Journal of Operations Management 52 (2017) 1–14

Contents lists available at ScienceDirect

Journal of Operations Management

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

Addressing the endogeneity dilemma in operations management research: Theoretical, empirical, and pragmatic considerations

Mikko Ketokivi [a, *], Cameron N. McIntosh [b]

[a] IE Business School, IE University, Spain
[b] Government of Canada, Public Safety, Canada

ARTICLE INFO

Article history:
Received 25 September 2016
Accepted 12 May 2017
Available online 27 May 2017
Handling Editor: Dr. V. Daniel R. Guide

ABSTRACT

In this paper, we examine the problem of endogeneity in the context of operations management research. Whereas the extant literature has focused primarily on the statistical aspect of the problem, a comprehensive treatment requires an examination of theoretical and pragmatic considerations as complements. The prevailing problem with the focus on statistical techniques is that the standards tend to be derived from idealizations: the correlation between a regressor and a disturbance term must be exactly zero, or the analysis will be invalid. In actual empirical research settings, such a knife-edge assumption can never be satisfied, indeed it cannot even be directly tested. Idealizations are useful in helping us understand what it would take to eliminate endogeneity, but when applied directly and unconditionally, they lead to unreasonable standards that may unnecessarily stifle substantive inquiry. We believe that it is far more productive and meaningful to ask: "What can we realistically expect empirical scientists to be able to achieve?" To this end, we cover and revisit some of the general technical material on endogeneity, paying special attention to the idiosyncrasies of operations management research and what could constitute reasonable criteria for addressing endogeneity in empirical operations management studies.

1. Introduction

Operations management (OM) researchers apply a variety of methods and research designs, but statistical modeling—analysis of variance, regression analysis, factor analysis, and structural equation modeling—remains by far the most commonly used tool we use to make inferences and to draw conclusions from empirical data. Despite decades of empirical OM research and much progress, there are a number of key assumptions underlying our statistical reasoning techniques that have not been sufficiently well explicated. These assumptions, when violated, can have fundamental implications for the credibility of our inferences and our theoretical interpretations. In this paper, we aim to explore some of the most insidious threats to trustworthy statistical inference, as well as the various ways in which researchers can tackle these threats.

To set the stage for our inquiry, consider the link between plant productivity and business unit profitability. Even a casual theoretical reflection suggests that as plant productivity rises, profitability rises as well; a simple regression analysis would confirm the relationship to be positive. But to what extent do increases in productivity actually *drive* increases in profitability? The answer to this question requires that we get not just the sign but also the magnitude of this effect right. In this paper, we discuss what is perhaps the most significant threat to getting the magnitude right. In the econometrics literature, this is dubbed *the problem of endogeneity*. Sometimes endogeneity causes so much bias that we may not even get the sign of the coefficient right.

In order to bring further conceptual and statistical clarity to the endogeneity problem, let us put the productivity-profitability relationship into a model by supposing a fairly typical OM research model, depicted in Fig. 1:

1) x_1-x_3 are aspects of a factory's production system, such as process choice or the degree of implementation of certain manufacturing principles and practices;
2) y_1 is a measure of the factory's operational performance, say, total factor productivity; and
3) y_2 is a measure of business success, say, business unit profitability.

* Corresponding author.
E-mail addresses: Mikko.Ketokivi@ie.edu (M. Ketokivi), Cameron.McIntosh@aadnc-aandc.gc.ca (C.N. McIntosh).

http://dx.doi.org/10.1016/j.jom.2017.05.001
0272-6963/

The causal argument proposed in the model is straightforward: the way the production system is designed has implications for how the factory performs, which in turn affects how the business unit performs. The idea that operational performance mediates the effect of the production system on business success is theoretically salient, and has received much empirical support in the research literature, starting with the seminal work of Skinner (1969).

In Fig. 1, *full mediation* is proposed, which means that once the value of y_1 is set, the values of x_1-x_3 no longer matter insofar as the value of y_2 is concerned: the direct effects—and all other possible effects not mediated by y_1—from x_1-x_3 to y_2 are zero, conditional on the mediator y_1. Mediational hypotheses typically have a theoretical basis, but in this paper, we will show that the full mediation hypothesis in Fig. 1 also turns out to be a crucial element in empirical attempts to tackle the endogeneity problem.

1.1. What is the problem?

In order to be able to inform theory and practice, the researcher interested in the model in Fig. 1 has one unambiguous, overriding objective: obtain good estimates of the key model parameters β_i. By *good*, we generally mean *unbiased* (the expected value of the estimate is the true parameter value) and *efficient* (the variance of the estimate is as low as possible), but at the very minimum, we want our estimates to be *consistent*, that is, tend to the true value of the parameter as sample size increases (Wooldridge, 2016: 150). In order to be practically relevant, we must be able to obtain reliable estimates of the magnitudes of the effects under scrutiny.

There are many challenges in getting magnitudes right. Perhaps the most critical challenge arises from the fact that practically all OM research that examines models like the one depicted in Fig. 1 is based on data where the researcher merely *observes* the variables in a statistical sample. In an experiment one would *manipulate* at least some of the x_1-x_3. When variables are merely observed, we do not know what the origins of their variances are. This also implies that we do not know whether or not they covary with one another. Consequently, in specifying the model we appropriately allow x_1-x_3 to correlate freely, which is indicated by the double-headed arrows connecting x_1-x_3. Variables x_1-x_3 are asserted to be *exogenous* variables.

It is important to bear in mind that the disturbance terms ε_1 and ε_2 are exogenous variables as well (see *endnote #1* in the Appendix). The fact that x_1-x_3 are measured variables and ε_1-ε_2 are latent is no reason to treat the two fundamentally differently in examining model specification. But we often do, typically by assuming the disturbance terms to be uncorrelated both with one another and with x_1-x_3. But since we do not know what drives the variances of the disturbance terms, we must not readily assume away these correlations. If this assumption of zero correlations is wrong, the model is misspecified, and all the estimates of the model parameters β_i are biased, inconsistent, and inefficient (see *endnote #2*).

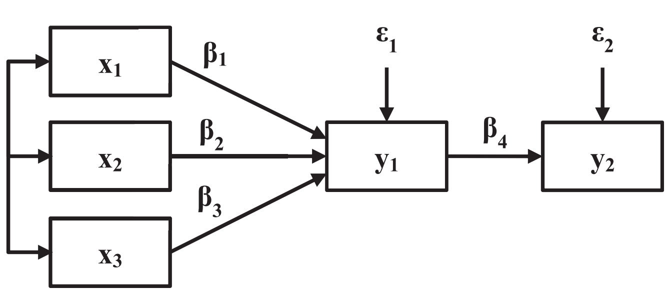

Fig. 1. A structural equation model linking production system characteristics (x1-x3) to performance (y1, y2).

