The Relative Validity of Inferences About Mediation as a Function of Research Design Characteristics
Eugene F. Stone-Romero and Patrick J. Rosopa
Organizational Research Methods 2008; 11; 326 originally published online Jul 23, 2007;
DOI: 10.1177/1094428107300342

The online version of this article can be found at:
http://orm.sagepub.com/cgi/content/abstract/11/2/326

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
The Research Methods Division of The Academy of Management

Additional services and information for Organizational Research Methods can be found at:

Email Alerts: http://orm.sagepub.com/cgi/alerts
Subscriptions: http://orm.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations http://orm.sagepub.com/cgi/content/refs/11/2/326
The Relative Validity of Inferences About Mediation as a Function of Research Design Characteristics

Eugene F. Stone-Romero  
*University of Texas at San Antonio*  
Patrick J. Rosopa  
*Clemson University*

Tests of assumed mediation models are common in the organizational sciences. However, the validity of inferences about mediation is a function of experimental design and the setting of a study. Regrettably, most tests of mediation have relied on the application of so-called “causal modeling” techniques to data from nonexperimental studies. As we demonstrate, inferences about the validity of assumed mediation models are highly suspect when they are based on the findings of nonexperimental research. One of the many reasons for this is the failure of the model being tested to be consistent with reality. Valid research-based inferences about mediation are possible. However, inferences from such tests are most likely to be valid when they are based on research that uses randomized experimental designs. Strategies for conducting research using these and other designs are described. Finally, we offer a set of conclusions and recommendations that stem from our analysis.

**Keywords:** causal inferences; structural equation modeling; tests of mediation; mediating effects; mediator variables

Tests of models that involve mediator variables are very common in research in a large number of disciplines, including management, organizational behavior, human resource management, and organization theory. A *mediator* is a variable that fully or partially transmits the effects of one or more independent variables to one or more dependent variables (James & Brett, 1984; Stone-Romero & Rosopa, 2004). Note that mediator variables are also referred to as intervening variables (Blalock, 1971; Stone, 1978).

In the simplest of models involving mediation (referred to hereinafter as mediation models), there is a single independent variable (*X*), a single mediator (*M*), and a single dependent variable (*Y*). Figure 1 shows such a model. In it, *X* has (a) a direct effect on *Y*, and (b) an indirect effect on *Y* through *M*. It is critical to note that mediation models are *causal models*. For example, in the case of a complete mediation model (described below) the causal sequence is *X* → *M* → *Y*.

**Purpose**

As is detailed below, the appropriateness of inferences about mediation varies primarily as a function of two important study design elements, i.e., experimental design and the
setting in which a study is conducted. In terms of experimental design, such inferences are most appropriate when the data are from studies that use randomized experimental designs, less justified when the data are from studies that use quasi-experimental designs, and least appropriate when the data are from studies that use nonexperimental designs. In addition, because of the way in which study settings influence control over both (a) the successful manipulation of independent variables and (b) confounding variables, inferences about mediation are typically most appropriate when studies are conducted in special purpose (SP) as opposed to non-special purpose (NSP) settings. The distinction between these is made clear below.

Regrettably, the authors of many articles published in the journals of organizational, social, and behavioral sciences appear to not recognize how study design influences the validity of inferences about mediation. As a result, unwarranted claims about cause, including causal models involving mediation abound (Stone-Romero & Gallaher, 2006). We document this problem below as it relates to tests of mediation models. In view of the very serious nature of the same problem, the major purpose of the present article is to detail why the legitimacy of inferences about cause differs as a function of both experimental design and study setting. In the interest of doing so, the sections that follow consider a number of important issues. First, we describe how such study design-related factors as experimental design and the setting in which research is conducted affect the validity of inferences about mediation. Second, we argue that unwarranted inferences about cause are all too common in mediation research that uses nonexperimental designs, and provide numerous illustrations of this problem from articles published in top-quality journals. Third, we comment on a number of issues that are important in inferences about mediation that are based on nonexperimental research (e.g., model specification). Fourth, we offer examples of several designs that can be used in testing for assumed mediation and specify their relative value in generating valid inferences about mediation. Fifth, and finally, we advance several conclusions that stem from the consideration of these issues.

Prior to dealing with the same issues, we consider notation conventions used in this article. In instances of a properly specified mediation model (i.e., one that corresponds to reality; Bollen, 1989), we use $X$, $M$, and $Y$, respectively, to denote the independent, mediator,
and dependent variables. However, when mediation models are assumed, but their correspondence with reality is unknown, we use the symbols $X_A$, $M_A$, $Y_A$, and $C_A$, respectively, to denote assumed independent, mediator, dependent, and confounding variables.

### Research Setting Issues

Prior to describing the three above-noted types of experimental designs, it is important to distinguish between experimental designs and research settings (Stone-Romero, 2002, 2006c). Thus, in this section we focus on two setting-related issues, i.e., the laboratory versus field distinction and the influence of setting type on internal validity. As is detailed below, taken together these two issues are very important in tests of assumed mediation models.

#### The Often Inappropriate Laboratory Versus Field Distinction

Typically, a distinction is made between two types of settings, i.e., laboratory and field (e.g., Kerlinger & Lee, 2000). However, the same distinction is not very informative (Campbell, 1986; Stone-Romero, 2002, 2006c). As noted by Stone-Romero (2006c), a more appropriate distinction is between SP and NSP settings. The former are created for the specific purpose of doing research (e.g., a simulated work setting in a laboratory complex in an industrial park or a laboratory room at a university), whereas the latter are created for purposes other than research (e.g., organizations such as General Motors, IBM, and Xerox). Settings of the SP variety have two major attributes. First, they are created for the purpose of doing research and cease to exist when the research has been completed. Second, they are designed so as to provide for the effective (unconfounded) manipulation of one or more independent variables. Typically, SP settings have only a subset of the features or elements that are present in NSP settings (Aronson, Ellsworth, Carlsmith, & Gonzales, 1990; Berkowitz & Donnerstein, 1982; Fromkin & Streufert, 1976; Runkel & McGrath, 1972).

#### Research Settings and Inferences About Mediation

Inferences about mediation are causal inferences. More specifically, in simple mediation models, two causal inferences are involved. The first is that $X \rightarrow M$, and the second is that $M \rightarrow Y$. Thus, factors that affect causal inference, in general, are highly relevant to inferences about cause in mediation models.

Because SP settings are created specifically for the purpose of conducting a study, research in such settings typically provides for a greater degree of control over confounding (nuisance) variables than research in NSP settings. To the degree that the influence of confounding variables can be controlled, a researcher can be more confident about the internal validity of a study. As such, ceteris paribus, inferences about mediation are more justified when research is conducted in SP, than NSP settings.
Experimental Design and Inferences About Cause in Mediation Models

Tests of mediation can be based on data from research that uses a variety of experimental designs (i.e., nonexperimental, quasi-experimental, randomized experimental). This is a crucial issue, because the validity of inferences about mediation is a function of the type of design used in a study (Cook & Campbell, 1979; Shadish, Cook, & Campbell 2002; Stone-Romero, 2002). As noted above, the reason for this is that mediation models make explicit assumptions about causal connections between variables (Mathieu & Taylor, 2006; Stone-Romero & Rosopa, 2004). As such, inferences about mediation hinge on the internal validity of a study (Shadish et al., 2002; Stone-Romero & Rosopa, 2004). In view of this, we next consider the validity of inferences about causal relations between variables (e.g., as in mediation models) that result from three general types of experimental designs, i.e., nonexperimental, quasi-experimental, and randomized experimental (Shadish et al., 2002; Stone, 1978; Stone-Romero, 2002, 2006a, 2006b, 2006c).

