CAUSALITY AND ENDOGENEITY: PROBLEMS AND SOLUTIONS

John Antonakis
Faculty of Business and Economics
University of Lausanne

Samuel Bendahan
Faculty of Business and Economics
University of Lausanne

Philippe Jacquart
The Wharton School
University of Pennsylvania

Rafael Lalive
Faculty of Business and Economics
University of Lausanne

Reference :

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D.V. Day (Ed.), The Oxford Handbook of Leadership and Organizations (pp. 93-117). New York: Oxford University Press.

Abstract

Most leadership and management researchers ignore one key design and estimation problem rendering parameter estimates uninterpretable: *Endogeneity*. We discuss the problem of endogeneity in depth and explain conditions that engender it using examples grounded in the leadership literature. We show how consistent causal estimates can be derived from the randomized experiment, where endogeneity is eliminated by experimental design. We then review the reasons why estimates may become biased (i.e., inconsistent) in non-experimental designs and present a number of useful remedies for examining causal relations with non-experimental data. We write in intuitive terms using nontechnical language to make this chapter accessible to a large audience.

Author supplied-keywords: Endogeneity, Causality, Randomized Experiments Quasi-Experimentation, Instrumental Variables, Common-Methods Bias.

> "*Man is impelled to invent theories to account for what happens in the world. Unfortunately, he is not quite intelligent enough, in most cases, to find correct explanations. So that when he acts on his theories, he behaves very often like a lunatic.*"
>
> — Aldous Huxley (*Texts and Pretexts*, 1932, p. 270)

Theory is the ultimate aim of science (Kerlinger & Lee, 2000). Contrary to the lay individuals described in Aldous Huxley's opening quote, scientists put their theories to the empirical test in order to determine whether or not the theories are plausible. As stated by Murphy (1997, p. 4), "the methods chosen should be appropriate to the research question and the inferences drawn should be consistent with what was actually attempted in [the] study". Given the importance of theory testing for understanding and predicting how the world works, the choice of research design and analysis method is of the utmost importance, particularly because research findings influence policy and practice.

As we will explain in detail, the randomized experiment is the gold standard to identify and test causal relationships. However, be it for practical or ethical considerations, it may not always be possible to conduct randomized experiments (see Cook, Shadish, & Wong, 2008; Rubin, 2008). Although most researchers undoubtedly know that the randomized experiment is the method of choice to infer causality, many researchers ignore two key issues:

1. experimental design is not the only method available to make valid causal inferences; that is, there are other designs available to make valid causal claims, which do not require manipulation of the exogenous variables on the part of the researcher.

2. nonexperimental designs that do not address problems of endogeneity are pretty much useless for understanding phenomena; that is, finding a relationship between an endogenous regressor x—that has not been purged from endogeneity somehow—and y does not help

leadership theory one bit!

The way in which we state the above two points, particularly the second is, admittedly rather harsh and blunt; however, going through the chapter will make readers realize that what we have said is actually an understatement. To help research advance in leadership (and other social sciences) more researchers must join the effort to stomp-out endogeneity; this problem is far bigger than we dared to imagine.

We recently conducted a review of leadership studies suggesting that the conditions and designs that allow to uncover causal relationship with non-experimental data are not well understood by the majority of leadership researchers (Antonakis, Bendahan, Jacquart, & Lalive, 2010). This problem is not isolated to the field of leadership. In fact, aside from the field of economics, which starting addressing this problem a couple of decades ago, many social-sciences disciplines face a similar situation (Bascle, 2008; Duncan, Magnusson, & Ludwig, 2004; Foster & McLanahan, 1996; Gennetian, Magnuson, & Morris, 2008; Halaby, 2004; Larcker & Rusticus, 2010; Shaver, 1998).

For example, a recent review has found that less than 10% of the papers published in the top strategy journal (i.e., *Strategic Management Journal*) properly analyzed the non-experimental data they presented (Hamilton & Nickerson, 2003). In our review, where we examined a random sample of 110 leadership papers published in top scientific journals, we found that researchers failed to correct between 66 % to 90 % of design and estimation conditions that threaten estimate validity (refer to Table 1 for a summary of the threats). We also found that 109 of the articles had at least one threat to validity and that 100 articles had three or more validity threats (which we discuss in more detail later). This sad state of affairs has to be changed because policy implications that stem from research that is incorrectly undertaken will be wrong.

[Table 1 here]

When we refer to causal analysis of nonexperimental data, we are referring to designs that will produce coefficients that capture the magnitude of the true (causal) relationship rather than just an association or a correlation (which could be spurious). True estimates are called *consistent*. To say that an estimate is consistent suggests that it will converge to the true population parameter as sample size converges to infinity (i.e., asymptotically).

The main threat to consistency is *endogeneity*; much of what we will discuss in this chapter focuses on explaining what it is and how to deal with it. If an estimate is inconsistent, it is purely and simply uninterpretable. A coefficient may appear to adequately reflect the hypothesized relationship—for example, it is the right direction and the effect is highly significant—but in presence of endogeneity it will be inconsistent and will not reflect the true population parameter. Reporting it is pretty much useless to help understand a phenomenon because the observed correlation may be far off from the true relation; that is, the true relation could be higher, lower, zero, or of a different sign from the observed association (correlation). This is why understanding the nature of causal designs is crucial.

Our goal in this chapter is to present some of the methods available to researchers for testing theory correctly. We begin this chapter by discussing what theories are and why causality is important to theory testing; we then present a simple example of endogeneity with simulated data and extend the problem to leadership research to show that models with endogenous regressors are simply not very useful (these data will prove to be very useful as a teaching aid for those teaching methods courses). Next, we present the randomized experiment as a failsafe way to make causal claims; an understanding of what, precisely, random assignment does is essential for understanding how endogeneity is engendered, why it renders estimates biased. We then present some methods that can be used to causally analyze nonexperimental data. We close this chapter by discussing future directions in leadership research.

What is causality?

A theory consists in a set of interrelated constructs and data connecting these constructs with the empirical world—within certain boundaries and under certain constraints (Antonakis et al., 2004); a theory is constructed so as to answers a number of questions: *What* elements are being studied and *how* do they relate? *Why* would this be so? *When* (*where* and to *whom*) does the theory apply? In order to be acceptable, a theory should be devoid of contradictions and be consistent with the empirical world—that is, it should have internal and external consistency; moreover, a theory should be testable, have both generality and parsimony (for in-depth treatment see Bacharach, 1989; Dubin, 1976; Kerlinger & Lee, 2000). More importantly, a theory should present a causally valid explanation of a phenomenon.

What causality is and how it should be tested has important implications for understanding natural phenomena (and the theories that explain them); it also has important implications regarding how scientific research should be conducted. Causality is a fascinating topic that has been examined in-depth by many philosophers and scientists (cf. Mulaik, 2009; Pearl, 2009). In this chapter, we steer clear from philosophical considerations and adopt a pragmatic and broadly accepted view on causality. Here, we focus on understanding how one can assess and quantify a causal effect. Classically, *x* is said to have an effect on *y* if the following three conditions are met (Holland, 1986; Kenny, 1979):

(a) *y* follows *x* temporally

(b) y changes as x changes (and this relationship is statistically significant)

(c) no other causes should eliminate the relation between *x* and *y*.

The first two conditions are quite straight forward; regarding first condition, caution is warranted in the case where *x* and *y* simultaneously affect each other; also, that *y* follows *x* by no mean suggests that *x* caused *y*. This latter point will become clear in the first simulation we

present. Also, we should note that from the second condition it follows that the constructs being studied should be operationalized (measured) and statistically analyzed. Although necessary, it is clear that these two first conditions are not sufficient to establish causality. They are however sufficient for one to fall prey to the *post hoc, ergo propter hoc* fallacy, which consists in wrongly interpreting causality by inferring that *x* is the cause of *y* precisely because it occurred before *y* (Kerlinger & Lee, 2000). The third condition has more to do with design and analysis issues than it has with theoretical arguments, though theory is important too (see also James, Mulaik, & Brett, 1982; Mulaik & James, 1995). It is also the more troublesome condition and the one with which much of our chapter will be concerned.

This third condition can be restated by simply saying that changes in *x* produce changes in y holding all other things equal. This is clearly the case if *x* varies randomly and independently from the system of variables under study; if *x* depended on some unmodeled causes that also drive other variables in the model then *x* would be said to be *endogenous*—hence the problem of *endogeneity*. As we alluded to in the introduction, the consequences of endogeneity are dire. If the necessary precautions are not taken to purge the endogenous variable of endogeneity then estimated coefficients are devoid of any meaning and cannot be interpreted.

Endogeneity: Two inconvenient demonstrations

We start-off with a very simple demonstration, presented by John Antonakis in the podcast *Endogeneity: An inconvenient truth* (available on Youtube), and previously discussed by Antonakis (in Fairhurst & Antonakis, in press): A philosopher is sent out on a field to observe a naturally occurring phenomenon and is required to piece together a theoretical account of what she saw. She observes 50 trials of the phenomenon, which consists of a disk streaking across the sky that almost always shatters soon after a loud "crack" is heard; the disk never shatters when the crack is not heard. She carefully gathers the data, including number of trials, whether the

crack was present or not (and how loud it was in decibels), and whether the disc disintegrated or not. Refer to Table 2 (Panel A) for a summary of the data regarding the relation between the presence of the "crack" and the disc.

[Table 2 here]

The *observed* (and it is important to highlight the word *observed*), correlation between the two variables, noise (heard or not) and disk (shattering or not) is very strong and statistically significant: $\phi = .92$, $\chi^2(1) = 42.32$, $p < .001$. The data spoke clearly: Thus, the philosopher concludes that the soundwaves from this loud crack—which emanate from some yet-to-be-established source—caused the disks to shatter into smithereens. She writes an extensive theory around this explanation; for the sake of argument several policy implications follow, which are military related (i.e., building jamming defenses against the "noise," which would potentially be a very dangerous weapon against her city state).

The research efforts of the philosopher are nothing more than futile as are the policy implications; unbeknown to the philosopher is the true causal model behind the data (refer to Table 2, Panel B). The noise is caused by a hidden shooter who fires a rifle shot at the disk. The noise and the disk shattering are both caused by the rifle that is being fired by the shooter. Thus, the sound is in no way related to the disk shattering; they both share a common cause. What relation is observed is coincidental—spurious. Reporting it and building policies around it pointless.

Knowing the true causal structure behind this data allows us to estimate whether there is a correlation between the noise and the disk; we will use the decibels as the variable of interest (because hearing the crack and whether the rifle fires are perfectly collinear). We thus estimate the following multivariate least-squares regression:

$$Disk = \beta_0 + \beta_1 Rifle + e \quad \text{Eq. 1}$$

$$Noise = \gamma_0 + \gamma_1 Rifle + u \quad \text{Eq. 2}$$

Where Disk = whether the disk shatters (=1) or not (=0); Noise = Noise measured in decibels; Rifle = whether the rifle was fired (=1) or not (=0). Note, Equation 1 is estimated using a linear probability model (OLS), which is perfectly fine to use, particularly in this case where rifle = 0, which is always associated with noise = 0 (Caudill, 1988). Given that standard errors might not be consistent for Eq. 1, we bootstrap the standard errors (using 1000 replications; we could also have used a robust variance estimator). We also bootstrap the standard error for the test of the significance of the residual correlation (Breusch & Pagan, 1980) between Disk and Noise to determine whether they are still related once the phenomenon is correctly modeled.

