Contents lists available at ScienceDirect

The Leadership Quarterly

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

Full length article

Instrumental variables estimation: Assumptions, pitfalls, and guidelines

Nicolas Bastardoz [a,*,1], Michael J. Matthews [b,1], Gwendolin B. Sajons [c,1], Tyler Ransom [d], Thomas K. Kelemen [e], Samuel H. Matthews [f]

[a] *Department of Work and Organisation Studies, Katholieke Universiteit, Leuven, Belgium*
[b] *Price College of Business, University of Oklahoma, United States*
[c] *ESCP Business School, Berlin Campus, Germany*
[d] *Dodge Family College of Arts and Sciences, Department of Economics, University of Oklahoma, United States*
[e] *College of Business Administration, Kansas State University, United States*
[f] *College of Business Administration, Gonzaga University, United States*

ARTICLE INFO

Keywords:
Instrumental variables
2SLS
Endogeneity
Causality

ABSTRACT

Researchers striving to ensure rigor in their scientific findings face a common pitfall: Endogeneity. To tackle this problem, scholars have increasingly adopted instrumental variables estimation (IVE). Although there are many published works showing how IVE should be used, many applied researchers still have trouble understanding how to use the method correctly. In this article, we provide a methodological overview of IVE by discussing the underlying conditions valid instruments must satisfy as well as common mistakes made in using IVE. Using simulated data, we further demonstrate the sensitivity of IVE to violations of its conditions. We then take stock of the literature in a social science discipline (i.e., leadership research) and provide insights regarding trends and shortcomings in the application of IVE. Based on our review, we categorize the different types of instruments used and discuss the potential appropriateness of each type. We conclude by providing non–technical guidelines targeted at the study design, analysis, and reporting phases, which will help applied social science researchers to ensure they use IVE correctly.

Scholars in the fields of leadership and management often assert causal claims even when the type of data collected and the estimation strategies used do not allow them to do so (Antonakis et al., 2010; Antonakis et al., 2021; Fischer et al., 2017; Sajons, 2020). This practice is untenable because such research will mislead other scholars and misinform policymakers. Given the availability of techniques that can bolster causal claims, scholars have increasingly called on management researchers to make use of these tools (e.g., Güntner et al., 2020; Hill et al., 2021; Jacquart et al., 2017; Konlechner & Ambrosini, 2019; Sajons, 2020; Shaver, 2020; Sieweke & Santoni, 2020).

One methodological approach to support causal arguments is instrumental variables estimation (IVE; Angrist & Pischke, 2009; Antonakis et al., 2010; Bollen, 2012; Greene, 2003; Hamilton & Nickerson, 2003; Kennedy, 2008; Podsakoff et al., 2012; Shaver, 2005; Wooldridge, 2019), which typically relies on a “two-stage least squares” (2SLS) estimator. When applied correctly, IVE is a powerful means to establish causal effects and thus derive meaningful policy implications. For this reason, many renowned scholars have propagated its use, such as 2021 Nobel laureates Joshua Angrist and David Card (Card, 1995, 2001; Angrist et al., 1996). In fact, in a recent review of 384 management articles that addressed endogeneity or used robustness checks, Hill et al. (2021) report that IVE is by far the most common approach to combat the effects of endogeneity. Furthermore, Wooldridge (2019) reports that 2SLS is the second most popular way to estimate linear equations in applied econometrics, behind only ordinary least squares (OLS). Leadership researchers have also repeatedly emphasized the potential power of IVE to rigorously explore leadership phenomena (Antonakis et al., 2010; Sajons, 2020). As a result, the number of studies applying this method is increasing. However, despite its promise, when applied inappropriately, IVE can yield inconsistent estimates that are equally or even more biased than estimates derived from classic estimation techniques such as OLS (see Angrist & Krueger, 2001). Thus, although IVE can be very help-

[*] Corresponding author.

E-mail addresses: nicolas.bastardoz@kuleuven.be (N. Bastardoz), michael.j.matthews-1@ou.edu (M.J. Matthews), gsajons@escp.eu (G.B. Sajons), ransom@ou.edu (T. Ransom), kelemen@ksu.edu (T.K. Kelemen), matthewss@gonzaga.edu (S.H. Matthews).

[1] Shared first authorship; first authors listed alphabetically.

https://doi.org/10.1016/j.leaqua.2022.101673
Received 18 October 2021; Revised 29 November 2022; Accepted 17 December 2022
Available online xxxx
1048-9843/

ful for increasing causal knowledge, it is paramount that scholars implement it with caution.

Leadership researchers may face several roadblocks that prevent them from adequately applying IVE. For example, they may feel uncertain as to how they can identify valid instrumental variables (IVs), how they should apply IVE, or whether their causal identification using IVE is credible. Moreover, scholars may misunderstand what specific tests need to be conducted and reported or how these respective test results should be interpreted. Each of these mishaps may result in biased and inconsistent estimates and incorrect interpretations, limiting the progress of knowledge in the leadership field. Given the increasing calls for IVE in organizational and leadership studies, we argue that it is high time for a systematic review to assess where the field stands in this regard, identify major pitfalls in current applications of IVE, and help scholars exploit IVE's full potential.

As such, this manuscript makes several contributions to the field. First, "one of the biggest challenges that researchers face when attempting to estimate instrumental variable models has to do with where to find instruments" (Antonakis et al., 2010, p. 1103). Thus, we pinpoint the conditions a valid instrument must satisfy and outline how to identify potential instruments. Based on our review of published studies applying IVE, we include an overview of the main categories of instruments used in leadership research to date and critically discuss to what extent (or in which contexts) these are appropriate. In this way, we provide a resource to help scholars find valid IVs and, just as importantly, avoid invalid IVs. Second, we contribute to leadership studies by highlighting the different empirical tests required for IVE and evaluating their application in leadership research. We supplement our discussion with a Monte Carlo simulation (available in an Online Appendix at https://osf.io/amqzh/) to demonstrate how invalid (i.e., weak, not (as if) random, or non-excludable) instruments affect the estimates provided by IVE. Third, we sketch the way forward by offering concrete guidance regarding the study design, analysis, and reporting phases when using IVE. In this way, we offer a practical resource for leadership scholars to leverage IVE in their studies.

Our contributions parallel other methodological papers in the area of leadership, such as articles focusing on experimental designs (Podsakoff & Podsakoff, 2019) and polynomial regression (Tsai et al., 2022) methodologies. Our review focuses on leadership because leadership scholars have been at the forefront of pushing for clean causal identification since the seminal paper by Antonakis and colleagues (2010). This high-level paper on causal claims triggered a stark focus on the importance of causal identification and has been highly influential in guiding leadership and management studies (see, for instance, Antonakis et al., 2016; Banks et al., 2021, Fischer et al., 2020; Günther et al., 2020; Martin et al., 2021; Sajons, 2020). Furthermore, *The Leadership Quarterly*—the highest-impact journal in the leadership field—has put causal identification at the center of its editorial policy (Antonakis et al., 2019) and has devoted several recent Special Issues to tackling issues of causality and endogeneity (Clapp-Smith et al., 2018; Garretsen et al., 2020; Fischer et al., 2020; Jacquart et al., 2020). Critically assessing how well the field is doing in employing one primary method that facilitates identifying causal relationships is key to establishing the nomological net of leadership constructs and improving the rigor of leadership theories. Thus, taken together, we argue that a methodological overview of IVE will improve the appropriate use of this approach and, subsequently, theory-building and theory-testing in leadership research.

The rest of the manuscript is organized in three main sections. First, we contextualize IVE and outline the three conditions for a valid instrument: (a) relevance, (b) (as if) randomness, and (c) the exclusion restriction. For each of these three conditions, we discuss the underlying assumptions and whether the conditions are empirically testable. Based on a Monte Carlo simulation, we demonstrate graphically how deviations from these conditions may result in biased and inconsistent IV estimates. Second, we critically review the leadership literature and reflect on high-level theoretical and empirical trends. As part of this process, we identify seven different IV categories used. The potential validity of each of these types of IVs as well as concrete examples are critically discussed. Third, we propose how the leadership field can move forward in designing, analyzing, and reporting strong IVE studies. We conclude with final thoughts.

Instrumental variables

It is often challenging to identify the causal effect of a predictor[2] x on an outcome variable y. Consider a simple regression such as in Eq. (1), where β_0 is the constant term, β_1 is the coefficient of x, and ε is the error term (also referred to as the disturbance term) capturing all factors influencing y other than variation attributable to x.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

One central assumption of regression analysis is that the predictor x is exogenous, meaning that x does not correlate with the y-equation's error term ε. If the predictor x correlates with the error term ε, it is said to be *endogenous* (implying that cov $(x, \varepsilon) \neq 0$), which means that the estimated coefficient of the predictor x is biased and inconsistent.[3]

Scholars have identified various possible sources of endogeneity, such as omitted variables (including common method variance), omitted selection, measurement error, and simultaneity or reverse causality (Angrist & Pischke, 2010; Antonakis, Bendahan, Jacquart, & Lalive, 2010; Cameron & Trivedi, 2005; Hill, Johnson, Greco, O'Boyle, & Walter, 2021; Kennedy, 2008). Although the threat of endogeneity has recently been discussed more intensely in the leadership and management literatures (e.g., Antonakis et al., 2010; Antonakis et al., 2021; Hamilton & Nickerson, 2003; Ketokivi & McIntosh, 2017; Reeb et al., 2012; Sajons, 2020; Shaver, 2005), scholars still often fail to account for it by taking appropriate empirical remedies. Case in point, in their review of 110 empirical leadership studies, Antonakis et al. (2010) found that for potentially affected studies 70 % did not correctly deal with measurement error, and 77 % did not appropriately take common method variance into account. More recently, Sajons (2020) examined 74 articles using perceptions as predictor or mediator variables and concluded that not a single study had empirically accounted for the predictor's potential endogeneity. Furthermore, a quick survey of reviews within the leadership domain (e.g., Antonakis et al., 2016; Arthur et al., 2017; Fischer et al., 2021) suggests that endogeneity concerns permeate many leadership studies.

However, there are various techniques available that can help scholars tackle or attenuate endogeneity threats through design and estimation strategies, such as randomized controlled trials, difference-in-differences (DiD) estimation, regression discontinuity designs (RDD), and IVE (see Table 5 in Hill et al., 2021). This review focuses on IVE, which has been around for nearly 100 years (Wright, 1928) and has been broadly adopted across various scientific disciplines (e.g., Bollen, 2012; Ketokivi & McIntosh, 2017; Greenland, 2000; see Stock & Trebbi, 2003 for a historical overview of IVE).

The mechanics of IVE are relatively simple. When x is endogenous, we can introduce an IV z to estimate the causal effect of x on y by applying IVE in two stages. In the first stage (Eq. (2)), we regress the predictor variable x on z. Based on this regression, we predict $\hat{x}$ based on the values of our IV z, essentially parsing the exogenous, 'clean' part of the variation in x. In the second stage (Eq. (3)), we use the exoge-

[2] In this manuscript, we do not use the term "independent variable" to avoid any possible confusion with the term "instrumental variable." However, we recognize that the terms predictor and independent variable are often used interchangeably in other contexts.

[3] Note that both of these terms reflect whether the coefficient displays the true effect, yet consistency is a large sample property, whereas unbiasedness is a small sample property.

nous variation in x, $\hat{x}$, to predict y. The coefficient α_1 then represents the causal effect of our predictor x on the outcome y.

$$\text{First stage equation}: \ x = \delta_0 + \delta_1 z + u \tag{2}$$

$$\text{Second stage equation}: \ y = \alpha_0 + \alpha_1 \widehat{x} + w \tag{3}$$

Note that it is critical to perform these two steps using a canned procedure existing in most major statistical packages (e.g., Stata, R) because performing these two steps manually (i.e., actually performing an OLS regression using $\hat{x}$ as predictor in the second stage) will yield incorrect standard errors.

IVE offers some flexibility. It can be used with one or more instruments in the first stage. It can also be used in cases where x is either continuous or discrete or when the instruments are continuous or discrete. Moreover, researchers can add control variables, which must then be included in both stages. More specifically, exogenous controls may safely be added as long as they do not share too much variance with the instrument z. If there is too large of an overlap in variance between z and the controls, the variability in $\hat{x}$ decreases, which inflates the standard errors and thus reduces the precision of the estimate (Angrist & Pischke, 2009). If controls are themselves endogenous, they may only be added when they are uncorrelated with the IV z; otherwise, they—against all intentions of IVE—will reintroduce bias into the estimated effect of x on y.

When interpreting IVE results, scholars need to be aware that IV estimates do not reflect the effect of x on y for the full sample. Rather, they reflect a so-called ‘local average treatment effect’ (LATE), that is, the effect of the predictor on the outcome for those individuals affected by the IV (i.e., those who reacted to the IV z in terms of their value of the instrumented predictor x; for an extensive discussion, see Angrist & Pischke, 2009).[4]

Conditions for instrument validity

First and foremost, it is essential to reiterate that IVE must meet the unbiasedness assumptions of a regular OLS regression in terms of linearity in parameters, random sampling, and no perfect multicollinearity (for a detailed description see Wooldridge, 2019). In addition to these foundational requirements, an IV z must also satisfy three conditions to qualify as a valid instrument: (1) relevance, (2) (as if) randomness, and (3) the exclusion restriction. As we show below, leadership researchers frequently misunderstand, ignore, or incorrectly test these three conditions. Thus, the following sections carefully explain each condition’s intuition, technical requirements, and the extent to which it can be empirically tested. The decision tree in Fig. 1 summarizes the core information for researchers in the quest for a valid IV.