Conditions Vital to Inferences About Cause in Mediation Models

Prior to describing the three types of experimental designs, it is important to consider the conditions that are vital to valid inferences about cause, including inferences about mediation (Cook & Campbell, 1979; Shadish et al., 2002; Stone, 1978; Stone-Romero, 2002, 2006a, 2006b, 2006c). The first is that the cause ($X$) precedes the mediator ($M$) in time, and $M$ precedes the effect ($Y$) in time (i.e., temporal precedence). The second is that the cause and effect are related to one another (i.e., covariation). The third is that there are no rival explanations of the observed relation between the cause and effect (i.e., absence of confounds). As should be obvious from the material that follows, ceteris paribus, the findings of research using randomized experimental designs afford the firmest basis for inferences about cause in mediation models, quasi-experiments provide a less firm basis for such inferences, and nonexperiments furnish the weakest basis for such inferences. The reason for this is that research that uses randomized experimental designs meets the just-noted three conditions to a far greater extent than research that uses alternative designs.

Randomized Experimental Designs

The nature of randomized experimental designs. There are four major attributes of a study that uses a randomized experimental design (Shadish et al., 2002; Stone, 1978; Stone-Romero, 2002, 2006c). First, studied units (e.g., individuals, groups, organizations) are randomly assigned to two or more treatment conditions. Second, units in any given treatment condition are exposed to a specific level of a treatment ($X_i; i = 1, 2, \ldots I$), with treatment level differing across the $I$ conditions. Third, care is taken to insure that the $I$ conditions are as alike as possible on all variables other than the treatment, helping to rule out confounding variables ($C_j; j = 1, 2, \ldots J$) as rival causes of observed effects. Fourth, and finally, the
researcher uses measures of one or more dependent variables \((Y_l; l = 1, 2, \ldots L)\) as a basis for inferring the effects of the differing levels of \(X\) on the units.

*Tests of assumed mediation models with randomized experimental designs.* Relative to studies using either quasi-experimental or nonexperimental designs, those that use randomized experimental designs have very high levels of internal validity (Shadish et al., 2002; Stone, 1978; Stone-Romero, 2002). Thus, the results of randomized experiments provide a very firm foundation for inferences about causal connections between variables in mediation models. As noted above, inferences about mediation require evidence about two causal paths. In the simplest mediation model, the paths are that \(X \rightarrow M \) and \(M \rightarrow Y\). Evidence on these causal paths can be generated effectively and easily with research that uses randomized experimental designs. We explain the strategy for doing so below in the section titled “Some Study Designs and Their Value in Demonstrating Mediation.”

When tests of mediation models are based on data from well-designed and properly conducted randomized experimental research, (a) causal paths are well-known, (b) confounds are not an issue, and (c) model misspecification is typically not a concern. Thus, when mediation models are tested by randomized experimental means, inferences about mediation rest on a very firm foundation.

The latter argument is predicated on the assumption that the randomized experiments are properly designed and executed. However, there may be instances when randomized experiments “break down.” This is more likely to be a problem in NSP than SP settings (Cook & Campbell, 1979). Among the many causes of this are such threats to internal validity as differential attrition across treatment conditions, history, maturation, regression, diffusion of treatments, compensatory equalization, resentful demoralization, testing, instrumentation, and the interaction of these threats. To the extent that these threats are present in a randomized experiment, the validity of inferences about causal connections between variables in mediation models will suffer. Thus, to the degree that such threats exist, they will have a negative impact on inferences about mediation.

**Quasi-Experimental Designs**

*The nature of quasi-experimental designs.* Studies that use quasi-experimental designs have four major attributes (Cook & Campbell, 1979; Shadish et al., 2002; Stone, 1978; Stone-Romero, 2002, 2006b). First, the studied units (e.g., individuals, groups, organizations) are not assigned to treatment conditions on a random basis. Second, the units in given conditions are exposed to specific treatment levels \((X_i; i = 1, 2, \ldots I)\), with treatment level varying across the \(I\) conditions. Third, although the researcher may strive to insure that the \(I\) conditions are as alike one another as possible on all variables other than the treatment, the absence of random assignment makes it difficult, if not impossible, to establish pretreatment equality of units in the conditions on possible confounds. Thus, statistical methods (e.g., partial correlation, multiple regression) are often used in the hopes of ruling out confounding variables \((AC_j; j = 1, 2, \ldots J)\) as rival explanations of observed differences in measures of assumed mediator variables \((AM_k; k = 1, 2, \ldots K)\) and assumed dependent variables \((AY_l; l = 1, 2, \ldots L)\) (Cook & Campbell, 1979; Shadish et al., 2002; Stone-Romero, 2002, 2006b). Fourth, and finally, the assumed effects of the differing levels of \(X\)
on the assumed mediator variables and dependent variables are assessed using measures of $AM$ and $AY$.

Tests of assumed mediation models with quasi-experimental designs. Relative to studies that use randomized experimental designs, those that use quasi-experimental designs tend to have considerably lower levels of internal validity (Shadish et al., 2002; Stone-Romero, 2002, 2006b). Thus, the results of research using quasi-experimental designs provide a much weaker foundation for inferences about causal sequences associated with mediation than is afforded by the findings of studies that use randomized experimental designs (Shadish et al., 2002; Stone-Romero, 2002, 2006b). There are several reasons for this. One is that it is typically impossible to assume the pre-treatment equality of units in various treatment conditions on confounding variables. Second, a researcher often has little or no knowledge of the confounds that operate in quasi-experimental studies. Third, even if a researcher measures one or more confounds and tries to control for them through statistical means (e.g., partial correlation, multiple regression), it is typically the case that statistical controls are far less effective in ruling out the operation of confounds than are experimental controls (Cook & Campbell, 1979; Shadish et al., 2002; Stone-Romero, 2002, 2006b). Overall, therefore, tests of mediation models based on research that uses quasi-experimental designs have considerably lower levels of internal validity than tests that use randomized experimental designs.

Nonexperimental Designs

The nature of nonexperimental designs. There are several defining characteristics of a study that uses a nonexperimental design (Cook & Campbell, 1979; Shadish et al., 2002; Stone, 1978; Stone-Romero, 2002, 2006a). First, the researcher measures (as opposed to manipulates) the assumed (a) independent ($AX_i; i = 1, 2, \ldots I$), (b) mediator ($AM_k; k = 1, 2, \ldots K$), and (c) dependent variables ($AY_l; l = 1, 2, \ldots L$). Second, he or she assumes that the measured levels of the $AX$ reflect the treatments to which the units were exposed at some point in time. However, there is no definitive way of ascertaining the correctness of this assumption. Third, assumed confounding variables ($AC_j; j = 1, 2, \ldots J$) are measured and statistical techniques (e.g., partial correlation, multiple regression) are used in the hopes of “controlling for” their influences. Fourth, the researcher correlates scores on measures of $AX_i$ with scores on measures of $AM_k$ and $AY_l$, and uses the resulting evidence of correlation as a basis for inferring the existence of hypothesized “effects,” including mediation “effects.” Fifth, and finally, to the degree that potential confounds are known and capable of being measured, their “effects” are controlled through such statistical means as multiple regression (Cohen, Cohen, West, & Aiken, 2003; Kenny, 1979; Stone-Romero, 2002, 2006a). As is noted below, however, all important confounds are not always known, and even if they are, it is unlikely that they can all be measured. Thus, statistical controls are typically a very poor substitute for experimental controls.

Tests of assumed mediation models with nonexperimental designs. Relative to research that uses randomized experimental designs, studies that use nonexperimental designs generally have very low levels of internal validity (Cook & Campbell, 1979; Shadish et al., 2002;
Stone-Romero, 2002, 2006a). As such, nonexperimental research provides a very weak foundation for inferences about causal connections between variables, including those in assumed mediation models.