The estimated parameters are: $\beta_1 = .93$, $SE = .04$, $z = 22.47$, $p < .001$ and $\gamma_1 = 121.74$, $SE = 1.81$, $z = 67.34$, $p < .001$. However, the test for the significance of the residual correlation is not significant: $\chi^2(1) = 2.94$, $SE = 2.34$, $p > .10$. Thus, once the correct causal structure of the data is accounted for, it is clear that the noise is unrelated to the disk shattering. Thus, these variables have been "*d*-separated" or directionally-separated (Hayduk et al., 2003; Pearl, 2009). The noise was endogenous; thus regressing anything on the noise is pretty much as useless endeavor unless the true model is being estimated (or some corrective procedures are undertaken). Granted, we admit that there may be some very limited use to studying correlations in the initial phases of understanding a phenomenon. However, after studying the phenomenon for some time, we sincerely hope that researchers will go beyond merely studying associations; we do not see this being the case in leadership research (or organizational behavior, management, and applied psychology research in general).

A direct analog in leadership research to the "crack" in the above example is any

endogenous variable that does not vary randomly or independently of the specified model variables or omitted variables (i.e., it has a theoretical cause or several causes that correlate with the modeled variables). Consider Leader-Member Exchange (LMX), that is, quality of leader-member relations. LMX has been linked to several outcomes (*y*). However, LMX does not vary randomly in organizations. It depends on some factors that may stem from the leader, the follower, and the organization, which may correlate with a supposed outcome of LMX. If these factors are omitted from the model and if they predict *y* too, the effects of LMX on *y* cannot be correctly estimated. LMX (i.e., the "crack") depends on something (i.e., the "rifle shot"); if this "something" is not modeled when using LMX as regressor then what correlations are reported are really not very useful in advancing leadership research.

To better understand this problem, and how LMX (or any another other endogenous variable that is studied in leadership) relates to the "crack" assume the causal structure as described in Equations 3 and 4, which we have simulated. This account of what drives the two endogenous variables is theoretically plausible; it is, however, a simple model and not necessarily an adequate model that will suffice for the demonstration (note, all the coefficients in the model are "1"; the intercepts are -250 and 150 respectively):

$$LMX = \beta_0 + \beta_1 L\_Extra + \beta_2 L\_Incent + \beta_3 L\_IQ + \beta_4 F\_IQ + \beta_5 F\_Consc - \beta_6 F\_Neuro + 3 * e \quad \text{Eq. 3}$$

$$Turnover = \gamma_0 - \gamma_1 L\_Incent - \gamma_2 L\_IQ - \gamma_3 C\_Policies + \gamma_4 F\_Neuro + 3 * u \quad \text{Eq. 4}$$

Where LMX = quality of leader-member relations; L_Extra = leader extraversion; L_Incent = leader use of incentives; L_IQ = leader IQ; F_IQ = follower IQ; and F_Consc = Follower conscientiousness; F_Neuro = Follower neuroticism; Turnover = Follower turnover intentions;

C_Policies = company policies (including pay, working conditions etc.); *e* and *u* are random independent variables that are normally distribute. Also, suppose that the modeled independent variables are random variables (i.e., exogenous with respect to the two endogenous ones LMX and Turnover), variables are measured without error and that model is a correct causal account of what drives LMX and follower turnover. The summary data are listed in Table 3 (we generated these data using Stata and random seed 1234; note, because a covariance matrix can be generated from the summary data, those who are interested can replicate this analysis using a Structural Equation Modeling program). Interesting to note is that the observed correlation between LMX and Turnover is high and significant, $r(1000) = -.50$, $p < .001$.

[Table 3 here]

We then estimated a multivariate regression model (saturated), where we predicted LMX and Turnover from the independent variables. What is interesting to observe in this case is whether the residual correlation between LMX and Turnover is significantly different from zero: It is not, whether we estimate the model using OLS or maximum likelihood ($r = .02$, $p > .10$). Thus, whatever observed correlation is found between LMX and turnover is a false account of the relation between LMX and turnover (refer to the similarity of this conclusion with that which we present later when discussing the two-stage least squares estimator).

Now, we estimate the following naïve model:

$$Turnover = \delta_0 + \delta_1 LMX + \delta_2 C\_Policies + \psi \quad \text{Eq. 5}$$

From the above specification, LMX appears to affect turnover intentions on the part of subordinates, $\delta_1 = -.46$, $p < .001$. However, because theoretically, LMX is endogenous, this coefficient is devoid of any meaning. This point—and again, we are using LMX as an example and leadership research is replete with such potentially endogenous regressors (e.g., authentic

leadership)—has not garnered much interest from leadership scholars and is not well understood. For instance, House and Aditya's (1997) noted that "While it is almost tautological to say that good or effective leadership consists in part of good relationships between leaders and followers, there are several questions about such relationships to which answers are not intuitively obvious . . . . A specification of the attributes of high-quality LMX—trust, respect, openness, latitude of discretion—is as close as the theory comes to describing or prescribing specific leader behaviors. The theory implies that any leader behavior that has a positive effect on LMX quality will be effective. However, precisely what these behaviors are is not explicitly stated, as the appropriate leader behavior is dependent on anticipated subordinate response" (pp. 431-432).

Meta-analyses have established correlates of LMX, both antecedents and consequences (Dulebohn, Bommer, Liden, Brouer, & Ferris, 2011; Gerstner & Day, 1997); interestingly, Gerstner and Day (1997) had noted: "we avoid discussing [the relationships found] in terms of causal inferences regarding the direction of these relationships. For purposes of the present analyses, we treat them all as correlates." Fifteen years later, Dulebohn et al. (2011) noted: "In addition, many of the studies included in our analysis were based on a cross-sectional correlation design, which prevents the establishment of causal direction." Yet, Dulebohn et al. conducted tests of mediation to establish whether LMX mediates the effects of certain regressors on outcomes. These tests, however, reported biased coefficients because as we become clear later mediation must be undertaken using the 2SLS estimator (in the case of an endogenous mediator). Also, the problem with "causal direction" does not have to do with finding a coefficient of, say -.30, while not knowing whether this effect captures how *x* influences *y* or how *y* influences *x*; that is not the point. If the regressor is endogenous, this coefficient capturing the true effect of *x* on *y* or of *y* on *x* could be higher, lower, zero or of a different sign!

As it has become clear in the introduction, establishing the true (causal) relationship between two (or more) variables is not a simple matter. We will show how one can establish the true relation even when the regressor is endogenous. We first discuss the workings of the experiment and how it eliminates endogeneity by design (i.e., by manipulating the regressors); then we will show methods can be used to recover causal estimates even if the regressor has not been manipulated.

Counterfactuals and the Randomized Experiment

The counterfactual argument

The counterfactual argument is at the heart of the experimental design and serves as a main foundation of causal analysis. Let us consider a simple experiment where individuals, in a treatment or control group (captured by the dichotomous variable *x*), are measured on an observed variable *y*. Assuming that *x* preceded *y* temporally, and that *x* and *y* are significantly correlated beyond chance, how could we establish that *x* has a *causal* effect on *y*? In other words, how can we rule out alternate explanations as to why *x* could affect *y*? In order to do so, we need to consider either one of the two possible counterfactual conditionals—only in this way can causality be determined (Morgan & Winship, 2007; Rubin, 1974; Winship & Morgan, 1999).

If we consider the situation from the standpoint of the individuals in the treatment group, the counterfactual conditional would ask, "what would we have observed on *y* for the individuals in the treatment group had they not received the treatment?" Alternatively, if we first consider individuals in the control group, the counterfactual conditional would ask, "what would we have observed on *y* for the individuals in the control group had they received the treatment?" Comparing two given states of the world (i.e., what currently is vs. the counterfactual condition) allows us to establish causality. This is precisely what is done in the randomized experiment, which is achieved by randomizing participants to treatment, which is a failsafe way to eliminate

endogeneity.

Kerlinger & Lee (2000) refer to the laboratory experiments as one of the greatest inventions in history because of its ability to precisely identify and test causal relationships in uncontamined conditions. The randomized experiment establishes causality through the counterfactual argument. By randomly allocating participants to treatment and control groups, the experimental design ensures that both groups of individuals are similar on all (observable and unobservable) characteristics. Thus, each group serves as the counterfactual conditional for the other and consequently, the causal effect of the experimental treatment can be observed as the difference between the treatment and control groups on the dependent variable.

Let us focus on how statistical analysis of experimental data produces causal *estimates*, which brings to light how the problem of endogeneity and why nonexperimental data could support causal claims. We here examine the Ordinary Least Squares (OLS) estimator—the estimator commonly used in regression (or ANOVA models)—which derives estimates by reducing the sum of squared residuals (hence its name) between observed and predicted values. We use a simple model in which two groups (i.e., an experimental and a control group) are measured on a dependent variable *y*. We use a dummy variable *x*, which is 1 if an individual receives the treatment and to 0 otherwise, to model the experimental effect. For this example, we also assume that participants are pre-measured on *z*, indicating participant sex (female = 1, else 0), which is a predictor of *y*. The inclusion of the covariate *z* serves to increase statistical power and should consequently make it easier for the researcher to identify the effect of *x* on *y* (Keppel & Wickens, 2004; Maxwell, Cole, Arvey, & Salas, 1991). In addition, the covariate may correct for small differences remaining between the control and treatment groups despite randomization (Shadish, Cook, & Campbell, 2002); including covariates (as in an ANCOVA) design is thus a very good idea, particularly if the sample will not be very large. We thus estimate:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \quad \text{Eq. 6}$$

Important to note is that the error (or disturbance) term *e* captures all unobserved sources of variance in *y* along with any other sources of error (such as measurement error for example). To avoid any confusion, note that the error or disturbance term is not the same as the residual term. The error term refers to all unobserved and unmodelled sources of variance in *y*, whereas the residual term is the difference between the predicted and observed values of *y* (and the OLS estimator is concerned with minimizing the sum of the squares of these differences). By design, the residual term is orthogonal to the independent variables, which is not necessarily the case with the error term.

There is a key assumption made by the OLS estimator which is central to understanding how and when causal analysis is possible with non-experimental data. The OLS estimator assumes that the error term *e* is uncorrelated with any of the independent variables. If we only consider *x*, the manipulated variable in our example, OLS assumes that *e* is uncorrelated to *x*, which is the same to say that *e* and *x* are orthogonal, or that *x* is *exogenous*. This brings us back to the problem of endogeneity, which refers precisely to the situation where *x* and *e* correlate. In the randomized experiment, *x* and *e* are uncorrelated by design, because of two conditions: (a) the researcher has total control over *x* and (b) because participants are randomly assigned to conditions. Because factors that explain *y* beyond the treatment, *e*, are unrelated with *x* (or other covariates) estimates are consistent. If these two conditions are not met, it will be very likely that *x* and *e* will correlate.

