degree of bias will depend on the degree of difference between

the prespecified weights of the xs in the index or scale versus the true coefficients of the xs in the model.

Effect Indicators

When constructing an index or scale using effect indicators researchers may select weights on an *ad hoc* basis as we discussed in the previous section. Researchers may also use more formal procedures such as factor score prediction to derive weights. We use the following model to illustrate the problems with these approaches when working with effect indicators:

$$\begin{aligned} \eta_{2i} &= \alpha_\eta + \beta\eta_{1i} + \zeta_i \\ \mathbf{y}_i &= \boldsymbol{\lambda}\eta_{1i} + \boldsymbol{\varepsilon}_i \end{aligned} \quad . \quad (31)$$

The first equation shows the latent variable η_{2i} dependent on another latent variable η_{1i} with β as the coefficient. The second equation relates a vector of effect indicators ($\mathbf{y}_i$) to the latent variable η_{1i} with factor loadings $\boldsymbol{\lambda}$ and unique factors $\boldsymbol{\varepsilon}_i$. (We assume but do not show another equation that contains sufficient number of indicators of η_{2i} to identify the model.)

Researchers sometimes form the predicted factor score to use in place of the latent factor (η_{1i}). The predicted factor score is the same as an index or scale except that its weights are determined by factor regressions, Bartlett scores, or some other method for predicting the values of the factor using the observed variables. Regardless of the factor scoring method, we can write the factor score as

$$\hat{\eta}_{1i} = \mathbf{b}\mathbf{y}_i, \quad (32)$$

where **b** are the weights for forming the factor scores. The latent factor is equal to

$$\eta_{1i} = \hat{\eta}_{1i} + e_i, \quad (33)$$

where e_i is the discrepancy between the latent variable and the factor score. If we substitute the predicted factor score in place of η_{1i} in the first line of equation (31), then we get

$$\eta_{2i} = \alpha_\eta + \beta\hat{\eta}_{1i} + \beta e_i + \zeta_i. \quad (34)$$

Since $\hat{\eta}_{1i} = \eta_{1i} - e_i$ the error term $\beta e_i + \zeta_i$ correlates with $\hat{\eta}_{1i}$, which will create a bias in estimating its coefficient with say, OLS. In other words, we cannot simply substitute the predicted factor score for the actual latent variable without incurring bias. The degree of bias depends on the degree to which the factor score exactly predicts the factor (latent variable).

Example with Model for Health

Returning to our running example involving health, suppose we construct an index as the sum of our three causal indicators (visiting the doctor, going to the hospital in the last year, going to hospital in the prior four years) and estimate the effect of the index on self-rated health and satisfaction with health. To make the model comparable to our MIMIC model, we scale the index to have the same variance as the latent variable from our MIMIC model and we constrain the regression coefficient of self-rated health on the index to be one. When we estimate the model using our index, we find a highly significant chi-square (202.5 with 2 degrees of freedom) indicating the model does not fit the data. We also estimate the effect of the index on satisfaction with health to be 0.86 as compared with our estimate of 1.82 from the MIMIC model. Using the index as a proxy for the latent variable leads to an estimate

that is less than half of the effect of perceived health on satisfaction with health from the MIMIC model. Though the biases need not always be this large, the example does illustrate that an index is not a good substitute for the latent variable.

Summary

Using indexes or scales when a latent variable with causal, composite, or effect indicators or covariates are present will in most situations result in biased and inconsistent estimators of coefficients. In our discussion we have focused on unidimensional concepts represented by a single latent factor. The bias is likely worse when working with multidimensional concepts represented by more than one latent factor. In the multidimensional situation, a separate latent variable for each dimension is necessary. A single index or sum of items cannot capture these different dimensions in a single number.

There may be some situations where scales or indexes might work. For instance, if the effects of individual causal or composite indicators are approximately equal and the sample size is small, then assuming equal weights might be a good enough approximation that reduces the number of parameters to estimate in a model. Alternatively, if the coefficients for the individual causal or composite indicators are not that important to the analysis, it may be that the biases associated with using a scale or index will not be that severe. With effect indicators, the situation is a little different. If the index or scale is very highly correlated (.95 or higher) with the latent variable and the indicators represent an underlying unidimensional concept, then the index or scale could be a reasonable substitute for the latent variable. Alternatively, if the index never appears as a covariate but only as a dependent variable in a model, then the biasing effects could be less severe. Nonetheless, as a general rule, one should avoid using indexes or scales in place of latent variables whenever feasible.

Multiple Latent Variables with Causal Indicators

MacCallum and Browne (1993) note that it is not clear how one should handle the covariances among causal indicators of separate latent variables. In general, SEMs allow for all exogenous variables, latent or observed, to be correlated. Alternatively, one might argue, based on the logic from effect indicators, that measures from one latent variable should not be correlated with measures from another latent variable unless specified by theory. Finally, one might prefer to allow one latent variable measured by causal indicators to affect the causal indicators of another latent variable.