There are several very serious problems in tests of mediation models that are based on data from nonexperimental research. One is that causal paths are assumed, but are virtually never known. Thus, the models tested are often misspecified (Stone-Romero & Rosopa, 2004). Another problem is that the potential for statistical controls to rule out rival causes is conditional on several factors: (a) the underlying causal model must be properly specified, (b) the researcher must be aware of all relevant confounds, and (c) the potential confounds must be measured reliably. If they aren’t, then parameter estimates (e.g., standard partial regression coefficients) for the other variables in a model will be biased (typically upwards). Typically, the nonexperimental researcher (a) does not have a properly specified model, (b) is unaware of all relevant confounds, thus only measures a few, and (c) uses measures of confounds that have reliabilities that are below 1.0. Therefore, statistical controls are almost always inferior to experimental controls. This results in further problems with model specification. As a consequence, inferences about mediation that stem from nonexperimental research typically rest on a very shaky foundation.

It deserves adding that the restriction of a well-specified model can be relaxed somewhat. That is, a researcher need not specify all of the causes of a mediator and/or dependent variable. Rather, he or she needs to include all of the important causes in a mediation model that is being tested. To the extent that he or she can do so, the bias in estimates of effects of various causes will be low. Of course, this assumes that the researcher has correctly specified the causal ordering of variables and that all important confounds have been controlled through either design features or statistical means.

Researchers who test assumed mediation models using data from nonexperimental or quasi-experimental studies often advance inappropriate causal inferences (as defined below in the section titled “Inferences About Cause and Mediation are Quite Common in Articles Reporting the Results of Nonexperimental Research”) on the basis of data analyses that show that the findings of such studies are consistent with the assumed models. Thus, the next section considers the inferences that are appropriate on the basis of consistency between a model and a study’s findings.

Consistency Between a Model and the Results of a Study

In many nonexperimental studies, researchers make inferences about mediation (and, thus, about cause) on the basis of consistency between a model and covariances among measured variables. The typical premise is that if data are consistent with a model, then inferences about the validity of the assumed causal model are warranted. This can be expressed in terms of the following symbolic sentence (Kalish & Montague, 1964) of the “if → then” variety: MC → FC, where MC = assumed mediation model is correct (the premise) and FC = study’s findings are consistent with assumed mediation model.

Operating on the basis of the just-noted premise, a researcher might hypothesize a mediation model, conduct a nonexperimental study to test it, analyze data from the study, and find that the covariances among the measured variables are consistent with the assumed
model; that is, the researcher establishes that FC is true. Next, he, or she, then combines the premise with the finding in the following manner.

Premise: MC → FC
Show: FC
Conclude, therefore: MC

Regrettably, the just-noted conclusion is not based on sound reasoning. More specifically, it is an instance of the fallacious reasoning of affirming the consequent.

To illustrate why the just-noted reasoning is inappropriate, consider the following example. Through intuition, a researcher posits that $A$ causes $B$ (where $A$ is an assumed cause and $B$ is a mediator). This serves as his, or her, premise (MC). The researcher then conducts a nonexperimental study and finds a very strong positive relation between measures of $A$ and $B$ (e.g., $r = .80$). This is used to support his, or her, inference of FC. Now, the researcher inappropriately concludes that the showing of FC implies that the original premise (MC) is correct. The invalidity of this conclusion is apparent from considering two other scenarios that could have produced the pattern of results (FC) that led the researcher to the unwarranted conclusion that the hypothesized model was correct. First, the pattern of results could have resulted from $B$ causing $A$. Second, it could have resulted from $A$ and $B$ not being causally related to one another, but sharing a common cause of $C$.

In short, the fact that a study’s results are consistent with an assumed mediation model does not allow for any valid conclusion about the correctness of the model, unless the results stem from a study that allows for ruling out rival models (Bollen, 1989; Brannick, 1995; Cliff, 1987; Freedman, 1987; Holland, 1986; Kelloway, 1998; Ling, 1982; Millsap, 2002; Rogosa, 1987; Stone-Romero & Rosopa, 2004, 2005). However, two well-designed and executed randomized experiments afford a firm basis for ruling out competing models. For example, if the researcher manipulates $A$ and observes expected changes in $B$, then he or she can be confident about the correctness of the view that $A$ causes $B$ because (a) $B$ could not have caused $A$, and (b) rival causes of $B$ can be ruled out by the study’s design.

It deserves adding that whereas findings of a nonexperimental study that are consistent with an assumed model do not provide a sound basis for concluding that the model is correct, negative findings can be used to rule out its correctness. This is accomplished via the symbolic logic-based inference rule of modus tollens (Kalish & Montague, 1964). The relevant symbolic sentences are as follows:

Premise: MC → FC
Show: ~FC
Conclude, therefore: ~MC

The reasoning is as follows. The researcher assumes that if his, or her, mediation model is correct (MC) then the findings of a study will be consistent with the model (FC). The study shows that the findings are not consistent with the same model (~FC). Therefore, the researcher can legitimately conclude that the assumed mediation model is not correct (~MC). It is assumed here that the study in question had high levels of both construct
validity and statistical conclusion validity (Shadish et al., 2002). If not, the finding of ~FC could be attributable to such rival explanations as invalid measures, unreliable measures, and inadequate statistical power (Shadish et al., 2002).

**Prediction Versus Causation**

In the case of virtually all nonexperimental research, there is a need to recognize a very important difference between prediction, in a statistical sense, and understanding causal connections between variables. Theories provide a useful basis for predicting phenomena. For example, Expectancy Theory (Porter & Lawler, 1968; Vroom, 1964) affords a basis for predicting how motivation to engage in an act is influenced by both (a) the valences of outcomes, and (b) expectancies of effort leading to outcomes. In addition, the Theory of Reasoned Action (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975) affords a basis for predicting relations among attitudes toward behavior and intentions to behave. Here the term prediction is used in a **statistical** sense. Note, for example, that prediction models (Blum & Naylor, 1968; Stone, 1978) may involve either (a) a variable measured at Time 1 “predicting” another variable measured at Time 2 (e.g., in a predictive criterion-related validity study), (b) a variable measured at Time 1 “predicting” another variable measured at Time 1 (e.g., in a concurrent criterion-related validity study), or (c) a variable measured at Time 2 “predicting” a variable measured at Time 1, a previous point in time (e.g., in a postdictive criterion-related validity study). For example, a nonexperimental, concurrent validity study may show a positive correlation between worker attractiveness and annual compensation.

Note, however, that the ability to predict, in a statistical sense, does not imply that the variables in prediction models are **causally** related to one another (Bollen, 1989; Brannick, 1995; Cliff, 1987; Freedman, 1987; Holland, 1986; Kelloway, 1998; Ling, 1982; Millsap, 2002; Rogosa, 1987; Stone-Romero & Rosopa, 2004, 2005). Early work on the use of biodata for personnel selection (e.g., Owens, 1976) provides a vivid illustration of prediction in the general absence of evidence about cause. Whereas research on biodata showed that a large number of biographical variables were correlated with measures of job success, there was no evidence to show that the variables measured by biodata forms were the cause of job success.

As noted above, experimental research is needed to provide credible evidence on causal connections between variables. Thus, for example, the hypothesized causal connection between valences of outcomes and motivation could be tested in a randomized experiment in which valence levels were varied to determine their effects on motivation. To the extent that experimental research showed support for this and other predictions of Expectancy Theory, there would be evidence to demonstrate the validity of assumptions about cause.

The distinction between prediction and understanding causal connections between variables is extremely important in the context of tests of assumed mediation models. Because observed covariances among variables considered by a study are consistent with an assumed mediation model, researchers who test mediation models using data from non-experimental research often believe that the results of such tests yield information on causal connections between variables (Baron & Kenny, 1986; Cohen & Cohen, 1983;
Cohen et al., 2003). Regrettably, the most that such studies afford is a basis for prediction. Thus, when researchers make inferences about mediation on the basis of data from nonexperimental research, their inferences have little or no value in terms of understanding causal connections between variables (Bollen, 1989; Brannick, 1995; Cliff, 1987; Freedman, 1987; Holland, 1986; Kelloway, 1998; Ling, 1982; Millsap, 2002; Rogosa, 1987; Stone-Romero & Rosopa, 2004, 2005).

