Journal of Personality and Social Psychology
2011, Vol. 101, No. 6, 1174–1188

0022-3514/11/$12.00 DOI: 10.1037/a0024776

The Trade-Off Between Accuracy and Precision in Latent Variable Models of Mediation Processes

Alison Ledgerwood
University of California, Davis

Patrick E. Shrout
New York University

Social psychologists place high importance on understanding mechanisms and frequently employ mediation analyses to shed light on the process underlying an effect. Such analyses can be conducted with observed variables (e.g., a typical regression approach) or latent variables (e.g., a structural equation modeling approach), and choosing between these methods can be a more complex and consequential decision than researchers often realize. The present article adds to the literature on mediation by examining the relative trade-off between accuracy and precision in latent versus observed variable modeling. Whereas past work has shown that latent variable models tend to produce more accurate estimates, we demonstrate that this increase in accuracy comes at the cost of increased standard errors and reduced power, and examine this relative trade-off both theoretically and empirically in a typical 3-variable mediation model across varying levels of effect size and reliability. We discuss implications for social psychologists seeking to uncover mediating variables and provide 3 practical recommendations for maximizing both accuracy and precision in mediation analyses.

Keywords: mediation, latent variable, structural equation modeling, regression, power

Understanding mechanisms and process is a central focus of social psychologists, and thus few of us are satisfied with a simple empirical claim that variation in some independent variable *X* is related to variation in some outcome variable *Y*, even if *X* is experimentally manipulated. Instead, we often ask, What is the underlying mechanism? The classic tools of mediation analysis proposed by Baron and Kenny (1986) have seemed to provide a good method for finding an answer, but choosing the most appropriate method for testing mediation can be a more complex and consequential decision than researchers often realize. Increasingly, the complexity of the assumptions and threats to validity of inferences based on the Baron and Kenny steps have become clear to substantive researchers (see Bullock, Green, & Ha, 2010). In this article, we add to that cautionary literature and provide tools for moving forward in the quest for studying mechanisms.

Baron and Kenny's (1986) well-known approach asks the researcher to identify and measure an intervening variable, *M*, which appears to be a consequence of *X* and a causal component of *Y*. For example, Hodson and Costello (2007) used mediation analyses to examine the relation between interpersonal disgust (*X*) and attitudes toward immigrants (*Y*). They hypothesized that interpersonal disgust would predict social dominance orientation (*M*), which would in turn predict more negative attitudes toward immigrants. Social psychologists have used such mediation analyses to elucidate underlying processes with respect to a wide range of variables, examining potential mechanisms for observed relations between attitude commitment and selective judgment, construal level and negotiation outcomes, cultural values and socially desirable responding, implicit prejudice and policy judgments, political ideology and valence asymmetries in attitude formation, and a host of other constructs (e.g., Henderson & Trope, 2009; Knowles, Lowery, & Shaumberg, 2010; Kross, Ayduk, & Mischel, 2005; Lalwani, Shrum, & Chiu, 2009; Pomerantz, Chaiken, & Tordesillas, 1995; Shook & Fazio, 2009).

When carrying out such analyses, psychologists frequently use observed variables (e.g., a simple average of scale items), which are typically entered into a series of regression analyses to test for mediation (e.g., Baron & Kenny, 1986; Kenny, Kashy, & Bolger, 1998; MacKinnon, 2008; Shrout & Bolger, 2002; Wegener & Fabrigar, 2000). However, mediation can also be tested with latent variables through structural equation modeling (SEM; see Kline, 2005). Given this choice, which analytic strategy should a researcher use?[1] Statisticians often recommend latent variable models because they allow researchers to adjust for measurement error in the measured variables (e.g., Cole & Maxwell, 2003; Kline, 2005, pp. 44, 70–77; MacKinnon, 2008, pp. 175–176). For exam-

This article was published Online First August 1, 2011.

Alison Ledgerwood, Department of Psychology, University of California, Davis; Patrick E. Shrout, Department of Psychology, New York University.

This research was partially supported by a Hellman Family Foundation Fellowship to Alison Ledgerwood and by National Institutes of Health Grant R01-AA017672 to Patrick E. Shrout. We thank Dave Kenny and Victoria Savalei for their advice on the development of this article.

Correspondence concerning this article should be addressed to Alison Ledgerwood, Department of Psychology, University of California, 1 Shields Avenue, Davis, CA 95616. E-mail: aledgerwood@ucdavis.edu

[1] Observed variable analyses are usually carried out with ordinary least squares regression, but they can also be analyzed with maximum likelihood estimation in structural equation modeling. These estimation methods produce identical results for the coefficients, and so we will treat them as exchangeable in our analyses of observed variables. The key distinction we make in this article is between observed and latent variable models, and in our simulation studies, we hold estimation strategy (maximum likelihood) constant across model types.

ple, Hoyle and Kenny (1999) showed that unreliability in the mediating variable of a three-variable model leads to an underestimation of the mediated effect when analyzing observed variables. Hoyle and Kenny therefore advocated using latent variable models to adjust for the bias produced by measurement error.

If adjusting for measurement error makes the effect size of the mediated path larger, one might expect that it would also increase the power to test this effect. However, the increased accuracy of latent variable models (which often, though not always, produces larger coefficient estimates than those given by observed variable models) is usually accompanied by a decrease in precision: The standard errors of estimates produced by latent variable models can be higher than those produced by observed variable approaches. Indeed, Hoyle and Kenny (1999) noted that the boost to power provided by latent variable models in their study was "minimal" (p. 219). In fact, in some cases, we suspect that latent variable models may actually yield less power than their observed variable counterparts. Somewhat paradoxically, then, latent (vs. observed) variable models could sometimes yield larger but less significant coefficients—increased accuracy but with reduced precision.

Researchers who wish to make an informed decision about which statistical approach to take need to know more about the nature of this accuracy–precision trade-off.[2] Do latent variable models provide much more accurate estimates than observed variable models or just slightly more accurate? Do the standard errors increase dramatically or hardly at all? Which technique is more likely to closely estimate the true relationship between the variables in a given study, and which is more likely to correctly detect an effect by meeting standard cutoffs for significance testing?

To address these questions, we examine a simple three-variable mediation model[3] with a latent variable approach as well as the typical observed variable approach, and we compare both accuracy and precision of estimates. We build on Hoyle and Kenny's (1999) work and extend it in several ways. First, we specifically investigate the relation of the power of mediation tests to the reduction of bias in latent variable models. Second, we consider the consequences of measurement error in both the independent variable (X) and the proposed mediator (M). For researchers studying individual differences or conducting survey-based research, unreliability of X is likely to be a concern because the primary causal variable of interest is often measured rather than manipulated. Moreover, even within experimental studies, researchers are frequently interested in examining mediation paths between measured variables (e.g., Dovidio, Kawakami, & Gaertner, 2002; Knowles et al., 2010; Pomerantz et al., 1995; Shook & Fazio, 2009; Tormala & Petty, 2002). In practice, then, researchers often examine mediation models in which the predictor variable is measured rather than manipulated, and it is therefore important to understand the effects of measurement error in X on accuracy and precision in estimating mediated effects. Finally, we recommend practical approaches for planning studies and analyzing mediation that help minimize the deleterious effects of measurement error.

Illustrating the Problem of Bias With an Empirical Example

To anchor our discussion in the reality of substantive research on attitudes and social cognition, we begin with an example adapted from the study by Hodson and Costello (2007) mentioned earlier. Figure 1 is a three-variable simplification of a mediation model that Hodson and Costello considered to explain the effects of interpersonal disgust on attitudes toward immigrants. Hodson and Costello reported that for every unit increase in interpersonal disgust (measured with a reliability of $\alpha = .61$), attitudes toward immigrants ($\alpha = .80$) became more negative by $-.31$ units—a statistically significant change. When they examined social dominance orientation (SDO; $\alpha = .89$) as a mediator, they found that interpersonal disgust significantly predicted SDO (path $a = .36$), which in turn significantly predicted attitudes toward immigrants (path $b = -.44$), and that this indirect path ($a{*}b = -.16$) was significant, indicating mediation. The amount of the total effect that is not explained by the indirect effect (calculated as $-.31 - (-.16)$) equals the direct effect of disgust on attitudes toward immigrants ($c' = -.15$).

Although Hodson and Costello (2007) reported that interpersonal disgust (the predictor, X) and SDO (the mediator, M) were not measured with complete reliability, they used observed variables that do not adjust for measurement error in their analyses. According to Hoyle and Kenny (1999), if the reliability of M is ignored, the obtained path from M to Y is $R_M b$ rather than b, where R_M is the reliability of the measure of the mediator after partialing out the X variable. Applying their adjustment, we can obtain a ballpark estimate of how biased the coefficients in Figure 1 might be. The partialed reliability of the mediator is .88, and the inferred unbiased estimate of the b effect would be $-.44/.88 = -.50$, instead of the obtained $-.44$. If we stopped with this informal analysis, we would conclude that the degree of bias is small but that the extent to which SDO mediated the link between disgust and attitudes may have been underestimated.

Hoyle and Kenny (1999) only discussed the effects of unreliability of the mediator, but a number of authors have considered unreliability in a system of equations (e.g., Cohen, Cohen, West, & Aiken, 2003, pp. 55–57; Duncan, 1975, Chapter 9; Heise, 1975, pp. 188–191; Kenny, 1979, Chapter 5). The impact of unreliability of X is complicated. On the one hand, it will generally lead to an underestimation of the effect of X on M (path a in Figure 1) and of the direct effect of X on Y (path c' in Figure 1). On the other hand, because it underestimates the direct effect, it also undercorrects the path from M to Y, and this can lead to an overestimation of the effect of M on Y (path b in Figure 1) in the system of equations.