In practice the simplest solution is to allow all exogenous variables to be correlated. If a researcher has a strong theoretical reason for treating the causal indicators of one latent variable as uncorrelated with the causal indicators for another latent variable, then this constraint can be imposed and tested with a likelihood ratio test. Similarly, if a researcher's theory suggests a more precise causal structure among a set of causal indicators (e.g., that the latent variable measured by one set of causal indicators has an effect on the causal indicators of another latent variable), then this also can be explicitly modeled and tested.

Measurement Error in the Causal Indicators

In the causal indicator model specified in equation (2) we assume that the observed variables are actual causes as opposed to indicators of the real causal influences on η_{1i}. In other words, we assume that the observed variables contain negligible measurement error. If this is not the case, then we need to allow for measurement error in the indicators themselves (Bollen, 1989, p. 312). We can accomplish this by rewriting equation (2) as

$$\begin{aligned}
\eta_{1i} &= \alpha_\eta + \gamma_{11}\xi_{1i} + \gamma_{12}\xi_{2i} + \cdots + \gamma_{1Q}\xi_{Qi} + \zeta_{1i} \\
x_{1i} &= \xi_{1i} + \delta_{1i} \\
x_{2i} &= \xi_{2i} + \delta_{2i} \\
&\vdots \\
x_{Qi} &= \xi_{Qi} + \delta_{Qi}
\end{aligned} \quad , \quad (35)$$

where each x_{qi} is an indicator of a corresponding ξ_{qi} with a random measurement error or unique factor δ_{qi}. In practice, models that allow for measurement error among the causal indicators are even more difficult to identify. Generally we would need two or more effect indicators for each ξ_{qi}.

Conclusions

Awareness of the distinction between causal and effect indicators has gradually grown over the last 20 years. But as the awareness has grown so have questions about their treatment. Part of the confusion concerns methods to distinguish causal from effect indicators. Other questions have emerged about the best methods to assess the quality of causal indicators.

The field has made progress in recognizing that sometimes a latent variable is a function of an indicator rather than vice versa. However, researchers have not appreciated that there is more than one type of variable that can influence a latent variable. Indeed, it is common to treat terms like causal indicators or formative indicators as if they were interchangeable. In essence, conventional practice dichotomizes indicators into effect (reflective) indicators and causal (formative) indicators. A primary purpose of our paper is to propose that the dichotomous categorization is not accurate. The latter indicator type consists of three different variables: causal indicators, composite (formative) indicators, and covariates. The blurring of these three Cs has led to contradictory statements and controversy.

This paper discusses effect, causal, and composite indicators as well as covariates and describes their similarities and differences. Furthermore, it reviews the identification, estimation, and assessment of fit of models with these types of variables. We propose methods by which we can determine the type of indicator or covariate we have. These include both conceptual and empirical methods. It also presents ways to assess the validity of causal indicators relying on definitions of validity from Bollen (1989). We also discuss the hazards of using indexes and scales in light of these distinctions between types of indicators. We illustrate many of our results using an empirical example on perceived health.

In brief, our recommendation is for researchers to give more thought to the type of indicators they are using and to formulate models of their relationships to their latent variables or composites. It is not accurate to act as if there are just two indicator types, one that depends on a latent variable and the other that the latent variable is a function of. Attending to the differences between causal indicators, composite indicators, and covariates has a number of potential benefits for applied research. First, when facing a severely misspecified model, a careful consideration of the types of indicators may provide one route towards finding a better model. Second, in some cases it may be that one can obtain a reasonable model fit even with an incorrect specification of the indicators (e.g., specifying effect rather than causal indicators). However, these models imply very different ways in which latent variables might be changed. Shifting a causal indicator alters the values of a latent variable whereas shifting the same variable if it is an effect indicator has no consequences for the latent variable. Third, the framework we propose helps clarify past debates in this area, particularly between causal and composite (formative) indicators, and will help researchers avoid erroneous advice based on conflating different types of

indicators. Causal indicators imply an invariance in effects on the same latent variable regardless of the outcomes of the latent variable whereas composite (formative) indicators will likely shift depending on the outcome. Finally, the scientific enterprise seeks to use accurate models of reality while recognizing the approximate nature of models. Accuracy demands that we use the appropriate model for the relation between our measures and latent variables. If we confuse covariates, causal indicators, composite (formative) indicators, and effect indicators, we will set back the measurement of latent variables. We see no reason not to use the most appropriate measurement models we are capable of building.