The Importance of Timing in the Measurement of Assumed Mediators and Effects

In studies of assumed mediation models, it is critical that assumed mediators and effects be measured at appropriate times (Mathieu & Taylor, 2006; Shadish et al., 2002). The reason for this is that the influence of causes on mediators and effects is typically not instantaneous. For example, it may take several weeks for the effects of performance feedback (a cause) to influence the effort that individuals exert on the job (a mediator) and job performance (an effect). Thus, it is critical that there be appropriate lags between the time at which causes are varied (either naturally or experimentally) and the times at which mediators and effects are measured.

Construct Validity Issues

Internal validity is not the only concern in studies that purport to test assumed mediation models. Construct validity is also a vital issue in studies dealing with such models.

The Construct Validity Concept

Construct validity refers to the appropriateness of making inferences from (a) the sampling particulars of participants, measures, manipulations, and settings of a study, to (b) the higher order constructs that are the study’s focus (Guion, 2002; Shadish et al., 2002; Stone-Romero, 1994). Ideally, the validity of each inference can be supported by evidence from appropriate construct validation studies (Cook & Campbell, 1979; Cronbach & Meehl, 1955; Guion, 2002; Shadish et al., 2002).

Construct Validity in Tests of Assumed Mediation Models

Construct validity is critical in tests of assumed mediation models. For example, in a test of the model shown in Figure 2, whatever the type of experimental design used, it is critical that the operational definitions of $X$, $M$, and $Y$ have construct validity. Thus, in an experimental study the researcher would have to insure that the manipulation of $X$ was a faithful representation of the underlying construct. This may prove problematic in randomized experiments or quasi-experiments if manipulations in either SP or NSP settings have low levels of correspondence to the target construct. It can also prove problematic in nonexperimental studies when measures of assumed causes have little or no correspondence with the focal construct.
It deserves adding that there is no sound a priori reason to believe that construct validity inferences are any more justified in studies that are nonexperimental as opposed to either quasi-experimental or randomized experimental in nature (Fromkin & Streufert, 1976; Shadish et al., 2002; Stone-Romero, 1994). In the case of the latter two types of designs, the important issue is that the study’s manipulations have both experimental realism and enough mundane realism to allow for valid construct validity inferences. However, experimental realism is far more important in research than is mundane realism (Fromkin & Streufert, 1976). Moreover, the often heard criticism that experiments conducted in contrived settings lack mundane realism pertains to the setting of a study (SP vs. NSP), not to the method of experimentation. Indeed, randomized experiments can be conducted in both SP (e.g., a laboratory complex at a university) and NSP settings (e.g., a factory).

Nevertheless, there are phenomena that can not be studied easily in contrived settings. For example, it is virtually impossible to study the effects of such phenomena as wildfires, floods, hurricanes, and tornados in contrived settings. In addition, in studies of some phenomena (e.g., the effect of stressors), ethical considerations preclude strong manipulations of independent variables. This may lead to the construct validity-related problem of confounding of constructs with levels of constructs (Cook & Campbell, 1979; Shadish et al., 2002). Moreover, the manipulations used in randomized experimental research may lead to such problems as demand characteristics and experimenter expectancy effects, both of which may detract from valid inferences about construct validity. Likewise, in nonexperimental research, hypothesis guessing and response-response bias may lessen the validity of inferences about construct validity.

As noted above, tests of mediation models involve tests of two causal paths, $X \rightarrow M$ and $M \rightarrow Y$. Thus, it is vital the operational definitions of the associated variables have construct validity. For example, a researcher might conduct two experiments. In the first, $X$ would be manipulated to assess its effects on $M$. In the second, $M$ would be manipulated to assess its effects on $Y$. Assuming that the manipulations of $X$ and $M$ in the respective experiments were construct valid, an inference that $M$ was a mediator of the relation between $X$ and $Y$ would be justified. However, if the manipulation of $X$ lacked construct validity, the same inference would not be justified.

A number of construct validity issues are critical in tests of assumed mediation models. One is that the manipulations in a study are free of demand characteristics. Unless they are, for example, an observed relation between $X$ on $M$, may be a spurious byproduct of demand. Another is that the manipulations used in a study not suffer from the problem of confounding constructs with levels of constructs (Shadish et al., 2002). To the degree that this problem exists, estimates of the effects of (a) $X$ on $M$ and $Y$, and/or (b) $M$ on $Y$ will be.
biased (typically downward). If the same estimates are used as a basis for causal modeling, estimates of effects derived from such modeling also will be biased.

The Relative Importance of Validity Types in Tests of Mediation

Having considered internal and construct validity in the preceding paragraphs, it is critical to comment on their relative importance in tests of mediation models. In our view, the most important issue in such tests is internal validity. The reason for this is that mediation models always deal with causal connections between variables. As such, it is vital that inferences about cause be based on research that is strong in terms of the internal validity criterion. Unless there is convincing evidence about cause, the construct validity of sampling particulars (e.g., operational definitions of assumed causes, mediators, and effects) is a moot issue. However, if a researcher is able to provide sound evidence about cause (i.e., conduct a study that is strong with respect to internal validity), he, or she, is then obligated to provide evidence on the construct validity of the sampling particulars.

Inferences About Cause and Mediation are Quite Common in Articles Reporting the Results of Nonexperimental Research

Inappropriate inferences about causal connections between variables, including such connections in mediation models, are very common in research that uses nonexperimental designs. In this context, an inappropriate inference is defined as an instance of an author using such terms as “causes,” “effects,” “influences,” or their equivalents in describing the results of a mediation study that uses either a nonexperimental or a quasi-experimental design. In such cases, inferences about cause do not rest on a solid empirical foundation. As noted above, the greater the degree to which a study uses randomized experimental designs, the stronger the empirical foundation, and the more appropriate (justified) are inferences about mediation.

Researchers who conduct nonexperimental research often use causal language inappropriately in the titles and abstracts of their articles. In addition, causal language is very frequently used in describing both the results and implications of nonexperimental studies, including studies concerned with tests of assumed mediation models. A recent study by Stone-Romero and Gallaher (2006) shows strong evidence of this. They randomly sampled 161 articles from the 1988, 1993, 1998, and 2003 volumes of Personnel Psychology, Organizational Behavior and Human Decision Processes, Academy of Management Journal, and the Journal of Applied Psychology. The articles that were based on research that used randomized experimental, quasi-experimental, or nonexperimental designs were searched for the inappropriate use of causal language in their title, Abstract, and Results and Discussion sections. Results of the study showed that causal language was used inappropriately one or more times in 58 of the 73 nonexperimental studies (79%) and 14 of the 18 quasi-experimental studies (78%). Many of these instances of inappropriate causal inference involved tests of assumed mediation models. In the interest of illustrating this problem, we offer examples from recently published articles based on studies that tested
assumed mediation models. Several of them illustrate the inappropriate use of causal language in (a) the titles of articles, (b) the Abstracts of articles, and (c) the Results and Discussion sections of articles.

The Relative Validity of Inferences About Mediation in Recent Publications

In this section, we provide examples of six studies that deal with mediation in which the authors make claims about cause that have differing degrees of validity. In the first two, inappropriate inferences (i.e., inferences that are not justified by study design factors) about mediation are offered with no caveats concerning study design. In the second two, inappropriate inferences are advanced, but some cautions related to study design are considered. In the last two, the authors used appropriate study designs to test assumed mediation models.