Suppose an organization wants to assess the effectiveness of a leadership training it offers to its employees. In order to do so, participants are randomly assigned either to the leadership training or to a control condition (who receive no training), and are measured on leader prototypicality at the outcome of the training. We are keeping the problem simple with two

groups; however, we may have alternative treatments and can also cross designs as in the typical 2x2 ANOVA. Of course, at the outset of the experiment some participants will possess characteristics that make them more prototypical of leaders (e.g., being more charismatic). But because of random assignment, the proportion of participants high (or low) on these characteristics will be roughly the same in both groups. Therefore, any difference we observe between groups on leader prototypicality at the end of the experiment can only be attributed to the experimental manipulation.

Following on the same example, let us imagine now that rather than randomly assigning participants to treatment and control conditions, the organization compares managers from one division (Division 1) who were chosen to complete the training, to a group of managers from another division (Division 2) who do not attend training. What if participants who were chosen to do the leadership training program differ from those participants who did not attending the training on some characteristics? For example, Division 1 might spend a lot of time in carefully choosing who they promote to positions of leadership (i.e., they use 360-degree ratings to take the best leaders); however, suppose that Division 2 does not have these mechanisms (and instead only the division boss is the person who appoints leaders based on production figures). If the characteristics on which Division 1 and 2 participants predict leader prototypicality, then the effect of the treatment is confounded. In other words, the treatment will correlate with the error term. Why? Because some of the characteristics (e.g., leadership styles) are higher in the treatment group (Division 1 leaders) and these factors correlate with *y* too; thus, because these factors were not randomly assigned and they are omitted from the model, their effects are pooled into the error term, which will correlate with *x* and induce endogeneity.

Such conditions violate the orthogonality assumption (of *e* with *x*) of the OLS estimate and also of the maximum likelihood estimator. As a consequence, the estimator, in an attempt to

satisfy this assumption, will "adjust" the estimate of the problematic (i.e., endogenous) variable. The estimate of the endogenous variable will become inconsistent meaning that it will not converge to the true population parameters as sample size increases. The estimate is therefore useless; furthermore, the endogenous variable will also render inconsistent all other variables in the model with which it correlates even if these are not endogenous (refer to Figure 1 for a graphic representation).

[Figure 1 here]

Whereas the randomized experiment is the failsafe design to test theoretical propositions, it may not always be possible for researchers to implement a randomized experiment—be it for practical or ethical reasons. Also, randomized experiments typically concern small and quite specific populations thus limiting their external validity. Consequently, researchers must often rely on non-experimental data to test their theories. In non-experimental settings, scientists neither have direct *control* over independent variables, nor do they have the possibly to use random assignment—the two elements which allow to make causal claims with the experimental method. Thus, the important question to ask is: How can causal claims be made on the basis of non-experimental data? In order to answer this question, we must understand the causes of inconsistent estimates.

The Pitfalls of Non-Experimental Research

Endogeneity and can stem from a plethora of situations wherein a regressor *x*, correlates with the model's error term *e* (thus violating one of the underlying assumptions of OLS of maximum likelihood). Below we present conditions that potentially cause endogeneity and threaten estimate validity. We also cover an additional area researchers should be wary of when testing theoretical models which is concerned, not with the consistency of estimates, but with the consistency of inferences (i.e., the validity of the standard errors). We also discuss proper model specification in

the context of simultaneous equations models. Following Antonakis et al. (2010) we use some basic algebra to show how endogeneity is engendered particularly for the case of an omitted regressor and common-method variance (which we discuss in more detail). The advantage of using algebra is that we show very specifically how *x* correlates with the error term. We cover the rest of the threats briefly; readers can refer to Antonakis et al. (2010) for further details.

Omitting a regressor

Consider the following model in which each individual *i* is measured on a dependent variable *y*, and on two independent variables *x* and *z*:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \quad \text{Eq. 7}$$

The assumption of the estimator is that *x* and *z* are exogenous; in other words, that they are not predicted by the workings of this specific model. Thus, neither *x* nor *z* should correlate with any of the unobserved sources of variances in *y* (i.e., they do not correlate with *e*). Now, suppose a researcher is interested in understanding whether *x*, one's ability to wait before obtaining a desired outcome (i.e., delayed gratification), predicts leader effectiveness (*y*). Because delayed gratification is correlated in part to cognitive ability *z*, and because cognitive ability predicts leader effectiveness, the researcher must control for *z*. If the researcher fails to do so, the estimate of *x* will be biased because *x* will correlates with *e*. This can be easily seen in the following equations. This is the misspecified model omitting *z*:

$$y_i = \varphi_0 + \varphi_1 x_i + v_i \quad \text{Eq. 8}$$

If *z* and *x* correlated (irrespective of the direction), then we can note that:

$$z_i = \gamma_1 x_i + u_i \quad \text{Eq. 9}$$

The endogeneity is evident when substituting Eq. 9 into Eq. 7:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2(\gamma_1 x_i + u_i) + e_i, \quad \text{Eq. 10}$$

Multiplying out gives (notice, the error term v_i, which is the error term of Eq. 8)*:*

$$y_i = \beta_0 + \beta_1 x_i + (\underbrace{\beta_2\gamma_1 x_i + \beta_2 u_i + e_i}_{v_i}) \quad \text{Eq. 11}$$

Or, rearranging as a function of *x* gives

$$y_i = \beta_0 + (\beta_1 + \beta_2\gamma_1)x_i + (\beta_2 u_i + e_i) \quad \text{Eq. 12}$$

In the presence of endogeneity, one does not estimate $\beta_1$of Eq. 7, but $\varphi_1$in Eq. 8; these two estimates will not be equal except under two conditions: If (a) $\beta_2 = 0$ or if (b) $\gamma_1 = 0$. In these cases, then v_i reduces to e_i and there is no omitted variable when excluding *z* from the model. Also, whether $\varphi_1$is increases or decreases when excluding *z* will depend on the signs and magnitudes of β_2 and γ_1.

Given the consequences of omitting a variable, when in doubt about whether a given variable should be included or not in a model, it is always best to stay on the safe side by including this additional variable (Cameron & Trivedi, 2005); this is not the advice that management methodologists usually provide (e.g., Spector & Brannick, 2011). Indeed, the cost of including additional variables is higher standard errors (i.e., reduced efficiency); if the sample is large enough to detect significant effects then this is a small cost to pay. If there is a choice to be made, we will always prefer consistency to efficiency. What is the value of inconsistent estimates having precise standard errors?

How does a researcher determine whether there are omitted variables? There is only one test to examine whether polynomial terms are omitted from the model. This is called Ramsey's (1969) regression-error-specification (RESET); however, this test *cannot* determine whether there are other types of omitted variables. Thus, the most important guide is "theory, theory, and more theory" (Antonakis & Dietz, 2011, p. 218); there are no direct tests to determine whether

there is an omitted variable, which could be a main effect or an interaction effect too in a particular model (apart from comparing random-effects to fixed-effects estimators, as discussed later). One way to suggest that there are omitted variables is to compare the target model to a model that is known to be consistent (e.g., from an instrumental-variable model, discussed later). The parameter(s) of interest is tested using a Hausman (1978) test. This test is a very versatile test that can be used to compare estimators. Basically, the tests shows that if an estimate (from the efficient but not consistent model) is different from that of the consistent model, this difference must come from the fact that the variable correlated with the disturbance in the efficient model.

In its simplest form, the Hausman test may be computed for one parameter, where δ is the element of β being tested (Wooldrige, 2002); the efficient estimate is compared to the consistent estimate using a *t* test that follows an asymptotic standard normal distribution. Note too that the Hausman test for one parameter is also useful in situations where the test for an overall model is not defined and the researcher is concerned about whether a specific variable may be endogenous (note that Hausman tests can be conducted using other ways, e.g., Wald tests in the context of "seemingly unrelated" regression models). The formula for the one-parameter test is:

$$z = \frac{(\hat{\delta}_{CONSISTENT} - \hat{\delta}_{EFFICIENT})}{\sqrt{SE(\hat{\delta}_{CONSISTENT})^2 - SE(\hat{\delta}_{EFFICIENT})^2}}$$

We discussed the basic case of omitted variable bias in depth using two examples (i.e., the "inconvenient demonstrations") and with simple algebra. We will now briefly discuss other types of endogeneity bias. For in-depth discussion and detailed explanation regarding these forms of endogeneity refer to Antonakis et al. (2010).

Measurement error (errors-in-variables)

Measurement error is a common problem in leadership research, yet it remains largely

unaddressed (with the exception of researchers using structural equation models which correct for measurement error as we will discuss in the following section of this chapter; however, these researchers make another critical error by ignoring the overidentification test). There are many examples of how estimates can be severely compromised by measurement error (for demonstrations see Fiori & Antonakis, 2011; Schulte, Ree, & Carretta, 2004; von Wittich & Antonakis, 2011).

Many constructs of interest in social sciences cannot be perfectly observed; consequently measurement of these constructs includes some degree of measurement error. For example, imagine that we want to measure the intelligence of leaders. Intelligence in a "pure" theoretical form, which we will call *x**, cannot be directly observed. Rather, what we observe is *x* which consists of the *x**, the pure construct, and an error term *u* reflecting measurement error (see Cameron & Trivedi, 2005; Maddala, 1977). So, if our goal is to understand the relationship between follower motivation (*y*) and leader intelligence *x**, we would consider the following model:

$$y_i = \beta_0 + \beta_1 x_i^* + e_i \quad \text{Eq. 13}$$

We cannot directly observe *x**, but what we observe is a proxy of *x**, *x* as follows:

$$x_i = x_i^* + u_i, \text{ or } x_i^* = x_i - u_i \quad \text{Eq. 14}$$

So, substituting Eq. 14 in Eq. 13 gives:

$$y_i = \beta_0 + \beta_1 (x_i - u_i) + e_i \quad \text{Eq. 15}$$

Which is equivalent to the following model:

$$y_i = \beta_0 + \beta_1 x_i + (e_i - \beta_1 u_i) \quad \text{Eq. 16}$$

As it is clear from the above equation, the rearranged error term correlates with *x* and therefore the estimates of the effect of *x* will be inconsistent. Thus, we see that if we do not explicitly

model *u*, we create endogeneity by omitting a source of variance in *x* and *y*. This omission results in an attenuated estimate of the effect of *x*. As with the omitted variable bias, measurement error in *x* will affect all variables correlated with the problematic variable (Bollen 1989; Kennedy 2003).

Measurement error can easily be modeled by constraining the residual variance of *x* to (1-reliability$_x$) * Variance$_x$ (Bollen, 1989). Estimate of the reliability of a measure can be obtained by, for example, using the test-retest reliability or Cronbach's alpha (which is a lower-bound correction). Alternatively, if reliability is not known, estimates can be derived theoretically to constrain the residual (Hayduk, 1996; Onyskiw & Hayduk, 2001).

