The Leadership Quarterly 21 (2010) 1086–1120

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

On making causal claims: A review and recommendations

John Antonakis [*], Samuel Bendahan, Philippe Jacquart, Rafael Lalive

Faculty of Business and Economics, University of Lausanne, Switzerland

ARTICLE INFO

Keywords:
Causality
Quasi-experimentation
Instrumental variables
Common-method bias
Difference-in-differences
Simultaneous equations
Monte Carlo simulations
Regression discontinuity
Mediation

ABSTRACT

Social scientists often estimate models from correlational data, where the independent variable has not been exogenously manipulated; they also make implicit or explicit causal claims based on these models. When can these claims be made? We answer this question by first discussing design and estimation conditions under which model estimates can be interpreted, using the randomized experiment as the gold standard. We show how endogeneity – which includes omitted variables, omitted selection, simultaneity, common-method variance, and measurement error – renders estimates causally uninterpretable. Second, we present methods that allow researchers to test causal claims in situations where randomization is not possible or when causal interpretation could be confounded; these methods include fixed-effects panel, sample selection, instrumental variable, regression discontinuity, and difference-in-differences models. Third, we take stock of the methodological rigor with which causal claims are being made in a social sciences discipline by reviewing a representative sample of 110 articles on leadership published in the previous 10 years in top-tier journals. Our key finding is that researchers fail to address at least 66% and up to 90% of design and estimation conditions that make causal claims invalid. We conclude by offering 10 suggestions on how to improve non-experimental research.

Social scientists make causal claims. Some come out and say it straight, using statements like "*x* causes, predicts, affects, influences, explains, or is an antecedent of *y*" or that "*y* depends on *x*." Others shy away from using such explicit language, choosing instead to couch their claims in suggestive language stating instead that "*y* is associated or related to *x*." Researchers must not hide from making causal claims (cf. Pearl, 2000; Shipley, 2000). Causal claims are important for society and it is crucial to know when scientists can make them.

The failsafe way to generate causal evidence is to use randomized experiments. Unfortunately, randomization is often infeasible in social science settings, and depending on the phenomenon under investigation, results might not generalize from the laboratory to the real world. However, many recent methodological advances have been made allowing social scientists to have their causal cake and eat it (in the field!). These methods, though, have been slow to reach social science disciplines. Unfortunately, methods are still being used to estimate explicit (or implicit) causal models in design situations where the assumptions of the methods are violated, thus rendering uninformative results.

Given the importance of understanding causality in non-experimental settings, the purpose of our paper was threefold, to (a) demonstrate the design and estimation conditions under which estimates can and cannot be causally interpreted (or indeed interpreted at all, even as associations), (b) review methods that will allow researchers to test causal claims in the field, particularly in situations where randomization is not possible, and (c) take stock of the methodological rigor with which causal claims are being made in leadership, which straddles the disciplines of management and applied psychology.

What we care to show in this review are the necessary design and estimation conditions for causal interpretation. Our central focus will be on the *consistency* of parameter estimates; by consistent we mean that the estimate regarding the presumed causal

[*] Corresponding author. Faculty of Business and Economics, University of Lausanne, Internef #618, CH-1015 Lausanne-Dorigny, Switzerland. Tel.: +41 21 692 3438; fax: +41 21 692 3305.
E-mail address: john.antonakis@unil.ch (J. Antonakis).

1048-9843/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.leaqua.2010.10.010

relationship converges to the correct population parameter as the sample size increases. We are concerned about the regression coefficient, β, of a particular independent variable x and whether β accurately reflects the true treatment effect in predicting y. After model estimation, the result might seem to look good, particularly if an advanced statistical modeling program was used, the p-value of the parameter estimate is below .0001 and the model fits well because of high r-squares and in the case of simultaneous equation models because tests of model fit cannot reject the model. However, if certain essential design and methodological conditions are not present the coefficient *cannot be interpreted*, not even in terms of an association or relation — even in the correlational sense. That is, the coefficient may have an allure of authenticity but it is specious.

As we will demonstrate, *correlation can mean causation* in non-experimental settings *if* some essentials design conditions are present and the appropriate statistical methods are used. Knowing the conditions under which causal claims can be made – and their resulting practical and policy recommendations – is one of the most important tasks entrusted to scientists. Apart from the obvious importance and implications of understanding causality in the hard sciences, correctly modeling the causal relations that explain phenomena is also crucial in the social sciences.

Calls have been made before to pay attention to the correct estimation of non-experimental causal models; the major culprit is *endogeneity*, where the effect of x on y cannot be interpreted because it includes omitted causes. This problem of endogeneity has been noted both in psychology (Foster & McLanahan, 1996) and management (Shaver, 1998), and these calls are being repeated (Bascle, 2008; Gennetian, Magnuson, & Morris, 2008; Larcker & Rusticus, 2010). Unfortunately, these calls have mostly fallen on deaf ears. The results of our review are similar to a recent review that found that more than 90% of papers published in the premier strategy journal (and one of the top journals in management), *Strategic Management Journal* (SMJ), were not correctly estimated (Hamilton & Nickerson, 2003)! Hamilton and Nickerson (2003, pp. 53–54) went on to say, "We believe that the low number of papers in SMJ that account for endogeneity may indicate a failure of empirical research in strategic management…. Yet, ignoring endogeneity is perilous; … the resulting parameter estimates are likely to be biased and may therefore yield erroneous results and incorrect conclusions about the veracity of theory." Economics went though the same difficult period a couple of decades ago and economists have improved many of their practices regarding causal inference. Nowadays in economics it is virtually impossible to publish a non-experimental study in a top general or field journal (e.g., *American Economic Review*, *Quarterly Journal of Economics*, *Review of Economic Studies*, *Econometrica*, *Journal of Econometrics*, and *Journal of Labor Economics*) without providing convincing evidence and arguments that endogeneity is not present.

Our paper is structured in three major sections, as follows: we first explain what causality is; we then introduce the counterfactual argument, and explain why it is important to have a control group so that causal conclusions to be made. We look at the randomized experiment as a point of departure showing precisely why it allows for causal claims. Although the randomized experiment is a very useful tool sometimes experiments are impossible to do (see Cook, Shadish, & Wong, 2008; Rubin, 2008). At other times, researchers may come across a "natural experiment" of sorts, whose data they can exploit. We review these designs and methods and show that when correctly implemented they allow for causal inference in real-world settings. Unfortunately, many of these methods are rarely utilized in management and applied psychology research (cf. Grant & Wall, 2009). In our review, we borrow mostly from econometrics, which has made great strides in teasing-out causal relations in non-experimental settings (try randomly assigning an economy or a company to a treatment condition!), though, the "natural experiment revolution" has debts to pay to psychology given the contributions of Donald T. Campbell to quasi-experimentation (see Campbell & Stanley, 1963, 1966; Cook & Campbell, 1979). Also, some of the approaches we discuss (e.g., regression discontinuity) that are popular in econometrics nowadays were originally developed by psychologists (Thistlethwaite & Campbell, 1960).

Next, we discuss the intuition and provide step-by-step explanation behind the non-experimental causal methods; we maintain statistical notation to a minimum to make our review accessible to a large audience. Although the context of our review is management and applied psychology research, the issues we present and the recommendations and conclusions we make are very general and have application for any social science, even the hard sciences.

Finally, similar to the recent Leadership Quarterly review of Yammarino, Dionne, Uk Chun and Dansereau (2005) who examined the state of research with respect to levels-of-analysis issues (i.e., failure to correctly theorize and model multilevel phenomena), we examined a subset of the literature published in top management and applied psychology journals, making explicit or implicit causal claims about a "hot" social-sciences topic, leadership. The journals we included in the review were top-tier (in terms of 5-year impact factor), including: *Academy of Management Journal*, *Journal of Applied Psychology*, *Journal of Management*, *Journal of Organizational Behavior*, *The Leadership Quarterly*, *Organizational Behavior & Human Decision Processes*, and *Personnel Psychology*. We coded studies from these journals to determine whether the method used allowed the researchers to draw causal claims from their data. Our results indicate that the statistical procedures used are far from being satisfactory. Most studies had several problems that rendered estimates suspect. We complete our review with best-practice recommendations.

1. What is causality?

We take a simple, pragmatic, and widely-shared view of causality; we are not concerned about the nature of causes or philosophical foundations of causality (cf. Pearl, 2000), but more specifically *how to measure the effect of a cause*. To measure causal effects, we need an effect (y) and a presumed cause (x). Three classic conditions must exist so as to measure this effect (Kenny, 1979):

1. x must precede y temporally
2. x must be reliably correlated with y (beyond chance)
3. the relation between x and y must not explained by other causes

The first condition is rather straightforward; however, in the case of simultaneity – which we discuss later – a cause and an effect could have feedback loops. Also, simply modeling variable *x* as a "cause" merely because it is temporal antecedent of *y* does not mean that it caused *y* (i.e., *x* must be exogenous too, as we discuss in detail later); thus temporal ordering is a necessary but not a sufficient condition. The second condition requires a statistically reliable relationship (and thus quantitative data). The third condition is the one that poses the most difficulties and has to do with the *exogeneity* of *x* (i.e., that *x* varies randomly and is not correlated with omitted causes). Our review is essentially concerned with the first and third conditions; these conditions, particularly the third one, have less to do with theoretical arguments and more to do with design and analysis issues (see also James, Mulaik, & Brett, 1982; Mulaik & James, 1995).

If the relation between *x* and *y* is due, in part, to other reasons, then *x* is *endogenous*, and the coefficient of *x* cannot be interpreted, not even as a simple correlation (i.e., the magnitude of the effect could be wrong as could be the sign). The limitations often invoked in non-experimental research that "the relation between *x* and *y* might be due to *y* causing *x* (i.e., reverse causality may be at play)," "common-method variance may explain the strong relationship," or "this relationship is an association given the non-experimental data" are moot points. If *x* is endogenous the coefficient of *x* simply has no meaning. The true coefficient could be higher, lower, or even of a different sign.

1.1. The counterfactual argument

Suppose that we have conducted an experiment, where individuals were assigned by some method to an experimental and a control condition (*x*). The manipulation came before the outcome (*y*) and it correlates reliably with the outcome. How do we rule out other causes? There could be an infinite amount of potential explanations as to why the cause correlates with the effect. To test whether a causal relation is real, the model's predictions must be examined from the counterfactual model (Morgan & Winship, 2007; Rubin, 1974; Winship & Morgan, 1999). The counterfactual asks the following questions: (a) if the individuals who received the treatment had in fact not received it, what would we observe on *y* for those individuals? Or, (b) if the individuals who did not receive the treatment had in fact received it, what would we have observed on *y*?

As will become evident later, if the experimenter uses random assignment, the individuals in the control and treatment groups are roughly equivalent at the start of the experiment; the two groups are theoretically interchangeable. So, the counterfactual for those receiving the treatment are those who did not receive it (and vice-versa). The treatment effect is simply the difference in *y* for the treatment and control group. In a randomized experiment the treatment effect is correctly estimated when using a regression (or ANOVA) model.

However, when the two groups of individuals are not the same on observable (or unobservable characteristics), and one group has received a treatment, we cannot observe the counterfactuals: the groups are not interchangeable. What would the treated group's *y* had been had they not received the treatment and what would the untreated group's *y* be had they received the treatment? The counterfactual cannot be observed because the two groups are systematically different in some ways, which obscures the effect of the treatment. To obtain consistent estimates, therefore, this selection (to treatment and control group) must be modeled. Modeling this selection correctly is what causal analysis, in non-experimental settings, is all about.

Also, in this review, we are exclusively focusing on quantitative research because when done correctly it is only through this mode of inquiry that counterfactuals, and hence causality can be reliably established. Proponents of qualitative methods have suggested that causality can also be studied using rigorous case-studies and the like (J. A. Maxwell, 1996; Yin, 1994). Yin (1994), for example, compares case study research to a single experiment — although what Yin states might be intuitively appealing, a case study of a social-science phenomenon is nothing like a chemistry experiment. In the latter the experimenters have complete control of the system of variables that are studied and can add or remove molecules or perform interventions at will (experimenters have complete experimental control). If the experiment works and can be reliably repeated (and ideally, this reliability is analyzed statistically), then causal inference can be made.

However in the post-hoc case study or even one where observation is real-time, there are a myriad of variables both observed or unobserved that cannot be controlled for and thus confound results. These latter problems are the same ones that plague quantitative research; however, quantitative researchers can control for these problems if the model is correctly specified, and thus accounts for the bias. Qualitative research can be useful when quantified (cf. Diamond & Robinson, 2010); however, matching "patterns" in observations (i.e., finding qualitative "correlations") cannot lead to reliable inference if sources of bias in the apparent pattern are not controlled for and the reliably of the relation is not tested statistically (and we will not get into another limitation of observer, particularly participant–observer, confirmation bias, Nickerson, 1998).

We begin our methodological voyage with the mainstay of psychology: the randomized field experiment. A thorough understanding of the mechanics of the randomized field experiment is essential because it will be a stepping stone to exploring quasi-experimental methods that allow for causal deductions.

2. The gold standard: the randomized field experiment

This design ensures that the correlation between an outcome and a treatment is causal; more specifically, the origin of the change in the dependent variable stems from no other cause other than that of the manipulated variable (Rubin, 2008; Shadish, Cook, & Campbell, 2002). What does random assignment actually do and why does it allow one to make causal conclusions?

We first draw attention to how the Ordinary Least Squares (OLS) estimator (i.e., the estimator used in regression or ANOVA-type models that minimizes the sum of squared residuals between observation and the regression line) derives estimates for a model. For simplicity, we will only discuss a treatment and control; however, the methods we discuss can be expanded to more than two conditions (e.g., we could add an alternative/bogus treatment group) and interactions among conditions.

Assume a model where we have a dummy (binary) independent variable *x* reflecting a randomly assigned treatment (a manipulation, leadership training, which is 1 if the subject received the treatment else it is 0) and a continuous independent variable *z*, which is a covariate (e.g., IQ of leaders). This model is the typical ANCOVA model for psychologists; in an ANCOVA model, including a covariate that is strongly related to *y* (e.g., leadership effectiveness) reduces unexplained variance. Thus, it is desirable to include such a covariate because power to detect a significant effect in the treatment increases (Keppel & Wickens, 2004; S. E. Maxwell, Cole, Arvey, & Salas, 1991). The covariate is also useful to adjust for any observed initial – albeit it small, that are due to chance – differences in the intervention and control groups that may have occurred due to chance (Shadish et al., 2002). Let:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \tag{1}$$

Where *y* is the dependent variable, *i* is from 1 to *n* observations, β_0 is a constant (the intercept, where $x=0$ and $z=0$, and the line – a two dimensional plane in this case given that the equation has two independent variables – cuts the *y* axis), β_1 and β_2 are unstandardized regression coefficients of the independent variables *x* and *z* and refer how much a change in one unit of *x* and *z* respectively affect *y* (i.e., $\beta_1 = \Delta y/\Delta x$ and $\beta_2 = \Delta y/\Delta z$ respectively), *e* is a disturbance term (also known as the error term), reflecting unobserved causes of *y* as well as other sources of error (e.g., measurement error). The error term, which is an unobserved latent variable must not be confused with the residual term, which is the difference between the predicted and the observed value of *y*. This residual term is orthogonal to the regressors regardless of whether the error term is or not.

Let us focus on *x* for the time being, which is the manipulated variable. When estimating the slopes (coefficients) of the independent variables, OLS makes an important assumption: That *e* is *uncorrelated* with *x*. This assumption is usually referred to as that of the orthogonality of the error term with the regressor. In other words, *x* is assumed to be *exogenous*. Exogenous means that *x* does not correlate with the error term (i.e., it does not correlate with omitted causes). When *x* is not exogenous, that is, when it is endogenous (hence the problem of *endogeneity*) then it will correlate with the error term and this for a variety of reasons. We discuss some of these reasons in the next section.

To better understand the problem of endogeneity, suppose that extraversion is an important factor for leadership effectiveness. Now, if we assign the treatment randomly there will be an equal amount of extraverts in the treatment and control conditions. If we find that the treatment group is higher than the control group on effectiveness, this difference cannot be accounted for by an unmodeled potential cause (e.g., extraversion). Thus, random assignment assures that the groups are equal on all observed or unobserved factors because the probability that a particular individual has to be assigned to the treatment and control group is equal. In this condition, the effect of *x* on *y* can be cleanly interpreted.

When *x* correlates with *e* (i.e., *x* is endogenous) then the modeler has a serious problem and what happens next is something very undesirable: In the process of satisfying the orthogonality assumption, the estimator (whether OLS or maximum likelihood) "adjusts" the slope, β_1 of *x*, accordingly. The estimate thus becomes inaccurate (because it has been changed to the extent that *x* correlates with *e*). In this case suppose that selection to treatment was not random and that the treatment group had more extraverts; thus, *x* will "correlate" with extraversion in these sense that the level of extraversion is higher in the treatment group and that this level is correlated with *y* too because extraverts are usually more effective as leaders. Now because extraversion has not been measured, *x* will correlate with *e* (i.e., all omitted causes of *y* that are not expressly modeled). The higher the correlation of *x* with *e* the more inaccurate (inconsistent) the estimate will be. In such conditions, finding a significant relation between *x* and *y* is *completely useless*; the estimate is not accurate because it includes the effects of unmeasured causes, and having a sample size approaching infinity will not help to correct this bias. The estimate not only includes the effect of *x* on *y* but also all other unobserved effects that correlate with *x* and predict *y* (and thus the coefficient could be biased upwards or downwards)!

We cannot stress how important it is to satisfy the orthogonality assumption because not only will the coefficient of the problematic variable be inconsistent; any variables correlating with the problematic variable will also be affected (because their estimate will also be adjusted by the regression procedure to the extent that they correlate with the problematic variable). Refer to Fig. 1, which demonstrates these points graphically as path models (we explain this problem in more detail later using some basic algebra).

In a randomized field experiment, causal inference is assured (Shadish et al., 2002); that is, it is very unlikely that there could be any confounds. The correlation of the treatment to the outcome variable must be due to the treatment and nothing else. Because subjects were randomly assigned to conditions, the characteristics of subjects (on the average) are approximately equal across conditions, whether they are measured or unmeasured characteristics; any differences that might be observed will be due to chance (and hence very unlikely). Having, subjects that are approximately the same in the treatment and control groups occur allows for solid conclusions and counterfactuals. If there is a change in *y* and this change is reliably (statistically) associated with the manipulated variable *x* then *nothing else* could possibly have provoked the change in *y* but the treatment. Thus, in a randomized field experiment, the *selection process* to treatment groups is correctly modeled (it is random) and the model is estimated in accordance with the assumptions of the OLS estimator (i.e., given the random assignment, the correlation of *x* with *e* is truly zero). In other words, the assumption that OLS makes about selection is met by random assignment to treatment.