Unwarranted Inferences About Mediation With No Cautions Related to Experimental Design

Bulger and Mellor (1997). Using data from a nonexperimental study of female union members, the authors tested an assumed causal model in which union self-efficacy was billed as a mediator of the relation between the magnitude of perceived union barriers and magnitude of union participation. The data were analyzed using the Baron and Kenny (1986) procedure. The title of the article was “Self-Efficacy as a Mediator of the Relationship Between Perceived Union Barriers and Women’s Participation in Union Activities” (p. 935). In the Results section of their article, the authors wrote “The Results of the analyses for predicting self-efficacy magnitude indicated that union self-efficacy mediated the relationship between perceived union barriers and union participation” (p. 941). And, in the Discussion section, they argued “As hypothesized, women’s self-efficacy did mediate the relationship between perceived barriers and participation . . .” (p. 941). Interestingly, the same section of the article did not offer a single caveat about advancing causal arguments on the basis of a nonexperimental study.

Brown, Jones, and Leigh (2005). Using data from a nonexperimental study of representatives of an office supply manufacturer, the authors tested an assumed causal model in which self-efficacy was viewed as a mediator of relations between (a) the assumed antecedents of previous performance and organizational resources, and (b) the hypothesized consequence of current performance. The data were analyzed using the Baron and Kenny (1986) regression-based strategy. In the Abstract of the article, the authors wrote “The results reveal a pattern of moderated mediation, in which goal level mediates the indirect effect of self-efficacy on performance when role overload is low, but not when it is high” (p. 972). Consistent with the same inference, in the Results section of the article, the authors stated that their findings indicated support for “the moderated mediation effect in Hypothesis 4. In other words, goal level mediates the effect of self-efficacy on performance, but only when role overload is low” (p. 977). And, in the Discussion section of the article, the authors specified that “The results indicate a pattern of moderated mediation, in which
goal level mediates the indirect effect of self-efficacy on performance when role overload is low, but not when role overload is high. When role overload is low, self-efficacy has both direct and indirect effects on performance (Wood & Bandura, 1989), but high role overload negates both types of effects” (p. 977). The Discussion section did not contain a single caution about using the findings of a nonexperimental study as a basis for making causal inferences.

Unwarranted Inferences About Mediation With Some Caveats Related to Experimental Design

Higgins and Judge (2004). Higgins and Judge (2004) used data from a nonexperimental study to test an assumed causal model involving several hypothesized mediators (e.g., applicant ingratiation, applicant self-promotion) of relations between applicant self-monitoring and job offers. The study’s data were analyzed using structural equation modeling (SEM) procedures. The authors used causal language inappropriately in the title of their article, i.e., “The Effect of Applicant Influence Tactics on Recruiter Perceptions of Fit and Hiring Recommendations: A Field Study” (p. 622). In addition, in the Abstract of the article they argued that “…perceived fit mediated the relationship between ingratiation and hiring recommendations” (p. 622). What’s more, they made numerous other causal arguments in the article. For instance, they asserted that “Because ingratiation had a positive effect on perceived fit and perceived fit had a positive effect on hiring recommendations, it seemed likely that fit evaluations mediated the relationship between ingratiation and recruiter hiring recommendations” (p. 628). In addition, they stated that “Furthermore, it is worth noting the differential effects of ingratiation and self-promotion on perceived fit in the present study. Whereas ingratiation had a strong, positive effect on fit, self-promotion had only a weak, nonsignificant effect” (p. 630). It deserves adding that even though they made more cautious claims about causal connections between the measured variables at various places in the Results and Discussion sections of their article, there were several instances in which their claims about cause were unqualified and unsupported by their study’s findings.

Stewart and Barrick (2000). The authors used data from a nonexperimental study to test an assumed causal model in which intrateam processes were the hypothesized mediator of relations between (a) the putative independent variables of team interdependence and team self-leadership, and (b) the assumed dependent variable of work team performance. The title of their article was “Team Structure and Performance: Assessing the Mediating Role of Intrateam Processes and the Moderating Role of Team Type” (p. 135). In the Abstract of their article, the authors wrote “Intrateam process mediation was found for relationships with interdependence but not for relationships with team self-leadership” (p. 135). In the Results section of their article, the authors stated “Process, thus, mediates 86 percent of the effect of interdependence on performance for teams engaged in conceptual tasks and 55 percent of the effect for teams engaged in behavioral tasks” (pp. 142-143). In addition, they wrote “Hypothesis 5, predicting process mediation of the interdependence-performance relationship is strongly supported for teams engaged in conceptual tasks and moderately supported for teams engaged in behavioral tasks” (p. 143). And, in the Discussion section of their
article, the authors stated “In particular, we found that intrateam processes mediated the relationship between interdependence and performance in these teams” (p. 144). To their credit, the authors cautioned that “… the data are correlational in nature. Future studies that directly manipulate interdependence and self-leadership are needed to clearly provide evidence on causal relationships” (p. 145). Although this is an appropriate caveat, it would not have been needed if the authors had been more careful about the way they described the study’s findings in the Results and Discussion sections of their article. For example, the authors could have reported that whereas the study’s findings were consistent with their assumed causal model, they did not demonstrate that the same model was the only one that was consistent with the results. Had they described the study’s findings in this manner, there would have been no need for the mea culpa.

**Appropriate Inferences About Mediation**

*Heilman and Chen (2005).* The authors argued that gender stereotypic prescriptions (i.e., norms about how men and women should behave) influence expectations about altruistic citizenship behavior. In addition, these expectations can affect both performance evaluations and reward recommendations differently for men and women. In two experimental studies, the sex of the employee being rated was manipulated along with altruistic citizenship behavior. When the same altruistic citizenship behavior was performed by the female and male employee, men received higher performance evaluations and reward recommendations than women. When altruistic citizenship behavior was withheld (by both sexes), ratings decreased (for both sexes). However, men still received higher performance evaluations and reward recommendations than women. In a third experiment, the sex of the employee was manipulated. The researchers found that altruistic citizenship behaviors were viewed as less optional and more required for female than for male employees. Taken together, this article suggests the following mediation model: Gender Stereotypic Prescriptions → Expectations About Altruistic Citizenship Behavior → Performance Evaluations and Reward Recommendations. Note that, in the first two studies, Heilman and Chen (2005) manipulated employees’ sex and measured (a) performance evaluations and (b) reward recommendations. In the third study, employees’ sex was manipulated and expectations about altruistic citizenship behavior were measured. Although Heilman and Chen (2005) did not use causal language in describing their findings, their randomized experiments provide credible evidence on the causal links.

*Meyer and Gellatly (1988).* Using two experiments, the authors investigated whether assigned goals affect personal goals and task performance through perceived performance norms. In the first experiment, assigned goals (i.e., the independent variable) were manipulated at three levels (i.e., easy, difficult, and impossible goals), and perceived performance norms (i.e., the mediator) and personal goals and task performance (i.e., two of four dependent variables) were measured. Using analysis of covariance and path analysis, the researchers provided support for their hypothesized mediation model. In the second experiment, assigned goals (i.e., easy, difficult, and no goal) and perceived performance norms (i.e., low, high, and no norm) were manipulated, and personal goals and task performance were the measured outcomes. Again, using analysis of covariance and path analysis, the
researchers found support for their hypothesized mediation model. Note that in this study, the research design, not the statistical analysis, provided support for the mediation inferences. The authors rightfully noted that “Conclusions based on the findings of Experiment 1 are limited by the fact that relations other than those involving assigned goal difficulty were correlational. The purpose of Experiment 2 was to examine further the mediating role of perceived performance norm by manipulating performance norm independently of goal difficulty” (Meyer & Gellatly, 1988, p. 415).

Why do Unwarranted Claims About Cause Appear in Publications?