Ultimately, subject matter expertise is essential to distinguishing between effect indicators, causal indicators, composite indicators, and covariates. When possible, we recommend supplementing these theoretical considerations with empirical tests that can help to adjudicate between structures. The current casual approach toward measurement can undermine our tests of substantive hypotheses. We have the methods to address many of these issues.

References

Arbuckle, JL. AMOS user's guide. Spring House, PA: AMOS Development Corporation; 1995–2008.

Atkinson MJ, Lennox RD. Extending basic principles of measurement models to the design and validation of patient reported outcomes. Health and Quality of Life Outcomes. 2006; 4 Manuscript #65.

Blalock, HM. Causal inference in nonexperimental research. Chapel Hill: University of North Carolina Press; 1964.

Blalock, HM. Measurement in the social sciences: Theories and strategies. Chicago: Aldine Publishing Co; 1974.

Bollen KA. Multiple indicators: Internal consistency or no necessary relationship? Quality and Quantity. 1984; 18(4):377–385.

Bollen, KA. Structural equations with latent variables. New York: Wiley; 1989.

Bollen KA. Outlier screening and a distribution-free test for vanishing tetrads. Sociological Methods & Research. 1990; 19(1):80–92.

Bollen KA. Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). Psychological Methods. 2007; 12(2):219–228. [PubMed: 17563174]

Bollen KA, Davis WR. Causal indicator models: Identification, estimation, and testing. Structural Equation Modeling. [1994] 2009a; 16(3):498–522.

Bollen KA, Davis WR. Two rules of identification for structural equation models. Structural Equation Modeling. 2009b; 16(3):523–536.

Bollen KA, Lennox RD. Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin. 1991; 110(2):305–314.

Bollen KA, Lennox RD, Dahly DL. Practical application of the vanishing tetrad test for causal indicator measurement models: An example from health-related quality of life. Statistics in Medicine. 2009; 28(10):1524–1536. [PubMed: 19266502]

Bollen KA, Ting K. Confirmatory tetrad analysis. Sociological Methodology. 1993; 23:147–175.

Bollen KA, Ting K. Bootstrapping a test statistic for vanishing tetrads. Sociological Methods & Research. 1998; 27(1):77–102.

Bollen KA, Ting K. A tetrad test for causal indicators. Psychological Methods. 2000; 5(1):3–22. [PubMed: 10937320]

Cohen P, Cohen J, Teresi J, Marchi M, Velez CN. Problems in the measurement of latent variables in structural equation causal models. Applied Psychological Measurement. 1990; 14(2):183–196.

DeVellis, RF. Scale development: Theory and applications. Newbury Park, CA: Sage Publications; 1991.

Diamantopoulos A, Winklhofer HM. Index construction with formative indicators: An alternative to scale development. Journal of Marketing Research. 2001; 38(2):269–277.

Edwards JR, Bagozzi RP. On the nature and direction of relationships between constructs and measures. Psychological Methods. 2000; 5(2):155–174. [PubMed: 10937327]

Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: An example from quality of life. Journal of the Royal Statistical Society Series A. 2002; 165(2):233–261.

Fayers PM, Hand DJ. Factor analysis, casual indicators and quality of life. Quality of Life Research. 1997; 6(2):139–150. [PubMed: 9161114]

Fornell C, Bookstein FL. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. Journal of Marketing Research. 1982; 19(4):440–452.

Glanville JL, Paxton P. How do we learn to trust? A confirmatory tetrad analysis of the sources of generalized trust. Social Psychology Quarterly. 2007; 70(3):230–242.

Goldberger, AS. Structural Equation Models: An Overview. In: Goldberger, AS.; Duncan, OD., editors. Structural Equation Models in the Social Sciences. New York: Seminar Press; 1973. p. 1-18.

Grace JB, Bollen KA. Representing general theoretical concepts in structural equation models: The role of composite variables. Environmental and Ecological Statistics. 2008; 15(2):191–213.

Hauser RM, Goldberger AS. The treatment of unobservable variables in path analysis. Sociological Methodology. 1971; 3:81–117.

Hipp JR, Bauer DJ, Bollen KA. Conducting tetrad tests of model fit and contrasts of tetrad-nested models: A new SAS macro. Structural Equation Modeling. 2005; 12(1):76–93.

Hipp JR, Bollen KA. Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. Sociological Methodology. 2003; 33:267–305.

Howell RD, Breivik E, Wilcox JB. Reconsidering formative measurement. Psychological Methods. 2007; 12(2):205–218. [PubMed: 17563173]

Jarvis CB, MacKenzie B, Podsakoff PM. A critical review of construct indicators and measurement model misspecification in marketing and consumer research. Journal of Consumer Research. 2003; 30(2):199–218.

Johnson TR, Bodner TE. A note on the use of bootstrap tetrad tests for covariance structures. Structural Equation Modeling. 2007; 14(1):113–124.

Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. Journal of the American Statistical Association. 1975; 70(351):631–639.

Kelley, TL. Crossroads in the mind of man. Stanford: Stanford University Press; 1928.

Land KC. On the estimation of path coefficients for unmeasured variables from correlations among observed variables. Social Forces. 1970; 48(4):506–511.

Lawton MP, Brody EM. Assessment of older people: Self-maintaining and instrumental activities of daily living. The Gerontologist. 1969; 9(3 Part 1):179–186. [PubMed: 5349366]

MacCallum RC, Browne WM. The use of causal indicators in covariance structure models: Some practical issues. Psychological Bulletin. 1993; 114(3):533–541. [PubMed: 8272469]

Marsden P. A note on block variables in multiequation models. Social Sicence Research. 1982; 11(2): 127–140.

Muthén, LK. Formative indicators. 2006. responsesMessages posted to http://www.statmodel.com/discussion/messages/11/1031.html?1270606184

Pearl, J. Causality: Models, reasoning, and inference. 2. New York: Cambridge University Press; 2009.

Perreira KM, Deeb-Sossa N, Harris KM, Bollen KA. What are we measuring? An evaluation of the CES-D across race/ethnicity and immigrant generation. Social Forces. 2005; 83(4):1567–1602.

Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement. 1977; 1(3):385–401.

Savalei V, Kolenikov S. Constrained versus unconstrained estimation in structural equation modeling. Psychological Methods. 2008; 13(2):150–170. [PubMed: 18557683]

Spearman C. “General intelligence,” objectively determined and measured. American Journal of Psychology. 1904; 15(2):201–292.

Ting K. Confirmatory tetrad analysis in SAS. Structural Equation Modeling. 1995; 2(2):163–171.

Williams LJ, Edwards JR, Vandenberg RJ. Recent advances in causal modeling methods for organizational and management research. Journal of Management. 2003; 29(6):903–936.

Appendix 1

In this appendix we provide a proof concerning the identification of parameters in a model with a latent variable measured by causal indicators that emits only one path (see Figure 4, Panel A). The equations for this model are:

$$\begin{aligned}\eta_i &= \alpha_\eta - \gamma_{11}x_{1i} + \gamma_{22}x_{2i} + \gamma_{13}x_{3i} + \zeta_1 \\ y_{1i} &= \alpha_y + \lambda_{11}\eta_i + \varepsilon_{1i}\end{aligned} \quad (36)$$

where η_1 is a latent variable measured by three causal indicators (x_is) that has an effect on one variable, y_i. The structural effects of the xs on η_1 are given by the γs and the effect of η_1 on y_i is given by λ_{11}.

We demonstrate the following four statements:

1. If the latent variable is scaled by setting $\lambda_{11} = 1$, the γs are identified, but $V(\eta_1)$ and $V(\varepsilon_1)$ are not.
2. If the latent variable is scaled by setting $\gamma_{11} = 1$, then λ_{11}, γ_{12}, and γ_{13} are identified, but $V(\eta_1)$ and $V(\varepsilon_1)$ are not.
3. If the latent variable is scaled by setting $V(\eta_1) = 1$, then none of the parameters are identified.
4. If the latent variable is scaled by setting $V(\zeta_1) = 1$, then none of the parameters are identified.

To prove these statements we begin with equation (18),

$$y_{1i} = (\alpha_y + \lambda_{11}\alpha_\eta) + \lambda_{11}\gamma_{11}x_{1i} + \lambda_{11}\gamma_{12}x_{2i} + \lambda_{11}\gamma_{13}x_{3i} + \lambda_{11}\zeta_{1i} + \varepsilon_{1i}, \quad (37)$$

and write the equations for the variance of y_1 and the covariances of y_1 with each of the xs. For this model we have

$$\begin{aligned}\sigma_{y_1y_1} &= \lambda_{11}^2 V(\eta_1) + V(\varepsilon_1) \\ &= \lambda_{11}^2 \left[\gamma_{11}^2\sigma_{x_1x_1} + \gamma_{12}^2\sigma_{x_2x_2} + \gamma_{13}^2\sigma_{x_3x_3} + 2(\gamma_{11}\gamma_{12}\sigma_{x_1x_2} + \gamma_{11}\gamma_{13}\sigma_{x_1x_3} + \gamma_{12}\gamma_{13}\sigma_{x_2x_3}) + V(\zeta_1)\right] + V(\varepsilon_1)\end{aligned}, \quad (38)$$

where $\sigma_{..}$ represents the variance or covariance of the observed variables in the subscript. We also have

$$\sigma_{y_1x_1} = \lambda_{11}(\gamma_{11}\varphi_{11} + \gamma_{12}\varphi_{12} + \gamma_{13}\varphi_{13}) \quad (39)$$

$$\sigma_{y_1x_2} = \lambda_{11}(\gamma_{11}\varphi_{12} + \gamma_{12}\varphi_{22} + \gamma_{13}\varphi_{23}) \quad (40)$$

$$\sigma_{y_1x_3} = \lambda_{11}(\gamma_{11}\varphi_{13} + \gamma_{12}\varphi_{23} + \gamma_{13}\varphi_{33}). \quad (41)$$