As we discuss later, if there has been systematic selection to treatment or any other reason that may affect consistency then estimates could still be consistent *if* the appropriate methodological safeguards are taken. Note, that there is one situation in

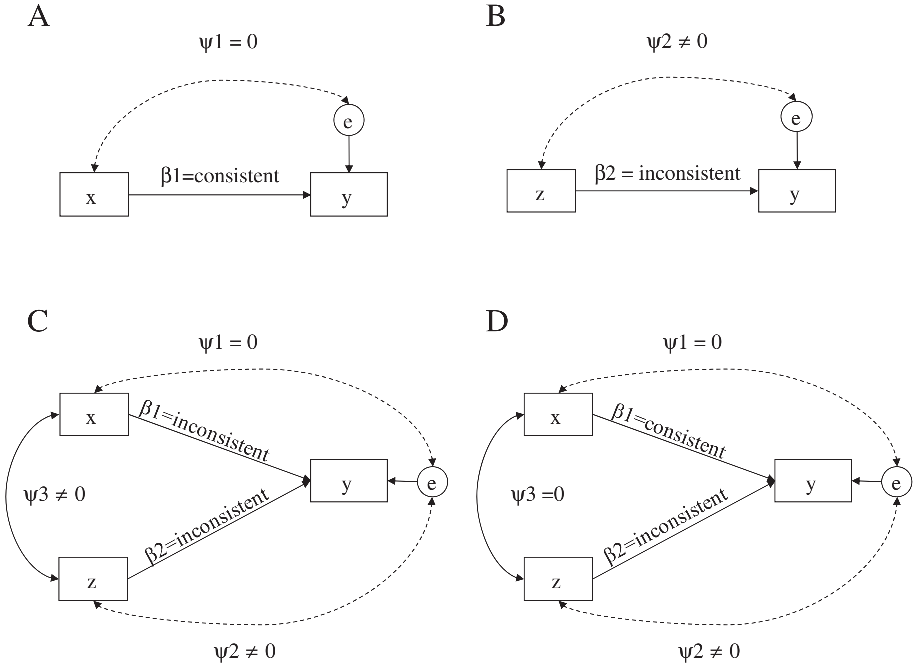

Fig. 1. How endogeneity affects consistency. A: β_1 is consistent because x does not correlate with e. B: β_2 is inconsistent because z correlates with e. C: Although x is exogenous β_1 is inconsistent because z, which correlates with e correlates with x too and thus "passes-on" the bias to x. D: β_1 is consistent even if β_2 is not consistent because x and z do not correlate (though this is still not an ideal situation because β_2 is not interpretable; all independent variables should be exogenous or corrected for endogeneity bias).

experimental work where causality can be confounded, which would be in the case where the modeler attempts to link the manipulation (x) to a mediator (m) in predicting y as follows: $x -> m -> y$. In this case, the mediator is endogenous (the mediator is not randomly assigned and it depends on the manipulation; thus m cannot be modeled as exogenous). This model can *only* be correctly estimated using the two-stage least squares procedure we describe later; the widely used procedure recommended by Baron and Kenny (1986), which models the causal mechanism by OLS will actually give *biased* estimates because it models the mediator as exogenous. We discuss this problem in depth later.

3. Why could estimates become inconsistent?

There are many reasons why x might be endogenous (i.e., correlate with e) thus rendering estimates inconsistent. We present five threats to what Shadish et al. (2002) referred to as "internal validity" (i.e., threats to estimate consistency). We introduce these five threats below (for a more exhaustive list of examples see Meyer, 1995); we then discuss the basic remedial action that can be taken. In the following section, we discuss techniques to obtain consistent estimates for more complicated models. We also address threats to inference (validity of standard errors) and model misspecification in simultaneous equation models. For a summary of these threats refer to Table 1.

3.1. Omitted variables

Omitted variable bias comes in various forms, including omitted regressors or omitted interaction terms or polynomial terms. We discuss the simplest case first and then more advanced cases later.

3.1.1. Omitting a regressor

Suppose that the correctly specified regression model is the following, and includes two exogenous variables (traits); y is leader effectiveness measured on some objective scale:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \tag{2}$$

Assume that a researcher wants to examine whether a new construct x (e.g., "emotional intelligence" measured as an ability) predicts leadership effectiveness. However, this construct might not be unique and suppose it shares common variance with IQ. Thus, the researcher should control for z (i.e., IQ) too, because x and z are correlated and, of course, because z predicts y as implied in the above model. Although one should also control for personality in the above equation, to keep things simple for the time

Table 1
The 14 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 statistical justification (i.e., Hausman test) (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) suffers 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-method 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 (i.e., multilevel hierarchical or longitudinal)
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: The 14 threats to validity mentioned are the criteria we used for coding the studies we reviewed.

being assume that both x and z are orthogonal to personality. Now, assume that instead of the previous model one estimated the following:

$$y_i = \varphi_0 + \varphi_1 x_i + v_i \tag{3}$$

This model now omits z; because x and z correlate and z also predicts y, x will correlate with v_i. In this case, instead of obtaining the unbiased estimate β_1 one obtains φ_1; these two estimates may differ significantly, as could be established by a what is referred to a Hausman (1978) test (see formula in Section 3.1.3). To see why these two estimates might not be the same, we use some basic algebra and express z as a function of x and its unique cause u. Note, the next equation does not necessarily have to be causal with respect to the relation between x and z; also, we omit the intercept for simplicity:

$$z_i = \gamma_1 x_i + u_i \tag{4}$$

Omitting z from Eq. (2) means that we have introduced endogeneity in the sense that x correlates with a new "combined" error term v_i. The endogeneity is evident when substituting Eq. (4) into Eq. (2):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2(\gamma_1 x_i + u_i) + e_i \tag{5a}$$

Multiplying out gives:

$$y_i = \beta_0 + \beta_1 x_i + \underbrace{(\beta_2\gamma_1 x_i + \beta_2 u_i + e_i)}_{v_i} \tag{5b}$$

Or, rearranging as a function of x gives

$$y_i = \beta_0 + (\beta_1 + \beta_2\gamma_1)x_i + (\beta_2 u_i + e_i) \tag{5c}$$

Whichever way we look at it, whereas the slope of x was correctly estimated in Eq. (2), it cannot be correctly estimated in Eq. (3) because as shown in Eq. (5c), the slope will include the correlation of x with z (i.e., γ_1). Thus, x correlates with the error term (as per Eq. (5b)) and is inconsistent. In the presence of omitted variable bias, one does not estimate β_1 as per Eq. (3), but something else (φ_1). Whether $\varphi_1$would go up or down when including z will depend on the signs of β_2 and γ_1. It is also clear that if $\beta_2 = 0$ or if $\gamma_1 = 0$ then v_i reduces to e_i and there is no omitted variable bias if z is excluded from the model.

Also, bear in mind that all other regressors that correlate with z and x will be inconsistent too when estimating the wrong model. What effect the regression coefficients capture is thus not clear when there are omitted variables, and this bias can increase or decrease remaining coefficients or change their signs!

Thus, it is important that all possible sources of variance in y that correlate with the regressor are included in the regression model. For instance, the construct of "emotional intelligence" has not been adequately tested in leadership or general work situations; one of the reasons is that researchers fail to include important control variables like IQ, personality, sex, age, and the like (Antonakis, 2009; Antonakis, Ashkanasy, & Dasborough, 2009; Antonakis & Dietz, 2010, in press a,b).

What if irrelevant regressors are included? It is always safer to err on the side of caution by including more than fewer control variables (Cameron & Trivedi, 2005). The regressors that should be included are ones that are theoretically important; the cost of including them is reduced efficiency (i.e., higher standard errors), but that is a cheap price to pay when consistency is at stake. Note, there are tests akin to Ramsey's (1969) regression-error-specification (RESET) test, which can be useful for testing whether there are unmodeled linearities present in the residuals by regressing y on the predicted value of polynomials of y (the 2nd, 3rd, and 4th powers) and the independent variables. This test is often incorrectly used as a test of omitted variables or functional form misspecification (Wooldridge, 2002); however, the test actually looks at whether the predicted value of y is linear given the predictors.

3.1.2. Omitting fixed effects

Oftentimes, researchers have panel data (repeated observations nested under an “entity”). Panel data can be hierarchical (e.g., leaders nested in firms; followers nested in leaders) or longitudinal (e.g., observations of leaders over time). Our discussion later is relevant to both types of panels, though we will discuss the first form, hierarchical panels (or otherwise known as pseudo-panels). If there are “fixed-effects” as in the case of having repeated observations of leaders (Level 1) nested in firms (Level 2), the estimates of the other regressors included in the model would be inconsistent if these fixed effects are not explicitly modeled (Cameron & Trivedi, 2005; Wooldridge, 2002). By fixed-effects, we mean the unobserved firm-invariant (Level 2) constant effects (or in the case of a longitudinal panel, the time-invariant panel effects) common to those leaders nested under a firm (we refer to these effects as u_j later, see Eq. (7)).

We discuss an example regarding the modeling of firm fixed-effects. By explicitly modeling the fixed effects (i.e., intercepts) using OLS, any possible unobserved heterogeneity in the level (intercept) of y common to leaders in a particular firm – which would otherwise have been pooled in e thus creating omitted variable bias – is explicitly captured. As such, the estimator is consistent if the regressors are exogenous. If the fixed effects correlating with Level 1 variables are not modeled, Level 1 estimates will be inconsistent to the extent that they correlate with the fixed effects (which is likely). What is useful is to conceptualize the error term e_{ij} in a fixed effects model as having two components: u_j, the Level 2 invariant component (that is explicitly modeled with fixed-effects), and e_{ij}, the idiosyncratic error component. To maintain a distinction between the fixed-effects model and the random-effects model, we will simply refer to the error term as e_{ij} in the fixed-effect model (given that the error u_j is considered fixed and not random and is explicitly modeled using dummy variables, as we show later).

Obtaining consistent estimates by including fixed-effects comes at the expense of not allowing any Level 2 (firm-level) predictors because they will be perfectly collinear with the fixed effects (Wooldridge, 2002). If one wants to add Level 2 variables to the model (and remove the fixed effects) then one must ensure that the estimator is consistent by comparing estimates from the consistent estimator to with the more efficient one, as we discuss in the next section.

Assume we estimate a model where we have data from 50 firms and we have 10 leaders from each firm (thus we have 500 observations at the leader level). Assume that leaders completed an IQ test (x) and were rated on their objective performance (adherence to budget), y. Thus, we estimate the following model for leader i in firm j:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{k=2}^{50} \beta_k D_{kj} + e_{ij} \tag{6}$$

The fixed effects are captured by 49 ($k-1$) dummy or indicator variables, D identifying the firms. Not including these dummy variables would be a big risk to take because it is possible, indeed likely, that the fixed effects are correlated with x (e.g., some firms may select leaders on their IQ) and they will most certainly predict variance in y (e.g., fixed effects would capture things like firm size, which may predict y). Thus, even though x is exogenous with respect to e_{ij} the coefficient of x will be consistent only if the dummies are included; if the dummies are not included then the e_{ij} term will include u_j and thus biases estimates of x. If the dummies are not included then the modeler faces the same problem as in the previous example: omitted-variable bias. Note, if x correlates with e_{ij}, the remedy comes using another procedure, which we discuss later when introducing two-stage least squares estimation.

Fixed-effects could be present for a number of reasons including group, microeconomic, macroeconomic, country-level, or time effects and researchers should pay more attention to these contextual effects because they can affect estimate consistency (Liden & Antonakis, 2009). Finally, when observations are nested (clustered), standard errors should not be estimated the conventional way (refer to the section later regarding consistency of inference).

3.1.3. Using random effects without meeting assumptions of the estimator

If the modeler wants to determine whether Level 2 variables predict y, the model could be estimated using the random-effects estimator. The random effects estimator allows for a randomly varying intercept between firms — this model is referred to as the “intercepts as outcomes” in multilevel modeling vernacular (Hofmann, 1997). Instead of explicitly estimating this heterogeneity via fixed effects, this estimator treats the leader level differences in y (i.e., the intercepts) as random effects between firms that are drawn from a population of firms and *assumed to be uncorrelated* with the regressors and the disturbances; the random effects are also assumed to be constant over firms and independently distributed. Failure to meet these assumptions will lead to inconsistent estimates and is tantamount to having omitted variable bias.

Also, prior to using this estimator, the modeler should test for the presence of random effects using a Breusch and Pagan Lagrangian multiplier test for random effects if the model has been estimated by GLS (Breusch & Pagan, 1980), or a likelihood-ratio test for random effects if the model has been estimated with maximum likelihood estimation (see Rabe-Hesketh & Skrondal, 2008); this is a chi-square with 1 degree of freedom and if significant, rules in favor of the random-effects model. We do not discuss the random-coefficients model, which is direct extension of the of the random-effects model and allows varying slopes across groups. Important to note is that before one uses such a model, one must test whether it is justified by testing the random-coefficients models versus the random-effects model (using a likelihood-ratio test); only if the test is significant (i.e., the assumption that the slopes are fixed is rejected) should the random-coefficients estimator by used (Rabe-Hesketh & Skrondal, 2008).

Now, the advantage of the random-effects estimator (which could simultaneously be its Achilles heel) is that then Level 2 variables can be included as predictors (e.g., firm size, public vs. private organization, etc.), in the following specification for leader *i* in firm *j*:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{k=1}^{q} \gamma_k z_{kj} + e_{ij} + u_j \tag{7}$$

In Eq. 7, we include regressors 1 to *q* (e.g., firm size, type, etc.) and omit the fixed-effects, but include a firm-specific error component, $u_{j.}$ The random effects estimator is more efficient than the fixed-effects estimator because it is designed to minimize the variance of the estimated parameters (loosely speaking it has fewer independent variables because it does not include all the dummies). But you guessed it; it comes with a hefty price in that it may not be *consistent* vis-à-vis the fixed-effects estimator (Wooldridge, 2002). That is, u_j might correlate with the Level 1 regressors. To test whether the estimator is consistent, one can use what is commonly called a "Hausman Test" (see Hausman, 1978) – this test, which is *crucial* to ensuring that the random-effects model is tenable – does not seem to be routinely used by researchers outside of econometrics, and not even in sociology, a domain that is close to economics (Halaby, 2004).

Basically, what the Hausman test does is to compare the Level 1 estimates from the consistent (fixed-effects) estimator to those of the efficient (random-effects estimator). If the estimates differ significantly, then the efficient estimator is inconsistent and the *fixed-effects estimator must be retained*; the inconsistency must have come from u_j correlating with the regressor. In this case estimates from the random-effects estimator cannot be trusted; our leitmotif in this case is consistency always trumps efficiency. The most basic Hausman test is that for one parameter, where δ is the element of β being tested (Wooldrige, 2002). Thus, the test examines whether the estimate of β of the efficient (RE) estimator differ significantly from that of the consistent (FE) estimator, using the following *t* test (which has an asymptotic standard normal distribution):

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

This test can be extended for an array of parameters. In comparing the fixed-effects to the random-effects estimator, an alternative to the Hausman test is the Sargan–Hansen test (Schaffer & Stillman, 2006), which can be used with robust or cluster-robust standard errors. Both these tests are easily implemented in Stata (see StataCorp, 2009), our software of choice. Again, because observations are nested (clustered), standard errors should not be estimated the conventional way (refer to the section later regarding consistency of inference).

One way to get around the problem of omitted fixed effects and to still include Level 2 variables is to include the cluster means of all Level 1 covariates in the estimated model (Mundlak, 1978). The cluster means can be included as regressors or subtracted (i.e., cluster-mean centering) from the Level 1 covariate. The cluster means are invariant within cluster (and vary between clusters) and allow for consistent estimation of Level 1 parameters just as if fixed-effects had been included (see Rabe-Hesketh & Skrondal, 2008). Thus, if the Hausman test is significant, we could still obtain consistent estimates of the Level 1 parameters with either one of the following specifications (given that the cluster mean will be correlated with the covariate but not with u_j):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \bar{x}_j + \sum_{k=1}^{q} \gamma_k z_{kj} + e_{ij} + u_j \tag{8}$$

$$y_{ij} = \delta_0 + \delta_1 (x_{ij} - \bar{x}_j) + \sum_{k=1}^{q} \varphi_k z_{kj} + w_{ij} + g_j \tag{9}$$

In the previous two equations, the interpretation of the coefficient of the cluster mean differs; that is, in Eq. (8) it refers to the difference in the between and within effects whereas in Eq. (9) it refers to the between effect (Rabe-Hesketh & Skrondal, 2008). In both cases, however, the estimate of β_1 or δ_1 is consistent (and equals that of the fixed-effects estimator, though the intercept will be different in the case of Eq. (9)). Note that if Level 2 variables are endogenous, the cluster-mean trick cannot help; however, there are ways to obtain consistent estimates by exploiting the exogenous variation in Level 2 covariates (see Hausman & Taylor, 1981).

3.1.4. Omitting selection

Selection refers to the general problem of treatment not being assigned randomly to individuals. That is, the treatment is endogenous. Assume:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \tag{10}$$

Here, x takes the value of 1 if the individual receives a treatment (e.g., attends a leadership-training program), else x is 0 (the individual has not received the treatment). Assume that y is how charismatic the individual is rated. However, assume that individuals have been *selected* (either self-selected or otherwise) to receive the training. That is, x, the binary variable has not been randomly assigned, which means that the groups might not be the same on the outset on observed or unobserved factors and these factors could be correlated with y and of course x. Thus, the problem arises because x is explained by other factors (i.e., the selection can be predicted) that are not observed in Eq. (10), which we refer to as x^* (subsumed) in e. That is, assume x^* is modeled in the following probit (or logistic) equation (Cong & Drukker, 2001)

$$x_i^* = \gamma_0 + \sum_{k=1}^{q} \gamma_k z_{kj} + u_i \tag{11}$$

Where k refers to regressors 1 to q and u to a disturbance term. We observe $x=1$ when $x^*>0$ (i.e., treatment has been received), else $x=0$. The problem arises because u will be correlated with e (this correlation is called $\rho_{e,u}$) and thus x will be correlated with e.

As an example, suppose that individuals who have a higher IQ (as well as some other individual differences that correlate with leadership) are more likely to attend the training; it is also likely, however, that these individuals are more charismatic. Thus, there are unmodeled sources of variance (omitted variables) subsumed in e that correlate with x. Suppose that one is highly motivated, so u_i is high, and attends training. If motivation is also correlated with charisma, then e_i will be higher too; hence the two disturbances correlate. As it is evident, this problem is similar to omitted variable bias in the sense that there are excluded variables, pooled into the error term, that correlate with endogenous choice variable and the outcome (Kennedy, 2003). If the selection is not explicitly (and correctly modeled), then using untreated individuals to estimate the counterfactual is misleading: they differ from treated individuals with respect to things we do not know; that is, the counterfactuals are missing (and the effect of the treatment will be incorrectly estimated).

Although this problem might seem unsolvable, it is not; this model can be estimated correctly if this selection process is explicitly modeled (Cong & Drukker, 2001; Maddala, 1983). In fact, for a related type of model where y is only observed for those who received treatment (see Heckman, 1979), James Heckman won the Nobel Prize in Economics! We discuss how this model is estimated later.

Another problem that is somewhat related to selection (but has nothing to do with selection to treatment) is having non-representative selection to participation in a study or censored samples (a kind of missing-data problem). We briefly discuss the problem here and suggested remedies, given that the focus of our paper is geared more towards selection problems. The problem of nonrepresentativeness has to do with affecting the observed variability, which thus attenuates estimates. Range restriction would be an example of this problem; for example, estimating the effect of IQ on leadership in a sample that is high on IQ will bias the estimate of IQ downwards (thus, the researcher must either obtain a representative sample or correct for range restriction). Another example would be using self-selected participants for leadership training (where participants are then randomly assigned to treatment); in this case, it is possible that the participants are not representative of the population (and only those that are interested in leadership, for example, volunteered to participate). Thus, the researcher should check whether the sample is representative of the population. Also, consider the case where managers participate in a survey and they select the subordinates that will rate them (the managers will probably select subordinates that like them). Thus, ideally, samples must be representative and random (and for all types of studies, whether correlational or testing for group differences); if they are not, the selection process must be modeled. Other examples of this problem include censored observations above or below a certain threshold (which creates a missing-data problem on the dependent variable). Various remedies are available in such cases, for example, censored regression models (Tobin, 1958) or other kinds of truncated regression models (Long & Freese, 2006) depending on the nature of the problem at hand.

3.2. Simultaneity

This problem is one that is tricky and which has given many economists and other social scientists a “headache”. Suppose that x causes y and this relation should be negative; you regress y on x but to your surprise, you find a non-significant relation (or even a positive effect). How can this be? If y also causes x it is quite possible that their covariation is not negative. Simultaneity has to do with a two variables simultaneously causing each other. Note, this problem is not necessarily the supposed simplistic “backward causality” problem often evoked by researchers (i.e., that the positive regression coefficient of x on y could be due to y causing x); it has to do with simultaneous causation, which is a different sort of problem.

Here is a simple example to demonstrate the simultaneity problem: Hiring more police-officers (x) should reduce crime (y), right? However, it is also possible too that when crime goes up, cities hire more police officers. Thus, x is not exogenous and will necessarily correlate with e in the y equation (see Levitt, 1997; Levitt, 2002). To make this problem more explicit, assume that x is a particular leadership style (use of sanctions) and y is follower performance (and we expect the relation, as estimated in β_1, to be negative):

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{12}$$

Because leader style is not randomly assigned it will correlate with e_i making β_1 inconsistent. Why? For one, leaders could also change their style as a function of followers' performance, leading to Eq. (13).

$$x_i = \gamma_1 y_i + u_j \tag{13}$$

We expect γ_1, to be positive. Now, because we do not explain y perfectly, y varies as a function of e too; y could randomly increase (e.g., higher satisfaction of followers because of a company pay raise) or decrease (e.g., an unusually hot summer). Suppose y increases due to e; as a consequence x will vary; thus, e affects x via y in Eq. (13). In simple terms e correlates with x, rendering β_1 inconsistent. Instrumental-variable estimation can solve this problem, as we discuss later.