Incorrectly assuming away these correlations amounts to erroneously assuming that the *x*-variables are exogenous, when they are in fact endogenous. In the statistics literature, this situation is dubbed *the problem of endogeneity*. In his definitive treatment of econometrics, Kennedy (2008: 373) noted that "forgetting about possible endogeneity" is one of the common mistakes in model specification. This poses an immediate, fundamental threat to both the theoretical validity and the practical relevance of research findings.

The adverse impact of the endogeneity problem may extend beyond simply over- or under-estimating the true effect. Tan and Staats (2016) examined the assignment of servers to tables at a restaurant, hypothesizing that servers with a heavier table load would be less likely to be assigned a new table even when the commonly-used "round-robin principle" of assigning tables suggested assigning the table to the server. The Ordinary Least Squares (OLS) model (where potential endogeneity of the table load variable was not taken into account) suggested, however, that servers with a heavy table load would be *more* likely to be assigned the next table. This means that despite a heavier work load, the already very busy server would be assigned yet another table, indicating a Matthew Effect (Merton, 1968) of sorts. This may be surprising, but is at the same time entirely plausible given our knowledge of the psychosocial processes involved in organizational reward systems. But the authors knew that table load was likely endogenous to other model variables through reverse causality. To address this, they ran a two-staged least squares (2SLS) model where endogeneity of the table load variable was taken into account. In stark contrast with the (incorrect) OLS model, the 2SLS model suggested that the busy server would indeed be *less* likely to be assigned the next table. This would lead to just the opposite, and again entirely plausible conclusion, that restaurants do indeed seek to ease the workload of busy servers. Of the two plausible conclusions, the latter is obviously the one supported by evidence: the OLS estimate is biased and cannot be trusted. The lesson here is that ignoring endogeneity would have led to seriously misguided conclusions: absent proper controls, we may not even get the sign right.

1.2. Is there a solution?

Researchers across many fields of science have struggled with the problem of endogeneity for decades. The earliest documented attempts at addressing the problem date back to the examination of supply and demand of animal and vegetable oils in the 1920s (Stock and Trebbi, 2003: 177). The problem is thus not novel, but it is as widespread as it is persistent. The reason is succinctly stated by Roberts and Whited (2013: 498): "*there is no way to empirically test whether a variable is correlated with the regression error terms because the error term is unobservable.*" This is why exogenous *latent* variables in particular—the disturbance term being the most common special case—are the cause of so much headache to empirical researchers. Because many key exogenous variables of concern are not measured, "there is no way to statistically ensure that an endogeneity problem has been solved" (Roberts and Whited, 2013: 498). This means that the problem of endogeneity is not so much a problem as it is a *dilemma*—hence the title of this paper. Dilemmas do not call for solutions, they call for choices. In the statistical sense, the dilemma boils down to trading one set of untestable assumptions for another. Our goal in this methodological note is to make these assumptions salient so as to enable an informed choice. There are no direct tests of endogeneity, and the consequences of this must be understood. But there are many indirect tests that give the researcher useful information to guide their decisions and conclusions. Our particular focus in this paper is on what could constitute reasonable criteria for assessing the choices researchers make on dealing with endogeneity in research manuscripts. But we also wish to examine the choices that peer-reviewers make with respect to potential endogeneity problems in the manuscripts they evaluate.

As far as choices regarding endogeneity are concerned, an important tension must be acknowledged. On the one hand, most rigorous academic journals require, more or less as part of their editorial policy, that endogeneity must be addressed somehow (e.g., Guide and Ketokivi, 2015; Reeb et al., 2012). Expectations are clearly rising. On the other hand, endogeneity can be a destructive tool in the hands of the skeptic: because endogeneity is not a problem that can be solved, the skeptic, if so inclined, can effortlessly dismiss as untenable any argument that is based on empirical data, particularly if the researcher has not sought to avoid the problem in the research design phase. This kind of "academic bullying" is frequent enough to warrant attention: it is not all that uncommon to see a passage in a peer review that simply *declares* the analysis suffers from an endogeneity problem and is therefore invalid. Strictly speaking, putative exogenous variables in a model are probably never truly exogenous. But to expect perfect exogeneity would set an unreasonable standard for empirical research. Therefore, this paper echoes the call for reasonable endogeneity standards found in the recent methods literature (Antonakis et al., 2010; Ashley and Parmeter, 2015; Conley et al., 2012). Among other things, these authors shift attention from *strictly* exogenous variables to *largely* (Antonakis et al., 2010), *approximately*, or *plausibly* (Conley et al., 2012) exogenous ones. The intent of these authors is not to give up rigor by introducing lenient standards, nor do they suggest that we should gloss over the endogeneity problem. Just the opposite, in fact: the conclusion that violating strict exogeneity constitutes a problem for inference is essentially *assuming* the conclusion, not *demonstrating* it. Thus, whether endogeneity is a cause for concern requires an explicit analysis of how sensitive inferences are to endogeneity in the specific research situation. To embrace plausible exogeneity is to commit to conducting such an analysis. Shifting from strict to plausible exogeneity is not in any way an attempt to replace formal statistical criteria with informal ones. Indeed, examining plausible exogeneity involves—and should involve—mathematically rigorous analysis. Ashley and Parmeter (2015: 71), for instance, sought to develop a procedure specifically for quantifying the "fragility" of inference with respect to endogeneity: express the *p*-value of a relevant null hypothesis partly as a function of the possibly non-zero correlations between the disturbances and predictor variables (see also Kiviet, 2016). For a given model, then, one can (and should) assess the minimum amount of endogeneity capable of reversing an initial statistical decision (under the naïve assumption of perfect exogeneity) about the existence of an effect. In the case of mediational models (such as that in Fig. 1), it is worth noting that similar types of sensitivity analysis have been developed—both in the social science and epidemiological literature—where it is also understood that perfectly satisfying all causal assumptions (e.g., no unobserved confounding) is an unattainable ideal in practice. Consequently, one should always investigate the consequences of varying degrees of violations of the required conditions (Imai et al., 2010; Imai and Yamamoto, 2013; Tchetgen Tchetgen and Shpitser, 2012; VanderWeele, 2010; VanderWeele and Chiba, 2014).

In the remainder of this paper, we examine endogeneity in the context of OM research. We start with the theoretical aspects, move on to the empirical, and then discuss some of the more pragmatic approaches to avoiding endogeneity in research design and analysis.

2. The theoretical aspect

Let us return to the model in Fig. 1 and express the endogeneity concern in the language of the research question and the model. Let us first examine the question: "What does it mean to assume x_1-x_3 are exogenous variables that do not correlate with ε_1?"