[2] Latent variable versus observed variable models are not the only contexts where statisticians confront bias–precision trade-off issues. Statistical texts on estimation (e.g., Mood, Graybill, & Boes, 1974) show how bias and mean square error are distinct features of alternate estimators.

[3] We examine these issues in the context of the three-variable model first considered by Judd and Kenny (1981; see also Baron & Kenny, 1986) and further explicated by Hoyle and Kenny (1999) because it simplifies the presentation, but note that we do not necessarily endorse this model for definitive mediation analyses. A number of commentators have noted that the assumptions of the simple model need to be examined seriously, particularly with regard to the independence of the error terms associated with M and Y, and with regard to the need for manipulation of both X and M (see, e.g., Bullock et al., 2010; MacKinnon, 2008; Shrout, 2011; Spencer, Zanna, & Fong, 2005). We take up the issue of how the accuracy–precision trade-off would generalize to more complicated mediation models later in the article.

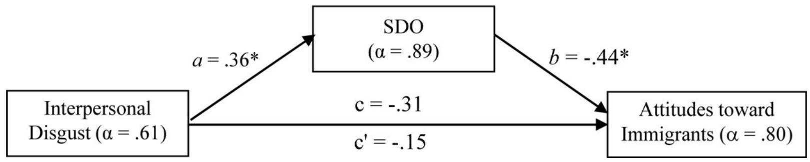

Figure 1. Hodson and Costello's (2007) observed variable model (averaged items contaminated by measurement error). Path *c* denotes the full effect, and *c*′ denotes the direct effect of interpersonal disgust on attitudes.

Because the indirect effect is the product $a{*}b$, the two biases can sometimes offset each other. An additional consideration is whether standardized or unstandardized coefficients are of interest. When effects are in standardized form, the estimates are affected by the reliability of the outcome as well as the predictors.[4] In the Appendix, we provide the equations for calculating the effects of unreliability in *X*, *M*, and *Y*, and we describe a tool that can be downloaded to allow convenient implementation of the equations.

When we use these equations to understand the impact of measurement error in the Hodson and Costello (2007) example, we estimate that the effect of interpersonal disgust on SDO (path *a*) might be a full third larger (.49 rather than .36) if interpersonal disgust had been measured more reliably. However, the overall bias in the estimated effect of SDO on attitudes toward immigrants is offset somewhat by the unreliability of interpersonal disgust as an adjustment variable. The equations in the Appendix suggest that the bias in the estimate of path *b* is slight, moving only from −.44 to −.49, when the full system of variables is taken into account. Both these equations affect the estimates of the indirect (mediated) effect and the direct effect. The indirect effect increases from −.16 to −.24 after adjustment, and the direct effect increases from −.15 to −.18. (Both are able to increase because the total effect increases from −.31 to −.44.) In this particular case, then, the biases created by unreliability of *X* happen to offset each other, and the interpretation is therefore similar before and after adjustment. There is an effect to be explained, and about half the effect is explained by the mediating variable.

This example illustrates the complexity and interdependence of these biases, but neither it nor the equations in the Appendix necessarily help yield an intuition about the impact of measurement error in *X* and *M*. It is possible to develop such intuitions by using a range of different values and making a plot. Building on the Hodson and Costello example, we contrasted two numerical examples that had effect sizes with similar magnitudes for path *a* ($X \rightarrow M$) and path *b* ($M \rightarrow Y$). We further elaborated these examples to have no direct effect of *X* on *Y* ($c' = 0$) or a medium direct effect of *X* on *Y* ($c' = .3$). The example representing a medium indirect effect had $a = .538$ and $b = -.538$ (so that $a{*}b = .3$), and the one representing a large indirect effect had $a = .707$ and $b = -.707$ (so that $a{*}b = .5$). Using the formulas in the Appendix, we constructed Figure 2 to show the impact of different reliability values for *X* and *M* for these four examples in terms of the standardized indirect effect $a{*}b$. Note that measurement error has the largest impact when it is the mediator that is poorly measured. The effect of measurement error for the initial *X* variable is largest when the mediator is measured reliably; this is apparent from the fact that the lines for different *X* reliability values are more widely spread when R_M is close to 1.0 and draw together as R_M decreases.

Correcting Bias With Latent Variable Models: Theory and Practice

In contrast to the usual observed variable approach to mediation analysis, SEM models that employ latent variables can adjust for measurement error and produce estimates of the direct and indirect effects that are unbiased in very large samples. Figure 3 illustrates the latent variable model. Instead of using a single indicator of the *X* and *M* constructs, the researcher defines multiple indicators of each construct, which are represented as latent variables by ovals rather than boxes in Figure 3. These indicators might be alternate forms of measures from the literature, or they might be defined as item parcels from a multi-item measure of a construct (Little, Cunningham, Shahar, & Widaman, 2002). Latent variable models account for measurement error by separating the variance common to all the indicators of a particular construct from the variance unique to a particular indicator (which includes measurement error). This separation allows the definition of the latent variables, which are free from measurement error if the model is properly specified. The latent variables can then be used in the mediation model to produce estimates of direct and indirect effects that in theory have no statistical bias (Hoyle & Kenny, 1999; Kline, 2005). Moreover, the results of some simulation studies seem to suggest that using latent variables will increase path coefficient estimates and therefore increase power (e.g., Coffman & MacCallum, 2005; Stephenson & Holbert, 2003).

However, as mentioned earlier, the standard errors of the unbiased latent variable estimates are often larger than those of the biased estimates produced by observed variable models—a fact that is typically overlooked in discussions of latent variable analyses (DeShon, 2006). This makes inference about mediation more complicated. Although mediation effect estimates will on average be more accurate when using latent variables (vs. observed variables), these estimates will tend to vary more across studies. Thus, in any one particular study, it is possible that the estimates produced by a latent variable approach could vary quite a bit from the average estimate. Moreover, although a latent variable approach can boost power by reducing the attenuation in estimates caused by measurement error, the larger standard errors will reduce power, potentially canceling or even outstripping the power boost provided by a larger estimated effect. This means that an investigator could observe an apparent significant indirect effect based on biased observed variable analyses (e.g., regression analyses) but then find that the larger, unbiased estimate is no longer statistically

[4] If the outcome *Y* is measured with error, its variance increases, and this will affect standardized but not unstandardized coefficients.

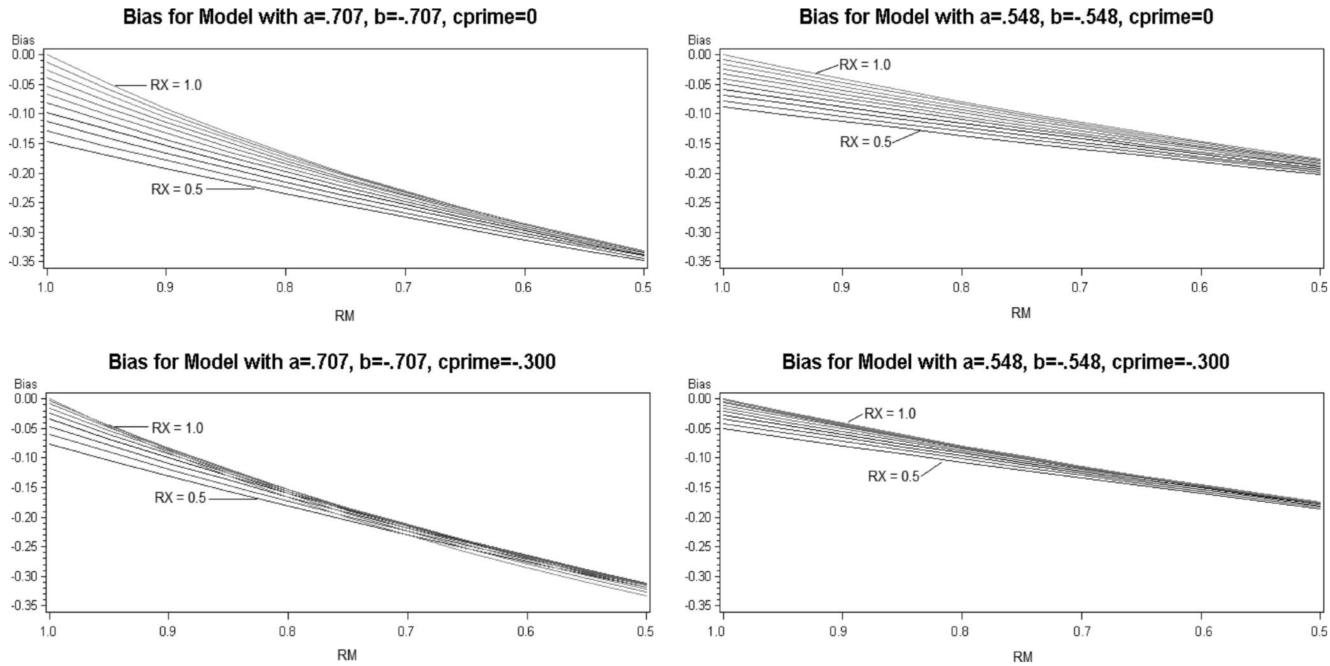

Figure 2. Bias in large and medium indirect effects (path $a*b$) at varying levels of reliability in X and M. RM = reliability of M, plotted along the horizontal axis; RX = reliability of X, ranging from 1.0 (top line) to 0.5 (bottom line).

significant when a SEM latent variable model is used with the same data.