3.3. Measurement error (errors-in-variables)

Suppose we intend to estimate our basic specification, however, this time what we intent to observe is a latent variable, x^*:

$$y_i = \beta_0 + \beta_1 x_i^* + e_i \tag{14}$$

However, instead of observing x^*, which is exogenous and a theoretically "pure" or latent construct, we observe instead a not-so-perfect indicator or proxy of x^*, which we call x (assume that x^* is the IQ of leader i). This indicator consists of the true component (x^*) in addition to an error term (u) as follows (see Cameron & Trivedi, 2005; Maddala, 1977):

$$x_i = x_i^* + u_i \tag{15a}$$

$$x_i^* = x_i - u_i \tag{15b}$$

Now substituting Eq. (15b) into Eq. (14) gives:

$$y_i = \beta_0 + \beta_1 (x_i - u_i) + e_i \tag{16}$$

Expanding and rearranging the terms gives:

$$y_i = \beta_0 + \beta_1 x_i + (e_i - \beta_1 u_i) \tag{17}$$

As is evident, the coefficient of x will be inconsistent given that the full error term, which now includes measurement error too, is correlated with x. Note that measurement error in the y variable does not bias coefficients and is not an issue because it is absorbed in the error term of the regression model.

The above discussion concerns a special kind of omitted variable bias because by estimating the model only with x, we omit u from the model; given that u is a cause of x creates endogeneity of the sort that x correlates with the combined error term in Eq. (17). This bias attenuates the coefficient of x, particularly in the presence of further covariates (Angrist & Krueger, 1999); the bias will also taint the coefficients of other independent variables that are correlated with x (Bollen, 1989; Kennedy, 2003) — refer to Antonakis (2009) for a example in leadership research, where he showed that "emotional intelligence" was more strongly related to IQ and the big five than some have suggested (which means that failure to include these controls and failure to model measurement error will severely bias model estimates, e.g., see Antonakis & Dietz, in press-b, Fiori & Antonakis, in press). In fact, using error-in-variables (with maximum likelihood estimation, given that he reanalyzed summary data), Antonakis (2009) showed that "emotional intelligence" measures were linearly dependent on the big five and intelligence, with multiple rs ranging between .48 and .76 depending on the measures used. However, this relation was *vastly underestimated* when ignoring multivariate effects and measurement error, leading to incorrect inference.

The effect of measurement error can be eliminated with a very simple fix: by constraining the residual variance of x to $(1 - \text{reliability}_x) * \text{Variance}_x$ (Bollen, 1989); if reliability is unknown, the degree of "validity" of the indicator can be assumed from theory and hence the residual is constrained accordingly (Hayduk, 1996; Onyskiw & Hayduk, 2001). What the modeler needs is a reasonably good estimate for the reliability (or validity) of the measure. If x were a test of IQ, for example, and we have good reason to think that IQ is exogenous as we discuss later (see Antonakis, in press), a reasonable estimate could be the test–retest reliability or the Cronbach alpha internal consistency of estimate of the scale. Otherwise, theory is the best guide to how reliable the measure is. Using this technique is very simple in the context of a regression model with a program like Stata and its eivreg (errors-in-variables) routine or most structural equation modeling programs using maximum likelihood estimation (e.g., Mplus). The advantage of using a program like Stata is that the eivreg least-squares estimator does not have the computational difficulties, restrictive assumptions, and sample size requirements inherent to maximum likelihood estimation so it is useful with single indicator (or index) measures (e.g., see Bollen, 1996; Draper & Smith, 1998; Kmenta, 1986); for multi-item measures of a latent variable a structural equation modeling program must be used.

Later, we also discuss a second way to fix the problem of measurement error – particularly if the independent variable correlates with e for other reasons beyond measurement error – using two-stage least squares regression.

3.4. Common source, common-method variance

Related to the previous problem of measurement error is what has been termed common-method variance. That *x* causes *y* could be because they both depend on *q*. For example, suppose raters rate their leaders on their leader style (*x*) and raters are simultaneously asked to provide ratings on the leaders' effectiveness (*y*); given that a common source is being used it is quite likely that the source (i.e., rater) will strive to maintain consistency between the two types of ratings (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Podsakoff & Organ, 1986) — suppose due to *q*, which could reflect causes including halo effects from the common source (note, a source could also be a method of data gathering). Important to note is that the common source/method problem does not only inflate estimates as most researchers believe; it could bias them upwards as well as downwards as we will show later. As will be evident from our demonstrations, common-method variance is a very serious problem and we disagree in the strongest possible terms with Spector (2006) that effects associated with common-method variance is simply an "urban legend."

Although Podsakoff et al. (2003) suggested that the common-method variance problem biases coefficients, they did not specifically explain why the coefficient of *x* predicting *y* can be biased upwards or downwards. To our knowledge, we make this demonstration explicit for the first time (at least as far as the management and applied psychology literature is concerned). We also provide an alternative solution to deal with common-method variance (i.e., two-stage least squares, as discussed later), particularly in situations where the common cause cannot be identified. An often-used remedy for common-methods variance problem is to obtain independent and dependent variables from different sources or different times, a remedial action which we find satisfactory as long as the independent variables are exogenous. In the case of split-sample designs where half the raters rate the leader's style and the other half the leader's effectiveness (e.g., Koh, Steers, & Terborg, 1995) precision of inference (i.e., standard errors) will be reduced particularly if the full sample is not large. Also, splitting measurement occasions across different time periods still does not fully address the problem because the common-method variance problem could still affect the independent variables that have been measured from the common source (refer to the end of this section).

One proposed way to deal with this problem is to include a latent common factor in the model to account for the common (omitted) variance (Loehlin, 1992; Podsakoff et al., 2003, see Figure 3A in Table 5; Widaman, 1985). Although Podsakoff, et al. suggested this method as a possible remedy and cited research that has used it as evidence of its utility, they noted that this method is limited in its applicability. We will go a step further and suggest that this procedure should never be used. As we will show, one cannot remove the common bias with a latent method factor because the modeler does not know how the unmeasured cause affects the variables (Richardson, Simmering, & Sturman, 2009). It is impossible to estimate the exact effect of the common source/method variance without directly measuring the common source variable and including it in the model in the correct causal specification.

Suppose that an unmeasured common cause, degree of organizational safety and risk, affects two latent variables, as depicted in Fig. 2; this context of leadership is one where team members are exposed to danger (e.g., oil rig). We generated data for a model where Ξ1 and Ξ2 measure subordinate ratings of a leader's style (task and person-oriented leadership respectively). The effect of the cause on Ξ1 is positive (.57), that is, in a high-risk situation the leader is very task-oriented because in these situations, violation of standards could cost lives; however, for Ξ2 the effect of the common cause is negative (−.57), that is, in high-risk situations, leaders pay less attention to being nice to subordinates. Thus, leadership style is endogenous; this explanation should make it clear why leader style can never be modeled as an independent variable. When controlling for the common cause the residual correlation between Ξ1 and Ξ2 is zero. The data are such that the indicators of each respective factor are tau equivalent (i.e., they have the same loadings on their respective factors) and with strong loadings (i.e., all λs are .96 and are equal on their respective factors). We made the models tau equivalent to increase the likelihood that the model is identified when introducing a latent common method/source factor. The sample size is 10,000, and the model fits the data perfectly, according to the overidentification test: $\chi^2(31) = 32.51$, $p > .05$ (as well as to adjunctive measures of fit – CFI = 1.00, RMSEA = .00 – which we do not care much for as we discuss later). Estimating the model without the common cause gives a biased structural estimate (a correlation of −.32 between the two latent variables), although the model fits perfectly: $\chi^2(25) = 28.06$, $p > .05$ (CFI = 1.00, RMSEA = .00); hence, it is important of theoretically establishing if modeled variables are exogenous or not because a misspecified model (with endogeneity) could still pass a test of fit. Finally, when including a latent common factor to account for the supposed common-cause effects, the model still fits well: $\chi^2(17) = 20.32$, $p > .05$ (CFI = 1.00, RMSEA = .00). However, the loadings and the structural parameter are severely biased. This method, which is very popular with modelers, is obviously not useful; also, as is evident, this misspecification is not picked up with the test of model fit. The correct model estimates could have been recovered when using instrumental variables (we present this solution later for the simple case of a path model and then extend this procedure to a full structural-equation model).

We first broaden Podsakoff et al.'s (2003) work to show the exact workings of common-method bias, and then present a solution to the common-method problem. We start with our basic specification, where a rater$_i$ has rated leader$_j$ ($n = 50$ leaders) on leader style *x* and leader effectiveness *y*, where we control for the fixed effects of firm (note, the estimator should be a robust one for clustering, as discussed later; also, assume in the following that we do not have random effects):

$$y^*_{ij} = \beta_0 + \beta_1 x^*_{ij} + \sum_{k=2}^{50} \beta_k D_{jk} + e_{ij} \tag{18}$$

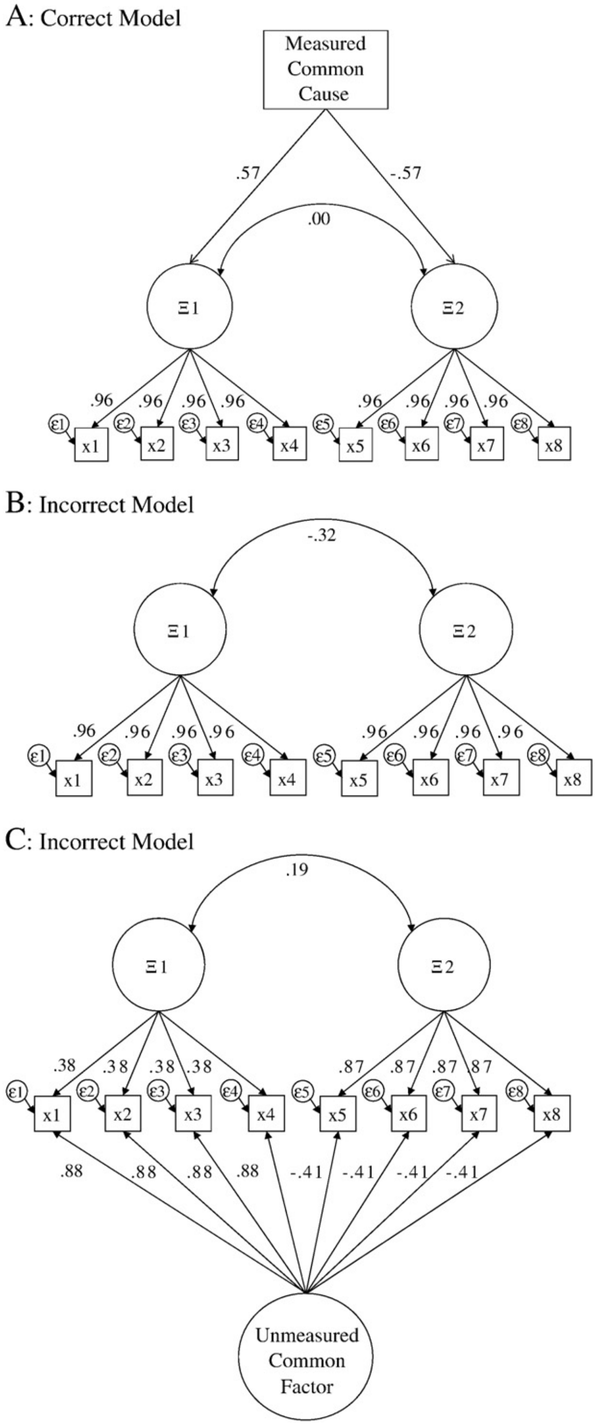

Fig. 2. Correcting for common-source variance: the common method factor fallacy (estimates are standardized). A: this model is correctly specified. B: failing to include the common cause estimates the correlation between Ξ1 and Ξ2 incorrectly (−0.32). C: including an unmeasured common factor estimates the loadings (which are also not significant for Ξ1) and the correlation between Ξ1 and Ξ2 (0.19, not significant) incorrectly.

Similar to the case of measurement error we cannot directly observe y^* or x^*; however what we do observe is y and x in the following respective equations (where q_i is the common bias):

$$y_{ij} = y^*_{ij} + \gamma_y q_{ij} \tag{19}$$

$$x_{ij} = x^*_{ij} + \gamma_x q_{ij} \tag{20}$$

Rearranging the equations gives:

$$y^*_{ij} = y_{ij} - \gamma_y q_{ij} \tag{21}$$

$$x^*_{ij} = x_{ij} - \gamma_x q_{ij} \tag{22}$$

Substituting Eqs. (21) and (22) into Eq. (18) shows the following:

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

Rearranging the equation gives:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{k=2}^{50} \beta_k D_{jk} + (e_{ij} - \beta_1 \gamma_x q_{ij} + \gamma_y q_{ij}) \tag{24}$$

As with measurement error, common-method variance introduces a correlation between x and the error term, which now consists of *three* components (and cannot be eliminated by estimating the fixed effects). Unlike before with measurement error, however, the bias can attenuate or accentuate the coefficient of x. Furthermore, it is now clear that this bias cannot be eliminated unless q is directly measured (or "instruments" are used to purge the bias using two-stage least-squares estimation).

Thus, as we alluded to previously, the problem is not one of inflation of variance of coefficients; it is one of *consistency* of coefficients. The coefficient β_1 is uninterpretable because it includes the effect of q on x and y. Assuming that the researcher has no other option but to gather data in a common-source way, and apart from measuring and including q directly in the model, which may be difficult to do because q could reflect a number of causes, there is actually a rather straightforward solution to this problem and one that, to our knowledge, will be presented for the first time to leadership, management, and applied-psychology researchers. This solution has been available to econometricians for quite some time, and we will discuss this solution in the section on two-stage least squares estimation.

Note, assume the case where only the independent variables (e.g., assume x_1 and x_2) suffer from common-method variance; in this case, the estimates of the two independent variables will be biased to zero and be inconsistent (though their relative contribution, $\frac{\beta_{x1}}{\beta_{x2}}$ is consistent), which can be shown as follows. Suppose:

$$y = \beta_0 + \beta_1 x^*_1 + \beta_2 x^*_2 + e \tag{25}$$

Instead of observing the latent variables x^*_1 and x^*_2, we observe x_1 and x_2, which are assumed to have approximately the same variance and are both equally dependent on a common variable q. Thus, by substitution it can be shown that both estimates will be biased downwards but equally so, suggesting their relative contribution will remain consistent:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (e - \beta_1 \gamma_q - \beta_2 \gamma_q) \tag{26}$$

3.5. Consistency of inference

We finish this section by bringing up another threat to validity, which has to do with inference. From a statistical point-of-view we mean whether the standard errors are consistent. There has been quite a bit of research on this area particularly after the papers by Huber (1967) and White (1980); this work is extremely technical so we will just provide a short overview of its importance and remedial action that can be taken to ensure correct standard errors.

In a simple experimental setting, regression residuals will usually be i.i.d. (identically and independently distributed). By identically distributed we mean that residuals are homoscedastic, that is, they have been drawn from the same population and have a uniform variance. By independently distributed we mean that they are not clustered or serially correlated (as when observations are nested under a Level 2 entity). It is always a good idea, however, to check whether residuals are homoscedastic. Whether they are clustered is certainly evident from the data-gathering design. Programs like Stata have nice routines for checking for heteroscedasticity, including White's test, and for the presence of clustering.

If residuals are heteroscedastic, coefficient estimates will be consistent; however, *standard errors will not be*. In this context, the variance of the parameters has to be estimated differently as the usual assumptions do not hold. The variance estimator is based on the work of Huber–White, and the standard errors are usually referred to as Huber–White standard errors, sandwiched standard errors, or just robust standard errors. We cannot stress the importance of having the standard errors correctly estimated (either with a robust variance estimator or using bootstrapping) and this concern is really not on the radar screen of researchers in our field. Consistent standard errors are just as important as consistent estimates. If standard errors are not correctly estimated, *p*-values will be over or understated, which means that results could change from significant to non-significant (and vice-versa); refer to Antonakis and Dietz (in press-a) for an example.

Similar to the previous problem of heteroscedasticity, is the problem of standard errors from clustered observations. A recent paper published in a top economics journal blasted economists for failing to correctly estimate the variance of the parameters and suggested that many of the results published with clustered data that had not corrected for the clustering were dubious (see Bertrand, Duflo, & Mullainathan, 2004), and this in a domain that is known for its methodological rigor! The variance estimator for clustered data is similar in form to the robust one but relaxes the assumptions about the independence of residuals. Note that at times, researchers have to correct standard errors for multiple dimensions of clustering; that is, we are not discussing the case of hierarchically clustered but truly independently clustered dimensions (see Cameron, Gelbach, & Miller, in press). Again, these corrections are easily achieved with Stata or equivalent programs.

4. Methods for inferring causality

To extend our discussion regarding how estimates can become inconsistent, we now review methods that are useful for recovering causal parameters in field settings where randomization is not possible. We introduce two broad methods of ensuring consistent estimates. The first is what we refer to as statistical adjustment, which is only possible when all sources of variation in y are known and are observable. The second way we refer to as quasi-experimentation: Here, we include simultaneous equation models (with extensive discussion on two-stage least squares), regression discontinuity models, difference-in-differences models, selection models (with unobserved endogeneity), and single-group designs. These methods have many interesting and broad applications in real-world situations, where external validity (i.e., generalizability) is assured, but where internal validity (i.e., experimental control) is not easily assured. Given space constraints, our presentation of these methods is cursory; our goal is to introduce readers to the intuition and the assumptions behind the methods and to explain how they can recover the causal parameter of interest. We include a summary of these methods in Table 2.

4.1. Statistical adjustment

The most simple way to ensure that estimates are consistent is to measure and include all possible sources of variance of y in the regression model (cf. Angrist & Krueger, 1999); of course, we must control for measurement error and selection effects if relevant. Controlling for all sources of variance in the context of social science, though, is not feasible because the researcher has to identify everything that causes variation in y (so as to remove this variance from e). At times, there is unobserved selection at hand or other causes unbeknown to the researcher; from a practical point-of-view, this method is not very useful *per se*. We are not suggesting that researchers must not use controls; on the contrary, all known theoretical controls must be included. However, it is likely that researchers might unknowingly (or even knowingly) omit important causes, so they must also use other methods to ensure consistency because of possible endogeneity.

4.1.1. Propensity score analysis (PSA)

Readers should refer back to Eqs. (10) and (11) so as to understand why PSA could recover the causal parameter of interest and thus approximate a randomized field experiment (Rubin, 2008; Rubin & Thomas, 1996). PSA can only provide consistent estimates to the extent that (a) the researcher has knowledge of variables that predict whether an individual *would* have received treatment or not, and (b) e and u in Eqs. (10) and (11) do not correlate. If e and u correlate, which may often be the case, a Heckman treatment effects model must be used to derive consistent estimates (discussed later).

The idea behind PSA is quite simple and has to do with comparing treated individuals to similar control individuals (i.e., to "recreate" the counterfactual). Going back to the randomized experiment: What is the probability, or propensity to provide an introduction to the term, that a particular individual is in the treatment versus the control group? If the treatment is assigned randomly, it is .50 (i.e., 1 out of 2). However, this probability is not .50 is the treatment was not assigned randomly. Thus, the essence of PSA is to determine the probability (propensity) that an individual would have received treatment (as a function of measured covariates). Then, the researcher attempts to compare (match) individuals from the treatment and control groups who

Table 2
Six methods for inferring causality in non-experimental settings.

Method	Brief description
1. Statistical adjustment	Measure and control for all causes of y (impractical and not recommended)
2. Propensity score analysis	Compare individuals who were selected to treatment to statistically similar controls using a matching algorithm
3. Simultaneous-equation models	Using "instruments" (exogenous sources of variance that do not correlate with the error term) to purge the endogenous x variable from bias.
4. Regression discontinuity	Select individuals to treatment using a modelled cut-off.
5. Difference-in-differences models	Compare a group who received an exogenous treatment to a similar control group over time
6. Heckman selection models	Predict selection to treatment (where treatment is endogenous) and then control for unmodeled selection to treatment in predicting y.

have the same probability of receiving treatment. In this way, the design mimics the true experiment (and the counterfactual), given that the researcher attempts to determine the treatment effect on y by comparing individuals who received the treatment to similar individuals who did not.