Although the possibility of endogeneity requires careful substantive reflection in any application of statistical modeling, it is not possible to prescribe universally applicable recommendations across all scientific fields. Rather, all examinations of endogeneity must be properly contextualized in order to understand where in the model the problem is, what could be causing it, and most importantly, what can be done about it. Suppose the population under examination in Fig. 1 is the automobile industry and the sample consists of final assembly plants. Striking similarities across plants notwithstanding, it is well established that there is variance in x_1-x_3, that is, different automakers use at least somewhat different production technologies, plant layouts, and outsourcing arrangements (Adler et al., 1999; Dyer, 1996; MacDuffie, 1995; Monteverde and Teece, 1982). The question of how this organizational and technological heterogeneity—heterogeneity, in short—links to factory productivity and business performance has been a relevant question for OM research for fifty years. It is also obvious that OM researchers neither design auto assembly plants nor are able to convince plant managers to modify plant layouts or try out new production technologies. The best we can realistically expect is to gain access to observe these systems and interview those who design and manage them.

The first question is whether the factors that give rise to heterogeneity are related to other, unmodeled factors that also drive productivity. If they are, then x_1-x_3 will likely correlate with ε_1, given that ε_1 represents the variation in y_1 that is due to these unmodeled factors. One obvious driver of both the choice of organizational technologies and productivity is strategic choice: if an automaker's strategy is to be responsive and flexible (both of which tend to compromise conventional productivity metrics), it chooses production technologies accordingly. Indeed, one of the first systematic prescriptions in the OM literature is that process choice must be consistent with strategic priorities (Skinner, 1974). But this directly implies that decision-makers—should they follow this commonsensical prescription—will also choose a whole host of other organizational arrangements that are consistent with the chosen strategy and will plausibly have implications for productivity. Yet, these factors are not modeled in Fig. 1. The researcher thus faces what is perhaps the most common source of the endogeneity problem in management research: omitted variables (e.g., Clougherty et al., 2016; Hamilton and Nickerson, 2003). Out of a hundred possible x's, the researcher has incorporated only a handful into the model. To the extent that the modeled variables correlate with omitted variables that also affect productivity, model estimates will be inconsistent (Fig. 2).

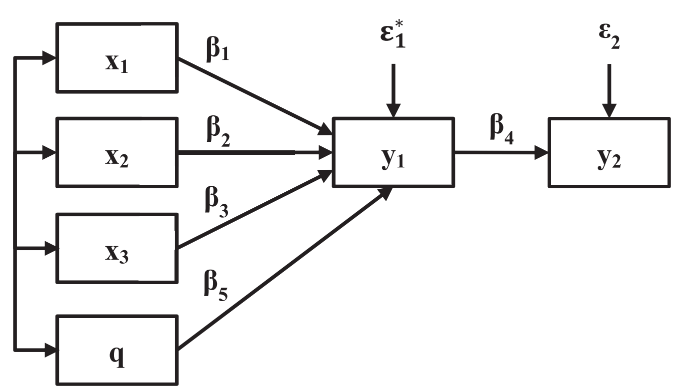

Fig. 2. The original model, with omitted variable q included.

Consider next the assumption that x_1-x_3 are uncorrelated with ε_2. Do factors that give rise to the heterogeneity at the plant level also affect business-unit profitability? We cannot possibly deny the existence of such factors, however, it is difficult to think of ones that would cause an immediate and considerable threat to model specification and estimation. Again, the objective is not to determine whether exogenous variables are actually endogenous; they *always* are, at least to an extent. Instead, we need to determine whether the problem is enough of a cause for concern in that it materially jeopardizes estimation. We do not see any compelling reasons for why x_1-x_3 would correlate strongly with ε_2. For one thing, profits are tallied and reported by business units that are separate from those making decisions about the design and implementation of the factory's production system, which would be expected to contribute to plausible endogeneity of the x's with respect to ε_2, due to reduced potential for omitted variable bias. Therefore, if we were to peer-review a manuscript presenting Fig. 1 in the context of automobile production, we would let the author assume the correlation is sufficiently low to be ignored. Letting authors ignore conceivably trivial correlations—even though they are known to be non-zero—is what we mean by reasonable standards.

How about the assumption that the two disturbance terms ε_1 and ε_2 are uncorrelated? Here it is easy to see how unmodeled factors that drive productivity may also significantly drive profitability, which would make this correlation non-zero. Numerous external factors that affect plant productivity also affect business-unit profitability. Energy prices are just one example out of many. The annual energy consumption of an auto assembly plant is hundreds of gigawatt hours, which means the added cost effect of a ten-percent increase in the price of electricity is in the order of a million dollars. This extra cost may directly affect both productivity and profitability. Whether this is the case obviously depends on how exactly productivity in particular is operationalized.

In summary, let us take stock of the two key concerns: (1) x_1-x_3 correlate with ε_1, and (2) ε_1 correlates with ε_2. That x_1-x_3 would correlate with ε_2 is in our view less of a concern. We must emphasize that these are judgment calls that stem from our own interpretation and experience; another person evaluating the same model might think differently (see *endnote #3*). The crucial thing to understand is that the conclusion about endogeneity—whatever it may be—is based not on statistical theory but on substantive, theoretical considerations, and that is why evaluation of endogeneity is ultimately a judgment call. As the Tan and Staats example showed, endogeneity has potentially lethal statistical consequences in applied research, but the task of actually gauging the nature and the extent of the problem falls first and foremost on the researcher, who is armed with subject matter knowledge. Statistical checks and solutions can certainly be valuable, but researchers who want to deal properly with endogeneity must, above all, continually remain abreast of theoretical developments and research findings that provide clues about potential sources of the problem in the models they are evaluating.

2.1. Theoretical prescription

The discussion above leads to a number of general prescriptions at the theoretical level. The most obvious one is to avoid specifying models that immediately lead to serious endogeneity problems. There are many fascinating questions and hypotheses that may be both academically and practically interesting but simply impossible to test empirically (Blalock, 1991: 326). Management researchers in particular must be cautious of this. In our attempts at being managerially relevant, we are understandably drawn to examining variables that can be subjected to managerial selection: if x is given, there may not be much managerial insight in showing that it links to performance (e.g., Bromiley and Rau, 2016). If the drivers of performance were exogenous to choice, what could we prescribe? But in embracing managerial relevance, we walk directly into an endogeneity trap: If x_1 is endogenous to strategic choice, so are x_2 through x_k, where k is a large number. It is more or less guaranteed that x_1 will correlate with many other x's that also correlate with performance (Clougherty et al., 2016; Hamilton and Nickerson, 2003). This constitutes an essential and fundamental endogeneity dilemma for empirical management research in particular.