Just as we can study the expected bias of observed versus latent variable model estimates using statistical theory, so can we study their expected precision. For example, Cohen et al. (2003, p. 87) provided expressions for the expected standard error of observed variable estimates that depend on sample size; the variances of X, M, and Y; and their degree of association. Bollen (1989, p. 134) described more complicated expressions for standard errors for latent variable estimates that are obtained from maximum likelihood methods. In the Appendix, we review these expressions and describe how the asymptotic standard errors can be computed and compared. These are the expected standard errors of the estimates if we knew the true underlying variances and covariances, rather than having to estimate them.

Assessing the Theoretical Accuracy–Precision Trade-Off

The calculations of bias and precision from the Appendix can be used to assess the expected (i.e., theoretical) trade-off between accuracy and precision as one moves from a regression analysis using observed variables to a SEM analysis using latent variables. As in Figure 2, we decided to focus on a few examples that were loosely based on the Hodson and Costello example. We varied the size of the mediated effect ($a*b$) to correspond to Cohen's (1988) approximations of small, medium, and large effect sizes ($\eta^2 = .1$, .3, and .5, respectively), and we considered a case in which all the constructs were measured with high reliability ($R = .9$) or moderate reliability ($R = .7$).

To examine the relative bias produced by an observed variable approach, we computed the ratio of the observed variable model's

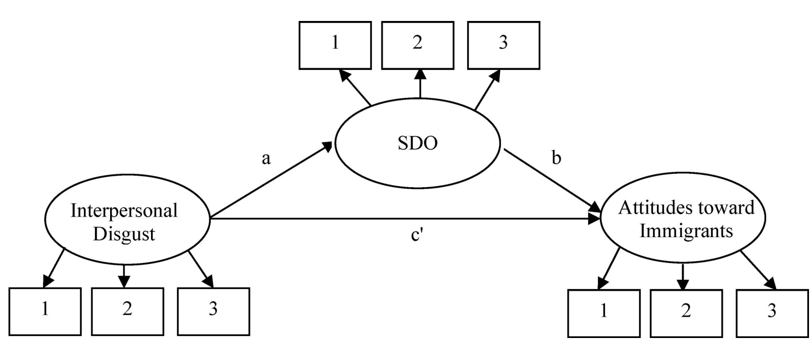

Figure 3. Latent variable model (items as indicators). SDO = social dominance orientation.

expected estimate to the unbiased values: $R_{\text{estimate}} = \text{Estimate}_{\text{observed}}/$ Population Value. A ratio of 1 would therefore indicate no bias (in other words, across samples, the estimates should accurately center around the true population parameters). The lower this ratio falls below 1, the more biased the observed variable approach that ignores reliability is expected to be (i.e., across samples, the estimates will center around a less and less accurate value). For large samples, R_{estimate} describes the ratio of $\text{Estimate}_{\text{observed}}$ to $\text{Estimate}_{\text{latent}}$.

To examine the relative precision of an observed versus latent variable approach, we computed the ratio of the expected standard error produced by an observed variable approach to that produced by a latent variable approach: $R_{SE} = SE_{\text{observed}}/SE_{\text{latent}}$. Insofar as R_{SE} falls below 1, it suggests that the estimates of the latent variable model are more affected by sampling variation (and are therefore more widely scattered across samples) than those estimated by the observed variable approach that ignores measurement error.

To portray the trade-off between these two aspects of the analysis—accuracy and precision—for estimates of the indirect effect $a{*}b$, we computed R_{estimate} and R_{SE} ratios for different effect sizes and levels of reliability and plotted these against each other in Figure 4A. If the points in this plot fell on the diagonal, we could infer that the difference in bias between the two approaches was exactly offset by the difference in imprecision. If this were the case, the power of each approach to detect the coefficient would be equal (assuming the test is the usual Wald test of the estimate divided by its standard error). Notably, the ratios in Figure 4A all fall below this diagonal line, indicating that the increased imprecision of the latent variable approach outweighs its reduction in bias. From a purely statistical standpoint then, we should expect latent (vs. observed) variable approaches to be more accurate but less powerful.

A Gap Between Theory and Practice: The Issue of Small Samples

Figure 4A is based on statistical theory that assumes sample size is not important. However, in practice, sample size can have substantial consequences for the validity of statistical inferences based on these two methods. Under certain assumptions, inferences for observed variable analyses (e.g., regression analyses) are valid for small sample sizes, but inferences for latent variable analyses assume large samples (Bollen, 1989; Kline, 2005). More specifically, significance tests for observed variable methods are appropriate for small samples if the residuals are independent and normally distributed; the conventional degrees of freedom in the t and F tests explicitly take the small samples into consideration (Cohen et al., 2003, pp. 88–90). Significance tests for latent variable methods, on the other hand, assume that the sample is large enough to make worrying about small sample adjustments unnecessary. This means that in latent variable models, ratios of parameter estimates (a, b, and c') to their standard errors are compared to standard normal (z) distributions rather than Student's t distribution.[5] The smaller the sample size, the larger the impact of this difference between observed and latent variable model assumptions.

Recognition that the SEM methods needed to analyze latent variable models assume large samples has led a number of methodologists to suggest rules of thumb for how large the sample must be. When analyzing simple path models (as in Figure 1) using SEM, the question turns on when the z and t statistics are basically the same. For example, if the sample is approximately 30 in each of two treatment groups, the degrees of freedom would be $60 - 3 = 57$, and the critical value for a two-tailed t test would be $t_{.05}(57) = 2.00$ rather than $z_{.05} = 1.96$. If an investigator found that the ratio of the estimate to its standard error was exactly 1.96, the test would be just significant at $p = .05$ with a z test, but it would have a $p = .0549$ if evaluated with a t test on 57 degrees of freedom. Most would agree that this difference would not open the floodgates of Type I error. From this perspective, a sample size of 60 or more can be considered large enough.

However, a different problem arises when one fits SEM models such as the one in Figure 3, which are called hybrid models because they include both latent variables and structural paths (Kline, 2005, pp. 208–233). When the sample size is small, the data are often inadequate to provide a stable numerical solution to the model. For example, Hoyle and Kenny (1999) simulated data sets with sample sizes of 25, 50, 100, and 200 and fit the data with a hybrid model similar to the one shown in Figure 3. In 522 simulated data sets (out of 24,000), the SEM analysis had estimation problems (failure to converge, impermissible values of estimates or standard errors), and nearly all of these occurred in samples with 25 or 50 observations. This result led Hoyle and Kenny to recommend that investigators aspire to obtain sample sizes of at least 100 when planning to use SEM latent variable models to adjust for statistical bias in mediation analysis.

Because latent variable analyses can become unstable when applied to the sample sizes that are often used in social psychology, we cannot be confident that the asymptotic results we just reviewed apply to real-world empirical studies. To determine whether the same trade-off is observed in smaller samples, we carried out a series of simulation studies that allowed us to look at the average size of the estimate and the average standard error in smaller samples. These studies extend the work of Hoyle and Kenny (1999) by considering bias correction when X and Y as well as M are measured with error and by considering the precision–accuracy trade-off.

Simulation studies are carried out with computer programs to create data that are consistent with a known population model. These data can be analyzed with different analytic methods to see how well each method can recapture the known true values used in the simulations. Because the data are constructed with a combination of known structure (the true values) and random numbers from random number generators (to reflect the kind of sampling fluctuation researchers encounter in the real world), it is possible to study both bias and precision of the estimates produced by each analytic method by creating a large number of samples and studying the distributions of each method's estimates. In our simula-

[5] When researchers wish to test the indirect effect, $\hat{a}{*}\hat{b}$, they are advised to use bootstrap methods of inference rather than the usual ratio of the estimate to its standard error (e.g., Shrout & Bolger, 2002). This is not because estimated standard errors of the indirect effect are inaccurate, but because the distribution of the estimated indirect effect is nonnormal (MacKinnon et al., 2002). This advice applies to tests of the indirect effect in both observed and latent variable models. Although the bootstrap is needed for accurate inference, the standard error still contains useful information about the relative precision of the estimator.

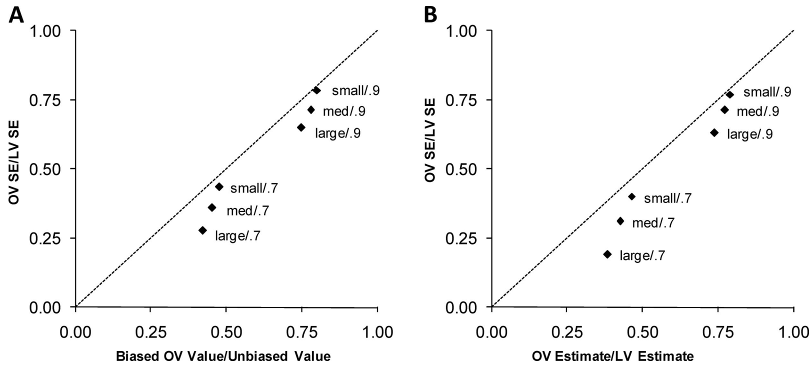

Figure 4. The accuracy–precision trade-off for testing the indirect effect in a three-variable mediation model with no direct effect (labeled by effect size and reliability). Dotted line indicates the border at which the increased accuracy provided by a latent variable (LV) approach ought to be exactly offset by an increase in standard error. (A) Theoretical trade-off based on computed expected values: on the x-axis, ratio of expected, biased observed variable (OV) estimates to unbiased value for the indirect effect (path $a*b$); on the y-axis, ratio of expected OV standard error to expected LV standard error. (B) Corresponding results based on actual values obtained in simulation studies (sample size = 100): on the x-axis, ratio of average coefficient estimates for the indirect effect ($a*b$) produced by OV versus LV approaches; on the y-axis, ratio of mean standard errors produced by OV versus LV approaches. Med = medium.

tions, we used the same models that we created for the evaluation of the asymptotic findings in Figure 2. As described in detail in the next section, we created models that had small, medium, or large effect sizes for the indirect effect; half of these also included a direct effect of X on Y, whereas half did not. Beyond the structure of the mediation model, we created indicators that had either modest reliability or high reliability. This was done by adding different amounts of random noise to the variables involved in the mediation. Each of the many simulated data sets was then analyzed twice, once with observed variable methods that ignored the problem of unreliability and once with latent variable models that attempt to adjust for unreliability.