In terms of technical implementation, measurement error can be modeled in a regression using, for example, the least-squares eivreg (errors-in-variables) command in Stata; one could also use maximum likelihood estimation in a structural equation modeling program. When measurement error concerns a measures with a single indicator, the eivreg routine should be chosen over structural equation modeling solutions as it much less restrictive in terms of assumptions and sample size (e. g., see Bollen, 1996; Draper & Smith, 1998; Kmenta, 1986). Structural equation modeling is the method of choice for treating measurement error in latent constructs with multiple indicators. Though in practice--and assuming item indicators are valid measures of the construct--if one uses a parcel (i.e., average of indicators) and models this as one indicator of a latent variable, structural estimates will be similar to those obtained from a full specification (e.g., Bandalos & Finney, 2001; Hall, Snell, & Singer Foust, 1999; Liang, Lawrence, Bennett, & Whitelaw, 1990). The instrumental variable method we discuss later also provides a solution to measurement error bias.

Common source, common method variance

Another cause for inconsistent estimates is common method variance; (cf. Podsakoff,

Mackenzie, & Podsakoff, 2010); this problem is related to measurement error. Common method variance refers to the situation where the relationship between *x* and *y* is dependent on a third variable *q*. At best, researchers acknowledge that common method variance can bias estimates, but with the erroneous assumption that estimates can only be biased upwards. At worse, some researchers suggest that this bias is exaggerated (e.g., Spector, 2006); unfortunately, it is not possible to know how exaggerated the bias is unless the correct procedures are used (e.g., instrumental-variable regression).

A prevalent example of common method variance is one in which subordinates are asked to provide ratings on independent and dependent measures on their leaders (there are 50 leaders in this sample)—for example, ratings of leader prototypicality (*x*) and ratings of leader ethical behavior (*y*). In this situation, subordinates will seek to maintain cognitive consistency between both ratings (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Podsakoff & Organ, 1986), which may be driven by a third variable(s) *q* (e.g., affect for the leader, knowledge of the effectiveness of the leader, and other biases). Assume we collected measures on leader_j from follower_i, in a model in which we control for leader fixed-effects too using $k - 1$ leader dummy variables (i.e., 49 dummy variables, refer to the section on fixed-effects below):

$$y_{ij}^{*} = \beta_0 + \beta_1 x_{ij}^{*} + \sum_{k=2}^{50} \beta_k D_{kj} + e_{ij} \qquad \text{Eq. 17}$$

As with measurement error, we do not directly observe *y** or *x**, rather we observe *y* and *x*, which can be modeled as a function of *q* and *y**:

$$y_{ij} = y_{ij}^{*} + \gamma_y q_{ij} \qquad \text{Eq. 18}$$

$$x_{ij} = x_{ij}^{*} + \gamma_x q_{ij} \qquad \text{Eq. 19}$$

The two later equations can be rearranged as follows:

$$y_{ij}^{*} = y_{ij} - \gamma_y q_{ij} \quad \text{Eq. 20}$$

$$x_{ij}^{*} = x_{ij} - \gamma_x q_{ij} \quad \text{Eq. 21}$$

We can substitute *y** and *x** in Eq. 17, which gives:

$$(y_{ij} - \gamma_y q_{ij}) = \beta_0 + \beta_1 (x_{ij} - \gamma_x q_{ij}) + \sum_{k=2}^{50} \beta_k D_{jk} + e_{ij} \quad \text{Eq. 22}$$

This equation can be rearranged to obtain:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{k=2}^{50} \beta_k D_{jk} + (e_{ij} - \beta_1 \gamma_x q_{ij} + \gamma_y q_{ij}) \quad \text{Eq. 23}$$

We now see that the expanded and rearranged error term correlates with *x*. Once again, this results in an inconsistent estimate of the effect of *x* (and of all covariates correlating with *x*). The resulting bias may cause *inflated* or *attenuated* estimates (and cannot be eliminated with fixed-effects estimation, see next section).

A common (but incorrect) solution to the common source bias is that common source bias can be eliminated by including an unmeasured latent method factor in the model (Podsakoff, et al., 2003). In order to work, this solution would require the researcher to know how the variables are affected by the unmeasured cause—which is not possible (cf. Antonakis, et al., 2010; Richardson, Simmering, & Sturman, 2009). Furthermore, simulations shown that this solution cannot recover correct model estimates (cf. Antonakis, et al., 2010; Richardson, et al., 2009).

Several solutions to the common source bias have been proposed (cf. Antonakis, et al., 2010; Podsakoff, et al., 2010; Richardson, et al., 2009). The most intuitive solution is to gather data on *q*; however, this solution is not practical because the researchers must know of all sources of *q*. Researchers could gather independent and dependent measures from difference sources (i.e., "objective" or hard measures of leader performance like profits). The independent variables must

of course be exogenous otherwise there will still be endogeneity in the model. Another solution is to use a split-sample design (e.g., Koh, Steers, & Terborg, 1995) where one half of the sample is used for ratings on the dependent measure and the other half is used for the independent measures; this solution is not ideal because with the split-sample design only half of the data is used and therefore the estimates of standard errors will be less precise (i.e., efficiency is reduced) and estimates will be less precise too given that fewer raters are used (Mount & Scullen, 2001; Scullen, Mount, & Goff, 2000). We later discuss another solution (i.e., instrumental variables models).

Omitting fixed effects

Researchers often have data on entities that are repeatedly measured over time (i.e., a longitudinal panel); data might also be hierarchically nested (i.e., a hierarchical or pseudo-panel) where entities under higher-level units are measured, as for example, companies nested under countries, leaders nested under companies, team members nested under leaders (cf. Liden & Antonakis, 2009). In either case, what we have are observations (Level 1) nested either in time or in a higher level entity of sorts (Level 2). Thus, with panel data, it is possible that Level 2 "fixed-effects" drive a part of the variance in the dependent variable and also correlate with other regressors. For example, when we observe leaders nested in organizations, leaders within organizations would share certain characteristics (e.g., as result of firm recruiting policies for example), which may affect the modeled variables. Thus, firm-level factors may predict performance outcomes; however, they may also correlate with leader-level characteristics (which were used by some firms to select leaders). Under the proviso that these Level 2 fixed-effects have an effect on the dependant variable and correlate with leader level (Level 1) characteristics, they will be pooled in the error term along with all other unmodeled sources of variance if they are not explicitly modeled; in this way, estimates of model become inconsistent (Cameron &

Trivedi, 2005; Wooldridge, 2002).

How can this situation be avoided? The easy solution is to explicitly model these fixed-effects by using *k*-1 company dummy variables. The difficult solution, which leads to what we call here the "*HLM problem*," is to include Level 2 predictors (i.e., company level regressors like firm size, etc.) and estimate the model using a random-effects estimator (e.g., HLM). The problem here is what if all sources of Level 2 variance are not included? If an important Level 2 variable is omitted, then endogeneity is engendered. This endogeneity can be tested for using a Hausman test (1978); that is, the random-effects estimates (efficient) are compared to the fixed-effects estimates (consistent). If there is a significant difference, it means that the efficient estimator is not consistent and must be rejected. If there is not a significant difference, the efficient estimator is not rejected.

Unfortunately, this point *is still not understood* by those who estimate HLM-type models (Antonakis, et al., 2010; Halaby, 2004); omitting fixed-effects is a major problem that must be taken seriously by researchers estimating HLM models. The only way to ensure that Level 2 fixed-effects are included is to model the Level 2 dummy variables. However, doing so precludes modeling Level 2 variables (because they will be perfectly collinear with the dummies). Researchers using HLM models can have their cake and eat it too, however. That is, it is possible to include both Level 2 fixed-effects and Level 2 variables by using the Mundlak (1978) procedure; refer to Antonakis et al. (2010) for intuitive explanations.

Omitting selection

Without random assignment, treatment is endogenous unless selection is explicitly modeled. Consider the equation below:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \qquad \text{Eq. 24}$$

Say that *x* is equal to 1 if the individual receives the treatment (i.e., a leadership-training

program), or is equal to 0 if the individual is in the control condition. The dependent variable, *y* is how prototypical of a leader the individual is and *z* is a dummy variable indicating participant sex (female = 1). Assume now that individuals have been self selected to the conditions. Because of this selection, both groups will differ on a number of characteristics on the outset (recall that they would have been roughly equivalent had *x* been randomly assigned). This is problematic because differences between both groups may correlate with the dependent variable causing *x* to correlate with *e*. Assume that the selection *x** can be modeled in the following probit (or logit) equation (Cong & Drukker, 2001):

$$x_i^* = \gamma_0 + \sum_{k=1}^{q} \gamma_k d_{kj} + u_i \qquad \text{Eq. 25}$$

Where we have *k* regressors and a disturbance term *u*. Individuals are selected (i.e., *x* = 1) if *x** > 0. The problem of omitted selection arises because *u* will be correlated with *e* (called $\rho_{e,u}$) and as a result of which *x* will be correlated with *e*.

For example, it is possible that individuals with high levels of extraversion are more likely to self-select to the leadership training, and it is also possible that because of their extraversion these individuals are perceived as more leaderlike than their counterparts. Therefore, we are here again in a situation where an unmodeled source of variance (pooled in *e*) correlates with *x*, thus creating endogeneity and yielding inconsistent parameters (Kennedy, 2003). The only way causality can be assessed with non-random assignment of participants to conditions is by explicitly modeling the selection process so as to create a clean counterfactual (Cong & Drukker, 2001; Maddala, 1983).

Researchers must not only be cautious of the selection process to treatment and control conditions, but also of the selection of their samples. Indeed, non-representative, or censored, samples will results in inconsistent estimates. For example, studying the effect of cognitive

abilities on leader effectiveness will produce misleading results in that there is little variance on cognitive abilities in the study sampled (e.g., because participants are all highly intelligent). In this later example, the researcher will find attenuated estimates of the effect of cognitive abilities on leader effectiveness. In the case where participants are non-randomly selected (either self-selected or selected on another basis), the researcher should ensure that participants are representative of the general population on relevant factors. If this is not the case, estimates could be misleading. Take for example the situation where leader performance ratings are obtained from followers who have been selected by the leader to provide these ratings. We can expect that the leader will select followers who are most likely to give positive feedback. If this is the case, again, selection must be explicitly modeled.

A final example of the problem of omitted selection is samples where a certain range of data is missing on the dependant variable (i.e., the dependent variable is censored). It is possible to deal with such problems by using censored regression models (Tobin, 1958) or truncated regression models or sorts (Long & Freese, 2006).

Simultaneity

The problem of simultaneity, although quite simple to understand, can be quite troublesome for researchers. Simultaneity happens when two variables *simultaneously* affect each other (hence the name). Note that this is different to what researchers sometimes name "backward causality" which is when the estimated effect of *x* on *y* is proposed to be caused by *y* affecting *x.*

A good example of simultaneous causality is the relation between levels of crime and number of police officers discussed by Steven Levitt (see Levitt, 1997; Levitt, 2002). The standard expectation is that that hiring more police officers will reduce crime. Thus, we would expect the estimate of the relationship between the number of police officers and crime to be negative. However, crime can also affect the number of police officers. Indeed, a response to

rising crime levels might be to increase the police force. Because of this simultaneous relation, the number of police officers will be endogenous. Such a type of simultaneity can be evident in leadership research too (e.g., a leader style of leading could depend on follower performance). Refer to Antonakis et al. (2010) for further discussion.