The endogeneity trap is difficult to avoid, but there are some theoretical steps that researchers can take in order to help both gauge and mitigate its potential impact. A logical starting point is to list all the potential endogeneity concerns implied by the model, and then try to determine how serious a threat each of these pose to empirical analysis. The ability to conduct such an analysis does not require statistical sophistication, only the ability to understand the structure of the model and to interpret the assumptions embedded in the structural restrictions of the model: Which variables (latent or measured) are assumed to be uncorrelated with one another and why? In fact, this exercise is purely theoretical and does not require any statistical training, it simply involves an examination of how and why the relevant factors of interest are interrelated (cf. Whetten, 1989). We are merely suggesting that in addition to the theoretically relevant variables, the disturbance terms be considered relevant variables as well. There are various visual inspection techniques that can be used to examine causal assumptions. Pearl (2009) and Morgan and Winship (2007) both offer useful expositions on the topic. As the figures in this paper suggest, we also find the graphical techniques useful to convey important points about endogeneity.

The second prescription is to think of omitted variables already at the model specification stage, not as an ex post concern. Potential omitted variables are implied by how one chooses the dependent variable. For instance, in modeling *system-wide* performance such as plant productivity, one must be prepared to measure a lot of variables. If the aim is to model the performance of a *sub-system*, such as a production cell, fewer variables will suffice. Complexity must be met with complexity: one cannot expect to be able to successfully model something complex by measuring just a few variables. This is why the most rigorous research approaches to modeling system performance in OM indeed measure hundreds of variables that probe different aspects of the operational system (Schroeder and Flynn, 2001). At the same time, the omitted-variable problem is merciless in the sense that even one omitted variable can cause problems. Indeed, Krause (2013: 3201) argued that *any* residual variance—a non-zero disturbance term—is not only cause for concern, but also pre-empts all causal claims about how the independent variables affect the dependent variable.

3. The empirical aspect

The vast majority of the methods literature across disciplines focuses on empirical remedies for endogeneity. They are an important part of the story, but we want to highlight the fact that theoretical and substantive concerns should always come first. Freedman (1991: 305) put it succinctly in writing that “technical fixes become relevant only when models are nearly right.” In a similar vein, Roberts and Whited (2013: 514) wrote in the context of corporate finance research that the only way to deal with endogeneity is “to understand the economics of the question at hand.”

Finally, Davis (1985: 11) put it bluntly: “computers cannot substitute for [the researcher] in analyzing data, because computers do not know anything about the real world.” Mechanistic, algorithmic approaches simply cannot serve as a substitute for substantive expert judgment, a sentiment best captured by Bollen (2000: 74): “The largely objective basis of statistical algorithms does not remove the need for human judgment in their implementation.” Yet, researchers tend to rely on algorithmic “if A then B”-type rule-following in addressing endogeneity in particular and in statistical reasoning more generally (Lance and Vandenberg, 2008; Ziliak and McCloskey, 2008). On the upside, algorithms can be useful tools in informing and complementing our judgment. But they are not substitutes: the question that calls for expert judgment is the degree of threat to the substantive conclusions in a specific research situation. In the following, we discuss the most common approaches, their strengths and their weaknesses.

3.1. Approach 1: run an experiment

The most obvious empirical remedy is to ensure that exogenous variables are truly exogenous. This is obviously not feasible with latent variables, but some observed variables may lend themselves to experimental manipulation, which means the researcher can control their variances to ensure they do not correlate with variables with which correlations are undesirable. The research model becomes correctly specified and estimates can be trusted. This is a huge advantage from an inferential point of view, because model misspecification is perhaps the biggest threat to valid inference.

But research is not just about getting the inference right, it is about explaining phenomena of theoretical and practical interest. To suggest that experimental research designs offer a general solution would be problematic in OM research in particular. Problems arise from the fact that experimental research tends to focus, understandably enough, on experimental manipulation at the individual or perhaps the small-group level. But how would one conduct experiments at higher levels of analysis than the individual or small group? As the title *operations management* readily suggests, many relevant OM research questions and theories reside at the system level. We looked at the ten most-cited articles (ISI Web of Science) that have appeared in JOM—every single one focused on a system-level phenomenon. What kind of an experimental research design can inform us on the productivity effects of outsourcing decisions, management of humanitarian relief operations, resource allocation for emergency room operations, or any large-scale production operation? System-level interventions tend to be infeasible in most contexts, because few managers and owners of businesses are willing to have their organizations manipulated by researchers (Clougherty et al., 2016).

3.1.1. The challenge of experimental research in (operations) management

Experimental research is not immune to endogeneity, particularly if experiments are conducted in the field instead of the lab: what the researcher is ultimately interested in examining may be difficult to make exogenous by manipulation. The notion of *experience* is the best example that comes to mind. In their experimental field study on employee productivity in a factory, Hossain and List (2012: 2162) noted that endogeneity of experience constitutes a “nagging issue” in experimental field research. What is relevant in the context of this paper is that empirical results from field experiments rely on exogeneity assumptions that are statistically identical to assumptions made in observational research. In an ambitious study, List (2011) sought to make experience exogenous in the context of a longitudinal field experiment in the sports card market context. List's objective was “to induce experience in a natural way in a natural environment” (List, 2011: 314). These kinds of longitudinal experimental designs are highly commendable, but they are as rare as they are challenging.

It is important to acknowledge that although experimental research designs do not preclude endogeneity, the problem is clearly not of the same magnitude as it is in observational research. However, insofar as the researcher's aim is to get the magnitude right, experimental research faces a serious handicap. The problem with the models estimated by experimental researchers is that it may be difficult to argue that something practically relevant is being estimated. Of course, one may legitimately argue that “getting the magnitude right” is not the primary objective in experimental research in the first place: often it is to determine whether a given intervention leads to at least some degree of measurable change or benefit. But if this is the case and the objective, then claims of practical relevance must be made with caution. An example from experimental research with student subjects elaborates: Bolton et al. (2012) examined the classic Newsvendor Problem using both students and managers as experimental subjects. They found that the direction of the effect was the same but the magnitudes were different. Therefore, if the experimental researcher is indeed interested in magnitudes, there are potentially serious problems. If getting the sign right satisfies the researcher's knowledge objective, then the problem may not exist.

In sum, the main challenge with experimental research is not endogeneity but external validity. At the same time, if a study lacks external validity, the effect is in many ways comparable to the effect associated with endogeneity: we do not get the relevant magnitude right. We also want to raise the concern that if we start emphasizing methodological aspects of endogeneity at the expense of substantive considerations, we may end up in a situation in which our desire to have a properly specified model trumps the idea of focusing on what is truly interesting and relevant. This is part of a larger problem called “method-driven research”, which unfortunately can often eclipse the more appropriate objective of “question-driven research” (MacCallum, 1998). Choosing models and explanatory variables based on whether they can be credibly manipulated in an experiment seems like putting the cart before the horse.