Before describing the three conditions, we would like to highlight several vital issues. First, whether an IV qualifies as valid depends on the model to be estimated. The exact same variable may qualify as a valid instrument in one model but not in another.[5] Second, if a researcher plans to collect primary data and use IVE, it is critical to reflect on a valid instrument in the ex-ante study design phase. Post-hoc, it is extremely difficult to find an IV that satisfies all three conditions. Third, an instrument *must satisfy every condition* to be valid. To drive this point home, we report in the online Appendix a Monte Carlo simulation showing what happens when a condition is not satisfied. Fig. 2 plots distributions of estimated coefficients derived from this Monte Carlo simulation, where the true effect of an endogenous variable x on y is 1 (the black, vertical, dot line). Each graph reports average quantities of interest across 500 simulation replications of a sample of size 1,000. At the peak of each distribution curve, we report the average estimate. Below each graph, we report information that indicates whether each of the three IVE conditions is satisfied: the first-stage F-statistic; the correlation between the instrument and an omitted variable; and the correlation between the instrument and the error term ε.

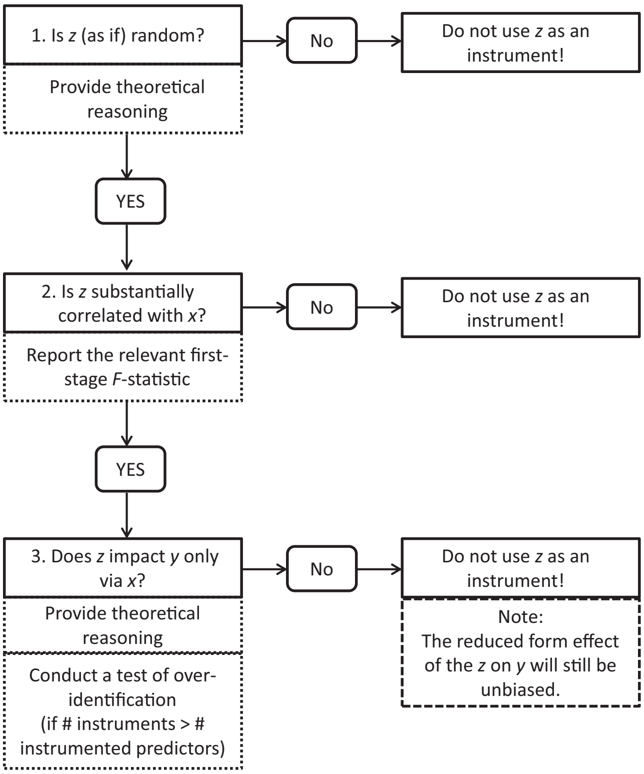

1. Is *z* (as if) random?
Provide theoretical reasoning
No
Do not use *z* as an instrument!
YES
2. Is *z* substantially correlated with *x*?
Report the relevant first-stage *F*-statistic
No
Do not use *z* as an instrument!
YES
3. Does *z* impact *y* only via *x*?
Provide theoretical reasoning
Conduct a test of over-identification (if # instruments > # instrumented predictors)
No
Do not use *z* as an instrument!
Note: The reduced form effect of the *z* on *y* will still be unbiased.

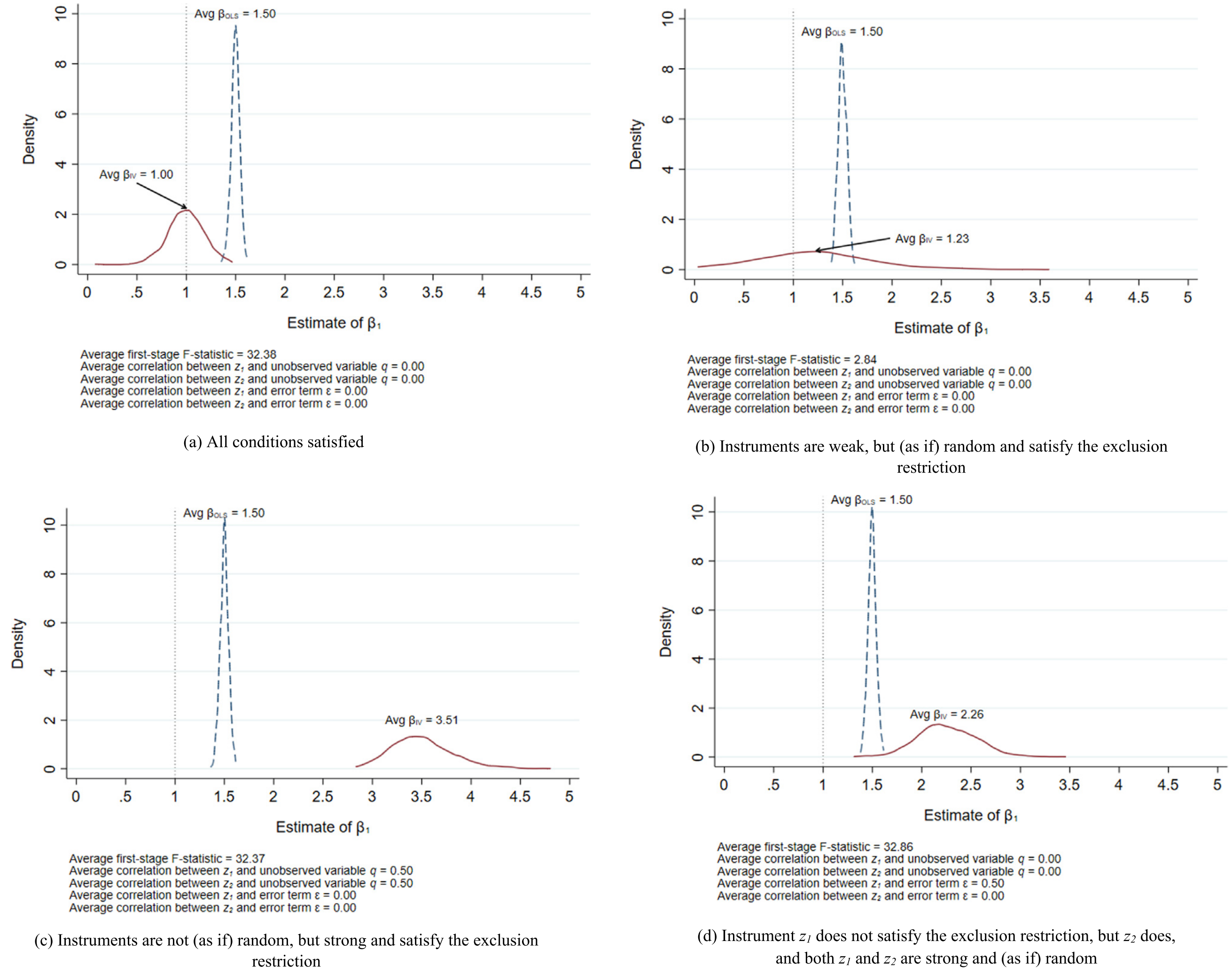
Fig. 1. Flow Chart for Identifying a Valid Instrument.

It is important to point out that the IV estimator is always biased, but is consistent when its assumptions hold (Wooldridge, 2019). Fig. 2 (Panel A) represents a situation whereby all three conditions are satisfied; it shows that IVE (the red, solid curve) leads to consistent estimates (estimated coefficients are centered around the true effect of 1) whereas the standard OLS estimator (the blue curve, dashed line) yields inconsistent estimates that have a bias of about 50 %. The distribution of the IV estimator is more spread out than that of OLS, demonstrating how consistency in IVE comes at the cost of efficiency. As we will describe in more detail below, the violation of just one of the three necessary conditions leads to inconsistent IV estimates (Fig. 2, Panels B, C, and D).

We also wish to address a common source of confusion when evaluating the conditions of IVE. Some scholars only demarcate between two conditions (Podsakoff et al., 2012), others, three (Canan et al., 2017; Hernán & Robins, 2006), others, four (e.g., Gennetian et al., 2008), and still others, five (Angrist et al., 1996). In particular, economists tend to group the (as if) random and exclusion restriction con-

[4] An interesting application of IVE is in randomized experiments where one is interested in the effect of the treatment on some outcome, but not everyone in the sample complies with the treatment allocation. Aside from the effect of the treatment allocation on the outcome (the so-called intent-to-treat effect represented by the simple effect of z on y, which is also the IV reduced-form effect; see later in the manuscript), one can via IVE also compute the effect of the treatment for those who actually complied with the treatment (i.e., those actually treated). The allocation to experimental treatment (z) here serves as an instrument for treatment compliance (x) to obtain the LATE, and IVE corrects for potential selection into (non–)compliance.

[5] For example, Flammer, Hong, & Minor, 2019 use the enactment of state-level constituency statutes as an instrument for corporate social responsibility (CSR) contracting. We consider this instrument to be valid in their analysis, yet the same instrument would hardly satisfy the relevance condition if it were used as an IV for corporate board members’ gender (Bechtoldt, Bannier, & Rock, 2019). Mellon (2022) provides an excellent illustration of the general idea that instruments in one situation may not be useful in another. He analyzes 217 different studies that use weather as an instrument and shows that exclusion restriction violations are common.



(a) All conditions satisfied

(b) Instruments are weak, but (as if) random and satisfy the exclusion restriction

(c) Instruments are not (as if) random, but strong and satisfy the exclusion restriction

(d) Instrument z_1 does not satisfy the exclusion restriction, but z_2 does, and both z_1 and z_2 are strong and (as if) random

Fig. 2. Sampling Distributions of OLS and IV Estimators Under Different Conditions.
Notes: Graphs depict the sampling distributions of the OLS and IV estimators using a sample size of 1000 and 500 simulation replications. The true value of β_1 is equal to 1 and is represented by the black vertical dotted line. The sampling distribution of OLS estimates is represented by the dashed blue curve. The sampling distribution of IV estimates is represented by the solid red curve. Below each graph we report several average values of interest, where the average is taken across all 500 simulation replications. These include the first-stage *F*-statistic value, as well as the correlations between the instrument and omitted variable q and error term ε. These latter two correlations are unknowable in practical settings because they rely on unobservables that are not observed but can be manipulated in a simulation. (For interpretation of the references to colors in this figure legend, the reader is referred to the web version of this article.)

ditions together (Kennedy, 2008). The primary reason for the variety of classifications of conditions are variations in models. As described in Angrist and Pischke (2009), in a model with constant causal effects (i.e., the causal effect of the predictor x on the outcome y is constant from observation to observation), the exclusion restriction and (as if) randomness conditions cannot be separated. However, in a more realistic model with heterogeneous causal effects (i.e., the causal effect of x on y differs from observation to observation), it is necessary to distinguish between the exclusion restriction and (as if) random conditions. Whereas both conditions formally reflect that the instrument z and the second-stage error term w are uncorrelated ($\text{Cov}(z, w) = 0$), the reasons for why this is the case differ (as outlined in our discussion of the conditions below). Throughout this paper, we follow trends in the leadership literature (e.g., Sajons, 2020) and demarcate three conditions to (1) build on the more general and less "highly stylized" (Angrist & Pischke, 2009, p. 111) framework of heterogenous causal effects, and (2) to emphasize the distinction between the (as if) random condition and the exclusion restriction, each of which is empirically not fully testable and which authors must thus also carefully evaluate at a theoretical level.

Relevance condition

The *relevance condition* indicates that the IV z needs to be substantially correlated with the endogenous predictor x. An IV must be "relevant" or "strong" because only when z predicts a sufficient part of the exogenous variation in x will it allow for the consistent estimation of the effect of x on y using IVE. If an instrument is weak, even if the other conditions hold, IVE can produce substantially biased and inconsistent estimates. Fig. 2 (Panel B) depicts such a situation based on our Monte Carlo simulation. In presence of a weak instrument (first-stage *F*-statistic is 2.84, averaged across the 500 replications), the mean IV estimated coefficient peaks at 1.23 (the population value is 1) and the distribution of estimated coefficients is very spread out. The distribution of IVE coefficients highlights how IV estimates may be even more biased than OLS coefficients when instruments are weak.

When looking for an appropriate IV, researchers often start by considering which variable could *theoretically* be a strong predictor of x. This step is useful to identify potentially relevant instruments; yet, what finally matters for instrument validity is whether z is *empirically* relevant. More specifically, the instrument (or set of instruments) must have a (joint) *F*-statistic exceeding a certain critical value in the first stage regression to be relevant. Traditionally, the critical threshold value for a single strong IV has been 10. However, more recently, Stock and Yogo (2005) have developed model-specific critical values that account for the number of instruments, the number of endogenous predictors, the maximum desired relative bias, and the estimator used. These model-specific critical values are more precise and can automatically be displayed with the IVE output in most common statistical software (e.g., using the post-estimation command *estat firststage* after *ivregress* in Stata).

When reporting the *F*-statistic from the regression output, researchers must take great care to report the *F*-statistic of only the instrumental variable(s), rather than the *F*-statistic of all first-stage covariates including the controls. To improve transparency in this regard, researchers need to report the *F*-statistic's degrees of freedom. Finally, it is essential for scholars to note that basing the argument for instrument relevance merely on the fact that z and x significantly correlate—rather than using the *F*-statistic for excluded instruments—is incorrect. Two variables may display a significant correlation even when having an *F*-statistic below the critical values either because the correlation is simply not high enough or because of a shared correlation between the instrument and controls with the endogenous predictor.

In contrast to the other conditions of IVE, it is straightforward to test instrument relevance empirically. However, because a researcher will only know whether an IV satisfies the relevance condition after data collection, we recommend pre-testing an instrument relevance for primary data collection efforts (whether experimental or survey-based) with a smaller sample size whenever it is possible to do so. Scholars can then perform power calculations to see whether the condition will likely hold in the full data set and will also be able to arrive at a reasonable estimation of the needed sample size for sufficient statistical power.