Consistency of inference

Up to now, we have discussed consistency only with regards to the consistency of estimates. However, consistency of standard errors (i.e., the consistency of inference) is also important but often overlooked. Work on this topic stems from the of Huber (1967) and White (1980).

Consistent standard errors can be derived from OLS estimation under the assumption that regression residuals are identically and independently distributed (or, simply, i.i.d.). This assumption regarding residuals is two-fold. Firstly, residuals are assumed to be homoskedastic (i.e., identically distributed), in other words, they are assumed to have been drawn from the same population and have a uniform variance. Secondly, residuals are assumed to be neither clustered (nested under a higher level entity) nor serially correlated (i.e., they are assumed to be independently distributed).

It is noteworthy that non i.i.d. residuals only affect the consistency of the standard errors and not the consistency of estimates. This problem is nevertheless a serious threat to validity because in presence of heteroskedasticity, standard errors will be biased and *p*-values will be either under or overstated. Thus, conclusions about the significance of parameter estimates will be wrong. The assumption of i.i.d. residuals, or lack thereof, can be tested using a number of tests readily available in programs akin to Stata.

If the homoskedasticity assumption is violated, variance has to be estimated using a variance estimator based on works of Huber and White. Using this estimator is akin to saying that

Huber-White standard errors have been used, or that that model has been estimated using sandwiched standard errors—or, robust standard errors (i.e., robust to heteroskedasticity). Alternatively, consistent standard errors can also be estimated using bootstrapping.

Standard errors could also become inconsistent because of clustered data, which directly violates the assumption that residuals are independently distributed. If this is case, standard errors must be cluster corrected using a specific variance estimator. Interestingly, this problem is often overlooked, and was so until recently even in economics (see Bertrand, Duflo, & Mullainathan, 2004). Note that, data may involve multiple (independent or hierarchical) dimensions of clustering which must be taken into account by researchers (Cameron, Gelbach, & Miller, 2011).

The assumption of i.i.d. residuals, or lack thereof, can be tested using a number of tests readily available in programs akin to Stata.

Quasi-Experimental and Structural Equation Methods

In this section we present methods available to researchers to test theoretical models (i.e., causal relationships) in non-experimental settings. We begin with and devote a greater part of this section to two-stage least square estimation. We then briefly discuss other methods too; for futher details refer to Antonakis et al. (2010).

Two-stage least square estimation

The two-stage least squares (2SLS), or instrumental-variable estimation, allows for *consistent* estimation of simultaneous equation with endogenous predictors. We have made reference to this method several times throughout this chapter as a means to treat endogeneity; 2SLS is one of the most potent and versatile tools available in this regard. This reason undoubtedly explains why this method is the workhorse of econometrics. Unfortunately, this method is scarcely used in other social sciences (see Cameron & Trivedi, 2005; Foster & McLanahan, 1996; Gennetian, et al., 2008). We hope that in the future researchers will reap the

benefits of this method: It truly is a cure to endogeneity resulting from omitted variables, measurement error, simultaneity, and common method bias (Cameron & Trivedi, 2005; Greene, 2008; Kennedy, 2003)! This seems almost too good to be true; but, 2SLS really is a clean and elegant way to purge models of endogeneity.

How does the 2SLS estimator correct for endogeneity? Recall that inconsistent estimates result from a regressor (*x*) correlating with the model's error term (*e*). If *x* and *e* did not correlate, we would obtain consistent estimates. This is precisely what the 2SLS estimation does: It removes the portion of variance in *x* that correlates with *e*. To do so, the 2SLS estimator relies on instrumental variables, which are exogenous regressors of the problematic (endogenous) variable. By definition, the instruments are uncorrelated with *e*, and therefore, they can be used in a first estimation stage to obtain predicted values of the endogenous variable that will be uncorrelated with *e* (for ideas about where to find instrumental variables refer to Antonakis et al., 2010). These predicted values can then be used in a second stage to predict the dependent variable. In essence, the instrumental variables purge the endogenous variable from variance that overlaps with the error term. In this way consistent estimates of the endogenous variable are obtained. However, consistency here comes at the cost efficiency, which is reduced given that less of the available information is used.

Note, endogeneity can also arise in an experiment where the causal effect of one dependent variable on another is estimated (i.e., the effect of *y1* on *y2*). If the 2SLS estimator is not used, then what is estimated will be biased. Thus, the 2SLS estimator is also useful to test mediation models (in the context of experiments or otherwise) to identify the causal effect of one endogenous regressor on another (stemming from the instrument/s). However, to estimate such models correctly one cannot use the simple mediation approaches that are popular in management and applied psychology research, that is, *x* → *y1* → *y2* (no matter how much standard errors are

bootstrapped). The problem is not necessarily with the standard errors of the indirect effect of *x* on *y2*; the problem has to do with the estimates and acknowledging that the mediator is endogenous (which is done by correlating the cross-equation disturbances of the endogenous variables: this is the 2SLS estimator). Refer to Figure 2 for a graphic depiction of this estimator. Failing to model the causal system correctly gives the same incorrect estimate that OLS gives.

[Figure 2 here]

In order to derive consistent estimates, the researcher needs to identify instruments to predict the endogenous variable *x*. There should be at least as many instruments as there are endogenous variables (this constitutes the order condition), although it would be desirable to have more instruments than endogenous variables to test overidentifying restrictions, as we will discuss later. Instruments should be significant and strong predictors of *x*, and only predict *y* through their effect on *x*; of course, instruments must also be uncorrelated with the model's error term—recall instruments must be exogenous. This condition, along with excluding one instrument from the second stage of the estimation, constitutes the rank condition. Only if both the order and rank conditions are satisfied can parameters be identified (Wooldridge, 2002). Note too that the instruments *may* correlate with *y* (in Antonakis et al., 2010, p. 1101, we had stated "the instruments must be related to *y*"; to be clear, the "*must*" is relevant in the case that *x* is truly a predictor of *y*--if *x* is not a true predictor of *y*, then the instruments need not correlate with *y* to be valid instruments).

The more instruments are included in a model, the more information will be used to obtain the predicted values of *x* (i.e., $\hat{x}$). It is consequently desirable to include all available exogenous variables as instruments. Also, all predictors must be used as instruments (even if they only are theorized to predict *y*) in case they correlate with first stage instruments; this can also avoid certain pitfalls which would otherwise results in inconsistent estimates (see Baltagi, 2002,

for more information).

As we explained earlier, consistency is ensured through the 2SLS estimation procedure, because only the "clean" (i.e., uncorrelated with the model error term) portion of *x* with is used to predict *y*. This clean portion is obtained by predicting *x* from the exogenous regressor which are uncorrelated with the model error term (see Kennedy, 2003). Furthermore, consistency can only be ensured when cross-equation error terms are correlated (see figure 2). Not estimating this correlation is akin to assuming that *x* is exogenous and will produce the same estimates as OLS (Maddala, 1977)—which will course be inconsistent (unless *x* is exogenous). A Hausman test can be conducted to compare the (consistent) 2SLS estimates with the (efficient) OLS estimates obtained without instrumenting. A significant difference indicates that the estimates obtained through OLS are inconsistent and that *x* should be instrumented.

The much valued consistency from 2SLS estimation comes, however, at the cost of lesser efficiency. Estimation by 2SLS is particularly imprecise if instruments only weakly predict the endogenous regressor *x*. Moreover, inference with weak instruments can be seriously biased (Bound, Jaeger, & Baker, 1995). Weak instruments can easily be detected. Instruments are weak if the *F*-test for joint significance of instruments falls below a rule of thumb threshold of 10. Stock and Yogo (2002) present exact threshold values and extensions for multi-variate models (see also Stock & Watson, 2007; Stock, Wright, & Yogo, 2002).

For increased efficiency, the researcher can use a three-stage least square (3SLS) estimator (e.g., Zellner & Theil, 1962) or a maximum likelihood estimator (Baltagi, 2002; Bollen, 1996; Bollen, Kirby, Curran, Paxton, & Chen, 2007). These estimators produce more precise standard errors because they are full information estimators. However, before retaining full information estimates, the researcher must make sure these are consistent. A Hausman test can be used to compare the estimates from the efficient (e.g., 3SLS) and the consistent (i.e., 2SLS)

estimators. A significant difference indicates that the consistent estimator must be retained.

The 2SLS estimator can also be used in the context of Structural Equations models too. The same logic applies as above. For more details refer to Antonakis et al. (2010).

Two-stage least squares in practice: An example

Turning back to the LMX example (from the data generated by Eq. 3 and 4, as presented in Table 3), to determine whether there is a causal effect of LMX on turnover, we should estimate a model where LMX is first purged from endogeneity bias by using the instruments to predict it (first stage model); then in the second stage, the predicted value of LMX is used as a regressor of turnover. That is, in the general form we estimate:

$$LMX = \lambda_0 + \lambda_1 L\_Extra + \lambda_2 F\_IQ + \lambda_3 F\_Consc + controls + \varpi \quad \text{Eq. 26}$$

$$Turnover = \mu_0 - \mu_1 LMX + controls + \xi \quad \text{Eq. 27}$$

Refer to Table 6 for model estimates. In Model 1 we estimated the system of equations using OLS, where the cross equation disturbances are not correlated; using maximum likelihood, where the correlation between the disturbances is constrained to be zero would give the same result (i.e., refer to Figure 2).

[Table 6]

The estimate of LMX on turnover is -.43, $p < .001$. Given the causal structure of the model that generated the data, this estimate is wrong. Thus, an important lesson to retain here is that simple estimating system of equations where the potentially problematic regressor (LMX) is modeled as an outcome of antecedents will not produce the correct estimates in predicting *y* if: (a) the problematic regressor is endogenous and (b) the 2SLS estimator is not used (i.e., the cross-equation disturbances are constrained to be zero, which is what the OLS estimator does). If, and only if, LMX is not endogenous would OLS estimates be consistent.

Model 2 is like Model 1 except that the 2SLS estimator is used; LMX is predicted in the

first-stage equation by Leader Extraversion, Follower IQ and Follower Conscientiousness and turnover is regressed on the predicted value of LMX. Here it is clear that LMX does not predict turnover. Models 3 and 4 provide similar estimates. Interesting to note is that even including omitted predictors of turnover does not significantly change the estimates of the relation between LMX to turnover. Of course, in this case, the data were simulated such that there was no relation between LMX and turnover (which is evident too from the fact that the instruments overlap very little with turnover too). For other examples where the sign of the endogenous regressor is flipped when instrumented see Antonakis et al., (2010).

A note on overidentification

When estimating simultaneous equation models, like that in the previous examples, and also with more complex path models or Structural Equation Models (SEM), it is important that the overidentifying restrictions hold. That is, are the constraints that are made valid? If the answer is "no" the model parameters are suspect and cannot be trusted (because it suggests that the exogenous variables correlate with the disturbance/s of the endogenous variable/s). This point is very important to understand, and we will show below why researchers must pay attention to the overidentification statistic (and not dismiss it as is often the case).

Important to note is that if the model is correctly specified this statistic will not be significant, no matter how large the sample size (Bollen, 1990; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; McIntosh, 2007). If researchers do not trust the chi-square test then they should not trust tests of parameter estimates either (as we will show below, the chi-square test uses the same statistical theory that researchers use to estimate parameters and their significance levels).