Method

To anchor our study in a real-world example, we based our simulations on Hodson and Costello's (2007) mediation model described earlier, in which SDO mediates the relation between interpersonal disgust and attitudes toward immigrants (see Figures 1 and 3). The data were generated to be consistent with the Figure 3 version of this model, so that the first latent variable (interpersonal disgust) positively influenced the second (SDO), which negatively influenced the third (attitudes toward immigrants). As we did in constructing Figures 2 and 4, we kept the absolute size of the a effect equal to that of the b effect (see Table 1), and we set these to create indirect effects ($a*b$) that varied in size from small ($a*b = -.1$) to medium ($a*b = -.3$) to large ($a*b = -.5$). To increase the generalizability of our findings, we created one mediation model (Model 0) in which the direct effect c' was set to 0 (consistent with a situation in which the association between X and Y was completely explained by the intervening mediator M) and one mediation model (Model 1) similar to the results reported by Hodson and Costello with a medium direct effect ($c' = -.3$). We defined the latent variables so that they would have expected variances of 1. We did this by fixing the population variance of the latent variable representing interpersonal disgust to 1, and then we computed the variance of the disturbance of the mediator (SDO) as $1 - a^2$ and the variance of the disturbance of the outcome variable (attitudes) as $1 - (b^2 + c'^2 + 2abc')$.

Next, we simulated three indicators for each of the three factors. When creating the indicators, we added either a small or moderate amount of random noise to manipulate the reliability of the measures between high ($R = .9$) and moderate ($R = .7$). The paths between the latent variables and the indicators were set equal to 1, and the reliability was manipulated by adding measurement error variance to the indicators. We calculated the item error variances for the desired values for the two levels of reliability R using the formula $3(1 - R)/R$, where 3 reflects the number of items in the scale.

Hodson and Costello (2007) reported a sample size of close to 100 ($N = 103$). Hoyle and Kenny (1999) suggested that latent

Table 1
Actual Values Used in Simulation

Effect size[a]	a	b	c' (direct)	$a*b$ (indirect)
Model 0 ($c' = 0$)				
Small (.1)	.316	−.316	0	−.100
Medium (.3)	.548	−.548	0	−.300
Large (.5)	.707	−.707	0	−.500
Model 1 ($c' = -.3$)				
Small (.1)	.316	−.316	−.300	−.100
Medium (.3)	.548	−.548	−.300	−.300
Large (.5)	.707	−.707	−.300	−.500

Note. a = direct effect of interpersonal disgust on social dominance orientation; b = direct effect of social dominance orientation on attitudes toward immigrants; c' = direct effect of interpersonal disgust on attitudes toward immigrants; $a*b$ = indirect effect of interpersonal disgust on attitudes toward immigrants.
[a] Indirect effect of disgust on attitudes.

variable models could be fruitfully employed with sample sizes of at least 100 (with smaller sample sizes, they found that estimates became less accurate and technical problems were more likely to surface, as noted earlier). Given that small sample sizes are common in the field of social psychology, we used a sample size of 100 in our simulations to allow for appropriate generalizations to the typical studies conducted in this field (see e.g., Paxton, Curran, Bollen, Kirby, & Chen, 2001). In our study, we created 1,000 data sets of 100 observations each using Mplus software (Muthén & Muthén, 2010) for each cell of a 3 (indirect effect size: small vs. medium vs. large) × 2 (reliability: moderate vs. high) × 2 (direct effect model: Model 0 vs. Model 1) design. (See http://sites.google.com/site/ledgerwoodshrout2011 or e-mail the authors for examples of the syntax used to generate a simulation.)

Next, we analyzed the data sets we had simulated for each of the 12 cells of the study design using two analytic strategies. In the first (observed variable analysis), we averaged the three items for each variable together to form composite scores and then used simple regression methods to estimate paths *a*, *b*, and c', as in Figure 1. Because we want to consider both accuracy and precision, we focused on unstandardized coefficients. In the second (latent variable analysis), we allowed the items to act as indicators for their respective factors and fit the data to a SEM model as in Figure 3.[6] We conducted both sets of analyses in Mplus (Muthén & Muthén, 2010) using maximum likelihood estimation.

The two sets of analyses allow us to examine the distribution of the estimates over the 1,000 data sets. If an estimate is unbiased, the mean of the 1,000 sample estimates will be close to the population value used to create the data (see Table 1). If an estimate from one method is more precise than the other, the standard deviation of the 1,000 sample estimates will be smaller for the one than the other. The standard error from a single analysis is designed to approximate the standard deviation from the multiple replications. Just as we can examine the standard deviations for estimates of *a*, *b*, and c', we can also examine the standard deviation of the product of the *a* and *b* estimates when comparing observed variable and latent variable estimation methods. These standard deviations are well estimated by the Sobel standard error formula (Sobel, 1982) and related estimates (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002).

In addition to comparing the precision of observed and latent variable estimates of the indirect effect using the standard deviations from the simulations, we considered carrying out bootstrap analyses (Efron & Tibshirani, 1993; Shrout & Bolger, 2002) on each of the 1,000 simulated data sets within each condition. Significance tests of the indirect effect based on bootstrapping have been shown to be more accurate than those based on normal theory tests (MacKinnon, Lockwood, & Williams, 2004), and bootstrapping is available for a single analysis in structural equation programs such as Mplus for both observed and latent variable models. However, bootstrapping is not currently available in Mplus simulations, and hence we could not provide an easily implemented method for readers to replicate and extend our results. For this reason, we rely on the standard error of the indirect effect as a useful method for comparing the relative precision of observed and latent variable models across our aggregated data sets, even as we recommend using the bootstrap method for ultimately reporting results from a single study.[7]

Results

The latent variable SEM model failed to converge for three of the 12,000 data sets (all in the moderate reliability, large effect size condition: two for Model 0 and one for Model 1), and so the estimates from these data sets were excluded from subsequent analyses. For each remaining data set, estimates of the indirect effects for the path and hybrid models were computed by multiplying $a*b$ (see Figure 3), and the associated standard errors were calculated in Mplus with a formula similar to the Sobel standard error recommended by Baron and Kenny (1986). Whereas the Sobel approach assumes that the estimates for paths *a* and *b* are uncorrelated, Mplus uses a multivariate delta method and does not make this assumption (see Bollen, 1989, p. 391).

Table 2 shows the mean unstandardized estimates of the direct and indirect effects, the mean standard errors associated with these effects, and the average magnitude of bias (computed as the difference between the average coefficient estimate and the true parameter value used in the simulation) produced by each analysis strategy at each level of reliability, indirect effect size, and direct effect model. The top portion of the table shows the results for the observed variable analysis, and the bottom portion shows results for the latent variable analysis.

Accuracy

The observed variable analyses yielded systematically biased estimates, consistent with predictions from psychometric theory represented in the Appendix. As expected, the estimates of path *a* in these analyses were attenuated by a factor equal to the underlying reliability of the composite measures (see, e.g., Kenny, 1979, Chapter 5). For example, in the medium effect size condition, the path *a* that was set to .548 in the simulation was estimated to be

[6] We modeled *Y* as a latent variable in these analyses simply to reflect what social psychologists often do in practice. Note, however, that leaving it as an observed variable will produce very similar results, as error in the outcome variable does not create bias in the coefficient estimates.

[7] To verify the assumption that the standard error of the indirect effect provides information about precision that is comparable to the bootstrap method, we sampled 54 simulated data sets (nine from each of the conditions shown in Table 1) to represent the range of precision conditions, and we carried out bias-corrected bootstrap analyses of the indirect effect within each individual data set using Mplus for both the observed variable model and the latent variable model. As a measure of precision, we calculated the width of the bootstrap confidence interval from the estimate to the bound relevant to the significance test (the upper bound in our example). We compared this bootstrap measure with the estimated standard error across the 56 samples. The two measures of precision were highly correlated: .97 for observed variable models and .94 for the latent variable models. We also found that the ratios of the bootstrap interval to the standard error were very similar for the two kinds of models: 1.63 for the observed variable indirect estimate and 1.60 for the latent variable indirect effect. Had the normal theory test of the indirect effect been equivalent to the bootstrap test, we would have expected these ratios to be 1.96 on average. Because the deviation from this value was comparable for both observed and latent variable models, we can conclude that comparing the precision of the two types of models based on standard errors will lead to the same conclusions about relative precision as would using an alternative measure based on bootstrap methods. We thank our reviewers for helping us to clarify this and other points.