3.2. Approach 2: seek instruments for troublesome variables

In order to understand the rationale for the most common statistical solutions for endogeneity, it is useful to revisit the model in Fig. 1. One way to address the potential correlation between ε_1 and ε_2—which we deemed was indeed cause for concern—is to do the obvious: add the correlation to the model (Fig. 3). Now, if x_1-x_3 are uncorrelated with ε_2 (or these correlations are negligible), consistent estimates of model parameters can be obtained.

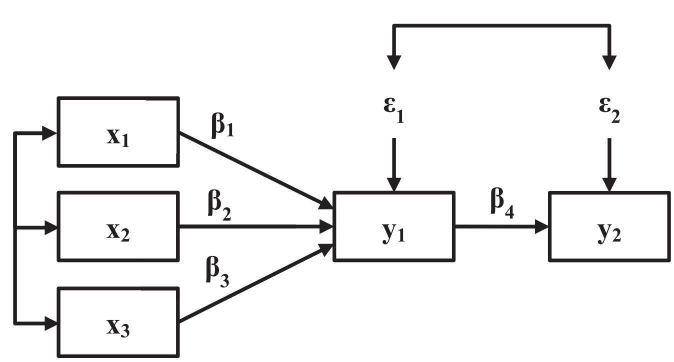

Fig. 3. The original model, with correlated disturbances added.

Fig. 3 is in fact a graphical depiction of what is known as *the instrumental-variable (IV) approach* (Antonakis et al., 2014; Bollen, 2012). Even though the links from x_1- x_3 to y_1 have a theoretical and substantive basis, x_1-x_3 can indeed be thought of as instrumental variables for y_1. Specifically, if the β_i are non-zero, the *relevance condition* is satisfied; if x_1- x_3 do not correlate with ε_2, the *exclusion condition* is satisfied (Roberts and Whited, 2013: 512). Therefore, x_1-x_3, as instruments of y_1, help us estimate β_4 consistently even when y_1 and ε_2 correlate; in SEM terms, the causal effect of productivity on performance, β_4, is mathematically identified. The underlying statistical logic here is that because x_1-x_3 are uncorrelated with ε_2, then the part of y_1's variance explained by x_1-x_3 is also uncorrelated with ε_2. Now, once the part of y_1's variance that does correlate with ε_2 is partitioned into ε_1, only the "good variance" in y_1 is used to estimate the path from y_1 to y_2, resulting in a consistent estimate of β_4 (see *endnote #4* for more statistical details on how instruments can be used to diagnose and deal with endogeneity). Note, however, that the correlation between ε_1 and ε_2 must still be included in the IV model in order to obtain consistent estimates (Antonakis et al., 2010, Fig. 3B and 3C).

One of the enduring questions in both strategic management (e.g., Rumelt et al., 1991) and OM (e.g., Bromiley and Rau, 2016; Hitt et al., 2016) research is what determines organizational performance and why some organizations perform better than others. We believe that OM researchers can contribute to these conversations in a unique way, because many grassroots organizational and technological factors that contribute to micro-level performance metrics are measured. To the extent that the relevance and exclusion conditions for exogeneity are plausibly satisfied, OM researchers can contribute to broader discussions in management by providing consistent estimates of the relationships between operational performance and business performance. Even though OM researchers do not think of x_1-x_3 in Fig. 3 as instruments of y_1, that is exactly how they can be interpreted and used to advantage. Indeed, we would not be at all surprised if strategy and economics scholars saw x_1-x_3 in Fig. 3 primarily as instruments of productivity.

Of course, if variables x_1-x_3 are themselves troublesome—there was indeed a concern that they correlate with ε_1—then obviously they require instruments as well. As is being increasingly recognized in the more general methodological literature on mediation models, causal interpretation of the parameters—both direct and indirect effects—rests on being able to adequately control for, among other things, unobserved confounding of the relationships between independent variables and the proposed mediators (Imai et al., 2010; Keele et al., 2015; Muthén and Asparouhov, 2015; VanderWeele, 2015). There are two reasons for why this is particularly complicated in the OM context: (1) it is typically not feasible to measure all relevant covariates in any given situation; and (2) it is difficult to think of how to instrument organizational variables at such a micro level of analysis. It may be possible, but at least we cannot think of useful instruments for factors such as process choice or implementation of certain manufacturing processes. This is a common obstacle in management and strategy research more generally: "it is difficult in many strategy data sets to find instrumental variables that affect strategy choice but not performance" (Hamilton and Nickerson, 2003: 67). Again, in management research generally, many variables of interest are endogenous to managerial choice, which means many practically relevant research questions suffer from an endogeneity problem by definition. At the very least, however, researchers can and should aim to include several theoretically-relevant independent variables in their models. Including only one variable or a very small number of variables—such as top management commitment—as the sole initiators of the chain of causal effects in a mediational model is risky. Consider the extreme case of just one exogenous variable in a mediational model, which is not at all uncommon in OM mediational models. The endogeneity of this variable would contaminate the parameter estimates of the entire model. Striving to include multiple sources of exogenous variation can potentially help to mitigate such biases.

3.2.1. *The problem with instruments*

The general problem with instrumental variables is that they often cause more problems than they solve (Murray, 2006). In employing an instrumental variable one is able to mitigate estimation inconsistency, but estimates lose efficiency (become less accurate) and their finite-sample bias increases. Even if the instruments are uncorrelated with the disturbances at the population level, there will be a "nuisance correlation" in any finite sample of the population, which results in a bias toward the OLS estimator (Jeong and Yoon, 2010). These nuisance correlations can be expected to be even worse with weak instruments, as small effects are subject to greater sampling variability. Bias-corrected IV-estimators do exist, however (see Bun and Windmeijer, 2011).

The Tan and Staats example discussed earlier illustrates the efficiency-bias tradeoff, as well as the tradeoff of one set of untestable assumptions for another. In their analysis, Tan and Staats found that the variance of the table load coefficient in the 2SLS model was *200-fold* compared to the variance of the OLS estimate. The lower the correlation of the instrument and the original troublesome predictor, the greater the efficiency loss. Efficiency loss can obviously be offset by having a large sample, but with typical sample sizes (in the hundreds or perhaps a thousand), the efficiency loss tends to render the IV approach infeasible. Tan and Staats had a sample size of over one million observations, and could thus easily afford the massive loss of efficiency, as well as more or less completely avoid the small-sample bias effect. A sample size of one million is obviously an exceptional circumstance in OM research.