In the case of the previous 2SLS examples, we will use Model 2 in Table 6 to demonstrate how the overidentification statistic is derived. Notice that there are three instruments that predict

LMX (Leader Extraversion, Follower IQ, and Follower conscientiousness); however, they are excluded from the turnover equation wherein only LMX is the predictor. By being excluded from the 2nd stage equation, the constraints that are made are that the relationships of the instruments with *y* are zero.

In this system of equations we have 5 variables. Thus, the variance-covariance matrix has v(v + 1)/2 bits of information (variance-covariance matrix) or 5(5 + 1)/2 = 15. We estimated the following parameters:

Regression coefficients	4
Correlations (between exogenous variables):	3
Correlations (between disturbances):	1
Variances (of endogenous variables):	2
Variances (of exogenous variables):	3
Total:	13

The overidentification test has 15-13 = 2 degrees of freedom (DF). For simple mediation models the DF are the number of instruments less mediators (i.e., endogenous variables). The overidentification statistic of interest is the Sargan chi-square statistic, which is also often called the Hansen-Sargan statistic or the J-Test (Hansen, 1982; Sargan, 1958). This test is a direct analog to the chi-square test of fit commonly used in structural equation modeling programs. What this test examines is whether the residuals of the turnover equation correlate with the instruments. If they do, it suggests that the model is misspecified (because there is systematic variance in the residuals that can be predicted by the instruments). Thus, parameters estimates are biased and cannot be trusted.

Using Stata's overidentification routine shows that the test is insignificant, $\chi^2(2) = 1.47$, *p*

= .48. To do this test manually one would first generate model predicted values of turnover ($\hat{t}$) from the second stage equation. That is, $\hat{t}$ = 44.46 + .02 * LMX (note, we have rounded the values of the intercept and slope, which originally were 44.4598 and .0184804 respectively). The residuals are simply calculated, for each observation, as turnover - $\hat{t}$ (i.e., the residual is the observed value of turnover minus the predicted value of turnover for each observation). Next, the residuals are regressed on the three instruments (Leader Extraversion, Follower IQ, and Follower conscientiousness); the Sargan statistic is simply $N*R^2$ (where N is the sample size and R^2 is the r-square from this regression model). This statistic is distributed as a chi-square statistic with DF as described above. The r-square is .0014733, which when multiplied by the N size (1,000), gives 1.4732984. At 2 DF, the p-value of this statistic is .47871531, which is precisely what the Stata program gave. Estimating the model with the Structural Equation Modeling program MPlus gives a chi-square value of 1.473 and a p-value of .4788. Also, with more complicated models the test is an omnibus test of model fit; this test may differ when using a full information (e.g., 3SLS, limited-information maximum likelihood or maximum likelihood) versus a limited information estimator. Also, misspecifications in full-information estimators may spread bias in the model, which is why it is always good to check estimates against a limited information estimator (Baltagi, 2002; Bollen, 1996; Bollen, et al., 2007).

Now, suppose we estimated the following wrong model:

$$LMX = \kappa_0 + \kappa_1 L\_IQ + \kappa_2 L\_Incent + \kappa_3 L\_Extra + \kappa_4 F\_IQ + \Xi \quad \text{Eq. 28}$$

$$Turnover = \omega_0 - \omega_1 LMX + \omega_2 L\_Extra + \omega_3 F\_IQ + \Upsilon \quad \text{Eq. 29}$$

We know that this model is wrong, because Leader Incentives are a direct cause of Turnover (as well as of LMX). Thus, Leader Incentives should also be used as regressor in Eq. 29. Estimating this model shows that the data do not fit the model, $\chi^2(1) = 3.89$, $p < .05$. Also, the coefficient of

LMX is now -.94, $p < .001$. Given the significant chi-square test, this coefficient cannot be interpreted. Researchers in management, organizational behavior, applied psychology, MIS (and probably other disciplines) but certainly not in economics, oftentimes ignore this chi-square test of fit test and use what they called approximate indexes fit (e.g., Comparative Fit Index, CFI, and the Root Mean Squared Error of Approximation, RMSEA, among others); in the case of the above misspecified model the approximate indexes show great fit: CFI = 1.00, RMSEA = .05 (which shows that they forgive really bad models). Such researchers argue that the chi-square test is too powerful and that even minute discrepancies in the model will cause the model to be rejected by the chi-square test; thus, chi-square test should be ignored in favor of approximate fit (i.e., the thinking here is that if the model is approximately good then it is still interpretable). However, using that line of thinking would suggest that if the sample size is too large for the chi-square test, it is also too large for the tests of the model parameters; thus, tests of model parameters should be summarily dismissed as well. This attitude is ludicrous and defeatist. If this chi-square test cannot be trusted then we cannot trust any other statistic either that uses the same statistical theory.

Other Methods for Inferring Causality

Other methods can be used to obtain consistent estimates in non-experimental settings (Cook, et al., 2008; Meyer, 1995; Shadish & Cook, 1999, 2009; Shadish, et al., 2002). We briefly introduce four methods below. For further details on the below, refer to Antonakis et al. (2010).

Propensity score analysis (PSA)

This method can be used to recover causal estimates in situations where treatment has been non-randomly assigned. Had treatment been randomly assigned (into 1 of 2 conditions), the probability of receiving treatment would be 50%. By determining the probability that an individual would have received treatment based on observable factors, a counterfactual can be

recreated by comparing individuals from the treatment and control groups who have the same propensity (i.e., probability) of being assigned to the treatment condition (D'Agostino, 1998).

In order to use PSA, the researcher must be able to know which variables determine the probability that an individual would have received treatment. Furthermore, unmodeled sources of variance in determining selection should be uncorrelated with unmodeled sources of variance in the main model (Cameron & Trivedi, 2005). If this later assumption is not met, a Heckman treatment effects model should be used.

Selection models (Heckman models)

Heckman-type two-step selection model (Heckman, 1979) or *treatment effects model* (see Cong & Drukker, 2001; Maddala, 1983) are two stage models which allow to recover causal estimates in presence of non-random assignment to treatment. In these models, in the first stage the probability being selected in the treatment group is predicted from exogenous instruments. In the second stage, the main model is estimated addition a control variable capturing the difference between treatment and control group resulting from unmodeled sources of variance in the selection process. Thus, the correlation between the error term and selection is removed and consistent estimates can be obtained.

Regression discontinuity models

These designs are useful when selection to treatment is non-random but is based on a known threshold or cut-off value. The idea behind the RDD is to explicitly model selection procedure. By doing so, the RDD very closely emulates the randomized experiment (Cook, et al., 2008). As with the randomized experiment, the error term is uncorrelated with the selection variable, which results from explicitly modeling the selection process. In this way, there are no unmodeled sources of variance in the selection variable that could otherwise correlate with the model’s error term.

The RDD is easy to implement and is an excellent design to test for policy effects. Another advantage of the RDD is that it allows the researcher to oftentimes give treatment to those individuals who need it the most (e.g., in terms of training needs). Lee and Lemieux (2009) provide a comprehensive review of the RDD.

Difference-in-differences models

Differences-in-differences models compare two similar groups before and after treatment is administered. The underlying idea is that by comparing two similar groups over time, it is possible to remove confounding factors affecting both groups, and thus recover causal estimates. Under a number of assumptions—the more important being that difference between groups remain stable over time and that the onset of the treatment is exogenous—causal estimates of the treatment effect can be correctly recovered (see Angrist & Krueger, 1999; Angrist & Pischke, 2008; Meyer, 1995). In psychology, the differences-in-differences design is known as an untreated control group design with pre- and post-test (Shadish, et al., 2002).

Conclusions

Lewin (1945, p. 129) once noted "nothing is as practical as a good theory." The point of all our research efforts is to develop theoretical models that explain natural phenomena. Doing so means undertaking different sorts of studies and ideally to do more fieldwork that emulates "natural experiments" (Meyer, 1995); the latter can be particularly useful for making strong and relevant causal claims *if* certain design and estimation conditions are respected. We trust that we have made it clear that that researchers in leadership and other applied areas must pay more attention to the problems of endogeneity and correct model estimation. Researchers have mostly been ignorant of these problems (Antonakis, et al., 2010) and graduate training in statistics has not been sufficient to ensure the needed methodological standards (Aiken, West, & Millsap, 2008). It is vital to understand endogeneity and how to deal with it; facing this "inconvenient

truth" will be difficult because many researchers have to break with past practices that produced specious estimates. Theories may have to be revamped and new ways of clean causal thinking and testing have to become the order of the day.

As the world economy hit crises after crises, and as research budgets get squeezed, it becomes all the more vital to ensure that research monies are well invested in approaches that can help make a real difference to practice. Relevance must go hand in hand with rigor (Vermeulen, 2005); for that to occur, models must be build around rigorous methods that can be applied to practical problems.

Future directions

Insofar as future directions in causal analysis are concerned, we hope that advances on par with those that allowed causal research to be done in the field will continue to be made (Heckman, 1979; Rubin, 1974; Thistlethwaite & Campbell, 1960). With respect to leadership research, our expectations are that researchers begin to use these tools (which are standard in other sciences, e.g., medical or economics) and teach them to their students. In our recent review (Antonakis, et al., 2010) we showed that these methods are foreign to leadership scholars; we also identified 10 best practices for ensuring valid causal claims, as noted in Table 7, which we hope leadership scholars will begin to adopt.

[Table 7 here]

Only when nonexperimental models are correctly tested we will be in a position to better evaluate current theories of leadership and better answer questions like:

1. Do potentially endogenous regressors (e.g., LMX, transformational leadership, authentic leadership, servant leadership, etc.) matter for organizational outcomes?

2. To what extent do multilevel leadership models estimated using HLM-type models (random effects/coefficients) still explain increment outcomes when controlling for omitted fixed effects?
3. Do leader individual differences (e.g., cognitive style, emotional intelligence, self-monitoring etc.) matter in predicting leader behaviors or outcomes beyond established personality (e.g., the big five) and cognitive ability models?
4. How would multifactorial models look (e.g., the Multifactor Leadership Questionnaire) when stronger modeling procedures are undertaken (e.g., using MIMIC models, see: Bollen, 1989; Muthén, 1989) while paying attention to real tests of overidentification?

At this time, we do not have enough well-designed studies to answer the above questions, as well as many other questions that are implied from the validity threats we have identified in Table 1. We hope that leadership scholars will rise to the challenge and test their causal models correctly.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology - Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*(1), 32-50.

Angrist, J. D., & Krueger, A. B. (1999). Empirical Strategies in Labor Economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. Volume 3, Part 1, pp. 1277-1366). Amsterdam: Elsevier.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086-1120.

Antonakis, J., & Dietz, J. (2011). More on testing for validity instead of looking for it. *Personality and Individual Differences, 50*(3), 418-421.

Antonakis, J., Schriesheim, C. A., Donovan, J. A., Gopalakrishna-Pillai, K., Pellegrini, E., & Rossomme, J. L. (2004). Methods for studying leadership. In J. Antonakis, A. T. Cianciolo & R. J. Sternberg (Eds.), *The Nature of Leadership* (pp. 48-70). Thousand Oaks: Sage.