We first note that in equations (39), (40), and (41), the ϕs are identified using the variances and covariances among x_1, x_2, and x_3. Suppose we choose to scale the latent variable by setting $\lambda_{11} = 1$. In this case, equations (39) to (41) have just three unknowns, the γs. Using

the three equations and three unknowns, we can identify the γs. We have only one remaining equation that includes both $V(\eta_1)$ and $V(\varepsilon_1)$, so we are unable to separately identify them. Now suppose that instead we scale the latent variable by setting $\gamma_{11} = 1$. We still have the situation just described, that is, three equations and three unknowns and we can identify λ_{11}, γ_{12}, and γ_{13}, but we still cannot separately identify $V(\eta_1)$ and $V(\varepsilon_1)$. This demonstrates our first two propositions.

A different situation arises if we choose to scale the latent variable by setting $V(\eta_1) = 1$. The first line of equation (38) reduces to an equation with just two unknowns [λ_{11} and $V(\varepsilon_1)$], but this doesn't help us. We cannot use (39) to (41) to identify the coefficients because we have three equations and four unknowns, and including equation (38) just adds another unknown, $V(\varepsilon_1)$. Therefore we are not able to separately identify any of the parameters in the model. We encounter the same problem if we try to scale the latent variable by setting $V(\zeta_1) = 1$. In this case, the second line of equation (38) reduces to a function involving five unknowns [λ_{11}, the γs, and $V(\varepsilon_1)$]. Once again, (39) to (41) do not provide enough information to identify the four unknowns and including the reduced form of equation (38) adds another unknown, so we are unable to separately identify any of the parameters. This demonstrates our third and fourth propositions. Our choice of three x variables was arbitrary. The same consequences follow whether we have fewer or more xs.