The illegitimate claims about causal relations between variables (e.g., as in models involving mediation) that appear in such outlets as journal articles, books, book chapters, and technical reports may have a number of roots. One is the desire on the part of authors to increase the seeming importance of the results of their research. Another possibility is that authors who make such claims have poor training in research design and statistical methods. They may actually believe that inferences about cause are warranted when “causal modeling” procedures are applied to data from nonexperimental studies. Yet another possibility is that the authors of such works believe the unfounded advice of individuals who are regarded as authorities in research design and statistics. For example, Cohen and Cohen (1983) wrote that “Causal model analysis provides a formal calculus of interference [sic] which promises to be as important to the systematically observing scientist as is the paradigm of the controlled experiment to the systematically experimenting scientist” (p. 14). They went on to argue that “Now, the major analytic tool of causal models analysis is MRC [multiple regression/correlation], and particularly regression analysis. Even the simplest regression equation states that \(Y\) is a linear function of \(X\) carries, in its asymmetry, the implication that \(X\) causes \(Y\), and not the other way around” (p. 14). Subsequently, they assert that “We find the old saw that ‘correlation does not mean causation,’ although well intentioned, to be grossly misleading. Causation manifests itself in correlation, and its analysis can only proceed through the systematic analysis of correlation and regression” (p. 15). Given their views on the use of covariance information to inform conclusions about cause, Cohen and Cohen (1983) devote an entire chapter of their book to testing causal models with multiple regression. Seemingly, the views in this and other works (e.g., Baron & Kenny, 1986; Blalock, 1964, 1971; Kenny, 1979) have led many authors to believe that the application of statistical techniques such as hierarchical multiple regression (HMR) to data from nonexperimental research allows for valid inferences about cause, including causal connections in mediation models. As we demonstrate below, such inferences are almost always unwarranted.

Critical Distinction Between Experimental Design and Statistical Methods

A review of the organizational sciences literature shows that many researchers do not seem to have an adequate appreciation for the difference between experimental design
and statistical methods. There are many examples of this. One is the high frequency with which studies that use nonexperimental designs are referred to inappropriately by their authors as “correlational” studies (e.g., Hui, Lee, & Rousseau, 2004; Stewart & Barrick, 2000). A second is that the authors of a large number of nonexperimental studies, including those dealing with mediation, make causal inferences because the data from such studies are analyzed with such “causal modeling” techniques as HMR, path analysis (PA), and SEM. Perhaps the authors of many such works believe that statistical methods that are billed as being of the “causal variety” provide an adequate basis for making inferences about cause. Regrettably, using “causal modeling” does nothing whatsoever to change the basic nature of a study. More specifically, analyzing data from a nonexperimental study with “causal modeling” methods can not change the study’s design to one that is of the randomized experimental variety. As such, claims about causality that stem from the application of causal modeling procedures to data from a nonexperimental study are no more warranted than those that could be made using less sophisticated data analytic procedures. Thus, “causal modeling” methods applied to data from nonexperimental or quasi-experimental research are incapable of providing a firm basis for inferences about mediation.

The language offered in articles written by researchers who have applied “causal modeling” procedures to data from nonexperimental research on mediation is troubling. More specifically, as noted above, many describe the results and implications of their research as if they were based on research that afforded a strong basis for causal inference (e.g., randomized experimental studies). Evidence in support of this argument can be found in numerous studies that have used causal modeling procedures to analyze data from nonexperimental studies, including the studies considered above in the section titled “The Relative Validity of Inferences About Mediation in Recent Publications.”

Model Specification Issues

In order for statistical techniques (e.g., HMR, PA, SEM) to provide valid information about causal connections between or among variables in mediation models, the models tested must be properly specified. There are several important aspects of model specification (Bollen, 1989; Millsap, 2002; Pedhazur, 1982). We consider them below.

Elements of Model Specification

First, proper specification requires that the researcher’s statistical model be consistent with the form of the relation being modeled. For example, if the actual relation between X and Y is quadratic (curvilinear), the regression of Y on X would have to include an $X^2$ term. This is seldom an issue in research in the organizational sciences because most relations that are considered by researchers are linear.

Second, correct specification entails the inclusion of all relevant causes in a model. Unless they are included, the model suffers from the omitted variables problem and its consequences, including biased estimates of path coefficients (Bollen, 1989; James & Brett, 1984). This is a major concern in research in the organizational sciences, for at least
two reasons. First, in studies of most phenomena, researchers seldom know all of the relevant causes. Second, even if they were known, the constraints that researchers face (e.g., time that study participants will devote to questionnaire completion) limit the number of variables that can be measured and included in models.

Third, proper specification requires that a model not contain irrelevant causes. This is often a serious problem in the organizational sciences because there are few well-developed and validated theories in this field. As a result, tested models often contain variables that have little or no true causal relations with measures of assumed effects.

Fourth, in a properly specified model, the direction of causal flow is consistent with reality, i.e., the true pattern of causal connections between or among constructs. This is an extremely important issue in research in the organizational sciences (Stone-Romero & Rosopa, 2004, 2005). As noted above, researchers seldom, if ever, know the actual pattern of causal connections between and among variables. Thus, model misspecification resulting from this may be a problem in virtually all nonexperimental studies in the organizational sciences. Model misspecification may also be an issue in experimental studies. For example, the correct causal model may be that \( S \) causes \( T \). However, a researcher may test an incorrect causal model, i.e., one in which it is assumed that \( T \) causes \( S \). Assuming that there are no confounds, the results of the study will show no relation between manipulated levels of \( T \) and measured levels of \( S \). The self-correcting nature of research conducted in accordance with the scientific method (Kerlinger & Lee, 2000; Stone, 1978) will reveal the flaw in the researcher’s thinking.

Fifth, it is critical that all relevant mediators be included in a mediation model. Thus, for example, the true causal model may involve multiple mediators (e.g., the model shown in Figure 3 below). Unfortunately, most tests of mediation models involve only a single mediator (e.g., models such as that shown in Figure 1). Indeed, virtually all of the tests that use the Baron and Kenny (1986) HMR procedure are examples of this problem. As a result, the same tests are very likely to produce biased estimates of regression coefficients.

The Need for a Completely Specified Model

Assuming that the causal flow in a model being tested is correct, the degree of bias in estimates of “effects” tested by mediation models is a function of the degree to which relevant and important causes (\( Xs \)) and mediators (\( Ms \)) are considered by a study. Thus, research that addressed the most important causes and mediators would yield estimates of effects that were not greatly biased. Again, however, to produce such estimates, a researcher would have to test a model that considered all important causes and mediators. The greater the degree to which the model was properly specified, the greater the validity of inferences about the magnitude of the effects revealed by a study.

Problems Stemming from Model Misspecification

Numerous problems result from model misspecification (Blalock, 1971; Bollen, 1989; Pedhazur, 1982). In the interest of brevity, we do not detail those problems here. However, we reiterate that one very important consequence of model misspecification is biased estimates of “effects” (e.g., regression coefficients, path coefficients, structural parameters).
The same biased estimates can result in (a) inferences about model tests that are invalid, and (b) recommendations for practice that rest on an invalid empirical foundation (Stone-Romero & Rosopa, 2004).

The implications of experimental design for model specification merit noting. This is especially true because of the fact that virtually all of the effect ($Y_l$) constructs studied by researchers in the social and behavioral sciences have multiple causes ($X_i$s), which often exert their effects on dependent variables through multiple mediators ($M_k$s). Figure 3 shows an instance of multiple mediation. In it, there is a single cause, a single effect, and multiple mediators. Note, however, that in most of the models in the social and behavioral sciences, there are likely to be multiple causes, multiple mediators, and multiple effects. The well-known Job Characteristics Model (Hackman & Oldham, 1976) is an apt example of this. It posits that several causes (i.e., the core job characteristics of skill variety, task identity, task significance, autonomy, and feedback) influence various outcomes (e.g., internal work motivation, work effectiveness, job satisfaction), through several mediators (i.e., the critical psychological states of experienced meaningfulness of work, experienced responsibility for the outcomes of work, and knowledge of the results of work activities). To test this model via randomized experiments or quasi-experiments, it would be vital to have appropriate manipulations of all of the causes. However, randomized experiments or quasi-experiments that considered only two levels of each of the five causes would require $2^5 = 32$ conditions. Conducting a randomized experiment with such a large number of
conditions would prove to be quite burdensome in SP settings and virtually impossible in NSP settings. Thus, it seems unlikely that any given pair of randomized experiments or quasi-experiments could provide a test of the complete model. So, it appears that for this and other reasons, most tests of the Job Characteristics Model have used nonexperimental designs. Regrettably, it is highly likely that all such tests suffer from model misspecification problems. The reason for this is simple. More specifically, the outcomes of internal work motivation, work effectiveness, and job satisfaction have many causes that are not considered by the same model. Thus, estimates of effect sizes (e.g., regression coefficients, path coefficients) that are based on tests that fail to include measures of all important causes of the outcomes will be biased. In addition, the lesser the construct validity of the measures of relevant constructs, the greater will be the bias.