Table 2
Observed and Latent Variable Results of Mediation Model in the Presence of Measurement Error Shown for Each Cell of the 2 (Reliability) × 3 (Indirect Effect Size) × 2 (Direct Effect Model) Design: Mean Coefficients, Mean Standard Errors, and the Average Magnitude of Bias (the Absolute Difference Between the Mean Coefficient and the Actual Parameter Values)

				a			*b*			*c′* (direct)			*a*∗*b* (indirect)		
Mediation	*R*	Effect size	*N*[a]	*B*	*SE*	Bias	*B*	*SE*	Bias	*B*	*SE*	Bias	*B*	*SE*	Bias
Observed variable results															
Model 0	.7	Small	1,000	.219*	.097	.097	−.213*	.100	.103	−.022	.100	.022	−.047	.032	.053
		Medium	1,000	.381*	.092	.167	−.352*	.100	.196	−.075	.100	.075	−.134*	.051	.166
		Large	1,000	.493*	.087	.214	−.422*	.099	.285	−.140	.099	.140	−.208*	.062	.292
	.9	Small	1,000	.283*	.096	.033	−.279*	.100	.037	−.010	.100	.010	−.079*	.040	.021
		Medium	1,000	.492*	.087	.056	−.473*	.100	.075	−.037	.100	.037	−.233*	.065	.067
		Large	1,000	.635*	.077	.072	−.585*	.100	.122	−.077	.100	.077	−.372*	.079	.128
Model 1	.7	Small	1,000	.219*	.097	.097	−.234*	.096	.082	−.227*	.096	.073	−.051	.033	.049
		Medium	1,000	.381*	.092	.167	−.392*	.090	.156	−.269*	.090	.031	−.150*	.051	.150
		Large	1,000	.493*	.087	.214	−.482*	.082	.225	−.321*	.082	.021	−.237*	.059	.263
	.9	Small	1,000	.283*	.096	.033	−.289*	.093	.027	−.278*	.093	.022	−.082*	.040	.018
		Medium	1,000	.492*	.087	.056	−.493*	.083	.055	−.297*	.083	.003	−.243*	.060	.057
		Large	1,000	.635*	.077	.072	−.618*	.065	.089	−.326*	.065	.026	−.393*	.064	.107
Latent variable results															
Model 0	.7	Small	1,000	.320*	.163	.004	−.323	.180	.007	.003	.167	.003	−.101	.080	.001
		Medium	1,000	.557*	.178	.009	−.569*	.239	.021	.011	.209	.011	−.313	.164	.013
		Large	998	.718*	.189	.011	−.762*	.380	.055	.037	.342	.037	−.542	.322	.042
	.9	Small	1,000	.316*	.110	.000	−.316*	.118	.000	.000	.115	.000	−.100	.052	.000
		Medium	1,000	.549*	.107	.001	−.549*	.131	.001	.000	.123	.000	−.301*	.091	.001
		Large	1,000	.709*	.102	.002	−.711*	.149	.004	.002	.138	.002	−.503*	.125	.003
Model 1	.7	Small	1,000	.318*	.162	.002	−.321	.173	.005	−.308	.171	.008	−.097	.075	.003
		Medium	1,000	.555*	.176	.007	−.563*	.218	.015	−.305	.196	.005	−.304*	.146	.004
		Large	999	.717*	.184	.010	−.757	.434	.050	−.277	.392	.023	−.535	.370	.035
	.9	Small	1,000	.316*	.110	.000	−.316*	.110	.000	−.302*	.110	.002	−.099*	.050	.001
		Medium	1,000	.549*	.107	.001	−.549*	.111	.001	−.302*	.105	.002	−.300*	.081	.000
		Large	1,000	.709*	.102	.002	−.710*	.106	.003	−.301*	.094	.001	−.502*	.096	.002

[a] Number of simulated studies used to calculate the statistics in the table.
* Average coefficient would be significant at the $p < .05$ level.

.381 in both Model 0 and Model 1 when reliability was moderate, which is approximately $a*.7$.

The amount of attenuation for effect *b* was not precisely proportional to the reliability, but the bias did increase as reliability decreased.[8] For example, in the medium effect size condition of Model 0, the average estimate for the path *b* that was set to −.548 in the simulation was −.473 when reliability was high and decreased to −.352 when reliability was moderate.

The bias associated with the direct path c' was affected by both the level of reliability and the sizes of the *a* and *b* effects. In Model 1, path c' was underestimated in some cases and slightly overestimated in others, and bias was generally reduced as reliability increased. In Model 0, where path c' was set to 0 in the simulation, the average estimates produced by the observed variable approach deviated from 0—an issue to which we return in our discussion of power and Type I error.

The bias associated with the indirect path ($a*b$) was approximately equal to the product of the reliabilities of the two composite measures. For instance, when the reliability of the two measures was moderate (.7), the estimated effect was approximately half (.49) the size of the effect used to generate the data.

In contrast, the latent variable analysis produced estimates that tended to be centered almost exactly at the actual parameter values (see Table 2, bottom portion). This pattern of differences between the two approaches occurred across all cells of the design but was most striking when scale reliability was moderate. For example, in Model 1, the medium indirect effect of −.300 was on average estimated as −.304 by a latent variable approach, but as only −.150 by the observed variable approach that ignored measurement error.

Precision

Table 2 reveals other remarkable differences between observed and latent variable approaches to mediation analyses, namely the sizes of the standard errors of the estimates for the two analytic strategies. Consistent with the results from large sample statistical theory, the typical standard errors produced by the observed variable analysis of the composite scores were fairly small for the samples with $N = 100$, but those produced by the latent variable approach were considerably larger. Again, this pattern occurred across all cells of the design but was exacerbated as reliability decreased and effect sizes increased, and was most pronounced for the indirect effects.

[8] The relation of attenuation to reliability is complicated in observed variable analyses because the measurement error affects both the direct path and the adjustment effects.

The standard errors produced by the latent variable approach were especially inflated for the moderate reliability, large effect size condition. A closer examination of the distributions of estimates in this cell reveals that the latent variable approach produced a handful of wildly inaccurate outliers. For instance, Figure 5 shows the distribution of estimates for path $a*b$. Although the majority of the estimates are accurately clustered around the true parameter value of $-.500$, a number of them deviate quite substantially (note also that the negatively skewed distribution pulls the mean estimate down below the true parameter value).

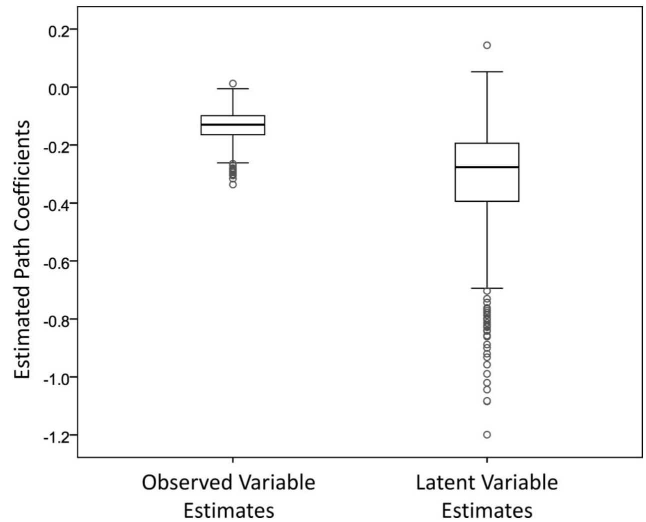

Figure 6. Comparison of the distribution of estimates produced by observed versus latent variable approaches for path $a*b$ in the moderate reliability, medium effect size condition for Model 0.

The Bias–Precision Trade-Off in Samples With $N = 100$

Just as we examined the accuracy–precision trade-off when considering statistical theory (see Figure 4A), we can plot the relative precision of the latent and observed variable approaches against their relative accuracy using the empirical small sample results from Table 2. Figure 4B shows this trade-off based on the average estimates of the indirect effect in our simulation study on samples with $N = 100$, as a function of effect size and reliability for Model 0. The actual results in Figure 4B closely mirror the theoretically expected values in Figure 4A. The points all fall below the dotted line, indicating that the latent (vs. observed) variable approach produced estimates that are more accurate but less powerful, especially as reliability decreases and effect size increases.

Taking the moderate reliability, medium effect size cell of Model 0 as an example, we can also compare the range of estimates produced by the two analysis strategies (see Figure 6 for a visual comparison of the estimates produced by each approach for the indirect effect in this cell). The sizes of the estimates produced by the observed variable approach that ignored measurement error were typically smaller than the actual parameter values for paths a, b, and $a*b$ and too large for path c'. In contrast, the estimates produced by the latent variable approach were typically close to the actual parameter values. However, they varied widely, and several fell substantially further from the true parameter value than the most extreme estimates produced by the observed variable approach. Overall, the latent variable approach yielded estimates that were better centered on the actual parameter value, and that usually—but not always—came closer to the actual parameter values than the estimates produced by an observed variable approach. Thus, the inflated variability in the estimates was caused by most of the estimates varying as much above the actual parameter values as below and by a number of outliers that (as noted earlier) tend to dramatically overestimate the effect. As Figure 6 illustrates for the indirect effect, the observed variable estimates almost never recovered the true parameter value, whereas the latent variable estimates averaged very close to it, and yet this latter approach also produced a substantial minority of estimates that were further from the true parameter value than any of the observed variable estimates. This finding underscores the importance of replication when using latent variable techniques.