Suppose in our example that we want to compare individuals who undertook leadership training (and were self-selected) versus a control group. In the first instance, we estimate a probit (or logistic) model to predict the probability that an individual receives the treatment:

$$x_i^* = \gamma_0 + \sum_{k=1}^{q} \gamma_k z_{ki} + u_i \tag{27}$$

And $x=1$ when $x^*>0$ (i.e., treatment has been received), else $x=0$. The predicted probability of receiving the treatment (x) as a function of q covariates (e.g., IQ, demographics, and so forth) for each individual (i) is saved. This score, which ranges from 0 to 1, is the propensity score. The point is to match individuals in the treatment and control with the same propensity scores. That is, suppose two individual having the same (or almost the same) propensity score but one is in the treatment group and the other in the control group. What they differ on is what is captured by the error term in the propensity equation: u_i. That is, what explains, beyond the covariates, whether a particular individual should have received the treatment but did not is the error term. In other words, if two subjects have the same propensity score and they are in different groups (assuming that u_i is just "noise"), then it is almost like these two subject were randomly assigned to the treatment (D'Agostino, 1998). As mentioned, the assumption of this method is that u_i is unrelated to the residual term e_i (in Eq. (10)); the unobserved factors which explain whether someone received the treatment must not correlate with unobserved factors in the y equation (Cameron & Trivedi, 2005). If this assumption is tenable, then in the simplest case we can match individuals to obtain the counterfactual. That is, a simple t-test for the individuals matched across the two groups using some matching algorithm (rule) will indicate the average treatment effect. For more information on this method, readers should consult more detailed exposés and examples (Cameron & Trivedi, 2005; Cook et al., 2008; D'Agostino, 1998; Rosenbaum & Rubin, 1983, 1984, 1985).

4.2. Quasi-experimentation

Next, we introduce quasi-experimental methods, focusing extensively on two methods that are able to recover causal parameters in rather straightforward ways: Simultaneous equation models, and regression discontinuity models. We discuss two other methods, which have more restrictive assumptions (e.g., difference-in-differences models, selection models) but which are also able to establish causality if these assumptions are met. We complete our methodological journey with a brief discussion on single group, quasi-experimental designs.

4.2.1. Simultaneous-equation models

We begin the explanations in this section with two-stage least squares (2SLS) regression. This method, also referred to also as instrumental-variable estimation, is used to estimate simultaneous equations where one or more predictors are endogenous. 2SLS is standard practice in economics – a workhorse – and probably the most useful and most-used method to ensure consistency of estimates threatened by endogeneity. Unfortunately, beyond economics, this method has not had a big impact on other social science disciplines including psychology and management research (see Cameron & Trivedi, 2005; Foster & McLanahan, 1996; Gennetian et al., 2008). We hope that our review will help correct this state of affairs particularly because this approach can be useful for solving the common-method variance problem.

The 2SLS estimator (or its cousin, the Limited-Information Maximum Likelihood estimator, LIML) is handy for a variety of problems where there is endogeneity because of simultaneity, omitted variables, common-method variance, or measurement error (Cameron & Trivedi, 2005; Greene, 2008; Kennedy, 2003). This estimator has some commonalities with the selection models discussed later because it relies on simultaneous equations and instrumental variables. Instrumental variables, or simply instruments, are exogenous variables and do not depend on other variables or disturbances in the system of equations. Recall, the problem of endogeneity makes estimates inconsistent because the problematic (endogenous) variable – which is supposed to predict a dependent variable – correlates with the error term of the dependent variable. Refer to Fig. 3, for a simplified depiction of the problem and the solution, which we explain in detail later.

In our basic specification we will assume that x is continuous. The endogenous variable could be dichotomous too, in which case the 2SLS estimation procedure must employ a probit model in the first stage equation, that is, in Eq. (29) (Greene, 2008). Other estimators are available too for this class of model (e.g., where the y variable is a probit but the endogenous is a continuous variable). The Stata *cmp* command (Roodman, 2008) can estimate a broad class of such single-indicator mixed models by maximum likelihood similar to the Mplus structural-equation modeling program (L. K. Muthén & Muthén, 2007).

Turning back to the issue at hand, let us assume we have a common methods variance problem, where x (leader behavior) and y (perceptions of leader effectiveness) have been gathered from a common source: boss$_i$ rating leader$_i$ ($n=50$ leaders; with q representing unobserved common-source variance, and c control variables). Here, following Eq. (24) we estimate:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{c} \gamma_k f_{ik} + (e_i - \beta_1 \gamma_x q_i + \gamma_y q_i) \tag{28}$$

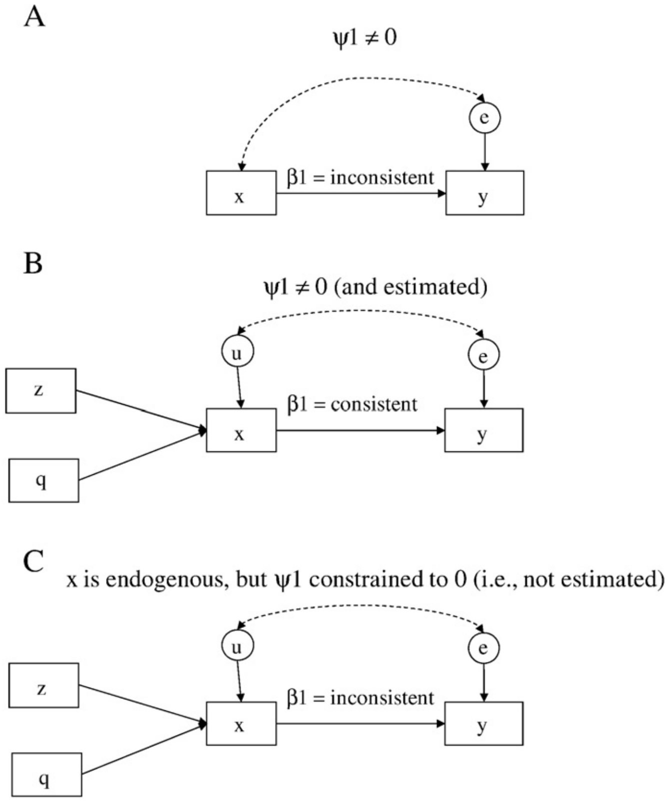

Fig. 3. Consistent estimation with a simultaneous equation (mediatory) model. A: β_1 is inconsistent because *x* correlates with *e*. B: β_1 is consistent because *z* and *d*, the instruments (which are truly exogenous), do not correlate with *e* (or *u* for that matter). C: β_1 is inconsistent because the common cause of *x* and *y*, which is reflected in the correlation between *e* and *u*, is not estimated (i.e., this is akin to estimating the system of equations using OLS, which ignores cross-equation correlations among disturbances).

The coefficient of *x* could be interpreted causally *if* an exogenous source of variance, say *z*, were found that strongly predicts *x* and is related to *y* via *x* only, and unrelated to *e* (the combined term). For identification of the parameters two conditions must be satisfied: We must have at least as many instruments as endogenous variables and one instrument must be excluded from the second-stage equation; also, the instruments should be significantly and strongly related to the endogenous variable *x* (Wooldridge, 2002). If we have more instruments than endogenous variables, then we can test the overidentifying restrictions in this system. If appropriate instruments are found, then the causal effect of *x* on *y* can be recovered by first estimating Eq. (29) (i.e., the first-stage equation) and then using the predicted value of *x* to predict *y*. Note, *all* exogenous variables (in this case *c*, the control variables) should usually be used as instruments of the endogenous variables, otherwise estimates may be inconsistent in certain conditions (for further information refer to Baltagi, 2002).

To illustrate the workings of 2SLS we use a theoretical example (later, we also demonstrate 2SLS with a simulated data set as well to should how solve the problem of common-method variance). Assume that *z* is IQ; given that IQ is genetically determined (i.e., has high genetic heritability, thus it is exogenous) it makes for an excellent instrument as would personality, and other stable individual differences (Antonakis, in press), as long as they do not correlate with omitted causes with respect to *y*. IQ affects how effectively a leader behaves (Antonakis, in press) and leader behavior affects leader outcomes (Barling, Weber, & Kelloway, 1996; Dvir, Eden, Avolio, & Shamir, 2002; Howell & Frost, 1986); note, these studies are not correlational but manipulated leadership. Also, the instruments must be related to *y* but less strongly than is the endogenous predictor. Assume that *d* is the distance of the rater from the leader (which is assigned by the company randomly), and which may impact how effective a leader can be with respect to that follower because it limits interaction frequency with followers (Antonakis & Atwater, 2002). We also include *c* control variables (e.g., leader age, leader sex, etc.). Thus, we model:

$$x_i = \gamma_0 + \gamma_1 z_i + \gamma_2 d_i + \sum_{k=1}^{c} \gamma_k f_{ik} + u_i \tag{29}$$

Because *z* and *d* (and *f*, which directly effects *y*) are exogenous, they will, of course not correlate with *u*, and more importantly with the error term in Eq. (28) (which consists of three components). Thus, the predicted value, x̂, will not correlate with the combined error term either. In the second stage we use x̂ to predict *y* as follows:

$$y_i = \lambda_0 + \lambda_1 \hat{x}_i + \sum_{k=1}^{c} \theta_k f_{ik} + e_i \tag{30}$$

How does 2SLS ensure consistency? What the 2SLS estimator does is very simple. Only the portion of variance that *z* and *d* (and the controls) predict in *x* that overlaps with *y* is estimated; given that *z* and *d* are exogenous, this portion of variance is isolated from the error term in the *y* equation (for an excellent intuitive explanation see Kennedy, 2003). Thus, the 2SLS estimate of β_1 is consistent, but less efficient than the OLS estimator given that less information is used to produce the estimate (Kennedy, 2003); this procedure must not be done manually but using specialized software to estimate the standard error correctly. The significance of each indirect (nonlinear) effect, that is $\gamma_1 * \beta_1$ and $\gamma_2 * \beta_1$ can also be tested using the traditional Sobel (delta) method (Sobel, 1982) or bootstrapping (Shrout & Bolger, 2002). Note that sums of indirect effects can be tested too in programs like Stata (i.e., $\gamma_1 * \beta_1 + \gamma_2 * \beta_1$).

What is very important to understand here about the estimation procedure is that, as we depict in Fig. 3, consistency can only occur if the cross-equation disturbances (*e* and *u*) are estimated. This procedure is standard practice in econometrics and the reason that it is done is quite straightforward. If the errors are not correlated then estimates will equivalent to OLS, which will be inconsistent if *x* is endogenous (Maddala, 1977). Estimating this correlation acknowledges the unmodeled common cause of *x* and *y*; it is a common unmeasured "shock" which affects both *x* and *y*, which must be included in the model. Failing to estimate it suggests that *x* is exogenous and does not require instrumenting.

How can we test if the errors *u* and *e* are correlated? The Hausman endogeneity test (see Hausman, 1978) or the Durbin–Wu–Hausman endogeneity test (or an augmented regression, wherein the residuals of the first-stage equation are included as a control in the second-stage equation) (see Baum, Schaffer, & Stillman, 2007) can tell us if the mediator is endogenous. Given that we have one endogenous regressor, this is a one degree of freedom chi-square test of the difference between the constrained model (the correlation of the disturbance is not estimated) and the unconstrained model (where the correlation of the disturbance is estimated); this procedure can be done in SEM programs. Thus, if the model where *x* is instrumented (the consistent estimator), generates a significantly different estimate from that where *x* is not instrumented (the OLS estimator, which is efficient), the OLS model must be rejected and *x* requires instrumenting.

A common mistake we see in management and applied psychology is the estimation of simultaneous equations without correlating the cross-equation disturbances as per the method suggested by Baron and Kenny (1986) or derivatives of this method. If the correlation is not estimated and if *x* is endogenous, then the estimate of β will change accordingly (and will not be consistent). Thus, most of the papers testing mediation models that have not correlated the disturbances of the two endogenous variables have estimates that are potentially biased. If, however, *x* is exogenous, then the system of equations could be estimated by OLS (or Maximum Likelihood) without correlating disturbances (refer to Section 4.2.1.4 for a specific example with data). This procedure we propose should not be confused with correlating disturbances of observed indicators in factor models, which addresses another issue to the one we discuss in mediation (or two-stage models). In principle, disturbances of indicators of factor models should not be correlated unless the modeler has *a priori* reason to do so (see Cole, Ciesla, & Steiger, 2007).

Systems of equations can be estimated using 2SLS, which is a limited information estimator (i.e., it uses information only from an "upstream" equation to estimate a "downstream" variable). This estimator is usually a "safe bet" estimator because if there is a misspecification in one part of the model and if the model is quite complicated with many equations, this misspecification will not bias estimates in other parts of the model as would full-information estimators like three-stage least squares (e.g., Zellner & Theil, 1962) or maximum likelihood, the usual estimator in most structural-equation modeling programs (Baltagi, 2002; Bollen, 1996; Bollen, Kirby, Curran, Paxton, & Chen, 2007). Thus, using a Hausman test, one could check whether the full-information estimator yields different model estimates (of the coefficients) from the limited-information estimator; if the estimates are significantly different, then the limited-information estimator must be retained (as long as the model fits).

4.2.1.1. Examining fit in simultaneous-equation models (overidentification tests). In the previous example, we can test whether the veracity of the model and the appropriateness of the instruments. For instance, one can examine whether the instruments are "strong" (Stock, Wright, & Yogo, 2002); these routines are implemented in the ivreg2 module of Stata (Baum et al., 2007). Also important, if not more important, is to test whether the overidentifying restrictions of the system of equations are viable (when having more instruments than mediators); this is a test of fit to determine whether there is a discrepancy between the implied and actual model. Essentially, what these tests examine is whether the instruments correlate with the residuals of the *y* equation. It should be now clear to readers that this undesirable situation is due to a model that is misspecified, which means that estimates are biased and cannot be interpreted. Thus, the model must fit before estimates can be interpreted.

In the previous example, Eq. (29) is overidentified (i.e., we have one more instrument that we do endogenous regressors); thus, the chi-square test of fit has 1 degree of freedom; if we had only one instrument, the model would be just-identified and a test of fit cannot be conducted (though the Hausman endogeneity test can still be done). In the context of regression models, these test of fit are chi-square tests and are usually called Sargan tests, Hansen–Sargan tests, or simply *J*-tests (see Basmann, 1960; Hansen, 1982; Sargan, 1958). These tests are direct analogs to the chi-square test of fit in the context of maximum likelihood estimation, as is usually the case with structural equation modeling software. A significant *p*-value for this test means that the model fails to fit (i.e., that the data rejected the model); this test is well-known in psychology and management but is often (and incorrectly so) ignored. Interestingly, economists pay attention to the test of fit. If it is significant, the model is no good, end of story (and one must refine the model or find better instruments); they do not use approximate indexes of fit, for instance the RMSEA (Browne & Cudeck, 1993), CFI (Bentler, 1990), or TLI (Tucker & Lewis, 1973), which are not statistical tests with known distributions (Fan & Sivo, 2005; Marsh, Hau, & Wen, 2004) or have arbitrary cut-offs, as in the case of RMSEA (Chen, Curran, Bollen, Kirby, & Paxton, 2008).

There are researchers (outside of economics) who are starting to seriously question the common practice in some social-sciences field of accepting models that fail the chi-square test of fit apparently because with a large sample even minute discrepancies will be detected and thus the *p*-value of the test will always be significant (see Antonakis, House, Rowold & Borgmann, submitted for publication; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; Kline, 2010; Marsh et al., 2004; McIntosh, 2007; Shipley, 2000). If the model is correct specified, it will not be rejected by the chi-square test even at very large samples sizes (Bollen, 1990); the chi-square test accommodates random fluctuations and "forgives" a certain discrepancy due to chance. Also, the chi-squared test is the most powerful test to detect a misspecified model, as Marsh et al. (2004) demonstrated in comparing the chi-square test to a variety of approximate fit indices. Thus, we urge researchers to pay attention to the chi-square test of fit and not to report failed models as acceptable.

Finally, it is essential to study samples that are causally homogenous (Mulaik & James, 1995); causally homogenous samples are not infinite (thus, there is a limit to how large the sample can be). Thus, finding sources of population heterogeneity and controlling for it will improve model fit whether using multiple groups (moderator models) or multiple indicator, multiple causes (MIMIC) models (Antonakis et al., submitted for publication; Bollen, 1989; B. O. Muthén, 1989).

4.2.1.2. The PLS problem. Researchers in some fields (particularly information systems and less so in some management subdisciplines) use what has been referred to as Partial-Least Squares (PLS) techniques to test path models or latent variable (particularly composite) models. We discuss this modeling method briefly, because it is quite popular in other fields yet PLS has no important advantages of regression or OLS. Because it seems to be slowly creeping into management research we feel it is important to warn researchers to not use PLS to test their models. PLS estimates are identical to OLS in saturated models with observed variables. Whether modeling composites in PLS or indexes/parcels in saturated regression models will not change estimates by much (Temme, Kreis, & Hildebrandt, 2006).

The problem with PLS, however, is that it cannot test systems of equations causally (i.e., overidentifying restrictions cannot be tested) nor can it directly estimate standard errors of estimates. Because the model's fit cannot be tested, the modeller cannot know if model estimates are biased. Also, its apparent advantages over regression-based (OLS and 2SLS) or covariance-based modeling (e.g., SEM) is rather exaggerated (see Hair, Black, Babin, Anderson, & Tatham, 2006; Hwang, Malhotra, Kim, Tomiuk, & Hong, 2010; Marcoulides & Saunders, 2006; McDonald, 1996); recently it has been also shown that PLS can experience convergence problems (Henseler, 2010). PLS users commonly repeat the mantra that “PLS is good for prediction, particularly in early phases of theory development whereas SEM models are good for theory testing;” this comment suggests that one cannot predict using SEM or 2SLS, which is obviously a baseless assertion. We really find it odd that those using PLS would knowingly not want to test their model when they could use more robust tests.

In a recent simulation PLS was found to perform worse than SEM (both in conditions of correct and misspecification); also, although a new approach, referred to as generalized structured component analysis, has been proposed as a better alternative to PLS (because it is similar to SEM in the sense that it can test model fit), it does not provide for better estimation when the model is correctly specified (Hwang et al., 2010). Interesting in this simulation is that the new method performed better under conditions of model misspecification (which makes sense given that it is a limited-information estimator); however, it is unclear as to whether this estimation approach is better than other limited-information (e.g., 2SLS) estimators (e.g., Bollen et al., 2007).

Other apparent advantages of PLS are that it makes no distributional assumptions regarding variables and does not require large sample sizes; however, regression or two-stage least squares analysis do not make any assumptions either about independent variables and can estimate models with small sample sizes. More importantly, there are estimators build into programs like MPlus, LISREL, EQS, and Stata that can accommodate a large class of models, using robust estimation and various types of variables, which might not be normally distributed or continuous (e.g., dichotomous, polytomous, ordered, counts, composite variables, etc.). Thus, given the advances that have been made today in statistics software, there is no use for PLS whatsoever (see in particular McDonald, 1996). We thus strongly encourage researchers to abandon it.

4.2.1.3. Finding instruments. Finally, one of the biggest challenges that researchers face when attempting to estimate instrumental variable models has to do with where to find instruments. In the case of an experiment, where the modeler wishes to establish mediation, the modeler will have the perfect instrument/s: the manipulated variables. As long as the model is estimated correctly (with the cross-equation disturbances of the endogenous variables correlated), then the causal mediation influence can be correctly identified. In the case of cross-sectional or longitudinal research, stable individual difference that are genetically determined could do (personality and cognitive ability), as would age, height, hormones (e.g., testosterone), or physical appearance (Antonakis, in press); geographic factors (distance from the leader as mentioned previously) could work. Time effects could be used as an exogenous source of variance as could exogenous “shocks” of from a particular event; there are contextual effects that could affect leadership, including laws or cultural-level factors (Liden & Antonakis, 2009). With panel data, fixed-effects of leaders or more simply cluster-means should also do the trick because they would capture all unobserved sources of variance in the leader that predict behavior (e.g., Antonakis et al., submitted for publication); this procedure will essentially purge rater *i*'s score from idiosyncratic bias, common-method bias, or other errors, given that the fixed-effect (i.e., the cluster-mean score) should mostly capture true variance (Mount & Scullen, 2001; Scullen, Mount, & Goff, 2000). Others have had ingenious ideas, estimating the effect of a change of leadership (presidents) on country-level outcomes using death in office as an exogenous source of variance (Jones & Olken, 2005); thus, the change of the handover of power is random (exogenous sources of variances such as this could be used to identify causal effects in two-stage models). Finding instruments is, at times, not easy; however, the time spent to find instruments is an investment that will serve science and society in good stead because the estimated parameters of the model will be consistent.