Some Study Designs and Their Value in Demonstrating Mediation

In the preceding sections, we detailed a number of problems that stem from tests of mediation models based on data from nonexperimental research. It is important to note that although the inferences about internal validity that stem from such tests are virtually always suspect, it is possible to conduct tests of mediation that yield inferences with greater levels of internal validity. In this section, we consider a number of design alternatives. Table 1 summarizes the major attributes of these designs. By way of introduction, it merits noting that these designs differ in terms of (a) the manipulation or measurement of the assumed independent variable ($A_X$), (b) the manipulation or measurement of the assumed mediator ($A_M$), (c) whether or not participants are assigned randomly to conditions, and (d) the number of studies used in testing mediation models.

Two Randomized Experiments

Mediation models can be tested with two or more randomized experiments. For example, assume that a researcher is interested in testing the model shown in Figure 2. One experiment can be conducted to demonstrate that $X$ causes both $M$ and $Y$. A second experiment can be used to show that $M$ causes $Y$. Taken together, the results of these experiments would provide convincing evidence that the path between $X$ and $Y$ is mediated by $M$. However, in and of themselves, the findings of these experiments would not allow for a determination of the degree to which mediation is partial, as opposed to complete. However, because they provide estimates of the degree of correlation between $X$ and $M$, $M$ and $Y$, and $X$ and $Y$, the correlation coefficients can be used to test a causal model that has both direct and indirect effects (Bollen, 1989). It deserves adding that the validity of inferences stemming from such a model would vary as a function of the degree to which the study’s manipulations and measures had construct validity. More specifically, the range of the manipulated variables would have to be sufficient to allow for unbiased estimates of the degree to which variables were related to one another. In addition, measures of the dependent variable(s) in each experiment would have to have construct validity.

In terms of the above-noted considerations, the results of the first experiment would be questionable if they were subject to the response-response consistency artifact; that is,
participants’ responses to measured levels of $M$ affected their responses to measured levels of $Y$. In such a case, the researcher could conduct three studies. One would test the causal connection between $X$ and $M$, a second would test the causal link between $M$ and $Y$, and the third would test the causal link between $X$ and $Y$. In summary, randomized experimental designs can be used to test models that posit mediation (e.g., Figures 1 and 2).

Note that if participants are randomly selected from the same population, the sample used to test the first prediction ($X \rightarrow M$) can differ from that used to test the second prediction ($M \rightarrow Y$). The use of separate samples of participants would be an especially important design feature if a researcher wanted to avoid the negative effects of such construct validity problems as hypothesis guessing and response-response bias.

It deserves adding that if the above-described two experiments produced expected findings, symbolic logic could be used to make a valid inference about mediation. More specifically, through a symbolic sentence derived from Theorem 26 of Kalish and Montague (1964, p. 80):

$$(X \rightarrow M) \land (M \rightarrow Y) \rightarrow (X \rightarrow Y)$$

where $X =$ changes in $X$, $M =$ changes in $M$, $Y =$ changes in $Y$, and $\land =$ and. In other words, if it can be demonstrated that (a) changes in $X$ cause changes in $M$ and (b) changes in $M$ cause changes in $Y$, then it is logical to conclude that changes in $X$ cause changes in $Y$.

The above-noted strategy of performing two or more randomized experiments can be used to show the effects of multiple independent variables (e.g., $X_1$, $X_2$, and $X_3$) on a hypothesized mediator ($M$). In addition, experimental research can be performed to show that the effects of a given independent variable (e.g., $X_1$) exert effects on $Y$ via several mediation paths (e.g., $M_1$, $M_2$, $M_3$, and $M_4$). Moreover, experimental tests of mediation can consider multiple dependent variables (e.g., $Y_1$, $Y_2$, and $Y_3$). Thus, through an appropriate set of experiments, a researcher can test very complex mediation models. In the process, he, or she, can develop highly credible evidence about the validity of such models.

Overall, the two randomized experimental design alternative provides for a test of an assumed mediation model that has a high level of internal validity. Thus, in Table 1, we rate two randomized experiments as very strong in terms of this criterion.

### Table 1

<table>
<thead>
<tr>
<th>Design of Study to Test Assumed Mediation Model</th>
<th>$\Delta X$</th>
<th>$\Delta M$</th>
<th>$\Delta Y$</th>
<th>Control Over Confounds</th>
<th>Internal Validity Inferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two randomized experiments</td>
<td>Manipulated</td>
<td>Manipulated</td>
<td>Measured</td>
<td>Design</td>
<td>Very strong</td>
</tr>
<tr>
<td>Two quasi-experiments</td>
<td>Manipulated</td>
<td>Manipulated</td>
<td>Measured</td>
<td>Statistical</td>
<td>Moderately strong</td>
</tr>
<tr>
<td>One randomized experiment</td>
<td>Manipulated</td>
<td>Measured</td>
<td>Measured</td>
<td>Design</td>
<td>Weak</td>
</tr>
<tr>
<td>One quasi-experiment</td>
<td>Manipulated</td>
<td>Measured</td>
<td>Measured</td>
<td>Statistical</td>
<td>Weak</td>
</tr>
<tr>
<td>Two nonexperiments</td>
<td>Measured</td>
<td>Measured</td>
<td>Measured</td>
<td>Statistical</td>
<td>Very weak</td>
</tr>
<tr>
<td>One nonexperiment</td>
<td>Measured</td>
<td>Measured</td>
<td>Measured</td>
<td>Statistical</td>
<td>Very weak</td>
</tr>
</tbody>
</table>

Overall, the two randomized experimental design alternative provides for a test of an assumed mediation model that has a high level of internal validity. Thus, in Table 1, we rate two randomized experiments as very strong in terms of this criterion.

Downloaded from http://orm.sagepub.com at UNIV OF SOUTH FLORIDA on August 3, 2009
One Randomized Experiment

A much weaker design, in terms of supporting claims about mediation, is one that involves a single experiment and one sample. In this design, $X$ is manipulated to determine its causal impact on $A_M$ and $A_Y$. Here, it is possible to make a strong inference about $X$ as a cause of both $M$ and $Y$. However, it is not possible to conclude validly that the causal chain is $X \rightarrow A_M \rightarrow A_Y$. It may well be the case that there is no mediation. That is, in reality, both $A_M$ and $A_Y$ are dependent variables. In addition, it may be the case that the causal chain is actually $X \rightarrow A_Y \rightarrow A_M$. In this case, the assumed mediator and dependent variables have roles that are opposite those specified by the model being tested. In view of these problems, we rate the one randomized experiment as weak in terms of internal validity (see Table 1). The principal reason for this is the inability of the study to establish the causal sequence between $A_M$ and $A_Y$.

Two Quasi-Experiments

An assumed mediation model can also be tested using two quasi-experimental studies. In the first, $X$ is manipulated to assess its effects on $A_M$. In the second, $M$ is manipulated to assess its effects on $A_Y$. It deserves noting that there are actually a large number and variety of quasi-experimental designs (Cook & Campbell, 1979; Shadish et al., 2002), each of which may lead to inferences about cause that have differing degrees of internal validity. Length considerations do not allow for a consideration of the many design alternatives.