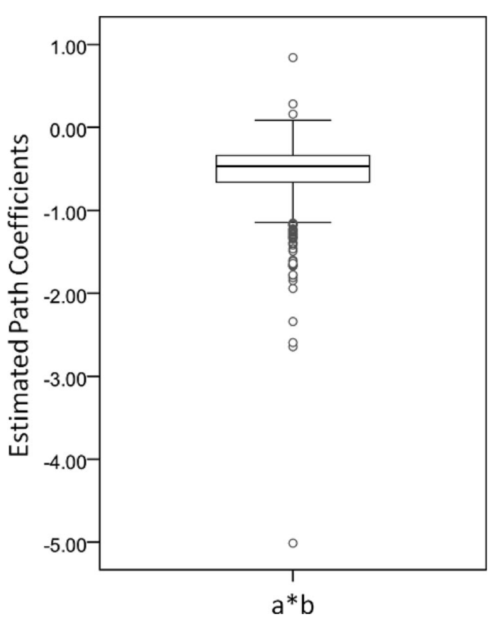

Figure 5. Box plot of the estimates produced by a latent variable approach for path $a*b$ in the moderate reliability, large effect size condition for Model 0.

Power

We also examined the bias–precision trade-off at a lower level by testing each coefficient in each data set for significance, to compare how frequently an observed versus latent variable approach produced significant results. This can be viewed as a power analysis for these approaches when the sample is set to 100 and the effect sizes are set to the designated values in Table 1.

For the most part, the two analytic strategies had comparable power for tests of path a (see Table 3), although the observed variable analysis had somewhat greater power to detect path a when reliability was moderate and the indirect effect size was small. More striking differences between the two approaches emerged for the tests of effect b and $a*b$: The observed variable

Table 3
Percent of Observed and Latent Variable Estimates That Reached Significance ($p < .05$) for Individual Studies Each With Sample Size 100

Mediation	R	Effect size	N[a]	a	b	c'	$a*b$
			Observed variable results				
Model 0	.7	Small	1,000	60.7	57.8	6.7	14.4
		Medium	1,000	97.7	92.6	12.6	84.7
		Large	1,000	100	98.1	30.1	97.5
	.9	Small	1,000	81.8	77.9	5.8	44.0
		Medium	1,000	100	99.2	7.3	98.6
		Large	1,000	100	99.9	13.6	99.9
Model 1	.7	Small	1,000	60.7	69.8	64.7	19.9
		Medium	1,000	97.7	98.2	83.7	93.0
		Large	1,000	100	99.9	97.4	99.7
	.9	Small	1,000	81.8	85.1	84.6	52.1
		Medium	1,000	100	99.9	94.0	99.8
		Large	1,000	100	100	99.6	100
			Latent variable results				
Model 0	.7	Small	1,000	52.7	44.9	3.8	5.9
		Medium	1,000	95.6	75.7	3.8	53.1
		Large	998	99.7	72.0	3.8	60.9
	.9	Small	1,000	81.2	76.1	5.5	41.0
		Medium	1,000	99.9	98.5	6.1	97.5
		Large	1,000	100	99.7	6.1	99.7
Model 1	.7	Small	1,000	52.7	49.0	45.8	8.8
		Medium	1,000	95.8	82.9	37.7	70.9
		Large	999	99.8	83.1	31.0	80.3
	.9	Small	1,000	81.1	80.9	80.7	48.2
		Medium	1,000	99.9	99.9	83.1	99.5
		Large	1,000	100	100	88.6	100

[a] Number of simulated studies used to calculate the percentages in the table.

approach had more power to correctly detect these effects across all levels of reliability and effect size. This difference was especially pronounced when reliability was moderate. For instance, in Model 0 with .7 reliability, the power for the observed variable test of the $a*b$ effect was about .85 for the medium effect size but dropped to only about .53 for the latent variable test of the same effect.

The power comparisons are most meaningful if the approaches have the same level of Type I error. However, Table 3 reveals a Type I error problem. For Model 0, one expects the percent of significant tests of effect c' to be 5%, the conventional Type I error rate. This is because the true value of c' in the simulation was set to 0 for this model. For the latent variable analyses, the percent of significant findings for Model 0 tests of path c' was below 5% when scale reliability was moderate and just slightly above (5.5%–6.1%) when reliability was high. However, when these same data were analyzed with an observed variable approach, the percent of significant findings for path c' ranged from 5.8% to as large as 30.1%, indicating an increased risk of Type I errors for the observed variable analysis of contaminated measures (for more on Type I errors in regression estimates when variables are measured with error, see Bollen, 1989; Cole & Maxwell, 2003, pp. 566–568; Kenny, 1979, pp. 79–83). The Type I error rate increased as reliability decreased and as indirect effect size increased. The fact that the observed variable approach has an inflated Type I error rate for Model 0 is problematic and calls into question its apparent greater power for detecting path c' in Model 1: When analyzing real data, it would be impossible to know whether an observed variable approach detected path c' because it was really there (as in Model 1) or simply because of an inflated Type I error rate (as in Model 0).

Discussion

The results of both the theoretical analysis and the simulation studies presented here suggest that the trade-off between accuracy and precision in latent variable modeling is both sizable and complex. In line with past research (e.g., Hoyle & Kenny, 1999; Stephenson & Holbert, 2003), we found that latent variable models provided coefficient estimates that were typically more accurate (and appreciably larger, for paths a, b, and $a*b$) than those produced by an observed variable approach that ignored measurement error. However, the adjustment for measurement error came at a cost. The precision of the latent variable models was noticeably reduced relative to observed variable models. Latent variable models have more parameter estimates than observed variable models and consequently are able to fit the data better. This means that these models fit sampling fluctuations more than observed variable models, with the result that the estimates are more influenced by sampling noise. As a consequence, in any one sample in our simulation study, latent variable estimates could substantially over- or underestimate the actual parameter value, and that risk was reflected in larger standard errors. In contrast, observed variable analyses yielded biased estimates that were sizably smaller than the actual parameter values for *a*, *b*, and *a***b*, but did so consistently and therefore had smaller standard errors.

Precision Versus Accuracy

The trade-off between precision and accuracy and the question of which to prioritize echoes an ongoing discussion in the field about alternative goals in conducting research (e.g., Abelson, 1985; Cohen, 1990, 1994; Cortina & Dunlap, 1997; Prentice & Miller, 1992; Schmidt, 1996). Whereas some studies focus primarily on establishing evidence that an effect exists (the purview of null hypothesis testing), others seek to shed light on the strength of the relations between variables. The latter goal calls for estimating parameter values and for representing precision with confidence intervals. Researchers concerned with the first goal are likely to be most concerned with maximizing precision, whereas researchers concerned with the second are likely to prioritize accuracy. Clearly, however, both are important, and we need to understand the costs of choosing an analytic strategy that achieves one at the expense of the other.

Because the observed variable approach yields estimates that are more precise, analyses that take this approach can apparently have more statistical power. Investigators using observed variables may therefore obtain results that seems worth reporting on the basis of the statistical significance of the mediated indirect effect. However, the seeming advantage of this approach for a single article is likely to be a disadvantage for the field. Our analysis shows that the observed variable approach is likely to underestimate the amount of variation that is explained by the mediating variable. Theory construction and testing can be hampered by these errors, especially if the biased results are replicated by other observed variable analyses.

From an accuracy standpoint, the latent variable approach tended to give better estimates on average across all levels of effect size and reliability. Both direct and indirect effects were more likely to be correctly estimated when latent rather than observed variables were employed. Thus, the best way to shed light on the true value of the relations between psychological constructs is to use latent variable methods, especially when employing measures that do not have high reliability. However, in such cases, it is particularly important to collect larger samples to offset the imprecision of the latent variable approach.

Moreover, a latent variable approach is less likely than an observed variable approach to incorrectly detect some null effects, such as path c' in Model 0 (when the true parameter value was 0; see also Cole & Maxwell, 2003, pp. 566–568). Whereas the latent variable approach produced Type I error rates for path c' that were quite close to the expected 5%, the observed variable approach yielded larger percentages of Type I errors, especially for the large effect size conditions. Thus, when the mediating variable is measured with error, observed variable approaches are too likely to incorrectly detect a direct effect of *X* on *Y* when none exists. Similarly, in supplementary analyses not reported here, an observed variable approach was sometimes too likely to detect an effect of *M* on *Y* (path *b*) when this path was set to 0 in the simulation.

Taken together, then, these findings suggest that ironically, an observed variable approach that ignores measurement error often yields incorrect estimates with little variability, whereas a latent variable approach tends to converge on the correct parameter values but with greater uncertainty. Importantly, this pattern was especially pronounced when the measures in our model had only moderate reliability. When reliability is high, the estimates produced by an observed variable approach are less attenuated (see also Hoyle & Kenny, 1999), and a latent variable approach yields smaller standard errors and is often as powerful as a model based on observed variable analyses. In other words, the better defined a factor is, the smaller the costs involved with either approach.

As reliability decreases, both approaches become more troublesome. Even with an alpha of .7, which is typically considered an acceptable level of reliability in our field, observed variable approaches greatly underestimate path coefficients and can produce highly inflated Type I error rates, and latent variable approaches lose considerable power and (especially with large effect sizes) can occasionally yield wildly inaccurate estimates. Thus, one key implication of the present study is that low or even "moderate" levels of reliability can be more problematic than social psychologists often assume in their day-to-day research.