Important to note, once again, is that the instruments must not correlate with *e*, omitted causes. For instance, if an omitted common cause of leader style and effectiveness is affect for the leader and if leader IQ is used as an instrument, the modeler must be sure that affect for the leader and IQ do not correlate. If they do correlate, the model will be misspecified; however, misspecification could be caught by the overidentification test (as long as true exogenous variables, in addition to the "bad" instrument are included). Thus, it is crucial to try and obtain more instruments than endogenous variables so that the overidentification test can be performed. Also, the instruments must first pass a "theoretical overidentification" test before an empirical one (if all the modeled instruments are not truly exogenous the overidentification test will not necessarily catch the misspecification, as we have shown).

4.2.1.4. Solving the common-method variance problem with 2SLS. We provide two examples next; one where we show how to recover causal estimates with instrumental variable and 2SLS. The second example is a full SEM causal model, where we recover the causal estimates with instrumental variables using maximum likelihood estimation.

Example 1 using 2SLS: The previous discussion has been a theoretical one and readers might be skeptical about how the 2SLS estimator can recover causal estimates. We thus generated data with a known structure where there is a strong common-method variance effect. Assume that we have an endogenous independent variable *x*, a dependent variable *y*, two exogenous and perfectly-measured variables *m* and *n*, and a common source effect, *q*. The true model that generated the data is (note that *e* and *u* are normally distributed and independent of each other):

$$x = \alpha_0 + q + 0.8m + 0.8n + e \tag{31}$$

$$y = \beta_0 + q - 0.2x + u \tag{32}$$

We generated this data for a sample size of $n = 10{,}000$. Refer to Table 3 for the correlation matrix and sample statistics of this data (note, these summary data can be inputted into a structural-equation modeling program to derive the same estimates with maximum likelihood).

Estimating the OLS model (or using Maximum Likelihood), where *y* is simply regressed on *x*, clearly gives a *wrong* estimate with the *wrong* sign (.11); the true estimate (−.20) is 281.82% lower! Here is an example of the sinister effect of the common method variable, which when omitted from the equation makes *x* endogenous; as we mentioned, the biased OLS coefficient could be higher, lower, of a different sign or not significant. We trust it is now clear that Spector's (2006) suggestion that common-method variance is an urban legend *is an urgent legend in itself.*

The estimates of this model are depicted in Panel A of Table 4. The known-model estimates, based on two OLS equations (i.e., not correlating cross-equation disturbances, which is not needed because of sources of variance in the endogenous variables are accounted for) reproduce the correct estimate precisely (−.20), as indicated Panel B in Table 4. However, in the real-world this model would not be estimated because it is highly probably that the common cause, *q*, cannot be measured directly.

Thus, the only correct solution that is available to address this problem is one that is straightforward to use, provided the modeler has instruments. Using the 2SLS estimator, which exploits the exogenous sources of variance from *m* and *n*, recovers the true estimate (see Panel C in Table 4); the exogenous variables do not correlate with *q* (and thus not with *e* when *q* is not included in the equation) nor with *u* because they vary randomly. They are strongly related to *x* and only affect *y* via *x*. Next, even though *q* is not included in the model, the 2SLS estimator recovers the estimate of interest exactly (−.20), though with a slightly larger confidence interval; as we said before, the price that is paid is reduced efficiency. In the case of two-equation models, and with strong instruments, the 2SLS estimator gives similar estimates to three stage least squares (3SLS), iterated 3SLS, maximum likelihood (ML), and limited information ML (LIML).

To demonstrate the stability of the 2SLS estimate, a Monte Carlo simulation of this data structure based on 1000 simulations provided a mean estimate of −.20, with a 95% confidence interval of between −.2007859 and −.1992456!). Finally, a Sargan chi-square test of overidentification (Sargan, 1958) suggests that the instruments are valid, $p = .30$ (the simulation results confirmed this finding too, mean $p = .32$).

Now, had we estimated this previous model using the standard approach (irrespective of the estimator) that is usually used in management and applied psychology where the cross-equation disturbance are not correlated would have given an incorrect estimate (i.e., .11, which is, in fact, that of the OLS model); not estimating the cross-equation disturbance suggests that there is no "common shock" that might predict *x* and *y*, which is unmeasured and not accounted for in the model. That assumption is too strong to make, and as we demonstrate, incorrect in the context of such mediation models.

Table 3
Correlation matrix for 2SLS demonstration.

Variable	Mean	SD	*q*	*m*	*n*	*x*	*y*
q	−.01	1.00	1.00				
m	−.01	1.01	−.01	1.00			
n	.02	1.00	.00	−.01	1.00		
x	−.01	1.82	.55	.44	.45	1.00	
y	.00	1.32	.62	−.13	−.12	.15	1.00

$N = 10{,}000$.

Table 4
Estimates for 2SLS demonstration.

Independent variables	Coef.	Std. err.	*t*	*p*-value	95% conf. interval	
Panel A: OLS (dependent variable is *y*) $F(1, 9998) = 237.47$, $p < .001$, $r^2 = .02$						
x	.11	.01	15.41	.00	.10	.12
Constant	.01	.01	.26	.79	−.02	.03
Panel B Two-equation model estimated with OLS (dependent variable is *y*) $F(2, 9997) = 3927.65$, $p < .001$, $r^2 = .44$						
x	−.20	.01	30.34	.00	−.21	−.18
q	1.02	.01	86.26	.00	1.00	1.04
Constant	.01	.01	.83	.41	.01	.03
Two-equation model estimated with OLS (dependent variable is *x*) $F(3, 9996) = 7706.09$, $p < .001$, $r^2 = .70$						
q	1.00	.01	10.20	.00	.98	1.02
m	.80	.01	81.41	.00	.78	.82
n	.81	.01	81.33	.00	.79	.83
Constant	−.01	.01	−1.07	.29	−.03	.01
Panel C Simultaneous equation model estimated with 2SLS (dependent variable is *y*) $F(1, 9998) = 263.32$, $p < .001$, $r^2 = .02$[a]; Sargan overidentification $\chi^2(1) = 1.07$, $p = .30$						
x	−.20	.01	−16.23	.00	−.23	−.18
Constant	−.00	.01	−.02	.98	−.03	−.03
Simultaneous equation model estimated with 2SLS (dependent variable is *x*) $F(2, 9997) = 3985.68$, $r^2 = .44$						
m	.80	.01	56.93	.00	.77	.82
n	.82	.01	57.75	.00	.79	.84
Constant	−.02	.01	−1.35	.18	−.04	.01

N = 10,000.

[a] Note, it is possible that the *r*-square in the *y* equation in simultaneous equations models is undefined; however, this is not a problem in simultaneous equation models and structural estimates will be correct (Wooldridge, 2009). As a measure of regression fit, the predicted value of *y*, $\hat{y}$ can be correlated to the observed *y* and then squared (which is one way that *r*-square is calculated). We used this calculation for *r*-square in this model.

Example 2 using ML: The previous demonstration should now explain further that if the effect of a common source/method is not explicitly modeled, true parameter estimates cannot be recovered (e.g., by attempting to model a method factor, because *how* the method factor affects the variable is unknown to the researcher). Thus, one defensible statistical way to control for this problem is the way we have demonstrated previously, by using instrumental variables. The same procedures can be extended to full SEM models. We provide a brief example later, following a similar specification in Fig. 2 where we include a dependent variable *y*, presidential leader effectiveness, and two independent variables. All measures were obtained from voters, who only have some knowledge of the leaders behaviors. We include a common cause – suppose it is affect for the leader or any other common-cause mechanism – as well as two instrumental variables *z*1 and *z*2 that do not correlate with the common cause (also assume no selection effects due to the instruments). The first instrument, *z*1 is the leader's IQ and *z*2 is the leader's neuroticism, which are orthogonal to each other. Ξ1 and Ξ2 are transformational and transactional-oriented leadership respectively (for simplicity these are the only styles of leadership that matter). Thus, the more subordinates like the leader the more they see her as charismatic and the less they see her as transactional; however, these styles vary too because of the leaders' personality and IQ. Given that the leader individual differences are largely exogenous (i.e., due to genes), they will vary independently of other factors in the model.

The correct model is depicted in Fig. 4, Panel A, which fits the data perfectly: $\chi^2(51) = 47.48$, $p > .05$, $n = 10{,}000$; all estimates are standardized. In Panel B, we estimate the model only using the instruments. The parameters estimates are correct as long as the disturbances are correlated; the model fits perfectly, despite omitting the common cause: $\chi^2(45) = 48.50$, $p > .05$. In Panel C the model is still correct. Given that the instruments are exogenous, they do not correlate with the common cause. Thus, omitting the instruments (or the common cause, as we showed in Panel B) will not bias the estimates; also, the model fits perfectly: $\chi^2(37) = 37.96$, $p > .05$. Finally, the model depicted in Panel D is incorrect because the latent variables are endogenous and are not purged from endogeneity bias. The structural estimates are incorrect although the model fits: $\chi^2(31) = 34.46$, $p > .05$. Again, this example not only demonstrates that instruments can purge the bias from endogenous variables but that it is imperative that the model be correctly specified. Note, we tried to recover the correct causal estimates by modeling a latent common factor; however, the model produced a "Heywood" case on *y* whose variance we had to constrain so that it could be estimated); doing so resulted in good model fit. However, the model estimates were still wrong.

Thus, we hope that our demonstrations will provide new directions in solving the common-method problem and in estimating mediation models correctly. Also, as is evident, the modeler must rely on theory as well as statistical tests when specifying models and ensure that they model exogenous sources of variance to obtain consistent estimates.

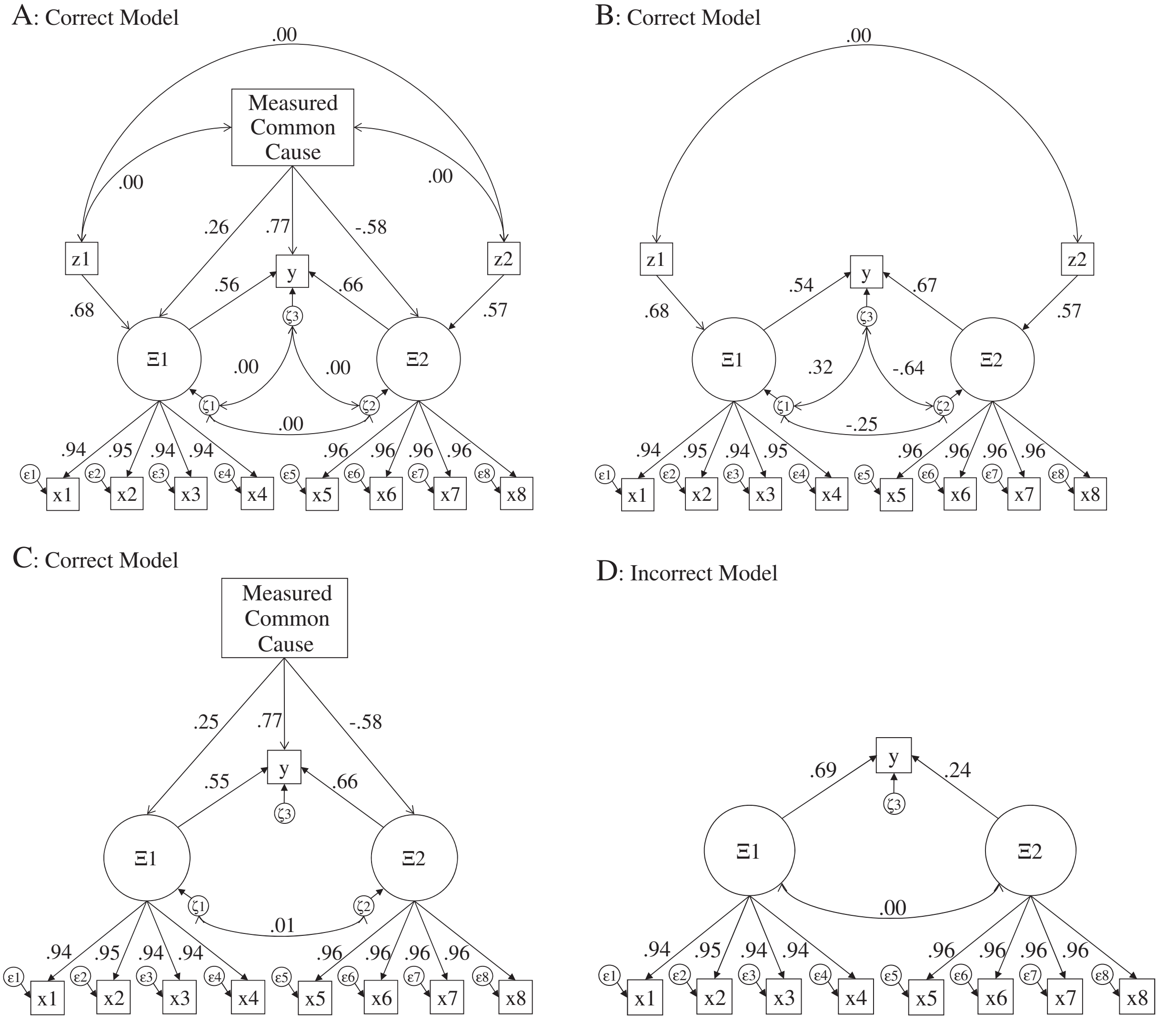

Fig. 4. Recovering causal estimates in a structural equation model with instrumental variables (estimates are standardized). A: This is the correctly specified model including the common cause and two instrumental variables $z1$ and $z2$; note, the instruments and the common cause do not correlated. Thus, omitting the common cause or $z1$ and $z2$ will not bias estimates. B: Despite omitting the common cause, this model is correctly specified given that the endogeneity bias is purged with instruments $z1$ and $z2$. C: This model is also correct; because the common cause does not correlate with the instruments $z1$ and $z2$. D: This model is incorrect, and estimates are biased because the latent variables $\Xi1$ and $\Xi2$ correlate with $\zeta3$.

4.3. Regression discontinuity models

The regression discontinuity design (RDD) is a deceptively simple and useful design. It was first proposed by Thistlethwaite and Campbell (1960) and brought to the fore by Cook and Campbell (1979). Interestingly, this design has been rediscovered independently in several fields (Cook, 2008). After the randomized experiment the RDD is the design that most closely approximates a randomized experiment (Cook et al., 2008); however, it is underutilized in social-sciences research and not well understood (Shadish et al., 2002). It is currently experiencing a renaissance in economics (Cook, 2008). We discuss this design extensively, because it is very useful in field settings.

The reason why the design is so useful is that like the randomized experiment, it *specifically* models the selection procedure. Whereas in the randomized experiment selection to treatment is random, in the RDD selection is due to a specific cut-off (or threshold) that is observed explicitly and modeled as such; this cut-off can be a pretest or any other continuous variable that does not necessarily have to be correlated with *y* (Shadish et al., 2002). From an ethical point of view, and when the cutoff is a pretest of *y*, this design is very useful in that the individuals who are most likely to benefit from the treatment obtain it; however, this non-randomization to condition is precisely why the RDD is difficult to grasp in terms of whether its estimates are consistent. The reason why the RDD yield consistent estimates is that selection to experiment and control group is based on an explicitly measured criterion that is included in the regression equation (thus, the disturbance term contains no information that might correlate with the grouping variable). The advantages of this design are many, given that it is relatively easy to implement in field settings to test the effectiveness of a policy (particularly when using a pretest threshold).

To explain the basic workings of this design assume that a company decides to give leadership training to its managers; however, the company CEO is not sure if leadership training works. A professor, eager to test the workings of the RDD, suggests that they could emulate a randomized experiment *and* simultaneously help those that need leadership training the most (i.e., provide the training only to the leaders who are below a certain threshold). Let s be the selection threshold for training, based on a pretest of a validated diagnostic test of the leaders' charisma (x_i). Leaders who score below a threshold – which will be the group mean (this choice maximizes power in the RDD Shadish et al., 2002) – are placed in the treatment condition. Thus, for leader i, selection is based on the following explicit rule:

$$s_i = 1 \text{ if } x_i \leq \overline{x}$$

$$s_i = 0 \text{ if } x_i > \overline{x}$$

Then, the following regression model is estimated, using the mean-centerd pretest score to set the intercept to the cut-off value (note, controls could be included in this equation to increase power):

$$y_i = \beta_0 + \beta_1 s_1 + \beta_2 (x_i - \overline{x}) + e_i \tag{33}$$

The treatment effect is β_1. The counterfactual is β_0 (what the treatment group would have had had it not been treated). Too good to be true, right? As mentioned by Shadish et al. (2002) many “will think it implausible that the regression discontinuity design would yield useful, much less unbiased, estimates of treatment effect” (p. 220). Below we thus show explicitly why RDD approximates the randomized experiment almost perfectly (we base our examples on the arguments and figures presented by Shadish et al., 2002). We use our own, simulated data, with a known structure, coupled with errors-in-variables regression as well as a Monte Carlo experiment to show how the RDD can provide consistent estimates (so we “kill three birds” with one stone given that we plead in the conclusions too for more use of the Monte Carlo simulation method).

We first begin with a simple example to show the parallel between the RDD and the randomized experiment. Five hundred participants were randomly assigned to a control and treatment condition ($x = 1$ if in treatment, else $x = 0$). We include a perfectly measured pretest (z) correlating .60 with the posttest (y); both variables are standardized, thus structural parameters are standardized too. We generated the data such that the treatment increases y by 2 points on the scale. The regression model we estimated (the typical ANCOVA model in psychology) was:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i \tag{34}$$

Results indicate a significant regression model, $F(2, 497) = 544.80$. The coefficient of $\beta_1 = 2.00$, standard error = .07, $t = 27.87$, $p < .001$. The coefficient of $\beta_2 = .60$, standard error = .04, $t = 16.71$, $p < .001$. The constant is 2.60. The regression lines are parallel given that the x^*z interaction was insignificant (see Fig. 5A).

Now, to understand how regression discontinuity works and to see its visual relation to the experiment (Shadish et al., 2002), suppose that we had given the treatment only to that part of the treatment group that scored below 9, which was the group mean, on the pretest. Also, suppose that those who score above the threshold do not receive the treatment. Using the same data as before, we obtain the two regression lines (see Fig. 5B).

The discontinuity can be seen at the mean of x (the threshold for assigning a participant to the treatment or control condition); this sharp-drop in the line suggests that those just left of the treatment cut-off benefitted greatly as compared to those just to the right of the cut-off in the control group who did not receive the treatment. Estimating Eq. (33) shows that the regression model was significant, $F(2, 243) = 82.21$. The coefficient of $\beta_1 = 1.99$, standard error = .17, $t = 6.78$, $p < .001$. The coefficient of $\beta_2 = .58$, standard error = .09, $t = 6.78$, $p < .001$. The constant is 8.03. The regression lines are parallel given that the x^*z interaction was insignificant; note, it is always good policy to include the x^*z interaction in case the experiment produces not only a change in the constant but also in the slope (Hahn, Todd, & Van der Klaauw, 2001; Lalive, 2008).

The treatment effect is almost precisely the same as before (1.99 now, versus 2.00). As we mentioned before, the counterfactual is the constant; thus, if the experimental group had not received the treatment, its mean would have been 8.03. Now, going back to the randomized experiment, the fitted model indicated that $\hat{y} = 2.60 + 2x + .60z$. Thus, at the mean value of z we predict y to be the following for the control group, which is the true counterfactual — or the estimated marginal mean: $2.60 + 2^*0 + .60^*9 = 8.00$!

This exercise never ceases to amaze us, but it is so obvious once one understands how the RDD works. As is evident from the graphs, the randomized experiment replaces the discontinuity with random assignment. Rather than allocating everyone using a cutoff to the treated condition, the randomized experiment assigns a random subgroup to either the treated or the control condition. Furthermore, readers should not fall into the trap of thinking that RDD is simply explained by regression to the mean, in the sense that when remeasuring participants with extreme values their post-scores regress to the mean. As mentioned by Shadish et al. (2002), any regression effects are already captured in the regression line. Of course, those initially scoring in the extremes will regress; however, this causes the slope of the regression line to become flatter, but it does not cause discontinuities.