(As if) random condition

Second, the IV needs to meet the *(as if) random condition* (Angrist & Pischke, 2009). This criterion formally describes that the instrument z must not itself be endogenous (i.e., must not be correlated with omitted variables q; cov $(z, q) = 0$).[6] When this condition is violated, IV estimates can be inconsistent with a substantial amount of bias. Fig. 2 (Panel C) demonstrates, based on our Monte Carlo simulation, that IV estimates center around a value of 3.51, which is far from the population value of 1 and considerably more so than OLS.

The (as if) random condition may hold when the instrument z is truly randomized by a researcher (e.g., experimentally randomized instrumental variables (ERIV); Sajons, 2020), by an external shock (e.g., an unexpected war), or by nature (e.g., gender composition of the first two children, Angrist & Evans, 1998, or damages caused by natural disasters, Deuchert & Felfe, 2015). Nevertheless, the condition can likewise be satisfied when z is "as if" random. An example from leadership research would be the (as if) random assignment of a female versus male leader to a mayoral position when considering only communities where a woman won or lost an election by a nominal margin. In this instance, one could reasonably assume that the community characteristics do not endogenously drive the gender of the political leader (see Arvate et al., 2018).

Importantly, it is impossible to empirically test whether the (as if) random condition holds because the error term of an equation (and thus its relation to z) is unobservable. However, researchers can provide reassuring evidence regarding the (as if) randomness of their IV. Scholars must theoretically describe—based either on fundamental "theory", strong intuition, or knowledge of the institutional environment—why they expect their instrument to satisfy this condition. That is, scholars should convincingly explain why their instruments can be considered (as if) random. Beyond this theoretical explanation, scholars should also provide evidence that an IV is balanced across observables, especially in the case of categorical IVs (e.g., experimental treatments). In that case, the balance can be investigated via a classic randomization check. In any case, even if z is balanced across observables, it does not guarantee that it is also balanced across unobservables.

Exclusion restriction

The third condition a valid instrument must satisfy is the *exclusion restriction* condition. This condition stipulates that the IV must only influence the dependent variable y via the instrumented predictor x, but not directly nor via other channels (formally, cov $(z, w \mid \boldsymbol{X}) = 0$), where $\boldsymbol{X}$ is a vector of included covariates and the endogenous predictor x. If the instrument correlates with the second-stage equation error term—either directly or indirectly through correlation with other predictors besides $\boldsymbol{X}$—the condition is not met, and the IV estimates will be biased and inconsistent. Fig. 2 (Panel D) from our Monte Carlo simulation depicts a situation in which one of the two instruments does not satisfy the exclusion restriction (but the other two conditions are satisfied). In this case, IV estimates center around 2.26 and are again considerably off the population value of 1, and substantially more so than OLS.

When considering a model with possibly heterogeneous causal effects, the (as if) random condition and the exclusion restriction must be satisfied separately (Angrist & Pischke, 2009), and thus both conditions need to be carefully discussed in the selection of IVs. As our review indicates, leadership researchers often incorrectly combine the two conditions in their argumentation, substantially weakening their instruments' credibility. Morevoer, this condition does not imply that z and y must not significantly correlate. In fact, if z is relevant and there is a true effect of x on y, then an instrument will mechanically correlate with the outcome even when the exclusion restriction is satisfied (Sajons, 2020). Our review finds that a required null correlation between z and y is another serious misconception among leadership scholars.

Empirically, when a model is overidentified (i.e., when there are more IVs than instrumented predictors), a test of overidentification such as the Hansen-Sargan test (Hansen, 1982; Sargan, 1958) or the χ^2 test of overall model fit (Bollen, 1989; Jöreskog, 1969) should be performed and reported. The Hansen-Sargan test indicates whether the second stage residuals correlate with n (number of IVs) linear functions of the instruments (Wooldridge, 2019).

When conducting and interpreting an overidentification test, it is vital to be aware of four important aspects. First, the test's fundamental assumption is that there is at least one valid IV; however, this is a non-testable assumption. If all instruments are invalid, the overidentification test may fail to detect that the exclusion restriction is not satisfied because the estimates of the different (combinations of) instruments may be similar but equally off (for a theoretical example, see Wooldridge, 2019). Scholars must thus be aware that there are situations where a set of invalid IVs would still pass the overidentification test. Therefore, an overidentification test should not be considered a definite test of instrument validity. Wooldridge (2019) points out that scholars should not be comfortable just because a test of overidentification is passed; instead, scholars must also explain, based on theory or strong intuition, why the chosen instruments would plausibly satisfy this condition (Antonakis et al., 2010).

Second, the overidentification test is only trustworthy under constant causal effects. With heterogenous effects, each instrument would estimate a different LATE, which could then result in a significant overidentification test even when the instruments are in fact valid.

[6] Or, more strictly speaking, must not be dependent on q in any linear or non-linear way.

Third, researchers need to be careful that the overidentification test is sufficiently powered; otherwise, it may fail to reveal existing statistical differences. An overidentification test may fail to indicate that something is amiss only because of a lack of statistical power when running the test. Fourth, the test of overidentification can be interpreted as a test of the instruments satisfying the exclusion restriction only under the assumption that all instruments plausibly satisfy the (as if) random condition. If one instrument may not be (as if) random, a failed test of overidentification instead indicates that the exclusion restriction and/or the (as if) random condition is violated for at least one instrument.

Overall, a non-significant overidentification test allows tentatively concluding that the exclusion restriction is satisfied when (a) the sample size is sufficiently large, (b) constant causal effects are plausible, (c) authors can convincingly argue that at least one instrument is valid, and (d) authors can provide a strong rationale for all their instruments satisfying the (as if) random condition. In such a case, it can thus be useful to be overidentified.

Estimators

We now turn to discuss the role of estimators in IVE. IVE can be conducted via different estimators (i.e., calculation methods). As our review indicates, the most common estimator is “two-stage least squares” (2SLS), which relies on least squares estimation in both stages (e.g., using the command *ivregress 2sls* in Stata). Note that 2SLS is an estimator and should not be amalgamated with the general IV approach. IVE can also be estimated via maximum likelihood (ML). Within the broad category of ML estimators, researchers frequently use limited information maximum likelihood (LIML; e.g., using the command *ivregress liml* in Stata). Because LIML tends to be more robust in the presence of weak instruments (Bascle, 2008; Hahn & Inoue, 2002; Greene, 2003), Stock and Yogo (2005) report different critical values for instrument relevance depending on whether a 2SLS or a LIML procedure is applied. Also, in the presence of heteroscedasticity, using a generalized method of moments (GMM; e.g., using the command *ivregress gmm* in Stata) estimator can be preferable (Baum et al., 2007). Again, it is important for scholars to remember that these estimators need to be applied via the respective statistical software (rather than manually) to obtain the correct standard errors.

There are also alternative estimators for IVE. For instance, structural equation models (SEM) can be estimated with ML, representing a full information estimator. In this procedure, IVE is performed by covarying the disturbance of all endogenous variables in the model (Antonakis et al., 2010; Shaver, 2005). Scholars can also perform different types of augmented regression. An example here is the control function approach (Heckman & Robb, 1985), in which the researcher estimates the first stage, saves the predicted residuals and includes these residuals together with the endogenous predictor x in the second stage. This procedure is also known as the two-stage residual inclusion procedure. It yields the same results as 2SLS for linear models but may show different results for non-linear ones (e.g., when the second stage is estimated via a Probit model). Note that the Heckman correction approach (Heckman, 1979) is a specific type of control function procedure.

Endogeneity tests

After performing IVE, researchers can perform a test of endogeneity. Such tests compare an efficient yet potentially inconsistent estimator (usually OLS) with an inefficient yet consistent estimator stemming from IVE (assuming the three conditions are met). If the estimates differ statistically such that the endogeneity test is significant, it is likely that the predictor x is indeed endogenous, and researchers should err on the side of caution by relying on the consistent estimator. The appropriate endogeneity test depends on the kind of IV estimator. After 2SLS, common options are to perform the Wu-Hausman test (Wu, 1973; Hausman, 1978) or the Durbin (score) χ^2 test (Durbin, 1954), which are also the default options displayed in Stata’s postestimation command (*estat endog* after *ivregress*). After 2SLS with robust standard errors, researchers should report Wooldridge’s (1995) robust score. After ML, one can test for endogeneity by performing a Wald test on the significance of correlated disturbances for all pairs of endogenous variables. For the control function approach, an endogeneity test is incorporated in the second-stage regression in the form of the p-value for the included residuals; if the p-value is significant, the endogeneity test is significant.

Notably, the outcomes of the test do not *prove* that endogeneity is absent; rather, these tests can provide *evidence* that the likelihood of endogeneity is low. Ultimately, a statistical test cannot completely rule out the potential presence of omitted variables, simultaneity, measurement error, or selection because endogeneity is a property involving the error term, which by definition is not observable. As a general rule, we thus urge researchers using IVE to always report a consistent estimator along with an efficient but potentially inconsistent estimator. Note that, for all endogeneity tests to be valid, the IV(s) must be valid. For instance, in the presence of weak IV(s), the endogeneity test will lack the power to detect differences between the two estimators (Hahn et al., 2011; Hausman et al., 2005). When IV(s) are not valid, researchers should not rely on an endogeneity test to conclude that a variable is likely exogenous.

The reduced form

In the context of IVE, it can also be very informative to estimate and interpret the so-called ‘reduced-form’ equation, which refers to estimating the direct effect of the instrument z on the outcome y as represented by coefficient π_1 in Eq. (4):

$$\text{Reduced form equation: } y = \pi_0 + \pi_1 z + \nu \tag{4}$$

Oftentimes, the direct effect of an IV on the outcome y is interesting. For instance, when an experiment or a policy intervention is used as IV, knowing the direct effect of such IV on the outcome may be valuable (e.g., see Jolly, Krylova, & Phillips, 2020; Liang et al., 2018; Lonati, 2020 as examples from our review). Importantly, of the three conditions for a valid instrument, estimating the reduced form equation only requires (as if) randomness to obtain consistent estimates.

So even when an instrument does not satisfy the relevance criterion and/or the exclusion restriction, the reduced form can still yield interesting policy implications (Chernozhukov & Hansen, 2008). Whereas the reduced form cannot speak to the mediating channel (i.e., the effect of z on y could go via x, but also via other channels), the effect of the instrument on the outcome will be causally identified. Also, note that the IV estimate technically reflects the ratio of the reduced form effect of z on y (π_1 in Eq. (4)) to the first-stage effect of z on x (δ_1 in Eq. (2)) as depicted in Eq. (5):

$$\alpha_1 = \frac{\pi_1}{\delta_1} \tag{5}$$

Review of IVE in leadership research

In the previous section, we detailed the underlying conditions of IVE and optimal ways of reporting; however, the question remains whether leadership scholars are adequately applying these principles. In this section, we take stock of the field to assess the strengths and weaknesses of current practices in comparison to best practices. To prepare for this review, we followed standard procedures for reviewing the literature. First, on August 5, 2021, we conducted a literature search to identify relevant articles without restricting the publication date. In particular, we searched the following top-tier journals for the words “instrumental” and “leader*” to appear anywhere (i.e., title, abstract, keywords, text, references) in the manuscript: *Academy of*

Management Journal, Administrative Science Quarterly, Journal of Applied Psychology, Journal of Management, Journal of Management Studies, Journal of Organizational Behavior, Organizational Behavior and Human Decision Processes, Organization Science, Personnel Psychology, Strategic Management Journal, and *The Leadership Quarterly*. These journals were selected based on recommendations for selecting journals included in a review (Short, 2009). This process resulted in an initial set of 337 articles. After reviewing the articles to determine their relevance, we eventually arrived at 68 articles containing 77 studies to be included in this review (a PRISMA diagram is available online at https://osf.io/amqzh/).

Once all the articles were identified, we jointly created a comprehensive coding sheet, and the first author coded the entire set of articles. To ensure the reliability of the coding, the third author coded five random articles independently (after a calibration phase of three articles). These five articles included 115 coding events (i.e., 23 coded variables per manuscript). Per chance, the expected agreement would be 21.44 %, yet coders agreed on 81.74 % of the coding events (94 out of 115), which is consistent with agreement statistics in methodological reviews. The kappa statistic is 0.77 (SE = 0.04, z = 17.1, p <.001), suggesting substantial agreement between the two coders (Landis & Koch, 1977). Disagreements were discussed and reconciled between the two authors. For transparency, the full final coding sheet is available in an online Appendix (https://osf.io/amqzh/).

High-level insights

Our review highlights several interesting, high-level insights regarding the state of the field with regard to IVE. Fig. 3 provides a chronological overview of published articles showing that the use of IVE in the leadership field has proliferated; indeed, over half of the studies in our review were published in the last three years. The journal that published the most studies using IVE was *The Leadership Quarterly* (34 studies), as seen in Table 1. Our review also indicates that IVE has been used across a wide array of macro and micro leadership areas, from understanding CEOs (e.g., McDonald, Khanna, & Westphal, 2008), boards (e.g., Gupta & Wowak, 2017) and top management teams (e.g., Carpenter & Sanders, 2004) to understanding managers and supervisors (e.g., De Vries, 2012).