An additional caveat with instrumental variables arises from the fact that all statistical analyses are erected on assumptions, at least some of which are untestable (see *endnote #4*). With instrumental variables, the relevance condition is testable (because it involves only measured variables), but the exclusion condition is not (because it involves the unobserved disturbance term). The exclusion condition obviously applies to instrumental predictor variables as well. Because consistent estimation under instrumental variables rests on this correlation being zero, there is no way to guarantee that the instruments have actually provided *any* protection against endogeneity. In an attempt to assure that the exclusion restriction is met, researchers sometimes incorporate covariates that correlate with the instruments, the troublesome independent variable, and the dependent variable. The logic in doing this is that by including such covariates the correlation between the instrument and the disturbance term would be eliminated. This property is referred to as *conditional exogeneity* of the instruments (White and Chalak, 2010). But once again, because conditional exogeneity involves the disturbance term, it remains an assumption that is not directly testable.

The literature on instrumental variables is massive and there are many other empirical approaches to address the problems they raise. The most realistic of these include ways to seek reliable inferences under conditions where the exclusion restriction in particular is violated (Conley et al., 2012; Kraay, 2012). These approaches remind us that endogeneity as a general problem and instrument quality as a particular concern are not yes/no issues. Consequently, applied researchers might ask not whether instruments are weak or strong (the relevance condition) or whether they are valid or invalid (the exclusion condition). Instead, they might ask *how strong* and *how valid* instruments are, and what *the severity* of the endogeneity problem is. There is a lot of contemporary literature that focuses specifically on the question of magnitudes, but a comprehensive review of these works is outside the scope of our inquiry. A key representative example that comes to mind is Conley, Hansen, and Rossi's (2012) notion of *plausible exogeneity*, which they apply specifically in the context of instrumental variables. Plausible exogeneity is based on the idea that relaxing the exclusion restriction may not materially jeopardize inference. After all, Conley et al. (2012: 270, emphasis added) argue, “researchers routinely provide informal arguments that their instruments satisfy the instrument exclusion criterion but recognize that this may only be *approximately true*.” Further, plausible exogeneity is not merely an idea, it is implemented in the PLAUSEXOG STATA module (the module can be installed simply by writing *ssc install plausexog* on the STATA command line).

In sum, instrumental variables can be useful and often serve an important purpose, but their application is based on a host of assumptions that are just as difficult to satisfy as those of classical regression. Applying instrumental variables amounts to trading one set of untestable assumptions for another, and using a bad instrument may well make things worse than sticking to OLS (Murray, 2006; Semadeni et al., 2014). In addition, one might end up being stuck in an endogeneity trap that is simply too severe to allow capitalization on the recent innovations that attempt to either loosen or to accommodate violations of classical assumptions. This observation offers a segue to the next candidate solution to tackling endogeneity: instead of trying to come up with better *models*, perhaps an actionable answer lies in getting better *data* (cf. Bettis et al., 2014).

3.3. Approach 3: examine variance and change over time

Panel datasets, where a large number of observational units are observed over time (e.g., Baltagi, 2005), is gaining more prominence in OM research. Longitudinal data do not of course solve the endogeneity problem (because it cannot be solved), but panel datasets have many notable advantages over cross-sectional data. In the following, we discuss two relevant variants of longitudinal data: observational and quasi-experimental data.

3.3.1. Longitudinal, observational data

With longitudinal data, one can often alleviate endogeneity concerns by eliminating at least some of the effects of one of the leading causes of endogeneity, unobserved heterogeneity (Roberts and Whited, 2013). The technical details can be found in published work (Baltagi, 2005), but in a nutshell, when longitudinal data are available on the observational units, researchers can add to their models unit-specific *fixed effects (FE)*, which will control for all unmodeled differences that are both stable over time and that may correlate with the predictors of interest. Again, a graphical depiction of the model is useful: Fig. 4 shows a four-wave panel model with one independent and one dependent variable, and a fixed effect μ (Bollen and Brand, 2010: 9). The FE can be thought of as a latent variable that has a constant effect on the outcome variable. Including this latent variable in the model effectively eliminates the unmodeled time-invariant effects from the dependent variable.

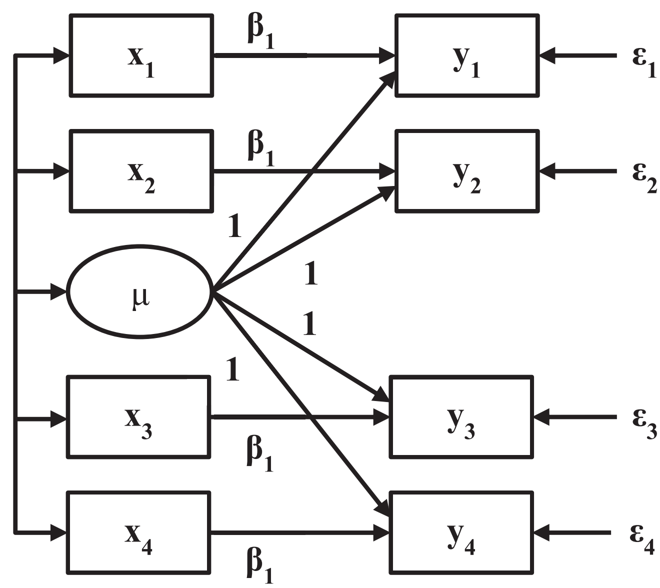

Fig. 4. A four-wave panel model with a fixed effect μ. Figure reproduced from Bollen and Brand (2010: 9) with permission from Oxford University Press.

Because adding the FE alleviates unobserved heterogeneity due to omitted variables, the FE model may already constitute a huge leap forward in addressing endogeneity. At the same time, other sources of endogeneity, such as measurement error and unobserved time-varying confounding, are not addressed. Measurement error could potentially be countered by the use of latent time-varying predictors, and known time-varying confounders could obviously be added to the model in Fig. 4, just like the unmodeled variable *q* was added into the model in Fig. 2. From Fig. 4, it is also easy to see how failing to include the FE in the model when it is non-zero constitutes a special case of the omitted variable problem, with the familiar consequence of a biased and inconsistent estimate of β_1, the parameter that links *x* to *y* in the model (Baltagi, 2005: 13).