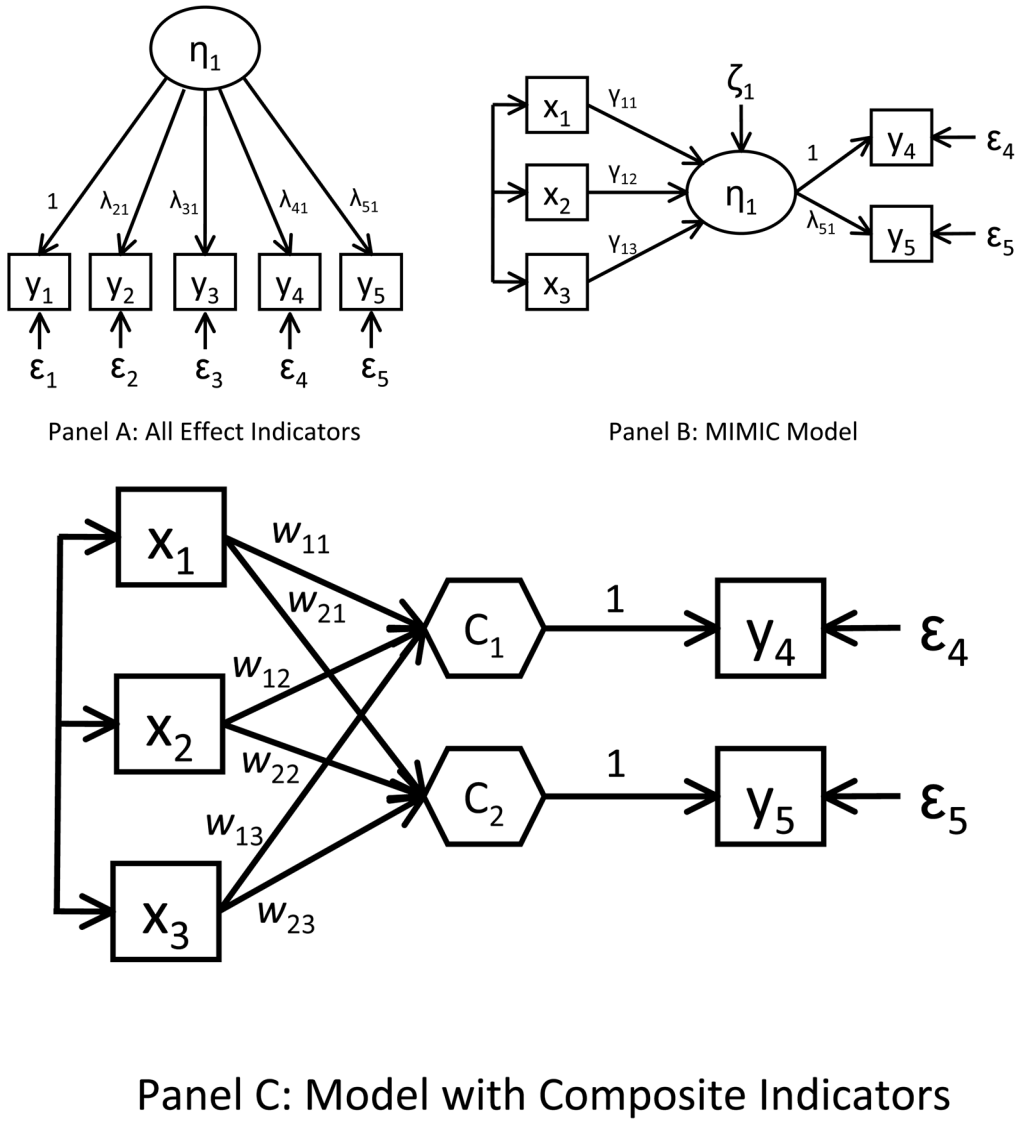

Figure 1. Measurement Models for Perceived Health Example

Notes: η_1 = perceived health. y_1 (x_1) = ill enough to go to a doctor in past year, y_2 (x_2) = hospitalized/disabled last year, y_3 (x_3) = hospitalized/disabled in prior 4 years, y_4 = self-rated health, y_5 = health satisfaction, C_1 and C_2 are the composites representing x_1 through x_3.