Implications and Recommendations

In this article, we built on Hoyle and Kenny (1999), who considered the bias associated with unreliable mediating variables but perfectly reliable *X* variables. Perfect reliability for the initial *X* variable is most conceivable when it represents level of an experimental manipulation. We extended this previous work by examining the relative precision of observed versus latent variable approaches and by examining models with unreliability in both the predictor and mediating variables. In separate analyses (not reported here but available from the authors), we replicated the biases reported by Hoyle and Kenny with a simulation study that treated *X* as binary (reflecting a two-condition experiment). These analyses also verified that the accuracy–precision principles we have described apply similarly (though slightly less dramatically) to mediation studies arising from experiments. Regardless of whether a predictor is manipulated or measured, the increased accuracy provided by a latent variable approach comes at the cost of reduced precision.

This pattern has important implications for social psychologists seeking to shed light on questions of mechanism and process via mediation methods. If we want to advance science, we need both unbiased estimates that converge on the true relations between variables in the population and stable results. Given the present findings, it is now clear that not only are observed variable analyses potentially biased, but latent variable analyses are potentially unstable. Researchers must therefore strive to avoid both of these pitfalls. Below we outline three recommendations for maximizing both accuracy and precision in mediation analyses. The first two are well known to quantitative researchers but bear repeating in this context. The third suggests a new, two-step analytic strategy that capitalizes on the benefits provided by each approach studied here while sidestepping drawbacks associated with relying exclusively on observed or latent variable analyses.

1. Invest in reliable measures. Attempting to adjust for unreliability will always be more complicated than simply using a reliable measure in the first place. Table 3 illustrates the dramatic boost in power to detect an indi-

rect effect shown by both analytic approaches when reliability increases. Instead of settling for measures with mediocre reliability on the assumption that an alpha of .7 is considered adequate, researchers may wish to take advantage of this power boost and seek out more reliable measures.

2. Plan mediation studies that have adequate power. The Mplus syntax used for our simulation studies (see http://sites.google.com/site/ledgerwoodshrout2011) can provide an important and useful tool when planning a mediation study that will be analyzed with a latent variable approach. It can be modified to estimate power to detect paths in any three-variable model and can be adjusted to see the effects of increasing reliability or sample size.

3. When highly reliable measures are unattainable, use a two-step strategy for testing and estimating the indirect effect in a three-variable mediation model. In Step 1, use observed variables to test the indirect effect $a*b$ for significance.[9] In Step 2, estimate $a*b$ using a latent variable approach that adjusts for measurement error.

Limitations and Extensions

We limited our theoretical and empirical analyses to mediation models in which the variables were normally distributed, with the exception of the supplemental analyses briefly mentioned above that treated the X variable as binary. It is possible that the results would be different for nonnormal data, such as skewed data or ordinal data, and future research should investigate this possibility. Moreover, the application of structural equation models to nonnormal data becomes increasingly problematic as the sample size gets small (Wang, Fan, & Willson, 1996). The Mplus syntax that we provide online can be modified to examine small sample studies when certain nonnormal data patterns are expected.

To describe the precision of the mediation effects, we focused on expected and observed standard errors of the estimates. Importantly, our simulation studies verified that the estimated standard errors gave values that approximated the observed variation of the a, b, c', and $a*b$ estimates. Although the standard errors provide a convenient index of precision, we are not recommending that normal theory tests based on the standard errors be used to test the indirect effect, $a*b$, in an individual study. The distribution of the indirect effect is not normally distributed, and normal theory tests of indirect effects are typically conservative (e.g., Shrout & Bolger, 2002). The absolute number of significant results reported in Table 3 for the indirect effect would be larger (for both observed and latent variable estimates) if the bootstrap were used for tests, but the overall result that latent variable models are less powerful than observed variable models is not expected to change (see Footnote 7).

Kenny (personal communication, July 1, 2008) pointed out that the precision of latent variable estimates can be increased if it is possible to constrain the paths from the latent variables to the observed indicators (i.e., loadings) to be equal. This constraint was appropriate in our simulation studies, which assumed constant confirmatory factor loadings and identically scaled variables (but note that in practice, researchers will want to choose constraints appropriate for their theoretical model and data set).[10] In analyses not reported here but available from the authors, we confirmed that the constraint did indeed improve the precision of the estimates from the latent variable approach, but the increase was small relative to the overall accuracy–precision trade-off described here.[11] When indicators are created from construct-based parcels of balanced items, it is advisable to take advantage of the improved precision of constrained factor loadings.

Although item parcels can be useful in defining latent variables, Sterba and MacCallum (2010) recently showed that the composition of the items into the parcels adds an extra source of variation into the analyses. The amount of variation due to parcel allocation decreases when sample size gets larger, when the number of items allocated to parcels increases, and when the internal consistency of the item set increases. In cases where parcel allocation variation is of concern, Sterba and MacCallum offer a programming tool to reduce its effects by averaging results from randomly allocated parcel groups.

In some cases, the sample size in a mediation study is not sufficient to consider latent variable models that can adjust for measurement error. At the very least, we recommend that investigators acknowledge the possibility of bias in these cases. Coffman and MacCallum (2005) described a way to adjust for unreliability with a structural equation program when the degree of reliability is considered to be a known constant rather than a quantity to be estimated from latent variable models. This method should be considered regardless of whether the measured variable is a simple sum of items or a weighted average of items such as a factor score. Even though factor scores have somewhat better reliability than simple averages (McDonald, 1999; p. 89), they are still contaminated by measurement error in the items. McDonald's (1999) omega statistic provides the appropriate reliability estimate if factor scores are used. The Coffman and MacCallum adjusted analysis can be fit with a somewhat smaller sample than a latent variable model; the final result depends on the assumed reliability value rather than a comprehensive model of the data.

Finally, we note that we deliberately avoided using the language of "partial" and "complete" mediation in this article. Although this distinction is clear in population models such as Model 0, for which the value of c' was exactly 0, the claim of complete mediation is problematic in practice. When observed variables are not perfectly reliable, our results show that investigators who analyze observed variables are likely to reject the null hypothesis, H_0: $c' = 0$, even if Model 0 is correct. On the other hand, when latent variable models are used, there is rarely sufficient power to definitively establish that the direct effect is 0. We therefore join others in recommending that investigators pursue mediation mod-

[9] In supplementary analyses (not reported here but available from the authors), we confirmed that the observed variable approach generally produces acceptable Type I error rates when there is in fact no real indirect effect (i.e., when path a and/or path b is set to 0 in the simulation).

[10] For instance, when variables are not scaled identically, communalities can be constrained to be equal across the indicators of a particular latent variable.

[11] The constraint was more successful at increasing the precision of the latent variable approach when X was binary and uncontaminated by measurement error.

els that help explain associations but refrain from attempting to conclude that no other mediation path need be considered (see, e.g., Rucker, Preacher, Tormala, & Petty, 2011).

Conclusion

Whereas an observed variable approach to mediation analysis produces biased estimates with little variability, a latent variable approach produces estimates that converge on the true value of the indirect effect, but at the cost of increased standard errors and reduced power to detect the effect in the first place. Researchers could capitalize on the advantages of each approach in a three-variable mediation model by using an observed variable analysis (e.g., regression) to test an indirect effect for significance and a latent variable analysis (e.g., SEM) to estimate the path coefficient more accurately.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133. doi:10.1037/0033-2909.97.1.129

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182. doi:10.1037/0022-3514.51.6.1173

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

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98,* 550–558. doi:10.1037/a0018933

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research, 40,* 235–259. doi:10.1207/s15327906mbr4002_4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312. doi:10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112,* 558–577. doi:10.1037/0021-843X.112.4.558

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2,* 161–172. doi:10.1037/1082-989X.2.2.161

DeShon, R. P. (2006, May). *Monkeys, ostriches and measurement error corrections.* Invited address at the annual convention of the Association for Psychological Science, New York, NY.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82,* 62–68. doi:10.1037/0022-3514.82.1.62

Duncan, O. D. (1975). *Introduction to structural equation models.* New York, NY: Academic Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall/CRC.

Heise, D. R. (1975). *Causal analysis.* New York, NY: Wiley.

Henderson, M. D., & Trope, Y. (2009). The effects of abstraction on integrative agreements: When seeing the forest helps avoid getting tangled in the trees. *Social Cognition, 27,* 302–417. doi:10.1521/soco.2009.27.3.402

Hodson, G., & Costello, K. (2007). Interpersonal disgust, ideological orientations, and dehumanization as predictors of intergroup attitudes. *Psychological Science, 18,* 691–698. doi:10.1111/j.1467-9280.2007.01962.x

Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks, CA: Sage.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5,* 602–619. doi:10.1177/0193841X8100500502

Kenny, D. A. (1979). *Correlation and causality.* New York, NY: Wiley.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). New York, NY: McGraw-Hill.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Knowles, E. D., Lowery, B. S., & Schaumberg, R. L. (2010). Racial prejudice predicts opposition to Obama and his health care reform plan. *Journal of Experimental Social Psychology, 46,* 420–423. doi:10.1016/j.jesp.2009.10.011

Kross, E., Ayduk, O., & Mischel, W. (2005). When asking "why" does not hurt: Distinguishing rumination from reflective processing of negative emotions. *Psychological Science, 16,* 709–715. doi:10.1111/j.1467-9280.2005.01600.x

Lalwani, A. K., Shrum, L. J., & Chiu, C. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology, 96,* 870–882. doi:10.1037/a0014622

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173. doi:10.1207/S15328007SEM0902_1

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Mahwah, NJ: Erlbaum.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7,* 83–104. doi:10.1037/1082-989X.7.1.83

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39,* 99–128. doi:10.1207/s15327906mbr3901_4

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York, NY: McGraw-Hill.