Panel-data regression models can be used to examine variance in the dependent variable over time: the model in Fig. 4 seeks to account for the variance in *y* in four different measurement points. However, in addition to explaining *variance* over time, one may also be interested in modeling *change* over time (see *endnote #5*). Indeed, when one thinks of the various theories of OM, it seems that many of them have implications specifically to how operational systems change over time. But change over time does not equal variance over time. To model change over time, one would need a model very different from that in Fig. 4. There is a rich literature on *latent change models* (e.g., Bollen and Curran, 2006; McArdle, 2009; Preacher et al., 2008) that discusses the methodological details of modeling change over time; the software applications are equally well-developed. An example of a latent change model is depicted in Fig. 5. In the model, β_0 and β_1 are not parameters but rather random variables, that is, observation-specific baselines and change trajectories. Specifically, β_0 represents the intercept and β_1 linear change over time. The paths from the latent variables are not factor loadings but regression weights that are fixed a priori to obtain the desired interpretations of the β's. The components of the change trajectories can then be predicted by independent variables (*x* in Fig. 5). Indeed, if the researcher wishes to examine whether, say, the implementation of a set of managerial

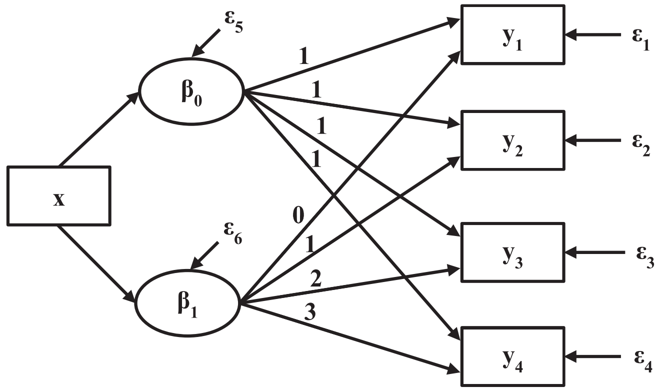

Fig. 5. A four-wave latent change model with a linear change trajectory β_1.

practices or a specific strategy has an effect on performance, a latent change model seems like the appropriate modeling method.

3.3.2. Longitudinal, quasi-experimental data

Sometimes the observed units in the sample are all subject to the same exogenous shock, such as the financial crisis or other unexpected changes. This exogenous shock is akin to an experimental treatment, but since there is no randomization or control versus treatment groups, the design is sometimes labeled a *quasi-experiment* (Cook and Campbell, 1979). However, quasi-experiments enable a before-and-after type analysis, which enables consistent estimation of the effect of the exogenous shock, given that subjects essentially act as their own controls in terms of stable observed and unobserved factors. In what are called *interrupted time series (ITS)* designs (e.g., Kontopantelis et al., 2015), the outcome is measured at multiple points before and after the intervention, and the objective to determine whether there is a discernible discontinuity in the time series at the intervention point. In fact, the latent change model in Fig. 4 could easily be turned into an ITS model. Suppose that there is an intervention or interruption sometime between t_2 and t_3. The effect of the intervention could be modeled by setting the regression weights of β_1 to zero at pre-intervention measurement points (t_1 and t_2) and to one at post-intervention measurement points (t_3 and t_4).

3.3.3. Problem with longitudinal research designs

There are a number of drawbacks to longitudinal research designs that are idiosyncratic to OM research. The obvious challenge is that panel data are difficult to obtain. In contrast to strategy or economics scholars who may simply *purchase* longitudinal accounting data, OM researchers often have to build their datasets from scratch. There are no commercial datasets that contain, say, longitudinal factory-level information about technology and productivity. Factory-level productivity is typically considered sensitive information, and very few companies will disclose it to outsiders. Barring a miracle that someone has already collected the data in which the OM researcher is interested, constructing a ten-year longitudinal dataset can indeed take ten years. Another OM-specific challenge is that as precious as exogenous shocks can be to a researcher, an OM researcher must be incredibly lucky to find a meaningful exogenous shock that has an impact of theoretical interest.

A third problem applies generally to all FE models, such as the one depicted in Fig. 4. Because there is a time-invariant FE in the set of predictors, any theoretically interesting predictor that is relatively stable over time will be highly collinear with the FE, which leads to estimation problems. The FE model therefore cannot include any predictors that are stable over time. In many OM research questions, there are variables whose effects on, say, performance, are of interest, but which are relatively stable over time—manufacturing process choice and the scale of operations are good examples. OM researchers tend to examine operational systems that do not undergo significant changes in the time window they are examined. In such cases, the FE model would be problematic. On the other hand, if one simply wishes *to control* for the effect of a time-invariant variable, say, industry or country of incorporation, this would be effectively captured by the FE. Problems arise when the time-invariant effect is of interest and needs to be *estimated*.

In the case where time-invariant effects are of theoretical interest, one would need to shift to a model where time-invariant regressors can be included. In econometrics, this is called the *random-effects (RE) model*, depicted graphically in Fig. 6 (Bollen and Brand, 2010: 7). The RE model allows for the inclusion of time-invariant regressors (z_1 in Fig. 6). But again, there is no free lunch: the challenge with the RE model is that the RE must be assumed to be exogenous to *all* of the predictor variables (Baltagi, 2005: 19). The choice of RE versus FE thus constitutes a tradeoff: FE relaxes the assumption of no correlation between the unobserved effect and the predictor variables, but simultaneously, preempts the modeling of time-invariant predictors. We conjecture that given the choice, in most research situations the OM researcher would prefer the FE model. The reasoning is aptly summarized by Wooldridge (2002: 265, emphasis added): "In many applications *the whole point* of using panel data is to allow for [the unobserved effect] to be arbitrarily correlated with the [regressors]." This applies directly to OM research (see *endnotes #4 and #6* for more details on the FE vs. RE distinction—confusion abounds).

Finally, it should be obvious from both Figs. 4 and 6 that FE and RE models both assume *all* regressors—including the unobserved effect—to be uncorrelated with the disturbances of the dependent variable (ε_1 through ε_4). The statistical assumption of exogeneity is therefore identical to that of cross-sectional models; the longitudinal aspect by itself does not eliminate the endogeneity problem. As Duncan (1972: 36) put it, "for some reason there is widespread, though not well articulated, opinion that in panel analysis the usual obstacles to inference and estimation are suspended for the benefit of the analyst." The only thing these models achieve is that adding the FE or the RE works toward alleviating endogeneity due to

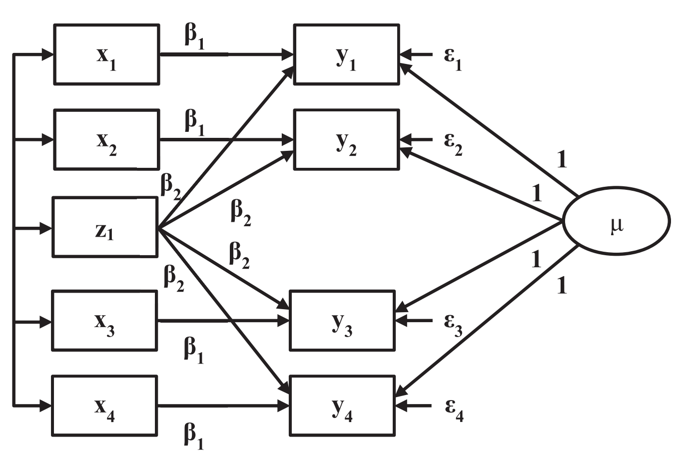

Fig. 6. A four-wave panel model with a random effect μ and time-invariant covariate z_1. Figure reproduced from Bollen and Brand (2010: 8) with permission from Oxford University Press.

omitted variables.