To test RDD a step further we then conducted a Monte Carlo experiment. To provide for a strong test, we made the correlation between y and x more realistic by adding error to x, and thus also show the workings of the errors-in-variables estimator: We add a

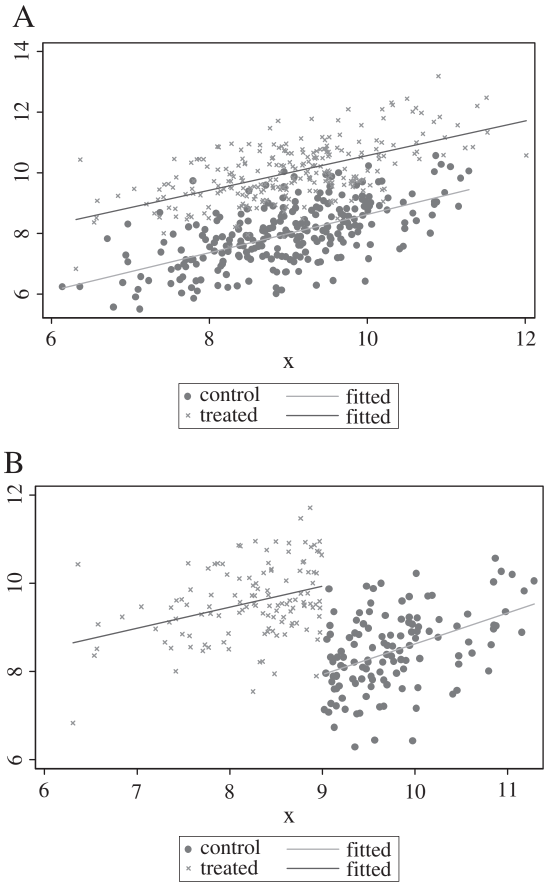

Fig. 5. Similarity between randomized experiment and regression discontinuity. A): Estimating causal effect using a randomized experiment. B): Estimating the causal effect using regression discontinuity.

normally distributed error term (e) to x (i.e., $.5*e$). The reliability of x is (Bollen, 1989): 1 − (error variance)/(total variance). Given that the original x (without error) had a variance of 1, and we observe the variance of *x-with-error* to be 1.28, the theoretical reliability of x is $1-(1.28-1)/1.28=.78$. We then ran a Monte Carlo experiment, estimating the same regression model as in the RDD shown previously using the mean of x as the cut-off. We simulated this process 1000 times to see how stable this estimator is (i.e., specifically to see how the causal parameters of interested were distributed).

The results showed that the RDD coupled with the errors-in-variables estimator recovered the true causal parameters almost precisely! The mean of the constant was 8.01 (95% confidence: 8.00 to 8.01). The mean of coefficient of $\beta_1=2.02$ (95% confidence: 2.01 to 2.03). Finally, the coefficient for of $\beta_2=.60$ (95% confidence: .59 to 60)!

We re-ran the Monte Carlo using OLS to demonstrate the effect of measurement error on the estimates. The mean of the constant was 8.21 (95% confidence: 8.20 to 8.21). The mean of coefficient of $\beta_1=1.62$ (95% confidence: 1.61 to 1.63). Finally, the coefficient for of $\beta_2=.34$ (95% confidence: .34 to .34). These estimates are way off the correct estimates; the treatment effect was underestimated by a large margin (−19.80%). The effect of the covariate was underestimated by a much larger margin (−43.33%). Finally, the counterfactual was slightly overestimated (+2.50%); however, the intercept seems to be less affected. Results with more than one ill-measured covariate would certainly create much more bias than what we have showed here with a very simple model.

To conclude, we trust that our demonstrations will create some interest in using RDD in leadership research as well as in related areas (management, applied psychology, strategy, etc.). This design is clean and simple to run. Because of space restrictions we have only covered the basics of RDD; readers should refer to more specialized literature for further details (e.g., Angrist & Krueger, 1999; Angrist & Pischke, 2008; Hahn et al., 2001; Lee & Lemieux, 2009). Finally, modelers can find creative ways to use the RDD. For instance, regression discontinuities could also be used where the modeled cut-off is an exogenous shock (e.g., war).

4.4. Difference-in-differences models

In the case where a treatment and a very similar control group are compared before and after a treatment, causal inference could be made provided certain assumptions are met (the most important being that in the absence of the treatment, the difference between the two groups is relatively stable over time). This type of model is called a difference-in-differences model in economics (see Angrist & Krueger, 1999; Angrist & Pischke, 2008; Meyer, 1995); in psychology, it is usually referred to as an untreated control group design with pre- and post-test (Shadish et al., 2002). We will discuss these models from the point of view of economics, given that the literature on estimation methods is more developed in this field.

The basic idea of the difference-in-differences model is to observe the effect of an exogenous "shock" on a "treatment" group; the treatment effect is the difference between the treated group and a comparable control group across time. Using a comparable group thus "differences-out" confounding contemporaneous factors. For a graphic depiction of this model see Fig. 6.

The following model, in panel form, is thus estimated:

$$y_{it} = \beta_0 + \beta_1 x_i + \beta_2 t + \beta_3 x_i \cdot t + e_{it} \tag{35}$$

Where person i is in a group ($x = 1$ for treatment group, else it is 0) in a particular time period ($t = 1$ if post treatment, else it is 0); for simplicity, suppose that data are on two periods, before and after the intervention, where $t = 1$ is post-treatment, else it is 0; the model should include control variables, which we have omitted for simplicity. The treatment effect is captured by the coefficient of the interaction term, β_3. Another way of looking at the treatment effect is to difference y across time and groups, which gives:

$$\{E[Y_{it}|x_i = 1, t = 1] - E[Y_{it}|x_i = 0, t = 1]\} - \{E[Y_{it}|x_i = 1, t = 0] - E[Y_{it}|x_i = 0, t = 0]\} = \beta_3 \tag{36}$$

That there may be differences between the groups prior to the intervention is captured by the fixed effect of group membership, that is, the coefficient of x (thus, random assignment is not of issue here, as long as the assumptions of the method are satisfied). Fixed effects of time are captured by the coefficient of t, that is, because changes in y might be due to time. Note, fixed effects of individual could be modeled as well in which case the between group differences will be captured by the individual fixed effects rather than by the parameter β_1. What is important for this model is that $x \cdot t$ is not endogenous, that is, that the difference between the groups is stable over time and that the timing of the treatment is exogenous (i.e., that differences in y are not due to unmeasured factors); this assumption can be examined by comparing data historically to see if differences are stable across the groups before (and after) the treatment (Angrist & Krueger, 1999). Also, given that the data are panel data, it is important to correct standard errors for clustering on the panel variables (Bertrand et al., 2004). Note, that $\beta_0 + \beta_1 + \beta_2$ provides for the counterfactual (i.e., $\overline{y}$ of the treatment group had it not been treated). Of course, the basic difference-in-differences model can be expanded in more sophisticated ways (Angrist & Krueger, 1999, 2001; Angrist & Pischke, 2008; Meyer, 1995).

Applied to leadership research, suppose that a CEO of a company that has two similar factory sites decides to hire a professor to conduct an experiment to see whether leadership training works. She decided this on the basis of yearly data the company has been gathering using a 360 leadership instrument, which showed that the mean level of charisma has been declining across the two sites (at a similar rate) and that it is now below a critical threshold in both sites. Trained as a medical researcher, she suggests to the professor that all the company's 500 supervisors should be randomized into a treatment or a control group. Instead of doing a randomized experiment within each factory, which could have spillover effects from the treatment to the control groups, the professor convinces the CEO to allow her to train the managers on one site only. Given the fact that they are separated by a distance of 2000 km and because they produce pharmaceutical products for different markets (which both have strong demand for their products), it is very unlikely that managers in the control site will get to know about the training that will be conducted in the experimental site. Furthermore, because demographic indicators regarding the managers and the workers are similar in the two sites (as are charisma trends in the managers), and socioeconomics about the same (as historical data indicates) the difference-in-differences would be an appropriate tool to use in this particular case.

4.5. Selection models (Heckman models)

As discussed previously, when there is unmodeled selection to treatment (i.e., participants attend leadership training, but training is not assigned randomly), estimates will be inconsistent because unobserved variance which affects selection in the selection equation (see Eq. (11)) could be correlated with unobserved variance that affects the dependent variable (see Eq. (10)) —

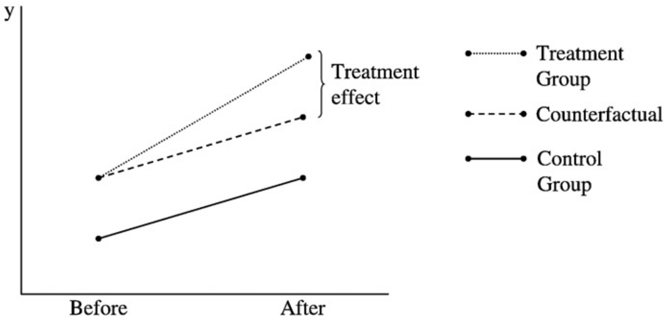

Fig. 6. Estimating causal effects using difference-in-differences.

this endogeneity can inflate or deflate the treatment effect. One way around this problem is to estimate a Heckman type two-step selection model (Heckman, 1979) or more specifically, what is referred to as a *treatment effects model* (see Cong & Drukker, 2001; Maddala, 1983).

The idea behind this model is to use instruments to predict participation in the treatment or control group (the probit first-step equation). Thereafter, a control variable, which captures all unobserved differences between the treatment and control groups due to selection, is added in the second step (the substantive equation). This control variable will remove the variance from the error term due to selection, so as the coefficient on the treatment term can be correctly estimated. This model is easily estimated in advanced statistics programs like Stata.

Note, there are other types of models that can be estimated having sample selection bias, for example, models where the dependent outcome could be binary instead of continuous. Another type of selection model is the classic *Heckman two-step model* for situations where one observes the dependent variable only for the selected group (i.e., there is missing data on the dependent variable). The example Heckman used is to estimate the effect of education on wages for women, with the selection problem being the fact that women choose to work depending on the offer wage and the minimum wage a woman would expect to have (i.e., her reservation wage); thus, simply regressing the wages on education will not provide a consistent estimate.

Applied to leadership, suppose we wish to estimate whether there are sex differences in leadership effectiveness. The selection problem is that individuals are appointed to positions of leadership based in part on their sex and not only on their competence (Eagly & Carli, 2004). That is, because of social prejudice mechanisms, stereotype threat, and self-limiting behavior, females may be less likely to be appointed to leader roles as a function of the gender typing of the context. Thus, in male-oriented environments, the sample of observed male leaders is biased upward in male-stereotypical settings and only the performance of very competent women would be observed (because women are held to a higher standard of performance and thus only the more competent women are observed); indeed, when comparing the effectiveness of women versus men in business settings that would reflect this selection, women are significantly more effective (Antonakis, Avolio, & Sivasubramaniam, 2003; Eagly, Johannesen-Schmidt, & van Engen, 2003). The effect of being a woman on leadership effectiveness could thus be overstated.

The Heckman model could be useful in this context. In the first step, we would predict the probability of being a leader using exogenous instruments (e.g., sex, competence, sex-typing of the job, cultural factors, etc). Then in the second step, we would include sex as a predictor and control for unobserved heterogeneity in the selection in predicting effectiveness of leaders who we observe to derive a consistent estimate of the effect of sex on effectiveness.

4.6. Other types of quasi-experimental designs

There are other ways to obtain causal estimates using very simple methods. Researchers should refer to Cook and Campbell (1979) and Shadish et al. (2002) for ideas. For instance, extending the idea behind the non-equivalent dependent variable design (see Shadish et al., 2002), suppose that a researcher wants to investigate the efficacy of a leadership training program; however, for whatever reason (e.g., restrictions imposed by an organization, ethical reasons, etc.) the researcher cannot have a control group. One way to obtain estimates that could be consistent is to pretest the participants on the measure of interest (e.g., charisma) as well as on a closely-related measure that the researcher did not intend to change (e.g., communication skills, see Frese, Beimel, & Schoenborn, 2003 for an example). The point of this design is to show a significant difference between the Time 1 and Time 2 measure of interest and no difference in the other measure that the researchers did not intend to manipulate. In the Frese et al. (2003) study, however, they did find differences too in communication skills, which can be interpreted as learning effects; however, they could have used this information to "unbias" their parameters of interest (though they did not). That is, a simple way to remove the variance due to learning effects is to include the non-equivalent measure as a control variable, particularly if one has pre and post measures as well as control variables (and can thus estimate a panel model). Of course, such methods will are not substitutes for the experiments, but if the right controls are included they may provide good enough estimates of treatment effect.

Next, we discuss the state-of-the-art of causal analysis in leadership research. We first explain the sample we used in this review and our coding method. Thereafter we present the findings and discuss their implications.

5. Review of robustness of causal inference in management and applied psychology

5.1. Sample

To gauge whether leadership research is currently dealing with central threats to causal inference (i.e., reporting estimates that are consistent), we reviewed and coded a random sample of articles appearing in top management and applied psychology journals. The initial sample from which the final set of articles was drawn was quite large ($n = 120$) and current — covering the last 10 years (i.e., between 1999 and 2008). We did not code any laboratory experiments given that their estimates would be consistent by design (because of randomization). We only coded empirical non-experimental papers and field experiments, because it is in these categories of research where potential problems would be evident. The population of journals we surveyed, including *The Leadership Quarterly* are all top-tier journals according to objective criteria (i.e., 5-year ISI impact factor reported in 2009) in the domain of management or applied psychology. These journals publish research on leadership and have a strong micro or psychology focus. We include the list of journals as well as their 5-year impact factor (IF) and rank in either management (MGT) where there are 89 journals listed and/or in applied psychology (AP) where there are 61 listed journals:

1. *Academy of Management Journal*: Management (IF = 7.67; MGT = 3rd)
2. *Journal of Applied Psychology* (IF = 6.01; AP = 1st)
3. *Journal of Management* (IF = 4.53; MGT = 9th)
4. *Journal of Organizational Behavior* (IF = 3.93; MGT = 14th; AP = 4th)
5. *The Leadership Quarterly* (IF = 3.50; MGT = 18th; AP = 5th)
6. *Organizational Behavior & Human Decision Processes* (IF = 3.19; MGT = 21st; AP = 10th)
7. *Personnel Psychology* (IF = 5.06; AP = 2nd)

We first identified the population of articles that met our selection criteria. We used ISI Web of Science to initially identify potential articles which included either "leader" or "leadership" in the "topics" field, which searches in the title, keywords, and abstract. We only examined studies that focused on leadership per se. We limited studies using the definition of leadership provided by Antonakis, Cianciolo, and Sternberg (Antonakis, Cianciolo, & Sternberg, 2004, p. 5), that is, "leadership can be defined as the nature of the influencing process – and its resultant outcomes – that occurs between a leader and followers and how this influencing processes is explained by the leader's dispositional characteristics and behaviors, follower perceptions and attributions of the leader, and the context in which the influencing process occurs." Thus, we coded only studies that examine the influencing process of leaders from a dispositional or behavioral perspective, where leadership could be either an independent or dependent variable.

We then determined how many papers were quantitative non-experimental studies or field experiments. The population of studies that met our criteria was 287 (i.e., 281 non-experiment and 6 field experiments). This population was distributed as follows across the journals: *Academy of Management Journal* (9.06%), *Journal of Applied Psychology* (24.04%), *Journal of Management* (3.14%), *Journal of Organizational Behavior* (13.24%), *The Leadership Quarterly* (42.16%), *Organizational Behavior & Human Decision Processes* (3.14%), and *Personnel Psychology* (5.22%).

We then randomly selected 120 studies using stratified (proportionate) sampling by journal and type of study (i.e., non-experimental or field experiment). From this sample of 120 studies, we dropped 10 which, although quantitative in nature, did not make any implicit or explicit causal claims as in the case of scale validation studies; we did though retain those that made, for example, comparison of factors across groupings like gender (e.g., Antonakis et al., 2003). Thus, the final sample was 110 studies, distributed as follows: *Academy of Management Journal* (8.18%), *Journal of Applied Psychology* (26.36%), *Journal of Management* (3.64%), *Journal of Organizational Behavior* (14.55%), *Leadership Quarterly* (38.18%), *Organizational Behavior & Human Decision Processes* (3.64%), and *Personnel Psychology* (5.45%). The final distribution of papers was the same as the original distribution, $\chi^2(6) = .67$, $p = 1.00$.

5.2. Coding

We evaluated studies on each of the sub-criteria of the seven categories listed later (i.e., in total there were 14 criteria). We coded each criterion, using a categorical scale: 0 = irrelevant criterion; 1 = relevant criterion for which the authors did not correct; 2 = relevant criterion for which we were unable to determine whether it was taken into account by the authors; 3 = relevant criterion which the authors addressed. Note, we not code for correct use of statistical tests, for example, use of the chi-square overidentification test. The criteria we coded included those listed in Table 1.

When papers reporting several studies, we only coded those which were non-experimental studies or field experiments even if they represent a small portion of the paper; for example, the coding of de Cremer and van Knippenberg (2004) is based solely on the one non-experimental study reported by the authors and does not take into account the experimental studies presented in the paper.

The coding was undertaken by the second and third authors of this study. The coders were first familiarized with the coding criteria. To ensure that the coders of the study were well calibrated with each other, they independently coded five randomly-selected studies from the eligible population of leadership studies we had identified (but which had not been selected in our random sample). Thereafter, differences were reconciled. The coders then independently coded 20 studies and we calculated agreement statistics (which indicated very high agreement, i.e., 80.51% agreement across the 280 coding events). After differences were reconciled, the coders then coded the rest of the studies independently.

Each study was then discussed between the coders and differences were reconciled. Finally, the first author crossed-checked a random sample of 10 studies from the total population of studies coded (and reconciled situations where either one or the other coder was unsure as to what to code). The final coding represents the agreed ratings of both coders.

6. Results

We first report results for the coding to examine whether it was undertaken reliably by the two coders. The total coding events were 1540 (14 criteria times 110 papers); however, we computed agreement statistics for 1519 coding events only given that for 21 of the coding events either one or the other coder was unsure about how the coding procedure should be applied. In this case, the first author reconciled the coding.

Initial agreement based on the first independent coding of the 110 studies was 74.39% (1130 agreements out a possible 1519 coding events) and the initial agreement κ coefficient was .60, $SE = .02$, $z = 33.51$, $p < .001$ (see Cohen, 1960). This result suggests that the coders did significantly better than chance (which would have generated an agreement of 36.41% given the coding events and coding categories). This level of agreement has been qualified as being close to "substantial" (Landis & Koch, 1977).

We present the results of the coding in Appendix A, summarized for the full sample and by journal. As a descriptive indicator of our findings, the data across the four coding categories for all journals indicated that: 43.83% (675/1540) received a code of 0 (irrelevant criterion), 37.21% (573/1540) received a code of 1 (relevant criterion for which the authors did not correct), 13.57% (209/540) received a code of 2 (relevant criterion for which we were unable to determine whether the authors undertook the necessary correction), and 5.39% (83) received a code of 3 (relevant criterion, which the authors addressed). The frequency distribution of coding categories across the journals (see Appendix) were very similar as indicated by a chi-square test: $\chi^2(18) = 14.88$, $p = .67$. Because the distribution of this chi-square test can be affected by small sample sizes in cells and given that we could not compute the Fisher exact test (Fisher, 1922) with so many permutations, we repeated this analysis only for the journals that had many observations (i.e., which regularly publish leadership research: *Academy of Management Journal*, *Journal of Applied Psychology*, *Journal of Organizational Behavior*, and *The Leadership Quarterly*). The result remained unchanged: $\chi^2(9) = 7.20$, $p = .62$.