In a study that uses two quasi-experiments, inferences about cause are conditional on the researcher’s ability to rule out rival explanations of detected effects. As noted above, it is often the case that the data needed to rule out rival causes are not available to the quasi-experimental researcher. As a result, inferences about mediation based on quasi-experiments rest on a weaker foundation than inferences stemming from two randomized experiments. Overall, we view a test of an assumed mediation model that uses two quasi-experiments as having the potential to provide moderately strong internal validity inferences (see Table 1).

One Quasi-Experiment

Another design option is to test an assumed causal model using a single quasi-experiment. In the simplest such study, a treatment would be administered to two groups that were not randomly assigned to conditions. After this, levels of $A_M$ and $A_Y$ would be measured. A study of this type would suffer from the same weaknesses just noted for the two quasi-experimental design. In addition, there would be ambiguity about the true causal chain. Thus, in such a study, it would not be possible to conclude validly that the actual causal chain is $X \rightarrow A_M \rightarrow A_Y$. In reality, there may be no mediation; both $A_M$ and $A_Y$ may be dependent variables. In addition, it may be the case that the causal chain is actually $X \rightarrow A_Y \rightarrow A_M$. If so, a researcher using this design would falsely infer that the causal sequence is $X_A \rightarrow M \rightarrow A_Y$. Overall, the single-study quasi-experimental design provides for a much lower degree of internal validity than a two-study quasi-experimental design. As a consequence, we regard the single-study quasi-experimental design as providing a weak basis for inferences about the validity of an assumed mediation model (see Table 1).
Two Nonexperimental Studies

Yet another design option is to conduct two nonexperimental studies. This sort of design might be used in several situations. One is when the test of mediation is based on the aggregation of research findings from separate studies (e.g., two meta-analyses). A second is when the researcher hopes to rule out the possible confounding effects of response-response bias.

In the first such study, the degree of relation between measured values of $AX$ and $AM$ would be assessed. Assuming a nonzero relation is found and it is in the expected direction, the researcher could infer that it was consistent with the hypothesized model of $AX \rightarrow AM$. However, he, or she, could not conclude validly that the assumed model was correct.

In a second nonexperimental study, the researcher would use data on measured levels of $AM$ and $AY$ to test the assumed model of $AM \rightarrow Y$. However, even if the expected relation were found, it would not provide a solid basis for inferring that the researcher’s assumed causal model was correct. Thus, for the reasons cited above, if research used the two nonexperimental studies strategy, an inference that the assumed causal model was correct would generally rest on a very shaky empirical foundation. Overall, therefore, we regard the two nonexperimental design strategy as affording a very weak basis for internal validity inferences (see Table 1). Moreover, as is noted in the following section, the validity of inferences about cause would be little improved by the use of so-called causal modeling techniques.

One Nonexperimental Study

In this type of study, the researcher measures all variables in an assumed mediation model. This design suffers from all of the problems of the two-study nonexperimental design just described. It also suffers from the possibility of participants’ responses being influenced by response-response bias. Overall, therefore, we view the one-study nonexperimental design option as providing a very weak basis for internal validity inferences (see Table 1).

Conclusions and Recommendations

In the foregoing sections, we considered a number of issues associated with tests of assumed mediation models. In this final section, we offer several conclusions about tests of such models. In addition, we provide recommendations for tests of such models and the presentation of findings that result from such tests.

Conclusions

We believe that our analysis provides a sound basis for the following conclusions. First, because of very important differences between research that uses nonexperimental, quasi-experimental, and randomized experimental designs, the internal validity of studies using such designs also differs markedly. As a result, claims about mediation are (a) almost never
justified when they are based on research that uses either of the two above-described non-experimental design strategies, (b) much more justified when based on research that uses the two quasi-experimental design approach, and (c) highly justified when based on the findings of research that uses the two randomized experimental design strategy.

Second, inferences about mediation are causal inferences. As such, they are most justified when they are based on the findings of research using the two randomized experimental design approach. In addition, as noted above, complex mediation models can be tested with studies using multiple randomized experiments.

Third, there is a clear distinction between experimental design and the statistical techniques that can be used to analyze data from empirical studies. Thus, in terms of the criterion of valid inferences about causal connections between or among variables (i.e., internal validity), the use of such “causal modeling” procedures as HMR, PA, and SEM are not an adequate substitute for research that uses randomized experimental designs.

Fourth, there is a nontrivial difference between (a) consistency between an assumed causal model and the results of a study, and (b) consistency between reality (a true causal model) and the results of a study. As we noted, the findings of nonexperimental research can only address the former issue. In addition, evidence of the former is not proof of the latter.

Fifth, model specification problems abound in nonexperimental research, resulting in the inability of researchers to make valid inferences about cause in tests of assumed causal models, including those concerned with mediation.

Sixth, although there are many strategies for conducting tests of assumed causal models (e.g., HMR, PA, SEM), the use of all such strategies leads to highly equivocal inferences about the validity of assumed causal models when the data being analyzed are from non-experimental research.

Seventh, and finally, a number of previous studies have shown very serious problems with techniques that purport to provide tests of assumed mediation models (e.g., HMR). Thus, tests of mediation models should not rely on the use of such techniques.

Recommendations

In view of the above, we offer several recommendations. First, we recommend the use of the two randomized experimental design approach when it is important to understand (explain) phenomena (e.g., mediation) of interest to a researcher. Second, we believe that individuals who analyze data from nonexperimental research should refrain from arguing that the results of their analyses provide valid evidence of mediation. All that they can argue legitimately is that patterns of covariances among measured variables are consistent with an assumed causal model. At the same time, they must directly and clearly acknowledge that (a) a number of other causal models are also consistent with the same pattern of covariances, and (b) their findings do not provide a valid basis for making causal inferences about relations between or among variables.

It deserves adding, however, that causal language is quite appropriate in talking about theories or models that are to be tested by research. More specifically, there is nothing, whatsoever, wrong with talking about an assumed causal model in the Introduction section of an article. However, in nonexperimental research, substantive (alternative, as opposed to
null) hypotheses must be talked about in terms of covariance (e.g., correlation). In addition, in the Discussion section of an article, an author should be free to say something to the effect of “Hypothesis 1 argued that there would be a positive correlation between X and Y. The test of this hypothesis showed that there was. This finding is consistent with the assumed causal model shown in Figure 1. However, it may also be consistent with a number of other causal models.”

Third, we believe that journal editors need to insure that authors of manuscripts do not make unwarranted claims about cause (including mediation) when their data are from non-experimental research. In addition, we think that members of the editorial boards of journals need to recognize the limitations of “causal modeling” procedures. As a result, they must advise authors to not make unwarranted claims about cause (including mediation) in their articles.

Fourth, we recommend that individuals who teach undergraduate and/or graduate courses in such areas as statistics, research design, research methods, and causal modeling, instruct their students on the inferences that can and can not be made on the basis of data from nonexperimental research. Moreover, they need to disabuse students of the baseless arguments that appear in various publications about the inferences that are appropriate on the basis of “causal modeling” procedures (e.g., Baron & Kenny, 1986; Cohen & Cohen, 1983; Cohen et al., 2003; Kenny, 1979).

To the extent that these recommendations are followed, we should see more responsibly written articles in the organizational sciences and other fields that use data from empirical research to show support for assumed causal models. In addition, to the degree that research uses randomized experimental designs it can facilitate the development of sound explanations about phenomena. Moreover, it will foster recommendations for practice that rest on a sound empirical foundation.

References


**Eugene F. Stone-Romero** (PhD, University of California–Irvine) is a professor of management at the University of Texas at San Antonio. He is a fellow of the Society for Industrial and Organizational Psychology, the Association for Psychological Science, and the American Psychological Association.

**Patrick J. Rosopa** (PhD, University of Central Florida) is an assistant professor at Clemson University. His current research interests include mentoring, organizational citizenship behavior, mediation inferences, and quantitative methods including testing for interactions, regression diagnostics, and weighted least squares estimation.