Muthén, L. K., & Muthén, B. O. (2010). Mplus (Version 6). Los Angeles, CA: Muthén & Muthén.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287–312. doi:10.1207/S15328007SEM0802_7

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69,* 408–419. doi:10.1037/0022-3514.69.3.408

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164. doi:10.1037/0033-2909.112.1.160

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass, 5,* 359–371. 10.1111/j.1751-9004.2011.00355.x

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129. doi:10.1037/1082-989X.1.2.115

Shook, N. J., & Fazio, R. H. (2009). Political ideology, exploration of novel stimuli, and attitude formation. *Journal of Experimental Social Psychology, 45,* 995–998. doi:10.1016/j.jesp.2009.04.003

Shrout, P. E. (2011). Causal thinking in psychopathology research. In P. E. Shrout, K. M. Keyes, & K. Ornstein (Eds.), *Causality and psychopathology: Finding the determinants of disorders and their cures* (pp. 3–24). New York, NY: Oxford University Press.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7,* 422–445. doi:10.1037/1082-989X.7.4.422

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (Vol. 13, pp. 290–312). San Francisco, CA: Jossey-Bass.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89,* 845–851. doi:10.1037/0022-3514.89.6.845

Stephenson, M. T., & Holbert, R. L. (2003). A Monte Carlo simulation of observable versus latent variable structural equation modeling techniques. *Communication Research, 30,* 332–354. doi:10.1177/0093650203030003004

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research, 45,* 322–358. doi:10.1080/00273171003680302

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology, 83,* 1298–1313. doi:10.1037/0022-3514.83.6.1298

Wang, L., Fan, X., & Willson, V. L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling, 3,* 228–247. doi:10.1080/10705519609540042

Wegener, D. T., & Fabrigar, L. R. (2000). Analysis and design for nonexperimental data: Addressing causal and noncausal hypotheses. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 412–450). New York, NY: Cambridge University Press.

Appendix

Calculating Expected Bias and Precision

Expected Bias of Ordinary Least Squares (OLS) When Unreliability Is Ignored

The path model from Figure 1 implies two regression equations. To simplify the analysis, we assume that the variables all have mean 0. The equations are

$$M = aX + e_M$$

$$Y = bM + c'X + e_Y.$$

Cohen et al. (2003) described the OLS estimates for unstandardized variables as follows:

$$\hat{\alpha} = r_{MX}*(S_M/S_X) \tag{A1}$$

$$\hat{b} = \left[\frac{r_{MY} - r_{XY}r_{XM}}{1 - r_{XM}^2}\right]\left(\frac{S_Y}{S_M}\right) \tag{A2}$$

$$\hat{c}' = \left[\frac{r_{XY} - r_{MY}r_{XM}}{1 - r_{XM}^2}\right]\left(\frac{S_Y}{S_X}\right). \tag{A3}$$

The quantities r_{MX}, r_{MY}, and r_{XY} represent the correlations among the X, M, and Y variables, and the quantities S_X, S_M, and S_Y represent the standard deviations of the variables. When the ratios of the standard deviations are removed, the equations give standardized regression estimates. If X and M are measured without error, then these estimates provide unbiased estimates of population values. However, if X has reliability R_X, M has reliability R_M, and Y has reliability R_Y, then the observed correlations become $r^*_{XM} = r_{XM}\sqrt{R_X R_M}$, $r^*_{XY} = r_{XY}\sqrt{R_X R_Y}$, and $r^*_{MY} = r_{MY}\sqrt{R_M R_Y}$. When these are used with Equations A1–A3, one obtains estimates of the biased values. Figure 2 was created by considering a range of values of R_M and R_X for mediation models with small, medium, and large effects. To make these calculation more convenient to carry out, we created an Excel sheet that can be downloaded (http://sites.google.com/site/ledgerwoodshrout2011).

Expected Precision of OLS Estimates and Structural Equation Modeling (SEM) Latent Variable Estimates

For the OLS estimates described by Equations A1–A3, the asymptotic standard errors[12] can be expressed explicitly. For the estimate of a, the standard error is

$$Se(\hat{a}) = \sqrt{\left(\frac{1 - \rho_{MX}^2}{n}\right)\frac{\sigma_M^2}{\sigma_X^2}}.$$

For Equations A2 and A3, the asymptotic standard errors are obtained by taking the square root of the diagonal elements of the matrix

$$V\begin{bmatrix} \hat{b} \\ \hat{c}' \end{bmatrix} = \left(\frac{1 - R_Y^2}{n}\right)S_I^{-1}\sigma_Y^2,$$

where $R_Y{}^2$ is the squared multiple correlation for Y and $S_I{}^{-1}$ is the inverse of the covariance matrix of the independent variables (M, X). As the sample size, n, gets larger, the standard errors get smaller by a factor of $1/\sqrt{n}$. The standard errors are also affected by the degree of correlation among the variables. Most notably, the

[12] Because we will be comparing the OLS standard error to the standard error of the SEM maximum likelihood estimates, we use asymptotic forms. We assume the variables have population mean 0 and variance 1, but the standard errors are for unstandardized effects.

(*Appendix continues*)

larger the effect a, the larger the correlation of X and M, and the larger the diagonal of ${S_I}^{-1}$, which leads to greater standard errors for $\hat{b}$ and $\hat{c}'$ (see Cohen et al., 2003, p. 86).

Instead of being based on the correlations of three variables—X, M, and Y—the SEM estimates of the model shown in Figure 3 depend on correlations among nine variables: $X1$, $X2$, $X3$, $M1$, $M2$, $M3$, $Y1$, $Y2$, $Y3$. Suppose the sample variance–covariance matrix is called S. It is a nine-by-nine matrix with the variances of the indicators on the diagonal and the covariances among pairs of indicators in the off-diagonal. In SEM, multivariate equations are written to create a fitted version of the variance–covariance matrix, which we call Σ. The fitted covariance matrix depends on the a, b, and c' parameters from Figure 3 as well as the paths between each latent variable and its indicators, and the variances of the residuals. There are 21 parameters in this model that will be estimated, and statisticians represent them as a list in a vector called θ. Although the nine-by-nine matrix has 45 distinct elements, the fit depends on the 21 elements in θ, leaving $45 - 21 = 24$ degrees of freedom to assess the quality of fit. Unlike OLS, the maximum likelihood (ML) estimation method does not lead to explicit equations like Equations A1–A3. Instead, SEM programs use numerical methods to find values of the θ estimates that make Σ resemble the sample variance–covariance matrix S. According to Bollen (1989, pp. 134–135), the degree of resemblance is quantified by the ML-fitting equation

$$F_{ML} = \ln|\hat{\Sigma}| + \mathrm{trace}[S\hat{\Sigma}^{-1}] - \ln|S| - 9. \tag{A4}$$

Values of a, b, c', and the 18 other parameters in θ are chosen by the numerical methods to make Equation A4 as small as possible. These methods essentially evaluate F_{ML} for different trial values of a, b, c', and the other parameters, and when there is no improvement for small additional changes, the result is considered to be the ML estimate. This method of estimation is the default in most SEM programs.

Like the OLS estimates, the precision of the ML estimates depends on the sample size and the pattern of associations among the variables in the model. When the ML estimate is relatively more precise, the rate of change in F_{ML} as a, b, c', and the other parameters approach the optimal values is high: Small changes in the trial estimates are associated with relatively large changes in F_{ML}. When the estimate is less precise, the rate of change of F_{ML} is not strongly related to the final ML values. Statistical theory (Bollen, 1989, p. 135) states that standard errors can be computed from the second derivative of F_{ML} with respect to the list of parameters being estimated. The second derivatives provide a measure of how quickly F_{ML} would change as the estimates of the parameters move slightly from the values that minimize Equation A4. In practice, the numerical algorithm that finds the ML estimates can also provide an estimate of the second derivatives of F_{ML} near the ML solution. The asymptotic variance–covariance matrix of the θ estimates is given by

$$\mathrm{Var}(\hat{\theta}) = \left(\frac{2}{n}\right)\left(\mathrm{E}\left[\frac{\partial^2 F_{ML}}{\partial\theta\partial\theta'}\right]\right)^{-1}. \tag{A5}$$

The second derivatives in this equation are conceived to be constant once a model such as that in Figure 3 is specified. However, the numerical methods for estimating it are influenced by sampling variation. One way to obtain values for the matrix that are not much affected by sampling variation is to use simulation provisions in SEM software such as Mplus (Muthén & Muthén, 2010) to create a very large simulated data set that follows the model in Figure 3. The standard errors computed by extremely large data sets will not be much affected by sampling error, and then Equation A5 can be used to rescale the obtained standard error to smaller sample sizes of interest. For example, if $N = 10{,}000$ for a large simulated data set, but one wishes to know the asymptotic standard error of $n = 100$, then one multiplies the result from Equation A5 by $N/n = 10{,}000/100 = 100$. This method was used to compute the asymptotic standard errors shown in Figure 4.

Received December 20, 2010
Revision received June 3, 2011
Accepted June 8, 2011 ■