We provide descriptive statistics regarding the use and implementation of IVs in Table 2. This table highlights some interesting information regarding the current use of IVE in leadership studies. First, Table 2 indicates that the main reported reasons for using IVE are omitted variables (N = 47), selection (N = 18), and reverse causality (N = 13). Authors rarely cited measurement error (N = 5). Second, a majority of studies used IVs only for the purpose of performing IVE (N = 50) whereas a minority of studies adopted theoretically meaningful IVs, typically with the intent to test mediation models (N = 21). Third, although it was common to apply only one (N = 24 studies) or two (N = 24 studies) IVs, some studies included up to thirteen (Kwok, Hanig, Brown, & Shen, 2018), with a mean number of 2.86 and a median of 2. Fourth, the IVs used were mainly only continuous (N = 45), with a relevant number of studies using both continuous and categorical (N = 17) or only categorical (N = 13) IVs. Fifth, IVE was equally used as a primary design (N = 38) as it was used as a supplementary analysis or robustness check (N = 39). Sixth, IVE was reported mostly in the main sections of the paper (N = 36); however, a substantial number of studies did not explicitly report the results (N = 11) or reported them only sparsely in text or a footnote (N = 8). In certain instances, scholars only stated that the hypotheses remained supported after also performing IVE without further details (e.g., Zhang & Gimeno, 2016). In a majority of cases (57 out of 77), IVE results were paired alongside other estimators such as OLS. Seventh, the estimation technique most commonly used was 2SLS (N = 34), but in nearly one-fifth of articles authors were unclear or did not report their estimation technique (N = 15). Finally, less than half of the studies (N = 36) included an endogeneity test. The most frequent endogeneity test took the form of a Durbin (1954) or Wu-Hausman (Wu, 1973; Hausman, 1978) test (N = 24). Most studies (N = 40) did not report any endogeneity test whatsoever.

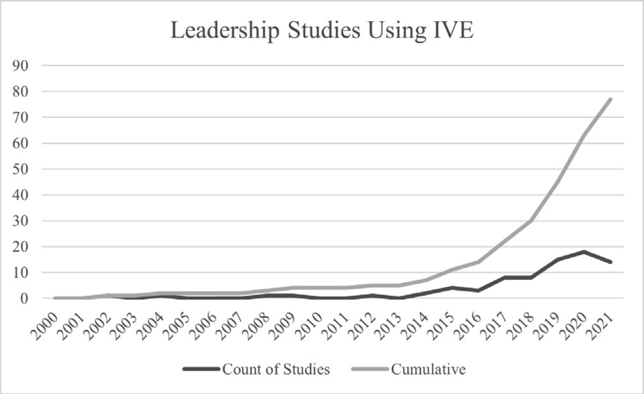


Fig. 3. Chronological Overview of Leadership Studies Using IVE.
Note. Of the 77 studies, over 94% came after Antonakis et al. (2010), which formally introduced IVE into the leadership literature.

Table 1
Overview of Leadership Studies Using IVE by Journal.

Journal Title	Number of Studies
The Leadership Quarterly	34
Strategic Management Journal	11
Journal of Management	9
Academy of Management Journal	9
Organization Science	6
Journal of Management Studies	3
Journal of Applied Psychology	2
Administrative Science Quarterly	2
Journal of Organizational Behavior	1

IVE conditions

One of the major questions is whether leadership scholars are adequately addressing the conditions that an IV must satisfy. Thus, we coded in our review how scholars discussed and justified each condition. In general, authors often failed to address the IV conditions sufficiently, with some manuscripts not discussing them at all (e.g., Gangloff, Connelly, & Shook, 2016; Gomulya, Wong, Ormiston, & Boeker, 2017; Klein, Chaigneau, & Devers, 2021).

Concerning the relevance condition, only 28 studies (38 %) clearly reported the correct F-statistic. However, many studies (34 %) did not at all report the first-stage F-statistic of the excluded instruments. Another concern is whether scholars are incorrectly reporting the F-statistic for all predictors included in the first-stage regression, as opposed to correctly reporting only the F-statistic for the excluded instruments. For instance, Bharanitharan, Lowe, Bahmannia, Chen, & Cui, 2021 reported an F-statistic with sixteen degrees of freedom. In contrast, they reported having only nine excluded instruments. Thus, the missing seven degrees of freedom likely result from control variables in the first and second stage equation, mistakenly included for the calculation of the F-statistic. Finally, in a substantial number of cases, authors did not report the relevant F-statistic's degrees of freedom (27 % of the studies), limiting readers' ability to evaluate its trustworthiness.

The (as if) random condition was the most ignored condition. Only 12 studies (16 %) theoretically justified the (as if) randomness of their IVs (e.g., Flammer, Hong, & Minor, 2019; Lonati, 2020; MacLaren

Table 2
Summary of IVE Application in Leadership Studies.

IV Study Characteristics and Usage	Amount	Percent
Study Design		
Secondary & archival	41	53 %
Cross-sectional survey	18	23 %
Experimental	15	19 %
Combination of survey & archival data	3	4%
Type of endogeneity concern stated in paper		
Omitted variables	47	61 %
Selection	18	23 %
Reverse causality	13	17 %
Measurement error	5	6 %
Broad discussion of endogeneity	4	5 %
Simultaneity	2	3 %
Number of IVs used		
One	24	31 %
Two	24	31 %
Three or more	27	35 %
Unclear	2	3 %
IV measurement: continuous, categorical, or both		
Continuous	45	58 %
Categorical	13	17 %
Both	17	22 %
Unclear	2	3 %
Type of IV		
Leader/follower individual differences	21	27 %
Internal organizational factors	20	26 %
External organizational factors	19	25 %
Leader/follower (non-trait) personal experiences	16	19 %
Experimental manipulations	14	18 %
Internally induced instruments	8	10 %
Natural experiments	2	4 %
Reporting of IV analyses		
Main analyses	36	47 %
Robustness check (in-text)	16	21 %
Robustness check (supplementary material)	6	8 %
Discussed briefly in-text	8	10 %
Not reported	11	14 %
IV: Theoretical relevance		
Purely empirical IV	50	65 %
Theoretically meaningful	21	27 %
Unclear	6	8 %
IV estimation technique		
Two-stage least squares (2SLS)	34	44 %
Maximum Likelihood using structural equation modeling (SEM) with correlated disturbances	14	18 %
Generalized method of moments (GMM)	4	5 %
Limited-information maximum likelihood (LIML)	3	4 %
Two-stage residual inclusion (2SRI)	3	4 %
Other	9	12 %
Unclear/unreported	15	19 %
First stage regression results reported		
No	40	52 %
Yes (main text)	26	34 %
Yes (not in the main text)	11	14 %
Relevant *F*-statistic reported		
Yes	28	36 %
No	26	34 %
Unclear	20	26 %
Irrelevant	2	3 %
Reported incorrect *F*-statistic	1	1 %
F-statistic benchmark used		
Unclear/unreported	22	45 %
Context-specific critical value	17	35 %
Rule of thumb of 10	10	20 %
Discussion of (as if) random condition		
No	52	68 %
Yes	12	16 %
Randomized experiment or internally constructed instruments	9	12 %
Unclear	4	5 %

Table 2 (*continued*)

IV Study Characteristics and Usage	Amount	Percent
Exclusion restriction theoretically or logically justified		
No	43	56 %
Yes	17	22 %
Only arguing it is satisfied because there is no theoretical reason to expect a direct relationship with the outcome	11	14 %
Falsely arguing it is satisfied because there is no empirical relationship between the IV and the outcome	6	8 %
Is the model over-identified?		
Yes	45	58 %
No	27	36 %
Part of the models	3	4 %
Unclear	2	3 %
Was an over-identification test reported (when a model was over-identified)		
Yes (Hansen-Sargan *J*-test)	28	58 %
Yes (χ^2 for SEM)	7	15 %
Yes (Basman χ^2)	2	4 %
Yes (Anderson & Rubin χ^2)	1	2 %
No	7	15 %
Unclear	3	6 %
Others	1	2 %
Results from efficient estimation technique reported		
Yes (in main analyses)	55	71 %
Yes (as comparison with IV)	2	3 %
No	20	26 %
Endogeneity test reported		
No	40	52 %
Yes, the Wu-Hausman statistic	18	23 %
Yes, Durbin (1954) and Wu-Hausman statistics	6	8 %
Yes, a Hausman (1978) test	5	6 %
Yes, Wald test for correlated disturbances in SEM	5	6 %
Yes, the *C*-Score (after GMM)	1	1 %
Yes, significance of residuals (with 2SRI)	1	1 %
Simply comparing average treatment effect (ATE) and local average treatment effect (LATE) estimates (no test)	1	1 %
Is the endogeneity test significant (when tested)?		
Not significant	16	44 %
Significant	15	42 %
Not significant for some, significant for others	5	14 %

et al., 2020). Sometimes, articles did not theoretically argue for the (as if) randomness of their IVs because this condition was likely satisfied by design such as when using randomized experiments or internally generated instruments (e.g., Meslec, Curseu, Fodor, & Kenda, 2020). A frequent phenomenon observed in our review is the tendency for authors to not distinguish between the (as if) random condition and the exclusion restriction in their discussion of instrument validity. This habit likely originates from the economics literature using the constant causal effect framework, which combines the two conditions into an overall exogeneity condition. However, such practice is problematic when authors fail to (1) provide a strong theoretical rationale for why their instruments plausibly satisfy each of the two conditions, and (2) justify why assuming constant causal effects is reasonable.

Furthermore, whereas arguing for the (as if) randomness is a theoretical exercise, certain scholars suggest having “performed appropriate statistical tests that confirmed both instrument relevance and exogeneity” (Kish-Gephart & Campbell, 2015, p. 1624). This type of language can mislead pretending that (as if) randomness is fully empirically testable, which is not true. Our review indicates that the lack of discussion of the (as if) randomness condition is one of the greatest weaknesses in the current state of the leadership literature and the most significant threat to the valid adoption of IVE in leadership research.

Lastly, the results were mixed with respect to the exclusion restriction, with most articles not discussing this condition at all ($N = 43$). Some authors discussed the exclusion restriction only under the “exogeneity” header and did not conceptually parse this condition from the (as if) random condition. Other scholars mentioned the criterion but remained vague regarding the theory or intuition of why their chosen IV would satisfy it. In the best examples, scholars drew upon well-established theories to contextualize their IVs (e.g., social learning theory; see Hopp & Pruschak, 2020; institutional theory; see Solal & Snellman, 2019). Note that for those who discuss the exclusion restriction ($N = 34$), six articles incorrectly argued that this condition is satisfied because there is no empirical relationship between the IV(s) and the outcome.

When a model is over-identified (47 studies had over-identified models), we also coded whether researchers performed tests for over-identification in these cases. Many reported a Hansen-Sargan *J*-test ($N = 28$) and/or the χ^2 test of model fit ($N = 7$), whereas seven articles did not report any over-identification test. Because the overidentification test requires more instruments than endogenous predictors, sometimes scholars added an additional instrument just to conduct this test (e.g., Audia, Rousseau, & Brion, 2022). We would, however, encourage scholars to be cautious when adding additional instruments solely for the purpose of overidentification. As we have already discussed, overidentification tests assume constant causal effects, that at least one IV is valid, and that all instruments are (as if) random. If these assumptions are not respected, the overidentification test may for instance spuriously indicate that the exclusion restriction is satisfied when it is not.

Breakdown of IV categories

As our review indicates, the leadership literature has used many different IVs, and we now discuss their relative appropriateness. Our discussion is at a high level, and readers should recall that the appropriateness of an IV always depends on the context and theoretical model. In other words, one variable can be an appropriate IV for one research question; however, the same variable may not be a good IV for another research question or within another context. To discuss the potential appropriateness or validity of different IVs, we categorize them and discuss each IV category concerning the three conditions for a good IV (i.e., relevance condition, (as if) random condition, and the exclusion restriction). Recall that all three must be satisfied for a valid IV; very convincingly satisfying one condition cannot make up for not satisfying another.

Our review highlights seven broad categories of IVs used in leadership research: (1) leader and follower personal experiences, (2) factors internal to the organization, (3) leader and follower individual differences, (4) factors external to the organization, (5) internally generated instruments, (6) natural experiments, and (7) experimentally randomized instrumental variables. Below, we describe these categories in more detail, provide applied examples from the literature, and assess how appropriate each category is likely to be. Fig. 4 summarizes the seven different IV categories and their level of potential appropriateness.

Leader and follower personal experiences

The personal experiences of leaders and followers include a broad range of variables. This category includes which year a leader graduated from college (Jung & Shin, 2019) and a leader's prior ties to other board members (Stevenson & Radin, 2009). These personal life experiences are often categorical variables. For example, Kish-Gephart and Campbell (2015) explored the interaction between CEOs' social class origins and elite education as antecedents to strategic risk-taking. Because of "the potentially endogenous nature of the elite education variable" (p. 1624), the authors instrumented this variable with a dummy variable indicating whether the CEO was awarded an honorary Ph.D.

Regarding the (as if) random condition, leader and follower personal experiences will rarely be (as if) random because life experiences have numerous antecedents (e.g., educational background, family network) that likely correlate with the outcomes of interest, such as strategic risk-taking in the example above.

Our review indicates that IVs in this category tend to score high on relevance, with *F*-values higher than the common threshold and critical values (e.g., Kish-Gephart & Campbell, 2015; Lu, Swaab, & Galinsky, 2021; O'Reilly, Doerr, Caldwell, & Chatman, 2014). Omitted variables likely drive these high relevance values by spuriously increasing the shared variance between IVs and endogenous predictors. Furthermore, the exclusion restriction for this category is hardly tenable. It is generally unrealistic to argue that personal experiences only affect outcome variables of interest via a single pathway. Researchers must thus avoid using leader and follower personal experiences as IVs.

Overall, we evaluate leader and follower personal experiences as IVs as inappropriate.

Factors internal to the organization

This category refers to organizational factors linked to an organization's history or nature. This category of IVs includes firm characteristics such as board meeting fees (Arora, 2018), change in CEO (Audia, Rousseau, & Brion, 2022), firm location (Chen, 2020), and board tenure (Chiu & Walls, 2019). As an example, Tang et al. (2015) investigate the effect of CEO hubris on firm innovation. Because CEO hubristic behaviors (and thus stakeholders' perception of CEO hubris) and firm innovation are likely affected by a range of omitted variables at the individual and organizational levels, Tang et al. (2015) use whether the CEO was among the firm's founders as an IV for CEO hubris.