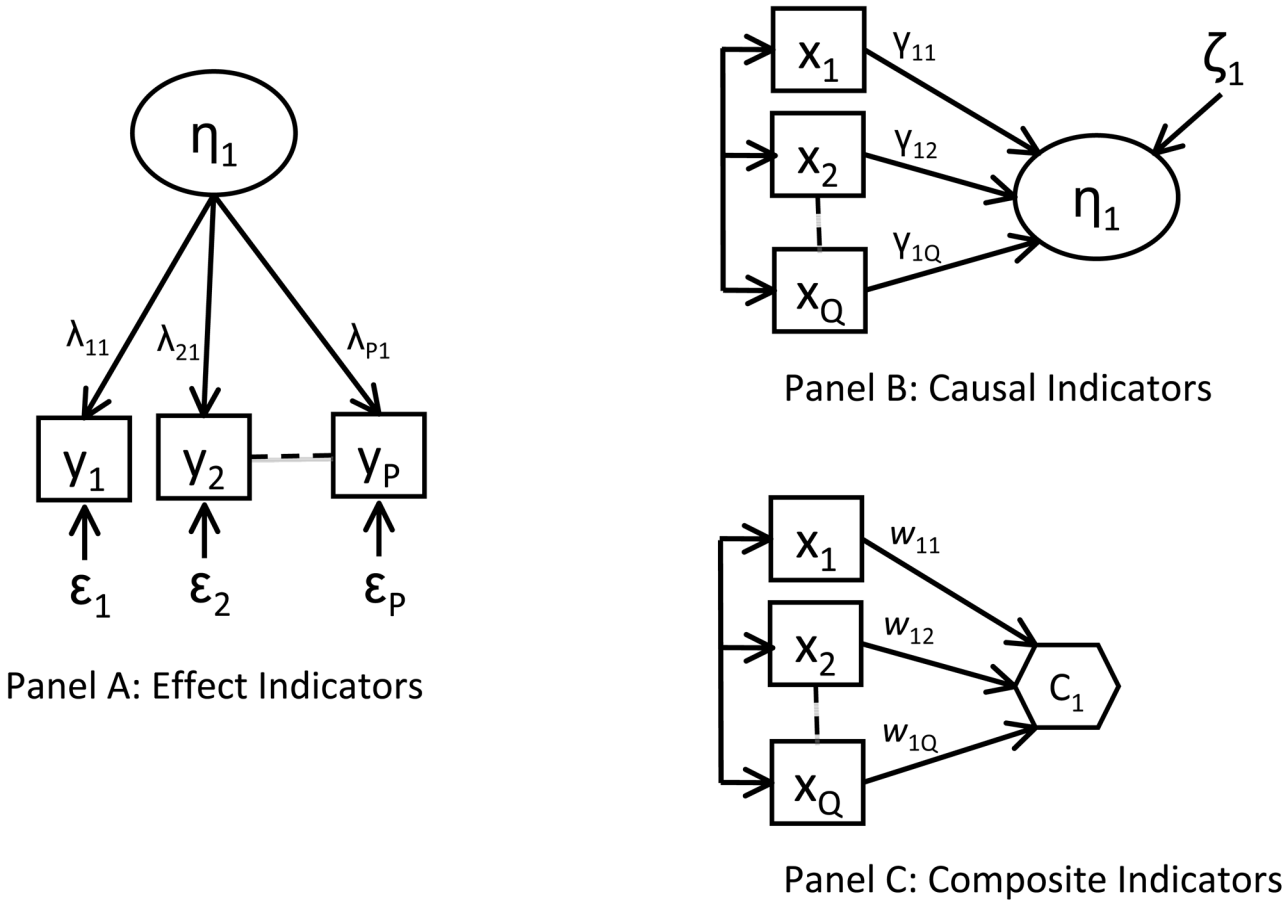

Figure 2.
Types of Measurement Models.

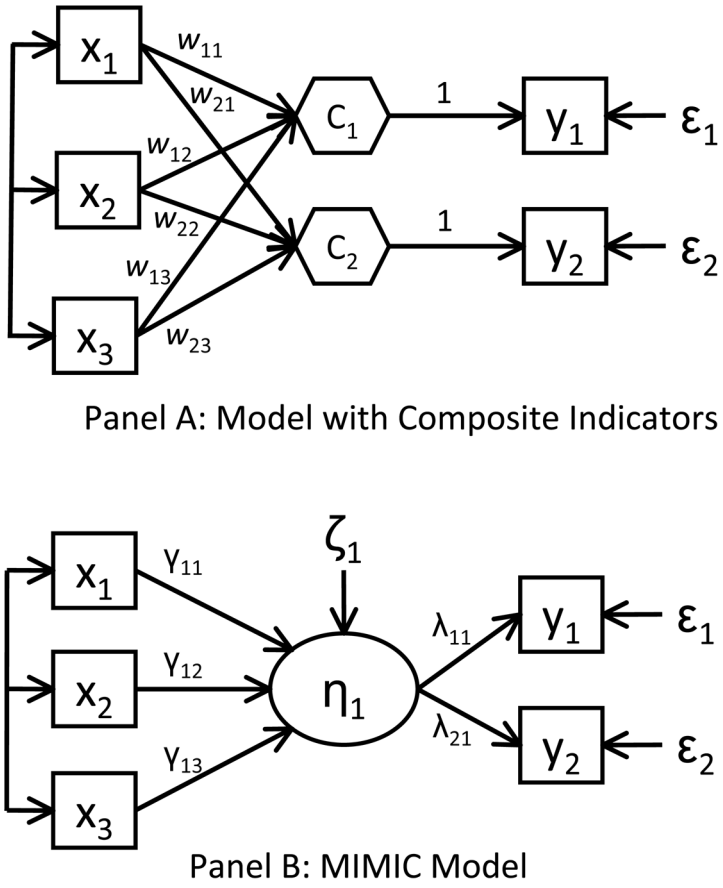

Figure 3.
Comparison of Composite and Causal Indicators.

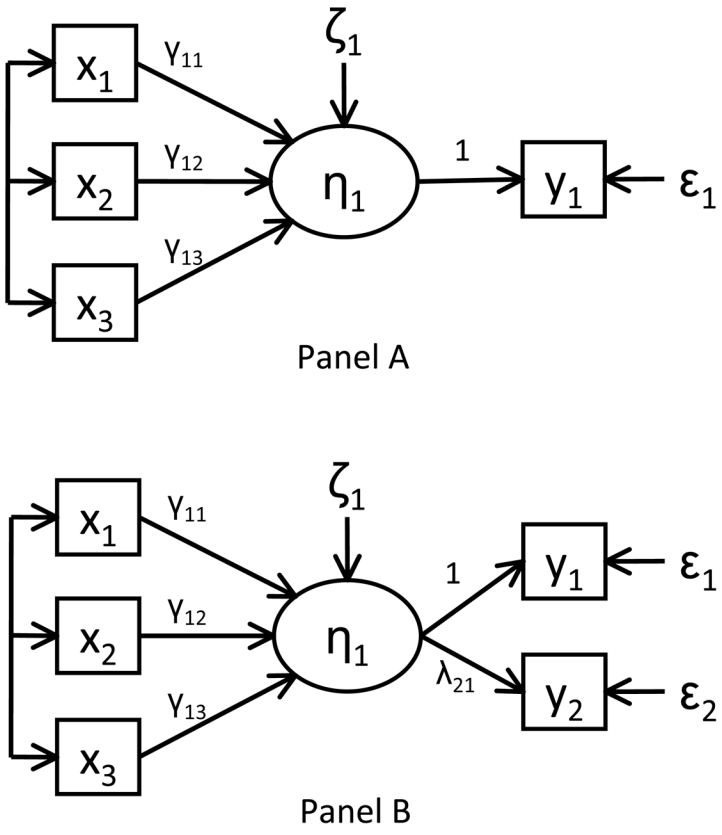

Figure 4.
Identification with Causal Indicators.