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14*(4), 496-515.

Baltagi, B. H. (2002). *Econometrics*. New York: Springer.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schmacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Academic Press.

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization, 6*(3), 285-327.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics, 119*(1), 249-275.

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

Bollen, K. A. (1990). Overall Fit in Covariance Structure Models - 2 Types of Sample-Size Effects. *Psychological Bulletin, 107*(2), 256-259.

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika, 61*, 109-121.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. N. (2007). Latent variable models under misspecification - Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research, 36*(1), 48-86.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association, 90*(430), 443-450.

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies, 47*, 239-253.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics, 29*(2), 238-249.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.

Caudill, S. B. (1988). An Advantage of the Linear Probability Model over Probit or Logit. *Oxford Bulletin of Economics and Statistics, 50*(4), 425-427.

Cong, R., & Drukker, D. M. (2001). Treatment effects model. *Stata Technical Bulletin, 10*(55), 25-33.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*(4), 724-750.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265-2281.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Dubin, R. (1976). Theory building in applied areas. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology* (pp. 17-40). Chicago: Rand McNally.

Dulebohn, J. H., Bommer, W. H., Liden, R. C., Brouer, R. L., & Ferris, G. R. (2011). A Meta-Analysis of Antecedents and Consequences of Leader-Member Exchange: Integrating the Past with an Eye toward the Future. *Journal of Management*.

Duncan, G. J., Magnusson, K. A., & Ludwig, J. (2004). The Endogeneity Problem in Developmental Studies. *Research in Human Development, 1*(1&2), 59-80.

Fairhurst, G. T., & Antonakis, J. (in press). A Research Agenda for Relational Leadership. In M. Uhl-Bien & S. Ospina (Eds.), *Advancing Relational Leadership Theory: A Conversation among Perspectives*. Greenwich, CT: Information Age Publishing.

Fiori, M., & Antonakis, J. (2011). The ability model of emotional intelligence: Searching for valid measures. *Personality and Individual Differences, 50*(3), 329-334.

Foster, E. M., & McLanahan, S. (1996). An Illustration of the Use of Instrumental Variables: Do neighborhood conditions affect a young person's change of finishing high school? *Psychological Methods, 1*(3), 249-260.

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44*(2), 381-394.

Gerstner, C. R., & Day, D. V. (1997). Meta-analytic review of leader-member exchange theory: Correlates and construct issues. *Journal of Applied Psychology, 82*(6), 827-844.

Greene, W. H. (2008). *Econometric Analysis* (6th ed.). Upper Saddle River, NJ: Prentice–Hall.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507-544.

Hall, R. J., Snell, A. F., & Singer Foust, M. (1999). Item Parceling Strategies in SEM: Investigating the Subtle Effects of Unmodeled Secondary constructs. *Organizational Research Methods, 2*(3), 233-256.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for Endogeneity in Strategic Management Research. *Strategic Organization, 1*(1), 51-78.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*, 1029-1054.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica, 46*(6), 1251-1271.

Hayduk, L. A. (1996). *LISREL issues, debates, and strategies*. Baltimore: Johns Hopkins University Press.

Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three - Testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841-850.

Hayduk, L. A., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., et al. (2003). Pearl's D-separation: One more step into causal thinking. *Structural Equation Modeling, 10*(2), 289-311.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*(1), 153-161.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

House, R. J., & Aditya, R. N. (1997). The social scientific study of leadership: Quo vadis? *Journal of Management, 23*(3), 409-473.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, 221-233.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal Analysis: Assumptions, Models, and Data*. Beverly Hills: Sage Publications.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge, MA: MIT Press.

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

Keppel, G., & Wickens, T. D. (2004). *Design and Analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson.

Kerlinger, F., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Forth Worth, TX: Harcourt Publishers.

Kmenta, J. (1986). *Elements of econometrics* (2nd ed.). New York: Macmillan Publishing Company.

Koh, W. L., Steers, R. M., & Terborg, J. R. (1995). The effects of transformational leadership on teacher attitudes and student performance in Singapore. *Journal of Organizational Behavior, 16*(4), 319-333.

Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics, 49*(3), 186-205.

Lee, D., & Lemieux, T. (2009). Regression Discontinuity Designs in Economics. *National Bureau of Economic Research, Working Paper 14723*.

Levitt, S. D. (1997). Using electoral cycles in police hiring to estimate the effects of police on crime. *American Economic Review, 87*(3), 270-290.

Levitt, S. D. (2002). Using electoral cycles in police hiring to estimate the effects of police on crime: Reply. *American Economic Review, 92*(4), 1244-1250.

Lewin, K. (1945). The research center for group dynamics at Massachusetts Institute of Technology. *Sociometry, 8*(2), 126-136.

Liang, J., Lawrence, R. H., Bennett, J. M., & Whitelaw, N. A. (1990). Appropriateness of composites in structural equation models. *Journal of Gerontology: Social Sciences, 45*(2), 52-59.

Liden, R. C., & Antonakis, J. (2009). Considering context in psychological leadership research. *Human Relations, 62*(11), 1587-1605.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). College Station, TX: StataCorp LP.

Maddala, G. S. (1977). *Econometrics*. New York: McGraw-Hill.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maxwell, S. E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subjects designs. *Psychological Bulletin, 110*(2), 328-337.

McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*(5), 859-867.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economics Statistics, 13*(2), 151-161.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Mount, M. K., & Scullen, S. E. (2001). Multisource feedback ratings: What do they really measure? In M. London (Ed.), *How people evaluate others in organizations* (pp. 155-176). Mahwah, NJ: Lawrence Erlbaum.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton: CRC Press.

Mulaik, S. A., & James, L. R. (1995). Objectivity and reasoning in science and structural equation modeling. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 118-137). Thousand Oaks, CA: Sage Publications.

Mundlak, Y. (1978). Pooling of Time-Series and Cross-Section Data. *Econometrica, 46*(1), 69-85.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*(1), 3-5.

Muthén, B. O. (1989). Latent variable modeling in heterogenous populations. *Psychometrika, 54*(4), 557-585.

Onyskiw, J. E., & Hayduk, L. A. (2001). Processes Underlying Childern's Adjustment in Families Characterized by Physical Aggression. *Family Relations, 50*, 376-385.

Pearl, J. (2009). *Causality : models, reasoning, and inference* (2nd ed.). Cambridge ; New York: Cambridge University Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology, 89*(5), 879-903.

Podsakoff, P. M., Mackenzie, S. B., & Podsakoff, N. P. (2010). Sources of Method Bias in Social Science Research and Recommendations on How to Control It. *Annu Rev Psychol*.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management, 12*(4), 531-544.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B, 31*, 350-371.

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A Tale of Three Perspectives: Examining Post Hoc Statistical Techniques for Detection and Correction of Common Method Variance. *Organizational Research Methods, 12*(4), 762-800.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics, 2*(3), 808-840.

Sargan, J. D. (1958). The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica, 26*, 393-415.

Schulte, M. J., Ree, M. J., & Carretta, T. R. (2004). Emotional Intelligence: Not much more than g and personality. *Personality and Individual Differences, 37*(5), 1059-1068.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Shadish, W. R., & Cook, T. D. (1999). Comment-Design Rules: More Steps toward a Complete Theory of Quasi-Experimentation. *Statistical Science, 14*(3), 294-300.

Shadish, W. R., & Cook, T. D. (2009). The Renaissance of Field Experimentation in Evaluating Interventions. *Annual Review of Psychology, 60*, 607-629.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shaver, J. M. (1998). Accounting for endogeneity when assessing strategy performance: Does entry mode choice affect FDI survival? *Management Science, 44*(4), 571-585.

Spector, P. E. (2006). Method variance in organizational research - Truth or urban legend? *Organizational Research Methods, 9*(2), 221-232.

Spector, P. E., & Brannick, M. T. (2011). Methodological Urban Legends: The Misuse of Statistical Control Variables. [Article]. *Organizational Research Methods, 14*(2), 287-305.

Stock, J. H., & Watson, M. W. (2007). *Introduction to econometrics* (2nd ed.). Boston: Pearson Addison Wesley.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics, 20*(4), 518-529.

Stock, J. H., & Yogo, M. (2002). Testing for weak instruments in linear IV regression. *NBER Technical Working Papers 0284*.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*(6), 309–317.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*, 24-36.

Vermeulen, F. (2005). On rigor and relevance: Fostering dialectic progress in management research. *Academy of Management Journal, 48*(6), 978-982.

von Wittich, D., & Antonakis, J. (2011). The KAI cognitive style inventory: Was it personality all along? *Personality and Individual Differences, 50*(7), 1044-1049.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*, 817-830.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-706.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldrige, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Zellner, A., & Theil, H. (1962). 3-Stage Least-Squares - Simultaneous Estimation of Simultaneous-Equations. *Econometrica, 30*(1), 54-78.

Table 1: Threats to Validity

Validity Threat	Explanation
1. Omitted variables:	(a) Omitting a regressor, that is, failing to include important control variables when testing the predictive validity of dispositional or behavioral variables (e.g., testing predictive validity of "emotional intelligence" without including IQ or personality; not controlling for competing leadership styles) (b) Omitting fixed effects (c) Using random-effects without justification (d) In all other cases, independent variables not exogenous (if it is not clear what the controls should be)
2. Omitted selection:	(a) Comparing a treatment group to other non-equivalent groups (i.e., where the treatment group is not the same as the other groups) (b) Comparing entities that are grouped nominally where selection to group is endogenous (e.g., comparing men and women leaders on leadership effectiveness where the selection process to leadership is not equivalent) (c) Sample (participants or survey responses) suffer from self-selection or is non-representative
3. Simultaneity:	(a) Reverse causality (i.e., an independent variable is potential caused by the dependent variable)
4. Measurement error:	(a) Including imperfectly-measured variables as independent variables and not modelling measurement error
5. Common-methods variance:	(a) Independent and dependent variables are gathered from the same rating source
6. Inconsistent inference:	(a) Using normal standard errors without examining for heteroscedasticity (b) Not using cluster-robust standard errors in panel data
7. Model misspecification:	(a) Not correlating disturbances of potentially endogenous regressors in mediation models (and not testing for endogeneity using a Hausman test or augmented regression), (b) Using a full information estimator (e.g., maximum likelihood, three-stage least squares) without comparing estimates to a limited information estimator (e.g., two stage-least squares).

Note: Reprinted from *The Leadership Quarterly*, 21(6), Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R., "On Making Causal Claims: A Review and Recommendations," pp. 1086-1120, 2010, with permission from Elsevier.

Table 2: Data from "Endogeneity: An inconvenient Truth" (simulated data)

Panel A: Summary data with omitted cause

	Disc not shattered	Disc shattered	Total
Noise not heard	19	0	19
Noise heard	2	29	31
Total	21	29	50

Panel B: Correlation matrix and descriptive statistics

Variable	Mean	Std. Dev.	1	2	3
1. Rifle (fired=1; else = 0)	.62	.49			
2. Disk shattered	.58	.50	.92		
3. Noise (db)	75.48	6.22	.99	.90	
4. Noise (heard =1; else = 0)	.62	.49	1.00	.92	.99

Note: The data can be downloaded at: http://www.hec.unil.ch/jantonakis/disk.xls.