IVs in this category are frequently not (as if) random. Consider board meeting fees or board tenure: These characteristics are affected by myriad factors (e.g., company performance or reputation) that also impact the potential outcomes of interest. For instance, using change in CEO as an IV is problematic because CEOs are not randomly removed from their position. Indeed, such a decision is often attributable to other factors (e.g., poor firm performance) that will likewise affect the outcome variables of interest.

Studies using internal organizational factors generally report high *F*-values for excluded instruments (e.g., Audia, Rousseau, & Brion, 2022; Chiu & Walls, 2019; Love, Lim, & Bednar, 2017). However, like the leader and follower personal experiences category, not being (as if) random suggests that such instruments are likely correlated with omitted common causes, inflating relevance scores. Similarly, it is theoretically unlikely that the exclusion restriction holds for this type of IV. For instance, in the Tang et al. (2015) example described above, the CEO founder may affect firm innovation through other means than executive hubris, such as legitimate power or network-like characteristics. Similar to the previous category, we urge researchers to avoid using factors internal to organizations as IVs.

Overall, we evaluate factors internal to organizations as IVs as inappropriate.

Leader and follower individual differences

Leader individual differences concern trait-like variables tied directly to the leader and/or follower. In our review, scholars used a leader's age (McDonald, Keeves, & Westphal, 2018), sex (Ronay, Oostrom, Lehmann-Willenbrock, Mayoral, & Rusch, 2019; Tskhay,

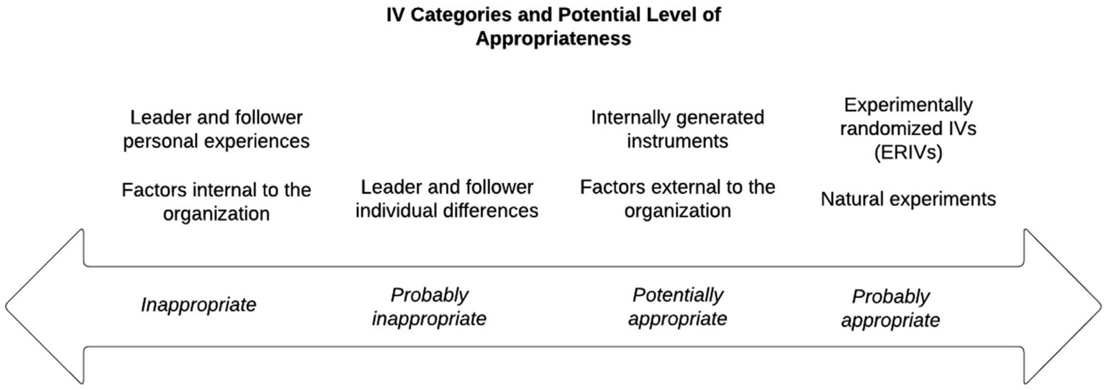

IV Categories and Potential Level of Appropriateness
Leader and follower personal experiences
Factors internal to the organization
Leader and follower individual differences
Internally generated instruments
Factors external to the organization
Experimentally randomized IVs (ERIVs)
Natural experiments
Inappropriate
Probably inappropriate
Potentially appropriate
Probably appropriate

Fig. 4. Summary of IV categories and potential level of appropriateness.
Notes. The appropriateness of an IV is highly context-dependent. An IV that is appropriate in one model can be inappropriate in another model. Beyond the potential appropriateness of each IV category, researchers must evaluate whether the three conditions for a valid instrument are likely satisfied in their specific model. For the 'leader and follower individual differences' category, intelligence seems a safer bet than other individual differences (e.g., personality, age, gender).

Zhu, & Rule, 2017), as well as stable traits such as personality (e.g., Bharanitharan, Lowe, Bahmannia, Chen, & Cui, 2021; De Vries, 2012; Klonek, Gerpott, & Parker, 2020) and intelligence (MacLaren et al., 2020). For example, Park, Chung, and Rajagopalan (2021) examined how the CEO's internal attribution of positive firm performance affects the degree of financial analysts' internal attributions of negative firm performance. Because the CEO's attributions of performance are endogenous (e.g., due to omitted variables at the leader or organizational level), the authors drew upon the political ideology literature and used the CEO's political ideology (considered to be a stable trait) to instrument the CEO's internal attribution of positive firm performance.

Stable individual differences may in certain contexts be used as IVs (Antonakis et al., 2010; Antonakis, 2011). Individual differences, and general intelligence in particular, are mainly inheritable, potentially (as if) random in the population, and tend to be stable (Bouchard & Loehlin, 2001). The contexts in which leader and follower individual differences can validly be considered (as if) random are when the study population is not a (self-)selective group. However, the (as if) random condition may not hold especially in organizational settings because individuals, along with their stable individual differences, are not randomly assigned to roles and organizations. Groups and organizations attract, select, and retain certain personality types (Schneider, 1987). Thus, there exists selection into groups, organizations, and leadership roles (Antonakis et al., 2010; Antonakis, 2011). Take sex as an example: although sex is randomly allocated at birth, the men and women at boardroom levels are not random men and women (Adams, 2016). Instead, it has been argued that women who make it to the top must be exceptionally competent and have a particular personality to overcome hurdles (Eagly & Karau, 2002). Thus, in this situation, sex will likely correlate with other dimensions (e.g., grit, perseverance, competence) that may also predict the outcomes of interest (e.g., job performance, effective decision making), implying that sex is not (as if) random in this context.

Regarding IV strength, our review indicates that personality dimensions such as the Big 5 or HEXACO were often weak IVs (e.g., Klonek, Gerpott, & Parker, 2020; Kwok, Hanig, Brown, & Shen, 2018). Even theoretically strong IVs were sometimes found to be empirically weak. For instance, in a registered report using leader Honesty-Humility as an IV for followers' perceptions of humble leadership, Bharanitharan et al. (2021) found that this IV was weak and failed to pass critical values. In contrast, our review indicates that intelligence was a fairly strong IV for speaking time (MacLaren et al., 2020). Intelligence could thus be relevant, particularly when instrumenting endogenous variables requiring strong cognitive skills (e.g., signaling charisma; Antonakis et al., 2016).

Considering the exclusion restriction, leader and follower individual differences have the potential to affect outcomes via multiple pathways, making them prone to violate the exclusion restriction. For instance, Kwok et al. (2018) investigated the effect of betweenness and in-degree centrality on leadership emergence using (among others) the Big 5 personality traits as IVs. The exclusion restriction requires that the IVs be excluded from the second stage model; in this case, the Big 5 personality traits should only affect leadership emergence via their effect on betweenness and in-degree centrality. This assumption is likely not tenable because the pathways to leader emergence are plentiful (Judge et al., 2002). Theoretically, conscientious individuals are more likely to be high performers (and thus emerge as leaders); similarly, individuals open to experience are more likely to come up with creative and innovative solutions, and thus emerge as leaders. Imposing that a wide array of personality dimensions affect outcomes via only one or two pathways is unlikely to be tenable. Here again, intelligence seems to fare better than personality dimensions or other trait-like variables. In an ERIV setting (see below), the model tested by MacLaren et al. (2020) using intelligence to instrument speaking time when predicting leader emergence did not violate the exclusion restriction as indicated by the overidentification test.

Overall, we evaluate leader individual differences as IVs as probably inappropriate.

Factors external to the organization

IVs may also originate from factors external to the organization. These factors can be separated into two main types. First, they may pertain to industry characteristics such as the number of firms in the industry that have a chief sustainability officer (e.g., Fu, Tang, & Chen, 2020), the average level of board diversity in the industry (Solal & Snellman, 2019), and the CEO option wealth of local firms (Zolotoy, O'Sullivan, Martin, & Veeraraghavan, 2019). Second, they may also reflect societal characteristics such as state-level constituency statutes (e.g., Flammer, Hong, & Minor, 2019), state-level gender equality (Gupta, Mortal, Chakrabarty, Guo, & Turban, 2020), and the relative proportion of immature female deaths in a geographical area (Kalnins & Williams, 2021).

In contrast to within-organization factors, external organizational factors can potentially be considered (as if) random. Unless firms significantly affect other firms in an industry (e.g., through strategic decisions, quasi-monopolistic behaviors), industry characteristics will likely be primarily independent of organizations. Similarly, unless firms significantly affect policies or laws through lobbying, political donations, or relocation (possibly leading companies to self-select into specific tax schemes or wage laws), societal characteristics will also be independent of organizations. Thus, leadership scholars may credibly argue for external organizational factors to satisfy this condition. A good example from our review is Kalnins and Williams (2021). Their study looked at whether geographic areas with a higher proportion of women business owners impact the survival duration of female-owned businesses. This study used the relative proportion of premature female deaths as the IV, and the authors specifically explain how their instrument is closely related to random chance, making it (as if) random. Another good example is Flammer et al. (2019). In their study, they use the enactment of state-level constituency statutes as an instrument of corporate social responsibility contracting. The authors argue that constituency statutes can be considered (as if) random because their enactment "does not reflect any firm's strategic decisions" (p. 1111).

Our review indicates that IVs in this category have mixed degrees of relevance. Some have high *F*-values, even higher than 100 (e.g., Kalnins & Williams, 2021; Solal & Snellman, 2019), whereas others have *F*-values below appropriate thresholds (e.g., Gupta, Mortal, Chakrabarty, Guo, & Turban, 2020; Li & Patel, 2019). However, one important concern for this category of IVs pertains to the exclusion restriction. In fact, factors as broad as laws, policies, or industry characteristics will often affect the behaviors of various actors in a market (e.g., competitors, firms offering substitute products and services) or stakeholders (e.g., customers, suppliers) and thus affect the outcome of interest also via other channels than the endogenous predictor. As an example, Gupta et al. (2020) used state-level gender equality to instrument CFO gender. It seems hardly tenable that state-level gender equality would affect financial misreporting (the outcome of interest in their study) only via CFO gender. For instance, state-level gender equality may affect financial misreporting because more honest and ethical individuals would tend to move to states with more gender equality. Having a higher proportion of such individuals in a state would likely affect organizational cultures that would be less prone to financial misreporting. Sometimes the exclusion restriction is more plausibly met. Researchers in financial economics, for example, used the existence and intensity of one-stop flight connections between the locations of potential directors' home addresses and firm headquarters as an IV for board gender diversity to estimate its effect on company outcomes such as performance or risk (Bernile et al., 2018).

Overall, we evaluate factors external to organizations as IVs as potentially appropriate.

Internally generated instruments

This category refers to approaches using IVs that are statistically generated to ensure that they are unrelated to the second-stage error term (Park & Gupta, 2012). These "instrument-free" approaches (Park & Gupta, 2012, p. 567) construct instruments from the variables present in the dataset, in the absence of external instruments. There are different approaches available,[7] but the only approach used so far in the leadership literature is to leverage the heteroskedasticity present in the endogenous predictor variables (Baum & Schaffer, 2021; Lewbel, 2000; 2012). More specifically, this procedure generates an instrument by relying on the variation in higher-order moments of the first-stage error distribution (Lewbel, 2012). To be valid, this approach requires certain assumptions related to heteroskedasticity to be satisfied in the first and second stages: (1) the error terms in the first and second stages have a common unobservable factor (i.e., an omitted variable); (2) this unobservable factor is homoscedastic; and (3) all or a subset of the covariates in the second stage are heteroscedastic. Assumptions (1) and (2) are untestable since they involve an error term. Assumption (3) is readily testable with a heteroskedasticity test (see Baum & Lewbel, 2019 for details).

The category of internally generated IVs has, for example, been used to instrument CEO-disaster experience (O'Sullivan, Zolotoy, & Fan, 2021), board member gender (Bechtoldt, Bannier, & Rock, 2019), and CEO proactivity (Kiss, Cortes, & Herrmann, 2021). Bechtoldt, Bannier, and Rock (2019) as well as O'Sullivan et al. (2021) only used internally generated instruments; yet, this procedure can also be applied in the presence of external instruments. For instance, MacLaren et al. (2020) initially relied on an experimental manipulation and stable individual differences as IVs. Given that these external IVs were not strongly related to their endogenous predictor, they additionally used internally generated IVs following Lewbel's (2012) procedure. These internally generated IVs were then much stronger and satisfied the relevance condition.

Internal IVs have no intuitive or theoretical meaning; thus, researchers do not need to theoretically argue for the (as if) randomness of such instruments. However, like external instruments, their statistical relevance and the exclusion restriction must be empirically tested. Our review indicates that internally generated instruments were at times strong (e.g., Bechtoldt et al., 2019) and at other times weak (e.g., Kiss, Cortes, & Herrmann, 2021). Furthermore, all the articles from our review that used internally generated IVs and reported an over-identification test passed this test, suggesting that the internally generated IVs were correctly excluded from the outcome equation.

Such a procedure may offer opportunities for causal identification in designs and research questions where good external IVs are complicated or almost impossible to find. Internally generated instruments may also be helpful in testing the validity of traditional external instruments. As an example of good practice, Hopp and Pruschak (2020) compared the estimated coefficients obtained using internally generated instruments with the estimated coefficients obtained using external instruments. Because both sets of estimation pointed in the same direction, the authors were slightly more confident in the validity of their results.

Yet, internally generated IVs are not a panacea, and we guard against researchers relying only on such a strategy prior to data collection. Because internally generated instruments cannot be theoretically justified and rely on untestable assumptions, researchers cannot assess their theoretical or empirical validity. Furthermore, Baum and Lewbel (2019, p. 757) caution against the overuse of this method, stating that "it is almost always preferable to use any available external instruments rather than constructed instruments." Because these approaches are relatively new in the management literature, we expect future scholarly work to refine and improve recommendations regarding its use.