Table 1

Descriptive Statistics for GSS Data; N = 483.

	Mean	SD
Age	45.37	17.98
Male	0.44	0.50
Black	0.14	0.35
Other race	0.04	0.20
Ill enough to see doctor	0.54	0.50
Hospitalized/disabled last year	0.18	0.38
Hospitalized/disabled prior 4 years	0.20	0.40
Self-rated health	3.02	0.82
Satisfaction with health	5.43	1.42
Government should help with medical care	2.33	1.14

Table 2

Characteristics of Types of Indicators and Covariates.

	Type of Indicator			
	Effect	**Causal**	**Composite**	**Covariate**
Conceptual unity	√	√		
VAR(ζ) = 0			√	
Stable γs or λs	√	√		
"Latent" variable function of obs. variable		√	√	√
Omitted variable leads to bias		√	√[1]	√
*X*s intercorrelated	√			

[1] See text for discussion of weights that are fixed in advance.

Table 3

Covariance/Correlation Matrix for Health Related Variables; N = 483.

	x_1	x_2	x_3	y_4	y_5
x_1	0.2491	0.3407	0.0148	−0.2308	−0.2731
x_2	0.0641	0.1420	−0.2276	−0.2727	−0.2559
x_3	0.0030	−0.0344	0.1609	−0.0001	−0.0324
y_4	−0.0929	−0.0829	0.0000	0.6505	0.5585
y_5	−0.1883	−0.1332	−0.0179	0.6223	1.9086

Notes: Variances and covariances on and below the diagonal, correlations above the diagonal. x_1 = ill enough to see doctor in past year, x_2 = hospitalized/disabled last year, x_3 = hospitalized/disabled prior 4 years, y_4 = self-rated health, y_5 = health satisfaction.

Table 4

Tetrad Test Results; N = 483.

	Chi-square	df	p-value
Models w/out Demographics			
Effect Indicators	38.76	5	0.000
MIMIC Model	1.66	2	0.436
Nested Test Effect vs. MIMIC	37.10	3	0.000
Models with Demographics			
Effect Indicators	43.19	21	0.002
MIMIC Model	11.64	6	0.071
Nested Test Effect vs. MIMIC	31.56	15	0.007

Table 5

Model Fit Statistics for Effect Indicator Model and MIMIC Model; N = 483.

	Chi-square	df	p-value	CFI	TLI	RMSEA	BIC
Models w/out Demographics							
Effect Indicators	61.39	5	0.00	0.82	0.65	0.16	30.49
MIMIC Model	1.66	2	0.44	1.00	1.01	0.00	−10.70
Composite Indicators	140.09	1	0.00	0.56	−3.35	0.55	133.91
Models with Demographics							
Effect Indicators	119.58	21	0.00	0.75	0.57	0.10	−10.20
MIMIC Model	25.21	6	0.00	0.95	0.70	0.08	−11.87
Composite Indicators	144.85	1	0.00	0.63	−12.29	0.56	138.67

Table 6

Parameter Estimates for Models; N = 483.

	Effect Indicators		MIMIC Model		Composite Indicators			
					C1		C2	
	Est	SE	Est	SE	Est	SE	Est	SE
Factor Loadings/Effects:								
Ill enough to see doctor	−0.32	0.05	−0.28	0.07	−0.24	0.08	−0.56	0.13
Hospitalized/disabled last year	−0.25	0.04	−0.45	0.09	−0.50	0.10	−0.75	0.18
Hospitalized/disabled prior 4 years	0.01	0.04	−0.13	0.08	−0.10	0.09	−0.26	0.16
Self-rated health[1]	1.00		1.00		1.00		-	
Satisfaction with health	1.77	0.21	1.82	0.25	-		1.00	
Variance of Latent Health	0.34	0.05	0.25	0.02	-		-	
R-Square:								
Ill enough to see doctor	0.15							
Hospitalized/disabled last year	0.16							
Hospitalized/disabled prior 4 years	0.00							
Self-rated health	0.53		0.53		0.11			
Satisfaction with health	0.56		0.59				0.10	
Latent Health			0.19					

Notes:

[1] Self-rated health used as the scaling indicator in the effect indicators model and the MIMIC model.

Table 7

Measures of Validity; N = 483.

	$\rho^2_{x_q\eta}$	β	Std(β)	$U_{\eta(xi)}$
See doctor	0.11	−0.28	−0.24	0.04
H/D last year	0.12	−0.45	−0.29	0.09
H/D prior 4 years	0.00	−0.13	−0.09	0.00

Notes: Based on model without covariates. The same pattern of results is observed in the model with covariates.