Considering only the codings that were applicable (i.e., excluding the 675 codings receiving 0), indicates that 66.24% received a code of 1, 24.16% a code of 2, and 9.60% a code of 3. Assuming that those codings given a 2 were not actually corrected by the authors indicates that 90.40% (66.24% + 24.16%) of coded validity threats were not adequately handled. We can consider this 90.40% as an upperbound percentage of codings that did not deal with the validity threat appropriately; thus, 66.24% is the lowerbound (assuming that those codings receiving a 2 were actually corrected for by the authors but that the correction was not reported). These results are depicted in Fig. 7. Again, the distribution of aggregate coding categories across the journals were very similar for the full sample, chi-square test: $\chi^2(12) = 8.37$, $p = .76$, as well as for the four journals that had sufficient observations: chi-square test: $\chi^2(6) = 1.18$, $p = .98$. Note that all articles, save one, had at least one threat to validity and most (90.91%) had three or more threats.

We compared the distribution of codings across the seven journals for the 14 coding criteria using the Fisher (1922) exact test and setting the overall Type I error to be less .05 across the 14 tests (Bonferroni correction). The distributions across the 14 criteria were the same across the journals, suggesting that practices and standards for these top-tier journals regarding leadership research were essentially the same.

What are the most frequent and important threats to validity (see bottom part of Panel A in the Appendix)? Criterion 4 (measurement error) and 6a (heteroscedastic errors) applies to more than 94% of all the studies we survey. Measurement error is not addressed by 70% of all studies to which the problem applies, and only 16.4% of the studies adequately deal with the problem. Heteroscedasticity is a potential pervasive problem but 92.3% of the studies possibly facing the problem do not report whether or not they dealt with it; only 6.7% reported using robust inference. Common-method variance (Criterion 5) is a threat to validity that is also very pervasive in leadership research (it applies to 83.6% of the studies). Yet 77% of the studies affected by the threat do not deal with it adequately, and only 20.7% of the studies adequately address it.

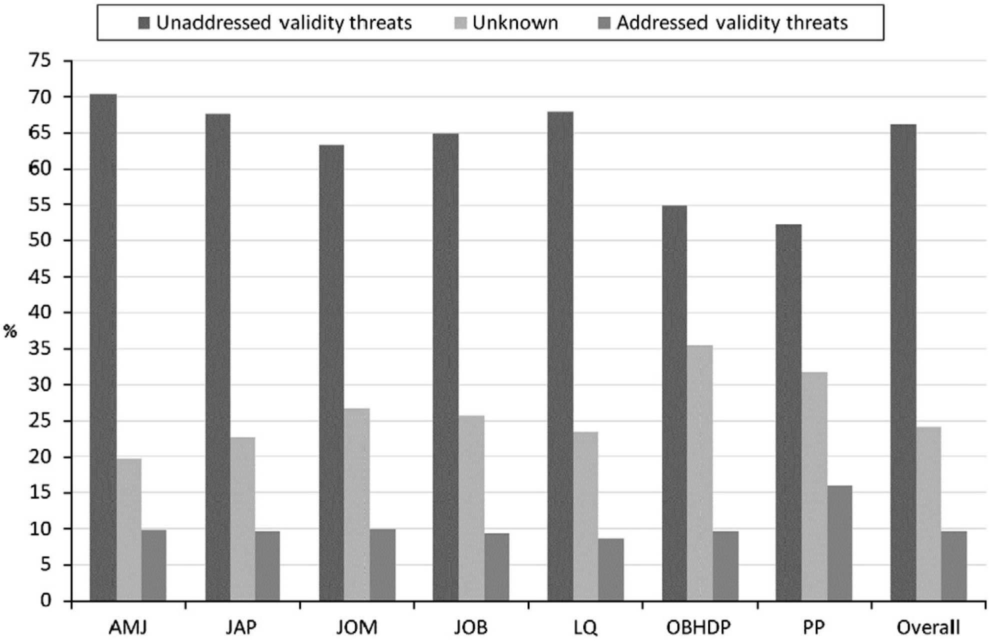

Fig. 7. Summary of coded validity threats by journal.

Are there threats to validity that are dealt with better than others? Threat 2c (sample is not representative) applies to 58.2% of all the studies we survey. Of the studies that face this problem, 32.8% address it adequately whereas 48.4% do not; leaving 18.8% for which we cannot assess whether sample selection has been addressed or not.

7. Discussion

ur review indicated that methodological practices regarding causal modeling in the domain of leadership are unsatisfactory. Our results essentially point to the same conclusions as do the recent reviews of the literature regarding endogeneity by Hamilton and Nickerson (2003) in the strategy domain, that of Halaby (2004) in sociology regarding panel models, and that of Bertrand et al. (2004) regarding the use of cluster-robust standard errors in econometrics. Although we looked at similar issues to those of the three reviews, the contribution of our review was unique in that we examined multiple validity threats (beyond those three reviews).

Except for *The Leadership Quarterly*, the articles we coded were published in general management and organizational behavior journals. Thus, we could assume that the practices of others disciplines publishing in those journals are very similar to the practices we identified; our findings may thus have implications for other areas and also for the meta-analytic reviews which may have used estimates that were inconsistent. We can only echo what Halaby (2004, p. 508) noted about for research in sociology using panel data "Key principles that ought to routinely inform analysis are at times glossed over or ignored completely."

Why is current practice not where it should be given that the methodological tools have been available for some time? We can only speculate as to why practice has been slow to follow the methodological advances that have been made. The most important reason probably has to do with doctoral training; in psychology at least, it appears that adequate training in field research and quantitative methods in general is not provided, even at elite universities (Aiken, West, & Millsap, 2008). We can assume that the level of training provided in non-experimental causal analysis in management is insufficient as well, particularly in econometrics training. As Aiken et al. (2008, p. 44) state "Psychologists must reinvigorate the teaching of research design to our next generation of graduate students, to bring new developments burgeoning in other fields into the mainstream of psychology."

We believe that coupled with the previous problem is the fact that users of statistical programs have been very slow to adopt software that can do the job correctly when causal analysis in non-experimental settings is concerned; as mentioned by Steiger (2001) statistical practice is, unfortunately, software driven and there are many "simplified" books that make it easy to use software to estimate complicated models (Alberto Holly, one of our econometrics colleagues, refers to this as the "push-button" statistics syndrome). We find it very unfortunate that easy-to-use programs (e.g., like SPSS now called PASW), which have very limited and at times inexistent routines to handle many of the challenging methodological situations we identified in our review, are firmly entrenched in psychology and business schools. In our experience, SPSS is sufficient for analyzing basic experimental data, but as soon as researchers venture out into the non-experimental domain we would urge them to migrate to other software (e.g., Stata, SAS and R) that will allow them to test models in robust ways and also to widen the research horizons on which they can explore. Of course, professors who teach methods and statistics classes should also seriously consider using more appropriate software (and in also providing more extensive training to their students).

We note the same concern regarding structural-equation modeling (SEM) software, where much of the market is using SPSS's AMOS software; this program makes it very easy to estimate models. However, this program has very limited capabilities as compared to MPlus (our SEM software of choice), LISREL or EQS, though even these programs have some catching-up to do concerning the estimation of certain types of models (e.g., selection models).

7.1. Recommendations: the 10 commandments of causal analysis

Our review and the coding criteria we identified can be used as a summary framework around which researchers should plan and evaluate their work to ensure that estimates are consistent and that inferences are valid. We briefly present these criteria next, grouped in the form of 10 best practices, implicating research design and analysis issues. Concerning these two aspects of research, simply put, *design rules* (Shadish & Cook, 1999); only when the design is adequate can appropriate statistical procedures be used to obtain consistent estimates.

7.1.1. Best practice for causal inference

1. To avoid omitted variable bias include adequate control variables. If adequate control variables cannot be identified or measured obtain panel data and use exogenous sources of variance (i.e., instruments) to identify consistent effects.
2. With panel (multilevel) data, always model the fixed effects using dummy variables or cluster means of level 1 variables. Do not estimate random-effects models without ensuring that the estimator is consistent with respect to the fixed-effects estimator (using a Hausman test).
3. Ensure that independent variables are exogenous. If they are endogenous (and this for whatever reason) obtain instruments to estimate effects consistently.
4. If treatment has not been randomly assigned to individuals in groups, if membership to a group is endogenous, or samples are not representative between-group estimates must be corrected using the appropriate selection model or other procedures (difference-in-differences, propensity scores).

5. Use overidentification tests (chi-square tests of fit) in simultaneous equations models to determine if the model is tenable. Models that fail overidentification tests have untrustworthy estimates that cannot be interpreted.
6. When independent variables are measured with error, estimate models using errors-in-variables or use instruments (well-measured, of course, in the context of 2SLS models) to correct estimates for measurement bias.
7. Avoid common-method bias; if it is unavoidable use instruments (in the context of 2SLS models) to obtain consistent estimates.
8. To ensure consistency of inference, check if residuals are i.i.d. (identically and independently distributed). Use robust variance estimators as the default (unless residuals can be demonstrated to be i.i.d.). Use cluster-robust variance estimators with panel data (or group-specific regressors).
9. Correlate disturbances of potentially endogenous regressors in mediation models (and use a Hausman test to determine if mediators are endogenous or not).
10. Do not use a full-information estimator (i.e., maximum likelihood) unless estimates are not different to that of limited information (2SLS) estimator (based on the Hausman test). Never use PLS.

Apart from addressing the previous guidelines and the methods we reviewed, researchers should also consider using Monte Carlo analysis more than they currently do. Monte Carlo analysis is very useful for understanding the working of estimators (Mooney, 1997); for example, when an estimator may be potentially unstable (e.g., in the case of high multicollinearity) a researcher could identify the sample size requirement to ensure that the estimator is consistent.

8. Conclusion

Research in applied psychology and related social sciences is at the cusp of a renaissance regarding causal analysis and field experimentation; there are many reasons for this push including, in part, for the need for evidence-based practice (Shadish & Cook, 2009). Researchers cannot miss this call; understanding the causal foundations of social phenomena is too important a function for society. Important social phenomena deserve to be studied using the best possible methods and in sample situations that can generalize to real-world settings; ideally our goals should be to improve policies and practices.

Although our review makes for telling conclusions we are hopeful and confident that research practice will change in ways that produces research that will be more useful to society. We conclude by referring to the problem of alignment of theory, analysis, and measurement: When not correctly aligned Schriesheim, Castro, Zhou, and Yammarino (2001, p. 516) noted that researchers "may wind up erecting theoretical skyscrapers on foundations of empirical jello." This warning is pertinent for a broader class of problems relating to causal modeling too; implicit or explicit causal claims must be made on concrete foundations.

Acknowledgements

We are grateful to the Yearly-Review Editor, Fran Yammarino and to the reviewers for their helpful comments. We also thank Marius Brulhart, Suzanne de Treville, Saskia Faulk, Lorenz Goette, Mikko Ketokivi, Cameron McIntosh, Edward Rigdon, and Laura Stapleton for their helpful suggestions. Any errors or omissions are our responsibility.

Appendix A

Coded studies and results.

	Coding criteria[a]															
Study coded	1a	1b	1c	1d	2a	2b	2c	3	4	5	6a	6b	7a	7b	Total	Total*
Panel A: All journals (*n* = 110)																
Summary statistics (by criterion)																
% of 0 (irrelevant)	19.1	20.0	86.4	20.9	97.3	95.5	41.8	25.5	5.5	16.4	5.5	21.8	68.2	90.0	43.83	
% of 1 (relevant not corrected)	80.9	64.5	12.7	76.4	1.8	4.5	28.2	73.6	66.4	64.5	0.9	4.5	31.8	10.0	37.21	66.24
% of 2 (relevant, unknown if corrected)	0.0	4.5	0.9	0.0	0.0	0.0	10.9	0.0	12.7	1.8	87.3	71.8	0.0	0.0	13.57	24.16
% of 3 (relevant, corrected)	0.0	10.9	0.0	2.7	0.9	0.0	19.1	0.9	15.5	17.3	6.4	1.8	0.0	0.0	5.39	9.60
Summary statistics excluding% of 0																
Relevancy percentage (100% − % of 0)	80.9	80.0	13.6	79.1	2.7	4.5	58.2	74.5	94.5	83.6	94.5	78.2	31.8	10.0		
% of 1 (relevant not corrected)	100.0	80.7	93.3	96.6	66.7	100.0	48.4	98.8	70.2	77.2	1.0	5.8	100.0	100.0		
% of 2 (relevant, unknown if corrected)	0.0	5.7	6.7	0.0	0.0	0.0	18.8	0.0	13.5	2.2	92.3	91.9	0.0	0.0		
% of 3 (relevant, corrected)	0.0	13.6	0.0	3.4	33.3	0.0	32.8	1.2	16.3	20.7	6.7	2.3	0.0	0.0		
Panel B: Academy of Management Journal (*n* = 9)																
Avolio, Howell, et al. (1999)	1	1	0	1	0	1	0	1	1	1	3	2	1	0		
Waldman, Ramirez, et al. (2001)	1	1	0	3	0	0	1	3	1	1	2	2	0	0		

Appendix A (*continued*)

	Coding criteria[a]															
Study coded	1a	1b	1c	1d	2a	2b	2c	3	4	5	6a	6b	7a	7b	Total	Total*
Panel B: Academy of Management Journal (*n* = 9)																
Shin & Zhou (2003)	1	1	1	1	0	0	0	1	1	1	2	2	1	0		
Wang, Law, et al. (2005)	1	1	1	1	0	0	1	1	3	1	2	2	1	1		
Rubin, Munz, et al. (2005)	1	1	0	0	0	0	0	0	1	0	0	2	0	0		
Agle, Nagarajan, et al. (2006)	0	1	0	3	0	0	1	0	1	0	2	2	0	0		
Sparrowe, Soetjipto, et al. (2006)	0	1	1	1	0	0	0	1	1	1	2	2	0	0		
Srivastava, Bartol, et al. (2006)	1	1	0	1	0	0	1	1	1	1	2	2	1	1		
Ling, Simsek, et al (2008)	1	1	0	1	0	0	3	1	3	3	2	2	1	1		
Summary statistics (by criterion)																
% of 0 (irrelevant)	22.2	0.0	66.7	11.1	100.0	88.9	44.4	22.2	0.0	22.2	11.1	0.0	44.4	66.7	35.71	
% of 1 (relevant not corrected)	77.8	100.0	33.3	66.7	0.0	11.1	44.4	66.7	77.8	66.7	0.0	0.0	55.6	33.3	45.24	70.37
% of 2 (relevant, unknown if corrected)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	77.8	100.0	0.0	0.0	12.70	19.75
% of 3 (relevant, corrected)	0.0	0.0	0.0	22.2	0.0	0.0	11.1	11.1	22.2	11.1	11.1	0.0	0.0	0.0	6.35	9.88
Panel C: Journal of Applied Psychology (*n* = 29)																
Hofmann & Morgeson (1999)	1	0	0	1	0	0	0	1	3	1	2	0	1	1		
Davidson & Eden (2000) †	1	0	0	0	3	0	0	0	3	1	2	0	0	0		
Liden, Wayne et al. (2000)	1	3	0	1	0	0	0	1	3	1	2	2	1	0		
Judge & Bono (2000)	1	0	0	1	0	0	1	1	3	1	2	0	0	0		
Lam & Schaubroeck (2000) †	1	0	0	0	1	0	0	0	1	0	2	0	0	0		
Martell & DeSmet (2001)	1	0	0	0	0	1	1	0	1	0	2	0	0	0		
Turner, Barling, et al. (2002)	1	3	0	0	0	0	2	0	3	0	2	2	0	0		
Sherony & Green (2002)	1	3	0	1	0	0	3	1	1	1	2	2	0	0		
Chen & Bliese (2002)	1	1	1	1	0	0	2	1	1	1	2	2	1	0		
Eisenberger, Stinglhamber, et al. (2002)	1	1	0	0	0	0	1	0	1	1	2	2	0	0		
de Cremer & van Knippenberg (2002)	1	0	0	1	0	0	1	1	1	1	2	0	0	0		
Offermann & Malamut (2002)	1	1	0	1	0	0	2	1	1	1	2	0	1	0		
Hofmann, Morgeson et al. (2003)	1	1	1	1	0	0	3	1	1	1	2	2	0	0		
Baum & Locke (2004)	1	3	0	1	0	0	3	1	3	1	2	2	1	1		
Lim & Ployhart (2004)	1	0	0	1	0	0	0	1	1	3	2	0	1	0		
Dineen, Lewicki et al. (2006)	1	1	1	1	0	0	0	1	1	1	2	2	0	0		
Judge, LePine et al. (2006)	0	1	0	1	0	0	0	1	1	1	2	2	0	0		
Aryee, Chen, et al. (2007)	1	1	0	1	0	0	0	1	1	1	2	2	1	1		
Tangirala, Green, et al. (2007)	1	1	1	1	0	0	3	1	1	3	2	2	0	0		
Liao & Chuang (2007)	0	1	1	1	0	0	0	1	1	3	2	3	1	0		
den Hartog, de Hoogh, et al. (2007)	1	3	0	1	0	0	3	1	1	1	2	2	1	0		
Mitchell & Ambrose (2007)	1	0	0	1	0	0	2	1	1	1	2	0	0	0		
Kamdar & Van Dyne (2007)	1	1	0	1	0	0	0	1	1	3	2	2	0	0		
Furst & Cable (2008)	1	1	0	1	0	0	0	1	1	1	2	2	0	0		
Ng, Ang, et al. (2008)	1	1	0	1	0	0	0	1	1	1	2	2	1	0		
Ozer (2008)	1	1	0	1	0	0	3	1	2	1	2	2	0	0		
Henderson, Wayne, et al. (2008)	0	1	1	1	0	0	1	1	1	1	2	2	1	0		
Hinkin & Schriesheim (2008)	1	1	0	1	0	0	0	1	2	1	2	2	0	0		
Eisenbeiss, van Knippenberg, et al. (2008)	1	1	0	1	0	0	0	1	1	1	2	2	1	0		
Summary statistics (by criterion)																
% of 0 (irrelevant)	10.3	27.6	79.3	17.2	93.1	96.6	48.3	17.2	0.0	10.3	0.0	31.0	58.6	89.7	41.38	
% of 1 (relevant not corrected)	89.7	55.2	20.7	82.8	3.4	3.4	17.2	82.8	72.4	75.9	0.0	0.0	41.4	10.3	39.66	67.65
% of 2 (relevant, unknown if corrected)	0.0	0.0	0.0	0.0	0.0	0.0	13.8	0.0	6.9	0.0	100.0	65.5	0.0	0.0	13.30	22.69
% of 3 (relevant, corrected)	0.0	17.2	0.0	0.0	3.4	0.0	20.7	0.0	20.7	13.8	0.0	3.4	0.0	0.0	5.67	9.66
Panel D: Journal of Management (n = 4)																
Ahearn, Ferris, et al. (2004)	1	0	0	1	0	0	1	1	1	0	2	0	0	0		
Elenkov & Manev (2005)	1	1	2	0	0	0	1	0	1	3	2	2	0	0		
Tepper, Uhl-Bien et al. (2006)	0	1	0	0	1	1	1	0	1	1	2	2	0	0		
Walumbwa, Avolio, et al. (2008)	1	1	0	1	0	0	0	1	3	3	2	2	0	0		
Summary statistics																
% of 0 (irrelevant)	25.0	25.0	75.0	50.0	75.0	75.0	25.0	50.0	0.0	25.0	0.0	25.0	100.0	100.0	46.43	
% of 1 (relevant not corrected)	75.0	75.0	0.0	50.0	25.0	25.0	75.0	50.0	75.0	25.0	0.0	0.0	0.0	0.0	33.93	63.33
% of 2 (relevant, unknown if corrected)	0.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	75.0	0.0	0.0	14.29	26.67
% of 3 (relevant, corrected)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	50.0	0.0	0.0	0.0	0.0	5.36	10.00
Panel E: Journal of Organizational Behavior (*n* = 16)																
Yukl & Fu (1999)	0	1	0	1	0	0	3	1	1	1	2	2	0	0		
McNeese-Smith (1999)	1	1	0	1	0	0	1	1	1	1	1	1	0	0		

(*continued on next page*)

Appendix A (*continued*)