Overall, we evaluate internally generated instruments as IVs as potentially appropriate.

Natural experiments

This category consists of IVs based on natural experiments at both micro (i.e., leader or follower) and macro (i.e., societal or organizational) levels. Natural experiments capitalize on (as if) random events or processes that researchers did not induce to identify causal effects (Dunning, 2012; for a comprehensive review of natural experiments in leadership research; see Sieweke & Santoni, 2020). Our review only uncovered one example that can be classified as a natural experiment. At a macro-level, Lonati (2020) used a geographical variable reflecting the potential productivity of agriculture within a country (specifically, the "proportion of land defined as suitable for cultivating six cereals", Lonati, 2020, p. 7) as an instrument for agricultural intensity. At a micro-level, no example emerged from our review. Yet, we want to point to an example from the realm of leadership outside of our review: Bennedsen et al. (2007) used the firstborn child gender of the departing family CEO as an instrument for CEO succession.

As suggested elsewhere (Antonakis et al., 2010) and used frequently in economics (e.g., Acemoglu et al., 2001), socio-historical and geographical factors (macro-level factors) can be well suited to serve as (as if) random IVs. At a micro-level, researchers can likewise reasonably argue that some events (such as child gender, Angrist & Evans, 1998; Bennedsen et al., 2007) are (as if) random. In all cases, the impetus is on scholars to justify the (as if) randomness of their suggested instrument(s).

Regarding the relevance condition, Lonati (2020) reported an *F*-value for the excluded IVs that indicated that the natural experiment IV was not weak. As for the exclusion restriction, natural experiment IVs may be prone to potential violations. Natural experiments are generally not under the researchers' control. As such, they are typically not tailored to instrument the endogenous variable and may also influence the outcome variable of interest directly or via other channels. Researchers thus need to theoretically explain why they deem the exclusion restriction to be satisfied and, in the multiple instrument case, accompany this argument with an empirical overidentification test.

We would like to highlight Lonati (2020) as a best practice example. In his paper, the author explicitly discusses all three conditions, why they should theoretically be satisfied, and provides empirical support that instruments are valid. For over-identification purposes, the author adds a second (potentially valid) instrument and finds no significant changes in estimated coefficients across the different specifications, providing further credence to his empirical strategy. Lonati also reports and compares IV estimates with OLS estimates, "visually"

[7] The four approaches listed in Park and Gupta (2012) are (1) use skewness in the endogenous predictors (Erickson & Whited 2002; Lewbel, 1997); (2) use heteroskedasticity in the endogenous predictors (Lewbel, 2012); (3) use discreteness of the endogenous predictors (Ebbes et al. 2005); and (4) use copulas to recover the joint distribution of the endogenous predictors and the second-stage error term (Park & Gupta, 2012). An additional approach, Model Implied Instrumental Variables (MIIV) proposed by Bollen (1996, 2019) generates valid internal instruments using assumptions within a structural model. MIIV constitutes an advanced use case for researchers in the leadership field, and we found only one article in our review using that approach (Tur et al., 2021); they use it to control for measurement error because the logic of MIIV does not solve for other endogeneity issues.

as well as with a Hausman (1978) endogeneity test. Regarding the Hausman test, Lonati clearly notes the limitations of the test given the small sample size ($N = 50$). Furthermore, Lonati provides reduced form estimates that can be causally interpreted and are policy relevant. Finally, the conclusion is cautious and tentative: “Evidence indicates that the instruments might satisfy the exclusion restriction; given that they are also theoretically exogenous, the results provide some tentative evidence that the instruments might be valid” (Lonati, 2020, p. 8). We encourage readers to follow this manuscript as a best practice example.

As our review argues, this category is currently under-explored, and we believe it is a missed opportunity given the benefits that natural experiments could provide for IVE. We thus encourage leadership researchers to look in other fields for inspiration and creative ideas regarding natural experiments (e.g., Angrist & Krueger, 2001; Diamond & Robinson, 2010; Dunning, 2012; Titiunik, 2021).

Overall, we evaluate natural experiments as IVs as probably appropriate.

Experimentally randomized IVs (ERIVs)

Researchers can also generate IVs in lab or field experiments by designing randomized experimental treatments that satisfy the three conditions for a valid instrument. Such ERIVs have been introduced to the leadership field and discussed at length by Sajons (2020). This type of IV is typically used in micro leadership studies. Examples include the manipulations of leader behaviors such as leader charisma (Antonakis, Fenley, & Liechti, 2011; Meslec, Curseu, Fodor, & Kenda, 2020), harm (Jolly, Krylova, & Phillips, 2020), ambidextrous leader behaviors (Klonek, Gerpott, & Parker, 2020) as well as leader anger expression and leader anxiety expression (Shao, 2019). Such experimental manipulations are primarily used to instrument measured endogenous variables such as perceptions, emotions, or behaviors. For instance, Liang et al. (2018) experimentally manipulated whether a participant was instructed to retaliate against a recalled abusive supervision behavior or not. This manipulation allowed them to study how subordinates’ retaliation behavior impacts their justice perceptions. Because retaliation behavior is potentially endogenous, they instrumented this variable via two IVs created out of the three experimental treatments (i.e., an “instruction to retaliate” condition; a “no instruction to retaliate” condition, and a control condition).

Instruments from this category tend to be able to satisfy the three IV conditions. By construction, such variables will be (as if) random. In fact, if randomization worked, which at least for observables can be confirmed via randomization checks, this condition will be valid by definition. Given that (as if) randomness is the only condition that is fully untestable, having it fulfilled by construction is a significant advantage of ERIVs.

Because researchers are free to design their ERIVs and tailor them to the potentially endogenous variable they wish to instrument, it will mostly be possible to ensure that the instrument is also relevant, at least if the researchers planned to use IVE before the design of their experimental study. Our review indicates some variability in the relevance for this category of IV, with some articles having low relevance (e.g., Klonek, Gerpott, & Parker, 2020; MacLaren et al., 2020) and others having appropriate *F*-values (Bharanitharan, Lowe, Bahmannia, Chen, & Cui, 2021; Meslec, Curseu, Fodor, & Kenda, 2020). However, our review also indicates that this category tends to suffer from low power due to low sample sizes (mean $N = 235$) in leadership experiments. It should be noted that smaller sample sizes reduce the strength of IVs, *ceteris paribus*.

Another important consideration for ERIVs is the exclusion restriction. Researchers designing experiments need to theoretically and empirically test that their endogenous predictor (sometimes a substantive mediator) is the only causal channel between the experimental manipulation and the outcome. For instance, Sajons (2020) investigated the effect of fairness perceptions on performance. Because fairness perceptions are de facto endogenous, she developed an experimental manipulation (i.e., an equitable vs. an inequitable boss) that should theoretically only affect performance via fairness perceptions. Other experimental manipulations, such as randomizing payouts to participants, also strongly predicted fairness perceptions. Yet, the valence of the payouts would have likely impacted participant performance (the outcome variable) via channels other than just fairness perceptions (e.g., participants’ desire to reciprocate) and thus not have constituted a suitable IV. Another example comes from Meslec et al. (2020), who manipulated the leader charisma signaling and the reward structure (i.e., fixed vs performance-based) to predict performance. Because they were interested in testing followers’ perceptions of the leader’s vision as a mediating mechanism (Study 2), they considered the leader charisma signaling condition as an ERIV that satisfied the (as if) random and relevance condition. Yet, they allowed their second experimental manipulation (i.e., the reward structure) to affect the performance outcome because performance pay affects performance at a task via other channels (e.g., increased effort, selection; Lazear, 2000).

Our review indicates that well-developed ERIVs offer some great potential. We urge researchers aiming to use ERIVs to carefully design their experimental manipulations in order to increase the chance that they pass the relevance condition and exclusion restriction. Researchers can also design more than one ERIV within the same study to allow for an overidentification test (see Liang et al., 2018). Also, having multiple potential ERIVs gives some flexibility in case one of the ERIVs turns out to be empirically weak.

Overall, we evaluate experimentally randomized IVs (ERIVs) as probably appropriate.

Moving forward

Despite the identified shortcomings of the current literature, IVE still holds much promise for answering causal leadership and management research questions. Thus, scholars in these fields may be curious about additional resources to increase their confidence in IVE. The econometric and applied statistical literature abounds in resources for researchers willing to apply this technique. For instance, practical textbooks include Angrist and Pischke (2009), Cunningham (2021), Kennedy (2008), Morgan and Winship (2015), and Wooldridge (2019). In the applied statistical literature, scholars have outlined flow charts (see Fig. 1 in Bascle, 2008), figures (Box 1 in Grosz et al., 2020; Figures 6.1 and 6.2 in Antonakis & House, 2014; Fig. 2 in Sajons, 2020), and tables (e.g., Table 1 in Sajons, 2020) to guide researchers in their conduct of IVE. Furthermore, researchers from universities such as Duke University[8] and the University of Pennsylvania[9] have promoted resources for learning the practical side of IVE. And the recent Nobel laureate Josh Angrist has developed a playful and intuitive online introduction to IVE.[10] We refer readers to these resources to further deepen their understanding of the method. Another resource is our Monte Carlo simulation investigating strong as well as moderate violations of the different instrument conditions (reported in the Online Appendix at https://osf.io/amqzh/).

[8] https://www.datacamp.com/community/open-courses/causal-inference-with-r-instrumental-variables-rdd.

[9] https://www.coursera.org/lecture/crash-course-in-causality/introduction-to-instrumental-variables-ueIMD.

[10] https://mru.org/courses/mastering-econometrics/introduction-instrumental-variables-part-one.

Checklist of IVE

In this final section, we aim to provide a checklist (see Table 3) and some best-practice recommendations intended to help leadership and management scholars better apply IVE. We provide concrete advice to guide researchers in the study design phase, the analysis phase, and the reporting phase. We hope this three-phase view of IVE will help scholars better plan their IVE and tailor its application to the overall research process.

Study design phase

The first step in applying the IV method is to assess the likelihood of endogeneity in the main predictors of interest (Antonakis et al., 2010; Hill et al., 2021). If the predictors of interest are exogenous within the purported study design, then IVE is unnecessary. When the predictor may, in contrast, be endogenous, scholars should first consider whether it is possible to conduct a randomized trial, the 'gold standard' of causal identification (Eden, 2021). If that is not possible (e.g., when the predictor variable is a perception or otherwise not manipulable; Sajons, 2020), quasi-experimental methods (e.g., IVE, RDD, DiD) offer a potential solution. Which of these methods holds the most promise depends on the study context and design (Adams, 2016; Hill et al., 2021; Sieweke & Santoni, 2020).

To consider whether IVE is a viable option, scholars need to carefully consider their theoretical model and invest extensive effort into finding instruments that are relevant, (as if) random, and satisfy the exclusion restriction. The design phase of IVE is critical, and it is difficult to overstate the importance of the instruments satisfying these conditions. Because the (as if) random condition and the exclusion restriction are first and foremost a theoretical exercise, scholars should engage in thought experiments and counterfactual reasoning to critically reflect on how likely their potential IVs will be to satisfy these conditions. Furthermore, scholars may consider discussing their IV design with colleagues well-versed in IVE for feedback and theoretical scrutiny.

Throughout this step, researchers have to be creative and persistent to find valid IVs, which is by far the most challenging aspect of IVE (Antonakis et al., 2010). Importantly, researchers must have good "institutional knowledge and ideas about the processes determining the variable of interest" (Angrist & Pischke, 2009, p. 117). Again, researchers would be well advised to look also into neighboring fields for relevant examples (e.g., Imbens, 2014; Kennedy, 2008). When it comes to leadership, interesting and creative IVs outside of our review include the use of (a) the firstborn child gender of the departing CEO's family as an instrument for CEO succession (to predict firm performance; Bennedsen et al., 2007); (b) students' (as if) random assignment to groups as an instrument for social network (to predict the attainment of leadership positions; Yang et al., 2019); and (c) traumatic youth experience as an instrument for neuroticism (both aggregated at the city level to predict economic growth; Garretsen et al., 2019). Beyond being creative instruments, the authors justify—both conceptually and empirically—why their chosen instrument is deemed valid within their specific context.

Whenever possible, researchers should consider using a natural experiment or an experimental manipulation as IV. Because suitable natural experiments tend to be harder to find, we recommend researchers to consider implementing an ERIV-design if the effect of the endogenous predictor can well be investigated in experimental settings. When performing such an ERIV procedure, researchers may want to pre-test the relevance of their ERIVs on a smaller sample (Sajons, 2020).

Another question pertains to the number of instruments to include. Researchers need at least one valid IV per endogenous predictor to causally identify their model and satisfy the order condition. However, if engaging in a prospective IV design (as opposed to seeking IVs in archival data), researchers may consider including multiple IVs so that their model is over-identified. It is important to remember that the goal here is not to bombard the design with unnecessary constructs or favor quantity over quality. The primary focus should be to ensure the integrity of every IV, and scholars should not assume that having more IVs is always superior. Nonetheless, it can be helpful for scholars to consider multiple IVs because this allows performing a test of overidentification and provides more options later down the road if some instruments happen to be invalid (e.g., when a theoretically strong IV turns out to be empirically weak).

We also encourage researchers using IVE to aim for substantially larger sample sizes. Even when all conditions are met, IVE produces less efficient estimates than OLS (Wooldridge, 2019). As a result, larger sample sizes are needed to obtain similar estimate precision. However, as our review indicates, sample sizes (particularly in the micro leadership literature) tend to remain small, which may create issues of power. We thus urge researchers using IVE to place sample size

Table 3
Checklist of Implementing IVE.