Table 3: Summary data showing regarding LMX-Turnover relation (simulated data)

	Mean	SD	1	2	3	4	5	6	7	8
1. Leader extraversion	50.04	2.93	1							
2. Leader use of incentives	9.86	2.99	-.03	1						
3. Leader IQ	110.07	2.90	.05	.03	1					
4. Follower IQ	105.00	2.94	.02	-.02	.03	1				
5. Follower conscientiousness	39.91	2.89	.01	.01	-.05	.01	1			
6. Follower neuroticism	35.04	3.08	.00	-.02	-.02	-.01	.02	1		
7. Company policies	19.98	3.12	.00	-.05	-.01	-.07	.00	.01	1	
8. LMX	29.78	7.88	.37	.40	.40	.39	.34	-.40	-.08	1
9. Turnover	45.01	6.70	-.02	-.43	-.45	.02	.03	.48	-.45	-.50

n = 10,000

Table 4: Regressions regarding LMX and turnover (simulated data)

	Coef.	Std. Err.	*t*	*p*
Panel A: Multivariate regression estimates				
Dependent variable: LMX				
Leader extraversion	.94	.03	29.05	.00
Leader use of incentives	1.05	.03	33.21	.00
Leader IQ	.99	.03	3.45	.00
Follower IQ	.99	.03	3.81	.00
Follower conscientiousness	.97	.03	29.62	.00
Follower neuroticism	-1.00	.03	-32.71	.00
Company policies	-.05	.03	-1.78	.08
Constant	-243.36	5.43	-44.82	.00
$F(8,991) = 857.15$, $p < .001$, $r^2 = .86$				
Dependent variable: Turnover				
Leader extraversion	-.04	.03	-1.17	.24
Leader use of incentives	-.98	.03	-31.85	.00
Leader IQ	-.98	.03	-31.03	.00
Follower IQ	-.02	.03	-.66	.51
Follower conscientiousness	.01	.03	.22	.83
Follower neuroticism	1.02	.03	34.31	.00
Company policies	-1.02	.03	-34.89	.00
Constant	151.04	5.26	28.71	.00
$F(8,991) = 626.61$, $p < .001$, $r^2 = .82$				
Panel B: Naïve regression estimates				
Leader-member exchange	-.46	.02	-23.75	.00
Company policies	-1.04	.05	-21.41	.00
Constant	79.52	1.18	67.49	.00
$F(2,997) = 474.74$, $p < .001$, $r^2 = .49$				

Note: the residual correlation for the Panel A model, between LMX and turnover, is zero.

Table 6: Predicting turnover from LMX using OLS and 2SLS regression (simulated data)

VARIABLES	Model (1) Turnover	Model (1) LMX	Model (2) Turnover	Model (2) LMX	Model (3) Turnover	Model (3) LMX	Model (4) Turnover	Model (4) LMX
LMX	-.43**		.02		-.00		-.02	
	(18.35)		(.42)		(0.03)		(0.94)	
Leader extraversion		.95**		.95**		.99**		.94**
		(14.32)		(14.35)		(22.17)		(29.17)
Follower IQ		1.01**		1.01**		1.02**		.99**
		(15.22)		(15.25)		(23.03)		(30.93)
Follower conscientiousness		.91**		.91**		.92**		.97**
		(13.42)		(13.45)		(20.22)		(29.74)
Leader use of incentives					-.95**	1.09**	-.96**	1.05**
					(14.74)	(24.81)	(26.82)	(33.35)
Follower neuroticism					1.03**	-1.02**	1.00**	-1.00**
					(16.51)	(24.08)	(28.83)	(32.84)
Leader IQ							-.97**	.99**
							(26.38)	(30.58)
Company policies							-1.02**	-.05
							(35.01)	(1.79)
Constant	57.72**	-160.19**	44.46**	-160.19**	18.41**	-138.82**	146.64**	-243.36**
	(80.56)	(19.82)	(33.85)	(19.86)	(6.98)	(24.58)	(38.35)	(45.00)
R-squared	.25	.39	.25	.39	.41	.73	.82	.86

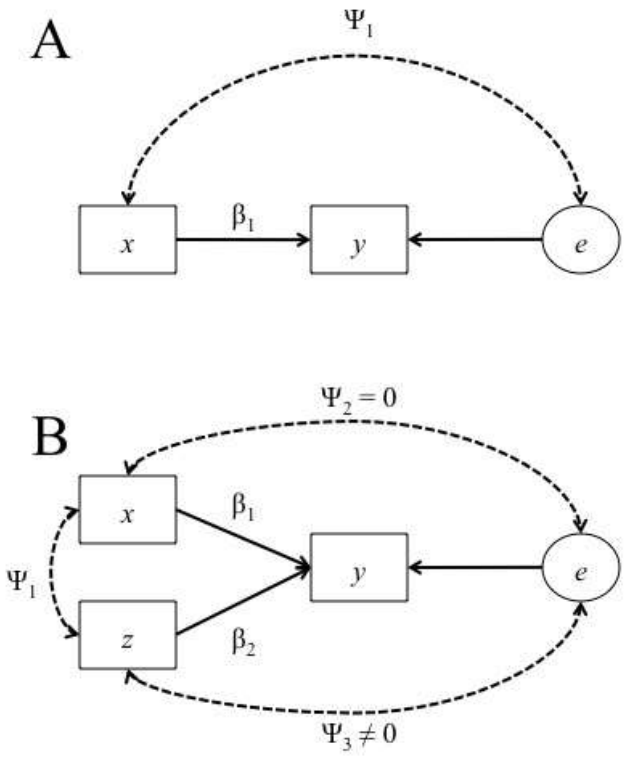
n=1,000; *t*-statistics in parentheses; **$p < 0.01$, *$p < 0.05$. Model 1 is estimated with OLS. Models 2, 3 and 4 are estimated with 2SLS. Note for the Turnover equation in Model 2 the r-square was negative (this is not necessarily a problem in simultaneous equation models, see: http://www.stata.com/support/faqs/stat/2sls.html): We calculated it by correlating (and squaring) the predicted and observed value of turnover; this correlation is the actually same as that in Model 1. The coefficients of LMX in Models 2, 3, and 4 are not significantly different from each other; the coefficient of LMX in Model 1 is significantly different from those in Models 2, 3, and 4. The Sargan chi-square test for Models 2, 3 and 4 is non-significant.

Table 7: The 10 Commandments of Causal Analysis

1. Avoid omitted variable bias by including appropriate control variables; if these are difficult to identify, using fixed-effects estimation or instrument the endogenous variables.
2. Prior to using HLM models, ensure that the estimator is consistent with respect to the fixed-effects estimator; test for differences between the estimators using a Hausman test.
3. Show that modeled independent variables are exogenous; if there is doubt, instrument them with truly exogenous variables.
4. Do not compare groups unless assignment to group was randomized or the selection procedure to group has been appropriately modeled.
5. Test overidentifying restrictions in simultaneous equations with a chi-square test of fit; if failed, do not interpret parameter estimates.
6. Use errors-in-variables regression, SEM or instruments (in 2SLS models) to correct estimates for measurement bias.
7. Avoid common-methods bias; use instrumental-variable models to correct for it if unavoidable.
8. Use robust variance estimators as the default (unless residuals are i.i.d). Use cluster-robust variance estimators with nested data.
9. Use the 2SLS estimator in mediation models (and correlate disturbances of endogenous regressors); examine endogeneity with a Hausman test
10. Use full-information estimators (i.e., maximum likelihood) if estimates are equivalent to limited information (2SLS) estimators. Never use PLS (which cannot test for overidentifying restrictions).

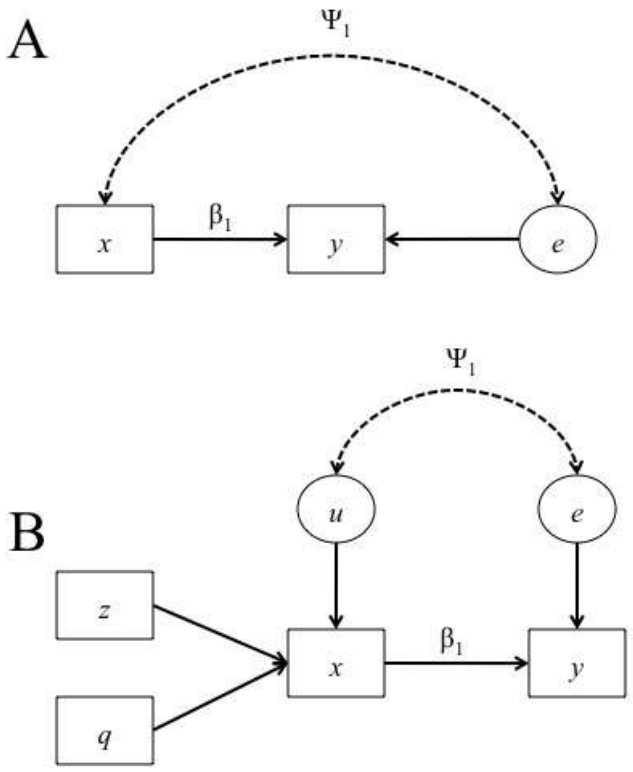
Note: Adapted from *The Leadership Quarterly*, 21(6), Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R., "On Making Causal Claims: A Review and Recommendations," pp. 1086-1120, 2010 with permission from Elsevier.

Figure 1: Endogeneity and the Consistency of Estimates



Panel	Condition	β_1	β_2	Explanation
A	$\Psi_1 = 0$	Consistent		*x* does not correlate with *e* thus β_1 is consistent.
A	$\Psi_1 \neq 0$	Inconsistent		*x* correlates with *e* and thus β_1 is inconsistent.
B	$\Psi_1 = 0$	Consistent	Inconsistent	*z* correlates with *e* and thus β_2 is inconsistent. β_1 is consistent because *x* is uncorrelated both with *z* and with *e*.
B	$\Psi_1 \neq 0$	Inconsistent	Inconsistent	*z* correlates with *e* thus β_2 is inconsistent. Although *x* is uncorrelated with *e*, β_1 it inconsistent because it is affected by the bias in *z* through *x's* correlation with *z*.

Figure 2: Endogeneity and the Consistency of Estimates in Simultaneous Equation (Mediatory) Models



Panel	Condition	Estimator	β_1	Explanation
A	$\Psi_1 \neq 0$	OLS	Inconsistent	*x* correlates with *e* thus β_1 is inconsistent.
B	$\Psi_1 \neq 0$	Instrumental Variable (e.g., 2SLS)	Consistent	Ψ_1 is estimated. β_1 is consistent because the instruments *z* and *q* are truly exogenous. In this case, β_1 (Panel A) $\neq \beta_1$ (Panel B)
B	$\Psi_1 \neq 0$	OLS	Inconsistent	Ψ_1 is constrained to zero; therefore β_1 is inconsistent. In this case, β_1 (Panel A) = β_1 (Panel B)