	Coding criteria[a]															
Study coded	1a	1b	1c	1d	2a	2b	2c	3	4	5	6a	6b	7a	7b	Total	Total*
Panel E: Journal of Organizational Behavior																
Wayne, Liden, et al. (1999)	0	1	0	1	0	0	1	1	1	1	2	2	0	0		
Crant & Bateman (2000)	1	0	0	0	0	0	3	0	1	1	2	2	0	0		
Cogliser & Schriesheim (2000)	0	0	0	1	0	0	1	1	1	1	2	0	0	0		
Conger, Kanungo, et al. (2000)	0	2	0	1	0	0	1	1	1	1	2	2	1	1		
Andrews & Kacmar (2001)	1	1	0	1	0	0	2	1	3	1	2	1	0	0		
Sparks & Schenk (2001)	1	0	0	1	0	0	3	1	3	1	2	0	0	0		
Sagie, Zaidman, et al. (2002)	0	1	0	1	0	0	0	1	2	1	2	2	1	0		
Cable & Judge (2003)	0	3	0	1	0	0	1	1	2	1	2	2	0	0		
Adebayo & Udegbe (2004)	1	0	0	0	0	0	0	0	0	0	0	0	0	0		
Spreitzer, Perttula, et al. (2005)	0	3	0	1	0	0	1	1	0	3	2	2	0	0		
Harris, Kacmar, et al. (2005)	0	1	0	1	0	0	0	1	1	3	2	2	0	0		
de Hoogh, den Hartog, et al. (2005)	1	0	0	0	0	0	1	0	1	3	2	0	0	0		
Liden, Erdogan, et al. (2006)	1	1	1	1	0	0	1	1	2	2	2	2	0	0		
Major, Fletcher, et al. (2008)	0	3	0	1	0	0	1	1	1	1	2	2	0	0		
Summary statistics																
% of 0 (irrelevant)	56.3	31.3	93.8	18.8	100.0	100.0	18.8	18.8	12.5	6.3	6.3	25.0	87.5	93.8	47.77	
% of 1 (relevant not corrected)	43.8	43.8	6.3	81.3	0.0	0.0	56.3	81.3	56.3	68.8	6.3	12.5	12.5	6.3	33.93	64.96
% of 2 (relevant, unknown if corrected)	0.0	6.3	0.0	0.0	0.0	0.0	6.3	0.0	18.8	6.3	87.5	62.5	0.0	0.0	13.39	25.64
% of 3 (relevant, corrected)	0.0	18.8	0.0	0.0	0.0	0.0	18.8	0.0	12.5	18.8	0.0	0.0	0.0	0.0	4.91	9.40
Panel F: The Leadership Quarterly (*n* = 42)																
Schneider, Paul, et al. (1999)	1	1	0	1	0	0	1	1	1	1	2	2	0	0		
Connelly, Gilbert, et al. (2000)	1	1	0	1	0	0	0	1	1	3	2	2	1	0		
Mumford, Zaccaro, et al. (2000)	1	1	0	0	0	0	0	0	1	1	0	0	0	0		
Zacharatos, Barling, et al. (2000)	1	1	0	1	0	0	1	1	3	1	2	1	1	1		
Hooijberg & Choi (2000)	1	1	0	1	0	0	1	1	3	1	2	1	0	0		
Murry, Sivasubramaniam, et al. (2001)	1	1	0	1	0	0	1	1	2	1	2	2	1	0		
Thomas, Dickson, et al. (2001)	1	1	1	1	0	0	3	0	1	1	2	2	1	0		
Deluga (2001)	1	1	0	0	0	0	0	0	1	0	2	2	0	0		
Shipper & Davy (2002)	1	2	0	1	0	0	2	1	2	1	2	2	1	1		
de Vries, Roe, et al. (2002)	1	1	0	1	0	0	1	1	1	1	2	2	0	0		
Sosik, Avolio, et al. (2002)	1	1	0	1	0	0	1	1	2	3	3	2	1	0		
Wong & Law (2002)	1	1	0	1	0	0	0	1	1	1	2	2	0	0		
Schneider, Ehrhart, et al. (2002)	1	1	0	1	0	0	1	1	1	1	2	2	0	0		
Vecchio & Boatwright (2002)	1	3	0	0	0	0	3	0	0	0	2	2	0	0		
McColl-Kennedy & Anderson (2002)	1	2	0	1	0	0	0	1	3	1	2	2	1	1		
Xin & Pelled (2003)	0	1	0	1	0	0	2	1	1	3	2	2	0	0		
Hedlund, Forsythe, et al. (2003)	0	1	0	1	0	0	0	1	1	0	2	2	0	0		
Antonakis, J., Avolio, et al. (2003)	1	1	0	0	0	1	1	0	3	0	0	0	0	0		
Dvir & Shamir, 2003)	0	0	0	1	0	0	0	1	1	3	2	0	0	0		
West, Borrill, et al. (2003)	1	1	0	1	0	0	0	1	1	1	2	2	1	0		
Krause (2004)	1	1	0	1	0	0	3	1	1	1	2	2	0	0		
Howell & Boies (2004)	1	1	0	1	0	1	1	1	1	1	3	2	0	0		
Bligh, Kohles, et al. (2004)	1	0	0	1	0	0	0	1	1	0	0	0	0	0		
Hirst, Mann, et al. (2004)	1	1	0	1	0	0	3	1	1	3	2	2	1	0		
Waldman, Javidan, et al. (2004)	1	0	0	0	0	0	0	0	1	1	2	2	1	0		
Tosi, Misangyi, et al. (2004)	1	0	0	0	0	0	3	0	1	0	3	0	0	0		
Whittington, Goodwin, et al. (2004)	1	1	0	1	0	0	3	1	2	1	2	1	0	0		
de Hoogh, den Hartog, et al. (2005)	1	0	0	3	0	0	3	0	1	3	2	2	1	0		
Rowe, Cannella, et al. (2005)	1	3	0	0	0	0	0	0	0	0	3	2	0	0		
Howell, Neufeld, et al. (2005)	1	1	0	1	0	0	0	1	1	3	3	2	0	0		
Epitropaki & Martin (2005)	0	1	0	1	0	0	0	1	1	1	2	2	0	0		
Arvey, Rotundo, et al. (2006)	0	0	0	0	0	0	2	0	0	3	0	0	0	0		
Ensley, Hmieleski, et al. (2006)	1	1	0	1	0	0	2	1	1	1	2	2	0	0		
Hiller, Day, et al. (2006)	1	1	0	1	0	0	2	1	1	1	2	2	0	0		
Paunonen, Lonnqvist, et al. (2006)	1	1	0	0	0	0	0	0	1	1	2	2	0	0		
Carmeli & Schaubroeck (2007)	1	1	0	1	0	0	3	1	1	1	2	2	1	0		
Schaubroeck, Walumbwa, et al. (2007)	1	1	1	1	0	0	0	1	1	1	2	2	0	0		
Harvey, Stoner, et al. (2007)	1	1	0	1	0	0	1	1	1	1	2	2	0	0		
Cole & Bedeian (2007)	1	1	1	1	0	0	0	1	1	1	2	2	0	0		
Luria (2008)	1	1	0	1	0	0	0	1	1	1	2	2	0	0		
Ligon, Hunter, et al. (2008)	1	1	0	0	0	0	0	0	1	0	2	0	0	0		
Campbell, Ward, et al. (2008)	1	1	0	1	0	0	0	1	1	1	2	2	0	0		
Summary statistics																
% of 0 (irrelevant)	11.9	14.3	92.9	23.8	100.0	95.2	45.2	28.6	7.1	19.0	9.5	16.7	71.4	92.9	44.90	
% of 1 (relevant not corrected)	88.1	76.2	7.1	73.8	0.0	4.8	23.8	71.4	73.8	61.9	0.0	7.1	28.6	7.1	37.41	67.90

Appendix A (*continued*)

	Coding criteria[a]															
Study coded	1a	1b	1c	1d	2a	2b	2c	3	4	5	6a	6b	7a	7b	Total	Total*
Panel F: The Leadership Quarterly (*n* = 42)																
% of 2 (relevant, unknown if corrected)	0.0	4.8	0.0	0.0	0.0	0.0	11.9	0.0	9.5	0.0	78.6	76.2	0.0	0.0	12.93	23.46
% of 3 (relevant, corrected)	0.0	4.8	0.0	2.4	0.0	0.0	19.0	0.0	9.5	19.0	11.9	0.0	0.0	0.0	4.76	8.64
Panel G: Organizational Behavior and Human Decision Processes (*n* = 4)																
de Cremer & van Knippenberg (2004)	1	2	0	1	0	0	0	0	2	1	2	2	0	0		
Brown, Trevino, et al. (2005)	1	2	0	1	0	0	0	1	3	3	2	2	0	0		
Martinko, Moss, et al. (2007)	1	3	0	1	0	0	0	1	1	1	2	0	1	0		
Giessner & van Knippenberg (2008)	1	0	0	1	0	0	2	1	2	1	2	0	1	0		
Summary statistics																
% of 0 (irrelevant)	0.0	25.0	100.0	0.0	100.0	100.0	75.0	25.0	0.0	0.0	0.0	50.0	50.0	100.0	44.64	
% of 1 (relevant not corrected)	100.0	0.0	0.0	100.0	0.0	0.0	0.0	75.0	25.0	75.0	0.0	0.0	50.0	0.0	30.36	54.84
% of 2 (relevant, unknown if corrected)	0.0	50.0	0.0	0.0	0.0	0.0	25.0	0.0	50.0	0.0	100.0	50.0	0.0	0.0	19.64	35.48
% of 3 (relevant, corrected)	0.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0	25.0	0.0	0.0	0.0	0.0	5.36	9.68
Panel H: Personnel Psychology (*n* = 6)																
Tierney, Farmer, et al. (1999)	1	3	0	1	0	0	3	1	2	2	3	3	0	0		
Ployhart, Lim, et al. (2001)	1	1	0	0	0	0	0	0	2	0	2	2	0	0		
Ehrhart (2004)	1	1	0	1	0	0	3	1	3	1	2	2	1	1		
Day, Sin, et al. (2004)	0	1	1	0	0	0	0	0	0	0	2	2	0	0		
Walker, Smither, et al. (2008)	1	0	0	1	0	0	2	0	2	0	2	0	0	0		
Walumbwa, Avolio, et al. (2008)	1	1	0	1	0	0	3	1	1	1	2	2	1	0		
Summary statistics																
% of 0 (irrelevant)	16.7	16.7	83.3	33.3	100.0	100.0	33.3	50.0	16.7	50.0	0.0	16.7	66.7	83.3	47.62	
% of 1 (relevant not corrected)	83.3	66.7	16.7	66.7	0.0	0.0	0.0	50.0	16.7	33.3	0.0	0.0	33.3	16.7	27.38	52.27
% of 2 (relevant, unknown if corrected)	0.0	0.0	0.0	0.0	0.0	0.0	16.7	0.0	50.0	16.7	83.3	66.7	0.0	0.0	16.67	31.82
% of 3 (relevant, corrected)	0.0	16.7	0.0	0.0	0.0	0.0	50.0	0.0	16.7	0.0	16.7	16.7	0.0	0.0	8.33	15.91

Note: †denote a field experiment; *total percentage less coding category 0; to save space we only include the names of the first two co-authors and add "et al." when there are more than two authors.

[a] We coded for the following criteria:

1. Omitted variables:
   - (a) omitting a regressor
   - (b) omitting fixed effects
   - (c) using random-effects without justification
   - (d) in all other cases, independent variables not exogenous
2. Omitted selection:
   - (a) comparing a treatment group to non-equivalent groups
   - (b) comparing entities that are grouped nominally where selection to group is endogenous
   - (c) sample is self-selected or is non-representative
3. Simultaneity:
   - (a) reverse causality
4. Measurement error:
   - (a) not correcting for imperfectly-measured independent variables
5. Common-method variance:
   - (a) independent and dependent variables that are correlated are gathered from the same source
6. Inconsistent inference:
   - (a) using normal standard errors in the potential presence of heteroscedastic residuals
   - (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 without comparing estimates to a limited information estimator.

The previous criteria were coded as follows:

0	Irrelevant criterion
1	Relevant criterion for which the authors did not correct
2	Relevant criterion for which we were unable to determine whether it was corrected by the authors
3	Relevant criterion which the authors addressed.

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*, *3*. (pp. 1277—1366) Amsterdam: Elsevier Part 1.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69—85.

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

Antonakis, J. (2009). "Emotional intelligence": What does it measure and does it matter for leadership? In G. B. Graen (Ed.), *LMX leadership—Game-changing designs: Research-based tools*, *VII*. (pp. 163—192)Greenwich, CT: Information Age Publishing.

Antonakis, J. (in press). Predictors of leadership: The usual suspects and the suspect traits. In A. Bryman, D. Collinson, K. Grint, B. Jackson & M. Uhl-Bien (Eds.), Sage handbook of leadership. Thousand Oaks: Sage Publications.

Antonakis, J., Ashkanasy, N. M., & Dasborough, M. T. (2009). Does leadership need emotional intelligence? *The Leadership Quarterly*, *20*(2), 247—261.

Antonakis, J., & Atwater, L. (2002). Leader distance: A review and a proposed theory. *The Leadership Quarterly*, *13*, 673—704.

Antonakis, J., Avolio, B. J., & Sivasubramaniam, N. (2003). Context and leadership: An examination of the nine-factor full-range leadership theory using the Multifactor Leadership Questionnaire. *The Leadership Quarterly*, *14*(3), 261—295.

Antonakis, J., Cianciolo, A. T., & Sternberg, R. J. (2004). Leadership: Past, present, future. In J. Antonakis, A. T. Cianciolo, & R. J. Sternberg (Eds.), *The nature of leadership* (pp. 3—15). Thousand Oaks: Sage.

Antonakis, J., & Dietz, J. (2010). Emotional intelligence: On definitions, neuroscience, and marshmallows. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(2), 165—170.

Antonakis, J., & Dietz, J. (in press-a). *Looking for Validity or Testing It? The Perils of Stepwise Regression, Extreme-Scores Analysis, and Heteroscedasticity*. http://dx.doi.org/10.1016/j.paid.2010.09.014.

Antonakis, J., & Dietz, J. (in press-b). More on Testing for Validity Instead of Looking for It. *Personality and Individual Differences*. http://dx.doi.org/10.1016/j.paid.2010.10.008.

Antonakis, J., House, R. J., Rowold, J., & Borgmann, L. (submitted for publication). *A fuller full-range leadership theory: Instrumental, transformational, and transactional leadership*.

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

Barling, J., Weber, T., & Kelloway, E. K. (1996). Effects of tranfsormational leadership training on attitudinal and financial outcomes: A field experiment. *Journal of Applied Psychology*, *81*(6), 827—832.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173—1182.

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

Basmann, R. L. (1960). On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association*, *55*, 650—659.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2007). Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal*, *7*(4), 465—506.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238—246.

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*(1), 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.

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.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136—162). Newbury Park, CA: Sage Publications.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (in press). Robust Inference with Multi-way Clustering. Journal of Business and Economic Statistics.

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

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171—246). Chicago: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: R. McNally.

Chen, F. N., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, *36*(4), 462—494.

Cohen, J. (1960). A coefficient agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37—46.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent variable covariance structure analysis. *Psychological Methods*, *12*, 381—398.

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

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636—654.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings.* Chicago, IL: Rand McNally.

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.

de Cremer, D., & van Knippenberg, D. (2004). Leader self-sacrifice and leadership effectiveness: The moderating role of leader self-confidence. *Organizational Behavior and Human Decision Processes*, *95*(2), 140—155.

Diamond, J. M., & Robinson, J. A. (2010). *Natural experiments of history.* Cambridge, Mass.: Belknap Press of Harvard University Press.

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

Dvir, T., Eden, D., Avolio, B. J., & Shamir, B. (2002). Impact of transformational leadership on follower development and performance: A field experiment. *Academy Management Journal*, *45*(4), 735—744.

Eagly, A. H., & Carli, L. L. (2004). Women and men as leaders. In J. Antonakis, A. T. Cianciolo, & R. J. Sternberg (Eds.), *The nature of leadership* (pp. 279—301). Thousand Oaks: Sage.

Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin*, *129*(4), 569—591.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 343—367.

Fiori, M., & Antonakis, J. (in press). The ability model of emotional intelligence: Searching for valid measures. *Personality and Individual Differences*. http://dx.doi.org/10.1016/j.paid.2010.10.010.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87−94.

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.

Frese, M., Beimel, S., & Schoenborn, S. (2003). Action training for charismatic leadership: Two evaluations of studies of a commercial training module on inspirational communication of a vision. *Personnel Psychology*, *56*, 671−697.

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.

Grant, A. M., & Wall, T. D. (2009). The neglected science and art of quasi-experimentation: Why-to, when-to, and how-to advice for organizational researchers. *Organizational Research Methods*, *12*(4), 653−686.

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

Hahn, J. Y., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201−209.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, N.J.: Pearson Prentice Hall.

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

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.

Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, *49*(6), 1377−1398.

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.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153−161.

Henseler, J. (2010). On the convergence of the partial least squares path modeling algorithm. *Computational Statistics*, *25*(1), 107−120.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, *23*(6), 723−744.

Howell, J. M., & Frost, P. J. (1986). A laboratory study of charismatic leadership. *Organizational Behavior and Human Decision Processes*, *43*, 243−269.

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

Hwang, H., Malhotra, N. K., Kim, Y., Tomiuk, M. A., & Hong, S. (2010). A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling. *Journal of Marketing Research*, *37*(4), 699−712.

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

Jones, B. F., & Olken, B. A. (2005). Do leaders matter? National leadership and growth since World War II. *The Quarterly Journal of Economics*, 835−864.

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.

Kline, R. B. (2010). *Principles and practice of structural equation modeling.* New York: Guilford Press.

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.

Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics*, *142*(2), 785−806.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159−174.

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.

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

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

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.

Marcoulides, G. A., & Saunders, C. (2006). PLS: A silver bullet? *MIS Quarterly*, *30*(2), III−IX.

Marsh, H. W., Hau, K. -T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320−341.

Maxwell, J. A. (1996). *Qualitative research design: An integrative approach.* Thousand Oaks, CA: Sage Publications.

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.

McDonald, R. P. (1996). Path analysis with composite variables. [Article]. *Multivariate Behavioral Research*, *31*(2), 239−270.

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.

Mooney, C. Z. (1997). *Monte Carlo simulation.* Thousand Oaks, Calif.: Sage Publications.

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., & 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.

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

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175−220.

Onyskiw, J. E., & Hayduk, L. A. (2001). Processes underlying children's adjustment in families characterized by physical aggression. *Family Relations*, *50*, 376−385.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* 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., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*(4), 531−544.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using stata.* College Station, TX: Stata Press.

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.

Roodman, D. (2008). Cmp: Stata module to implement conditional (recursive) mixed process estimator. http://ideas.repec.org/c/boc/bocode/s456882.html

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41−55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516—524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*(1), 33—38.

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.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*(1), 249—264.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, *26*, 393—415.

Schaffer, M. E., & Stillman, S. (2006). Xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg, xtivreg, xtivreg2 and xthtaylor. http://ideas.repec.org/c/boc/bocode/s456779.html

Schriesheim, C. A., Castro, S. L., Zhou, X., & Yammarino, F. J. (2001). The folly of theorizing "A" but testing "B": A selective level-of-analysis review of the field and a detailed Leader–Member Exchange illustration. *The Leadership Quarterly*, *12*, 515—551.

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.

Shipley, B. (2000). *Cause and correlation in biology: A user's guide to path analysis, structural equations, and causal inference.*Cambridge, UK: Cambridge University Press New York, NY. USA.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422—445.

Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290—312). San Francisco: Jossey-Bass.

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

StataCorp. (2009). Stata Statistical Software: Release 11. College Station, TX: StataCorp LP.

Steiger, J. H. (2001). Driving fast in reverse — The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, *96*(453), 331—338.

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.

Temme, D., Kreis, H., & Hildebrandt, L. (2006). PLS Path Modeling—A Software Review. SFB 649 Discussion Paper 2006-084, Institute of Marketing, Humboldt-Universität zu Berlin, Germany.

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.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1—10.

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

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, *9*(1), 1—26.

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.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). Mason, OH: South Western, Cengage Learning.

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

Yammarino, F. J., Dionne, S. D., Uk Chun, J., & Dansereau, F. (2005). Leadership and levels of analysis: A state-of-the-science review. *The Leadership Quarterly*, *16*(6), 879—919.

Yin, R. K. (1994). *Case study research: Design and methods.* Thousand Oaks, CA: Sage Publications.

Zellner, A., & Theil, H. (1962). 3-Stage least-squares — Simultaneous estimation of simultaneous-equations. *Econometrica*, *30*(1), 54—78.