Study Design Phase

- Assess the likelihood of endogeneity present in the predictor of interest.
- Pre-check validity of potential IV via thought experiments, counterfactual thinking, and discussions with experienced IV users.
- When possible, include multiple theoretically driven IVs.
- Ideally use experimentally randomized IVs (ERIVs).
- To be adequately powered, use larger sample sizes than for an efficient model (e.g., OLS).

Analysis Phase

- Use canned procedures in statistical software to estimate the model with correct standard errors.
- Include controls in both 1st and 2nd stages (done automatically by canned software).
- Exclude IVs from the 2nd stage equation (done automatically by canned software).
- Assess the relevance of IV(s) by comparing the excluded instruments' *F*-statistic to the appropriate critical values (e.g., Stock & Yogo, 2005).
- If the model is overidentified, perform an overidentification test (e.g., *J*-test; see Hansen, 1982 and Sargan, 1958).
- Compare results with an efficient estimator (e.g., OLS) and perform a test of endogeneity (e.g., Hausman, 1978).
- Estimate the reduced form model.

Reporting Phase

- Justify theoretically why the IV is (as if) random and satisfies the exclusion restriction.
- Report the relevance of each set of IV(s) and the respective *F*-test's degrees of freedom.
- If the model is overidentified, report the overidentification test.
- Report in a Table the estimated coefficients for all variables included in each specification for the 1st and 2nd stages.
- Report the exact estimator used (e.g., 2SLS, ML, LIML), as well as the software and command(s) used for the estimation.
- Report and compare IV estimates with estimates from an efficient (e.g., OLS) estimator, as well as report a test of endogeneity.
- Report robustness checks (e.g., adding internally generated instruments).
- Report the reduced form model.
- Critically discuss the limitations of the causal identification strategy.

among one of their core priorities when designing new studies. Finally, researchers should be aware that IVE is a versatile technique that can also be applied to time series equations, pooled cross sections, and panel data (Wooldridge, 2019).

Analysis phase

Once theoretically valid IVs have been identified and the data has been gathered, researchers should proceed with performing the IVE and assessing the empirical validity of their IVs. First, recall that researchers should use canned procedures in statistical software because a manual computation of IVE would yield incorrect standard errors. Second, in presence of appropriate controls, researchers should include them in the first and second stage equations. Third, scholars must exclude their IVs from the second stage equation. Note that these three steps are completed automatically when using canned procedures in statistical software (e.g., commands such as *ivregress 2sls* in Stata or the *AER* package in R). We also encourage authors to estimate models with an efficient but potentially inconsistent estimator (e.g., OLS) for comparison purposes.

With respect to the conditions for valid IVs, researchers should empirically assess the relevance of their IVs. The first stage *F*-statistic for the excluded IVs only (i.e., excluding the controls) should surpass the Stock and Yogo (2005) critical values. Higher *F*-statistics represent stronger and thus more relevant instruments, leading to more precise parameter estimates. In the presence of heteroscedasticity, autocorrelation, and clustering, researchers should rely on the weak instrument test developed by Olea and Pflueger (2013), which is more robust in these situations than the standard critical values of Stock and Yogo (2005). Also, in the presence of multiple endogenous predictors, Maydeu-Olivares et al. (2018) recommend assessing the different instruments' strengths by computing Cragg and Donald's (1993) smallest eigenvalue statistic.

When a model has more IVs than instrumented predictors, scholars should further perform an overidentification test. The most common test is the Hansen (1982)-Sargan (1958) *J*-test. Alternatively, researchers can evaluate the overall χ^2 test of model fit in SEM (Bollen, 1989). For testing the exclusion restriction, recall that these tests require three assumptions: At least one IV is valid, constant causal effects are plausible, and all IVs are (as if) random. If researchers can make a strong argument for the first two of these assumptions, but are less certain about the (as if) randomness of one of their instruments, the test could still be informative. In such a case, a significant overidentification test indicates that at least one instrument violates the (as if) random condition and/or the exclusion restriction. Along similar lines, different models can be estimated with single IVs. If instruments are valid and the second stage coefficient does not significantly vary across different specifications, results will be more credible (Murray, 2006; see Lonati, 2020 as an applied example). When a model is "exactly identified" (i.e., there are as many IVs as endogenous predictors), an overidentification test cannot be performed.

After estimating their IV model and providing empirical evidence for their IVs to satisfy the relevance condition and pass the test of overidentification, researchers should perform an endogeneity test (e.g., Hausman, 1978) to see whether the instrumented variable (the potentially endogenous predictor x) was indeed endogenous. If IVs are valid, the sample size is sufficiently large, and the endogeneity test is not significant, then researchers can build on the efficient estimator for their inferences (e.g., OLS). If the sample size is relatively small or the endogeneity test is significant, researchers should rely on a consistent estimator for their inferences. In the presence of weak instruments or instruments that do not pass the overidentification test, researchers may still consider estimating models with internally generated IVs. Researchers should finally estimate and interpret the reduced form model, which gives the unbiased direct effect of the instrument on the outcome.

Reporting phase

In organizational science, there has been an increased call for methodological transparency (e.g., Aguinis et al., 2018), and we particularly echo this admonition for the use of IVE. Without repeating the conditions of IVE, we cannot stress enough that it is incumbent on the authors to convince editors, reviewers, and readers of the theoretical and empirical validity of their IVs. Regarding theory, scholars must in their manuscript explicitly provide a strong rationale of why the (as if) random condition and the exclusion restriction should be satisfied from a theoretical perspective. Regarding the empirics, researchers must report the *F*-value of the excluded IVs, an endogeneity test, and an overidentification test when appropriate. For all tests, it is critical that the degrees of freedom are reported.

We also urge researchers to transparently report in tables the correlations and estimated coefficients of all variables included in the different specifications for the first and second stages (for best practice examples in our review, see Flammer, Hong, & Minor, 2019; Jolly, Krylova, & Phillips, 2020; Kalnins & Williams, 2021; MacLaren et al., 2020). If space within a manuscript is limited, researchers should report the estimated coefficients in an online supplement. Researchers should also report the exact estimator used, as well as the software and command(s) used for estimation. Moreover, estimates from an efficient and consistent estimator should be reported and compared. We would also encourage researchers to perform robustness checks (e.g., adding internally generated instruments) and report the reduced form model. Finally, authors need to critically discuss and reflect on the limitations of their IV strategy.

Final thoughts

Like all other statistical techniques, IVE has its nuance and—if not applied correctly—will cause more problems than it solves (Ketokivi & McIntosh, 2017; Murray, 2006). If the conditions for a valid instrument are not met, IV estimates will be inconsistent and likely exhibit even more bias than OLS (Hahn & Hausman, 2002; Larcker & Rusticus, 2010; Staiger & Stock, 1997). Also, invalid IVs may yield underpowered tests of endogeneity (Semadeni et al., 2014). Still, a recent review of the economics literature shows that weak instruments are frequently used in practice (Andrews et al., 2019). Based on the current state of the leadership field, we call for better standards and more stringent requirements in the application of IVE.

We also want to emphasize that endogeneity is a matter of degree (Ketokivi & McIntosh, 2017). Whereas leadership researchers will not always be able to eradicate endogeneity, tackling and attenuating it via the proper use of IVE will be a big leap towards rigorous causal identification.

Conclusion

About a decade ago, Antonakis and colleagues (2010) introduced the notion of IVs to leadership and management research. Since then, we have seen a significant uptick in the use of IVE, especially within the last few years, and there is a growing recognition that this methodology can help leadership scholars in establishing causal claims (e.g., Adams, 2016; Wu et al., 2021; Güntner et al., 2020). In this review, we have taken a critical stance as to how IVE has been applied in the field so far in an attempt to invite scholars to better use IVE in future research. Moving forward, we hope that our methodological overview and guide on best practices will help increase truly rigorous tests of causal arguments in leadership and management research.

Data availability

All data has been made available on an online repository at https://osf.io/amqzh/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.leaqua.2022.101673.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review, 91*(5), 1369–1401.

Adams, R. B. (2016). Women on boards: The superheroes of tomorrow? *The Leadership Quarterly, 27*(3), 371–386.

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals, 12*(1), 83–110.

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics, 11*, 727–753.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444–455.

Angrist, J. D., & Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review, 88*(3), 450–477.

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. (2009). *Mostly harmless econometrics: An empiricists' companion*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J. S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives, 24*(2), 3–30.

Antonakis, J. (2011). Predictors of leadership: The usual suspects and the suspect traits. *Sage handbook of leadership*, 269–285.

Antonakis, J., Banks, G., Bastardoz, N., Cole, M., Day, D., Eagly, A., ... Weber, R. (2019). The leadership quarterly: State of the journal. *The Leadership Quarterly, 30*(1), 1–9.

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods, 24*(2), 443–483.

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An Ill-Defined and Ill-Measured Gift. *Annual Review of Organizational Psychology and Organizational Behavior, 3*(1), 293–319.

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

Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *Academy of Management Learning & Education, 10*(3), 374–396.

Antonakis, J., & House, R. J. (2014). Instrumental leadership: Measurement and extension of transformational–transactional leadership theory. *The Leadership Quarterly, 25*(4), 746–771.

Arora, P. (2018). Financially linked independent directors and bankruptcy reemergence: The role of director effort. *Journal of Management, 44*(7), 2665–2689.

Arthur, C. A., Bastardoz, N., & Eklund, R. (2017). Transformational leadership in sport: Current status and future directions. *Current Opinion in Psychology, 16*, 78–83.

Arvate, P. R., Galilea, G. W., & Todescat, I. (2018). The queen bee: A myth? The effect of top-level female leadership on subordinate females. *The Leadership Quarterly, 29*(5), 533–548.

Audia, P. G., Rousseau, H. E., & Brion, S. (2022). CEO power and nonconforming reference group selection. *Organization Science, 33*(2), 831–853.

Banks, G. C., Fischer, T., Gooty, J., & Stock, G. (2021). Ethical leadership: Mapping the terrain for concept cleanup and a future research agenda. *The Leadership Quarterly, 32*(2) 101471.

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

Baum, C. F., & Lewbel, A. (2019). Advice on using heteroskedasticity-based identification. *The Stata Journal, 19*(4), 757–767.

Baum, C., & Schaffer, M. E. (2021). IVREG2H: Stata module to perform instrumental variables estimation using heteroskedasticity-based instruments.

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.

Bechtoldt, M. N., Bannier, C. E., & Rock, B. (2019). The glass cliff myth?–Evidence from Germany and the UK. *The Leadership Quarterly, 30*(3), 273–297.

Bennedsen, M., Nielsen, K. M., Pérez-González, F., & Wolfenzon, D. (2007). Inside the family firm: The role of families in succession decisions and performance. *The Quarterly Journal of Economics, 122*(2), 647–691.

Bernile, G., Bhagwat, V., & Yonker, S. (2018). Board diversity, firm risk, and corporate policies. *Journal of Financial Economics, 127*(3), 588–612.

Bharanitharan, D. K., Lowe, K. B., Bahmannia, S., Chen, Z. X., & Cui, L. (2021). Seeing is not believing: Leader humility, hypocrisy, and their impact on followers' behaviors. *The Leadership Quarterly, 32*(2) 101440.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Bollen, K. A. (1996). An Alternative Two Stage Least Squares (2SLS) Estimator for Latent Variable Equations. *Psychometrika, 61*(1), 109–121.

Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology, 38*, 37–72.

Bollen, K. A. (2019). Model Implied Instrumental Variables (MIIVs): An Alternative Orientation to Structural Equation Modeling. *Multivariate Behavioral Research, 54*(1), 31–46.

Bouchard, T. J., & Loehlin, J. C. (2001). Genes, evolution, and personality. *Behavior Genetics, 31*(3), 243–273.

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

Canan, C., Lesko, C., & Lau, B. (2017). Instrumental variable analyses and selection bias. *Epidemiology, 28*(3), 396–398.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, & R. Swidinsky (Eds.), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. Toronto: University of Toronto Press.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica, 69*(5), 1127–1160.

Carpenter, M. A., & Sanders, W. G. (2004). The effects of top management team pay and firm internationalization on MNC performance. *Journal of Management, 30*(4), 509–528.

Chen, J. (2020). A juggling act: CEO polychronicity and firm innovation. *The Leadership Quarterly, 101380*.

Chernozhukov, V., & Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. *Economics Letters, 100*(1), 68–71.

Chiu, S. C. S., & Walls, J. L. (2019). Leadership change and corporate social performance: The context of financial distress makes all the difference. *The Leadership Quarterly, 30*(5) 101307.

Clapp-Smith, R., Carsten, M., Gooty, J., Connelly, S., Haslam, A., Bastardoz, N., & Spain, S. (2018). Special registered report issue on replication and rigorous retesting of leadership models. *The Leadership Quarterly, 29*(2).

Cragg, J. G., & Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory, 9*(2), 222–240.

Cunningham, S. (2021). *Causal inference*. In Causal Inference: Yale University Press.

De Vries, R. E. (2012). Personality predictors of leadership styles and the self–other agreement problem. *The Leadership Quarterly, 23*(5), 809–821.

Deuchert, E., & Felfe, C. (2015). The tempest: Short-and long-term consequences of a natural disaster for children′ s development. *European Economic Review, 80*, 280–294.

Diamond, J., & Robinson, J. A. (Eds.). (2010). *Natural experiments of history*. Harvard University Press.

Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge University Press.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute, 22*, 23–32.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598.

Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income. *Quantitative Marketing and Economics, 3*, 365–392.

Eden, D. (2021). The science of leadership: A journey from survey research to field experimentation. *The Leadership Quarterly, 32*(3) 101472.

Erickson, T., & Whited, T. M. (2002). Two-Step GMM Estimation of the Errors-in-Variables Model Using High-Order Moments. *Econometric Theory, 18*(3), 776–799.

Fischer, T., Tian, A. W., Lee, A., & Hughes, D. J. (2021). Abusive supervision: A systematic review and fundamental rethink. *The Leadership Quarterly, 32*(6) 101540.

Fischer, T., Dietz, J., & Antonakis, J. (2017). Leadership process models: A review and synthesis. *Journal of Management, 43*(6), 1726–1753.

Fischer, T., Hambrick, D. C., Sajons, G. B., & Van Quaquebeke, N. (2020). Beyond the ritualized use of questionnaires: Toward a science of actual behaviors and psychological states. *The Leadership Quarterly, 31*(4) 101449.

Flammer, C., Hong, B., & Minor, D. (2019). Corporate governance and the rise of integrating corporate social responsibility criteria in executive compensation: Effectiveness and implications for firm outcomes. *Strategic Management Journal, 40*(7), 1097–1122.

Fu, R., Tang, Y., & Chen, G. (2020). Chief sustainability officers and corporate social (Ir) responsibility. *Strategic Management Journal, 41*(4), 656–680.

Gangloff, K. A., Connelly, B. L., & Shook, C. L. (2016). Of scapegoats and signals: Investor reactions to CEO succession in the aftermath of wrongdoing. *Journal of Management, 42*(6), 1614–1634.

Garretsen, H., Stoker, J. I., Soudis, D., Martin, R., & Rentfrow, J. (2019). The relevance of personality traits for urban economic growth: Making space for psychological factors. *Journal of Economic Geography, 19*(3), 541–565.

Garretsen, H., Stoker, J. I., & Weber, R. A. (2020). Economic perspectives on leadership: Concepts, causality, and context in leadership research. *The Leadership Quarterly, 31*(3), 101410.

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.

Gomulya, D., Wong, E. M., Ormiston, M. E., & Boeker, W. (2017). The role of facial appearance on CEO selection after firm misconduct. *Journal of Applied Psychology, 102*(4), 617–635.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology, 29*(4), 722–729.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science, 15* (5), 1243–1255.

Güntner, A. V., Klonek, F. E., Lehmann-Willenbrock, N., & Kauffeld, S. (2020). Follower behavior renders leader behavior endogenous: The simultaneity problem, estimation challenges, and solutions. *The Leadership Quarterly, 31*(6) 101441.

Gupta, A., & Wowak, A. J. (2017). The elephant (or donkey) in the boardroom: How board political ideology affects CEO pay. *Administrative Science Quarterly, 62*(1), 1–30.

Gupta, V. K., Mortal, S., Chakrabarty, B., Guo, X., & Turban, D. B. (2020). CFO gender and financial statement irregularities. *Academy of Management Journal, 63*(3), 802–831.

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica, 70*(1), 163–189.

Hahn, J., & Inoue, A. (2002). A Monte Carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators. *Econometric Reviews, 21*(3), 309–336.

Hahn, J., Ham, J. C., & Moon, H. R. (2011). The Hausman test and weak instruments. *Journal of Econometrics, 160*(2), 289–299.

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: Journal of the Econometric Society, 46*(6), 1251–1271.

Hausman, J., Stock, J. H., & Yogo, M. (2005). Asymptotic properties of the Hahn-Hausman test for weak-instruments. *Economics Letters, 89*(3), 333–342.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society, 47*(1), 153–161.

Heckman, J. J., & Robb, R. Jr, (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics, 30*(1–2), 239–267.

Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology, 17*(4), 360–372.

Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management, 47*(1), 105–143.

Hopp, C., & Pruschak, G. (2020). Is there such a thing as leadership skill?–A replication and extension of the relationship between high school leadership positions and later-life earnings. *The Leadership Quarterly, 101475*.

Imbens, G. (2014). *Instrumental variables: an econometrician's perspective. Statistical Science, 29*(3), 323–358.

Jacquart, P., Cole, M., Gabriel, A. S., Koopman, J., & Rosen, C. C. (2017). *Studying leadership: research design and methods, No. hal-02311084*.

Jacquart, P., Santoni, S., Schudy, S., Sieweke, J., & Withers, M. C. (2020). Harnessing Exogenous Shocks for Leadership and Management Research. *The Leadership Quarterly, 31*(5). Special Issue on 101464.

Jolly, P. M., Krylova, K. O., & Phillips, J. S. (2020). Leader intention, misconduct and damaged relational follower identity: A moral decision making perspective. *The Leadership Quarterly*, 101425.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780.

Jung, J., & Shin, T. (2019). Learning not to diversify: The transformation of graduate business education and the decline of diversifying acquisitions. *Administrative Science Quarterly, 64*(2), 337–369.

Kalnins, A., & Williams, M. (2021). The geography of female small business survivorship: Examining the roles of proportional representation and stakeholders. *Strategic Management Journal, 42*(7), 1247–1274.

Kennedy, P. (2008). *A guide to econometrics*. John Wiley & Sons.

Ketokivi, M., & McIntosh, C. N. (2017). Addressing the endogeneity dilemma in operations management research: Theoretical, empirical, and pragmatic considerations. *Journal of Operations Management, 52*, 1–14.

Kish-Gephart, J. J., & Campbell, J. T. (2015). You don't forget your roots: The influence of CEO social class background on strategic risk taking. *Academy of Management Journal, 58*(6), 1614–1636.

Kiss, A. N., Cortes, A. F., & Herrmann, P. (2021). CEO proactiveness, innovation, and firm performance. *The Leadership Quarterly, 101545*.

Klein, F. B., Chaigneau, P., & Devers, C. E. (2021). CEO gender-based termination concerns: Evidence from initial severance agreements. *Journal of Management, 47* (3), 567–596.

Klonek, F. E., Gerpott, F. H., & Parker, S. K. (2020). A conceptual replication of ambidextrous leadership theory: An experimental approach. *The Leadership Quarterly, 101473*.

Konlechner, S., & Ambrosini, V. (2019). Issues and trends in causal ambiguity research: A review and assessment. *Journal of Management, 45*(6), 2352–2386.

Kwok, N., Hanig, S., Brown, D. J., & Shen, W. (2018). How leader role identity influences the process of leader emergence: A social network analysis. *The Leadership Quarterly, 29*(6), 648–662.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363–374.

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.

Lazear, Edward P. (2000). Performance Pay and Productivity. *American Economic Review, 90*(5), 1346–1361.

Lewbel, A. (1997). Constructing Instruments for Regressions With Measurement Error When No Additional Data Are Available, With an Application to Patents and R&D. *Econometrica, 65*(5), 1201–1213.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics, 97* (1), 145–177.

Lewbel, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics, 30*(1), 67–80.

Li, M., & Patel, P. C. (2019). Jack of all, master of all? CEO generalist experience and firm performance. *The Leadership Quarterly, 30*(3), 320–334.

Liang, L. H., Brown, D. J., Lian, H., Hanig, S., Ferris, D. L., & Keeping, L. M. (2018). Righting a wrong: Retaliation on a voodoo doll symbolizing an abusive supervisor restores justice. *The Leadership Quarterly, 29*(4), 443–456.

Lonati, S. (2020). What explains cultural differences in leadership styles? On the agricultural origins of participative and directive leadership. *The Leadership Quarterly, 31*(2) 101305.

Love, E. G., Lim, J., & Bednar, M. K. (2017). The face of the firm: The influence of CEOs on corporate reputation. *Academy of Management Journal, 60*(4), 1462–1481.

Lu, J. G., Swaab, R. I., & Galinsky, A. D. (2021). Global Leaders for Global Teams: Leaders with Multicultural Experiences Communicate and Lead More Effectively, Especially in Multinational Teams. *Organization Science*.

MacLaren, N. G., Yammarino, F. J., Dionne, S. D., Sayama, H., Mumford, M. D., Connelly, S., ... Ruark, G. A. (2020). Testing the babble hypothesis: Speaking time predicts leader emergence in small groups. *The Leadership Quarterly, 31*(5) 101409.

Martin, R., Hughes, D. J., Epitropaki, O., & Thomas, G. (2021). In pursuit of causality in leadership training research: A review and pragmatic recommendations. *The Leadership Quarterly, 32*(5) 101375.

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 389–402.

McDonald, M. L., Khanna, P., & Westphal, J. D. (2008). Getting them to think outside the circle: Corporate governance, CEOs' external advice networks, and firm performance. *Academy of Management Journal, 51*(3), 453–475.

McDonald, M. L., Keeves, G. D., & Westphal, J. D. (2018). One step forward, one step back: White male top manager organizational identification and helping behavior toward other executives following the appointment of a female or racial minority CEO. *Academy of Management Journal, 61*(2), 405–439.

Mellon, J. (2022). Rain, Rain, Go Away: 192 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable. Working Paper.

Meslec, N., Curseu, P. L., Fodor, O. C., & Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. *The Leadership Quarterly, 31*(6) 101423.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives, 20*(4), 111–132.

Olea, J. L. M., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics, 31*(3), 358–369.

O'Reilly, C. A., III, Doerr, B., Caldwell, D. F., & Chatman, J. A. (2014). Narcissistic CEOs and executive compensation. *The Leadership Quarterly, 25*(2), 218–231.

O'Sullivan, D., Zolotoy, L., & Fan, Q. (2021). CEO early-life disaster experience and corporate social performance. *Strategic Management Journal, 42*(11), 2137–2161.

Park, S. H., Chung, S. H., & Rajagopalan, N. (2021). Be Careful What You Wish For: CEO and Analyst Firm Performance Attributions and CEO Dismissal. *Strategic Management Journal, 42*(10), 1880–1908.

Park, S., & Gupta, S. (2012). Handling Endogenous Regressors by Joint Estimation Using Copulas. *Marketing Science, 31*(4), 567–586.

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly, 30*(1), 11–33.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.

Reeb, D., Sakakibara, M., & Mahmood, I. P. (2012). From the editors: Endogeneity in international business research. *Journal of International Business Studies, 43*(3), 211–218.

Ronay, R., Oostrom, J. K., Lehmann-Willenbrock, N., Mayoral, S., & Rusch, H. (2019). Playing the trump card: Why we select overconfident leaders and why it matters. *The Leadership Quarterly, 30*(6) 101316.

Sajons, G. B. (2020). Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables. *The Leadership Quarterly, 31*(5) 101348.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society, 26*(3), 393–415.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*(3), 437–453.

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal, 35*(7), 1070–1079.

Shao, B. (2019). Moral anger as a dilemma? An investigation on how leader moral anger influences follower trust. *The Leadership Quarterly, 30*(3), 365–382.

Shaver, J. M. (2005). Testing for mediating variables in management research: Concerns, implications, and alternative strategies. *Journal of Management, 31*(3), 330–353.

Shaver, J. M. (2020). Causal identification through a cumulative body of research in the study of strategy and organizations. *Journal of Management, 46*(7), 1244–1256.

Short, J. (2009). The art of writing a review article. *Journal of Management, 35*(6), 1312–1317.

Sieweke, J., & Santoni, S. (2020). Natural experiments in leadership research: An introduction, review, and guidelines. *The Leadership Quarterly, 31*(1) 101338.

Solal, I., & Snellman, K. (2019). Women don't mean business? Gender penalty in board composition. *Organization Science, 30*(6), 1270–1288.

Staiger, D., Stock, J. H., & Watson, M. W. (1997). The NAIRU, unemployment and monetary policy. *Journal of Economic Perspectives, 11*(1), 33–49.

Stevenson, W. B., & Radin, R. F. (2009). Social capital and social influence on the board of directors. *Journal of Management Studies, 46*(1), 16–44.

Stock, J. H., & Trebbi, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives, 17*(3), 177–194.

Stock, J. H., and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg, ed. Donald W.K. Andrews and James H. Stock*, Chapter 5, 80–108. Cambridge University Press.

Tang, Y., Li, J., & Yang, H. (2015). What I see, what I do: How executive hubris affects firm innovation. *Journal of Management, 41*(6), 1698–1723.

Titiunik, R. (2021). Natural experiments. *Advances in Experimental Political Science*, 103–129.

Tsai, C. Y., Kim, J., Jin, F., Jun, M., Cheong, M., & Yammarino, F. J. (2022). Polynomial regression analysis and response surface methodology in leadership research. *The Leadership Quarterly, 101592*.

Tskhay, K. O., Zhu, R., & Rule, N. O. (2017). Perceptions of charisma from thin slices of behavior predict leadership prototypicality judgments. *The Leadership Quarterly, 28* (4), 555–562.

Tur, B., Harstad, J., & Antonakis, J. (2021). Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. *The Leadership Quarterly, 101476*.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics, 68*(1), 115–132.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach* (7th ed.). Cengage Learning.

Wright, P. G. (1928). *Tariff on animal and vegetable oils*. New York: Macmillan Company.

Wu, D. M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: Journal of the Econometric Society, 41*(4), 733–750.

Wu, Y. L., Shao, B., Newman, A., & Schwarz, G. (2021). Crisis leadership: A review and future research agenda. *The Leadership Quarterly, 32*(6) 101518.

Yang, Y., Chawla, N. V., & Uzzi, B. (2019). A network's gender composition and communication pattern predict women's leadership success. *Proceedings of the National Academy of Sciences, 116*(6), 2033–2038.

Zhang, Y., & Gimeno, J. (2016). Earnings pressure and long-term corporate governance: Can long-term-oriented investors and managers reduce the quarterly earnings obsession? *Organization Science, 27*(2), 354–372.

Zolotoy, L., O'Sullivan, D., Martin, G. P., & Veeraraghavan, M. (2019). The role of affect in shaping the behavioral consequences of CEO option incentives. *Journal of Management, 45*(7), 2920–2951.