We have summarized the different empirical approaches and the inherent dilemmas in Table 1. The key point is that every empirical approach builds on assumptions that researchers either do not test or that are simply impossible to test. Which approaches are deemed acceptable in a specific research situation depends on what those evaluating the analysis—the peer reviewers—find satisfactory. But it is becoming clear that the assumption of plausible exogeneity in cross-sectional research designs is no longer considered credible, particularly if the researcher has done nothing to establish plausibility. Both the OM research community and management research community more generally are now taking more seriously the notion that the predictor variables in our models are endogenous to managerial choice.

There are numerous examples in empirical OM and management research more generally that apply the techniques listed in Table 1. This underscores the fact that these techniques are neither methodologically novel, nor does their use need to be justified or prescribed. In terms of experimental research, for example, there is a sizable behavioral operations management community, where scholars conduct primarily experimental research using regression- and ANOVA-based models (Bendoly et al., 2015). Application of instrumental variables is also well established in the OM literature: a simple keyword search on “instrumental variables” yields dozens of hits to recently published articles in major OM outlets, most notably *Manufacturing & Service Operations Management* and *Management Science*. At the same time, it is worth observing that in two leading OM journals—*Journal of Operations Management* and *Production and Operations Management*—roughly 80 percent of articles applying instrumental-variable techniques have been published in the past five years. There are also OM journals that have published only a few, if any, articles applying instrumental variables, or more generally, articles that consider the endogeneity problem in the first place. Hence, even though the methods themselves are established, we observe massive variance across OM journals in their application. It is a cause for concern that in a number of rather influential journals, application remains either marginal or is a very recent development. This same observation applies to longitudinal research designs applying panel-data econometrics.

The only approach in Table 1 that remains virtually unused in OM research across all the entire journal base is latent change modeling. Common in other fields of social science, (e.g., Bollen and Curran, 2006), its applications in management research is almost non-existent. Indeed, we found no OM examples, and even examples from the general and strategic management research are scarce (see Jokisaari and Nurmi, 2009; for an example). At the same time, the idea of examining the change trajectories of observational units—employees, teams, projects, factories, buyer-supplier relationships—over time is a perfect fit for many classic and contemporary research questions in OM. The highly popular fixed-effects panel-data regression model can be used to model within-unit *variance*, but this is not the same as within-unit *change*, which is a special and a particularly relevant case of variance. It is relevant, because in many research questions, addressing how observational units change over time is a central concern. The latent change approach to the analysis of longitudinal data is both intuitive and theoretically relevant, but it remains largely unexploited and indeed, ignored even in the most recent literature on longitudinal methods, which focus exclusively on variance over time, or at best, changes from one time period to the next in a *difference-in-difference* model of some kind (e.g., Certo, Withers and Semadeni, 2016).

4. The pragmatic aspect

If the problem of endogeneity cannot be solved, the challenge for the empirical scientist seems almost overwhelming. How do we avoid unreasonable standards? How do we avoid the pitfall of turning endogeneity into a destructive tool that can be summoned at will to dismiss any argument based on empirical data? In this section, we discuss the pragmatic aspect of addressing the endogeneity dilemma: What can we reasonably expect from our colleagues as we review one anothers' work? This third substantive section is directed to both those who author arguments and those who evaluate them.

A straightforward rule based on a simple principle of equity and reciprocity is that we should always review the work of our peers using the same standards we wish they would use when they review ours. The point is not trivial, because volumes of experimental psychosocial research have shown that we apply different standards to our own behavior than we do to the behavior of others. If we demand that authors seek multiple instruments for each troublesome variable so as to enable the use of the Sargan-Hansen *J*-test, are we prepared to follow suit in our own research? If we call for an experimental research design, are we in return willing to let our peers tell us what kinds of research designs to adopt ourselves?

Table 1
The endogeneity dilemma in different empirical approaches.

Approach	Advantage	Endogeneity Dilemma
Regression or ANOVA model (experimental)	Experimentally manipulated regressors are *plausibly exogenous*	Many variables of theoretical and practical interest may be impossible to make exogenous by experimental manipulation (e.g., experience)
Cross-sectional regression model (observational)	Provides trustworthy inference when regressors are *plausibly exogenous*	Requires the assumption that regressors do not correlate with the disturbance term
Cross-sectional IV regression model (observational)	Introduces additional variables called *instruments* to take into account the fact that regressors may correlate with the disturbance term	Requires the assumption that instruments do not correlate with the disturbance term
Cross-sectional multiple-IV regression model (observational)	Enables a test of whether instruments correlate with the disturbance term	Requires multiple instruments per troublesome regressor and the assumption that at least one instrument per troublesome regressor is valid
Panel regression model, fixed effects (observational)	Mitigates unobserved heterogeneity, alleviating endogeneity arising from omitted variables	Requires the assumption that all regressors and fixed effects are uncorrelated with the disturbance term, and disallows the inclusion of time-invariant regressors of theoretical interest
Panel regression model, random effects (observational)	Mitigates unobserved heterogeneity, alleviating endogeneity due to omitted variables; allows inclusion of time-invariant regressors	Requires the assumption that all regressors are uncorrelated with the random effect, and that all regressors and random effects are uncorrelated with the disturbance term
Latent change model (experimental or observational)	Focuses on individual-level change over time, which is consistent with most theories of development and change	Requires the assumption that all regressors and the latent change components are uncorrelated with the disturbance term

We dare conjecture most of us want to decide on research designs independently, using our own judgment instead of advice given in peer review.

We call this *the reflexiveness consideration*: in reviewing an argument, engage in self-reflection of what you would find reasonable if it were your argument under scrutiny. As Maner (2014: 350) put the matter: “Adhering to arbitrary and outdated reviewer standards that demand perfection is likely to perpetuate a vicious cycle, such that authors of the manuscript we are reviewing may then go on to demand perfection when they review other manuscripts (possibly our own). With this point in mind, it would be sagacious to live by these words: Review unto others as you would have others review unto you.” However, in addition to a vicious cycle we feel compelled to point out the flipside of the problem—call it a *lowered expectations cycle*—where the community is tempted to put convenience before rigor. This cycle is fueled by the fact that ignoring the problem is obviously the easiest solution: If I ignore endogeneity in my own work, then maybe my peers will not challenge me in fear of being one day challenged themselves. It should go without saying that convenience cannot possibly be a viable option in a community of scientists—if anything, expectations must be constantly raised, just not to an impassable degree.

Another straightforward answer that has actionable implications is to focus on remediable problems. In reviewing a manuscript, we must avoid simply declaring that the model suffers from endogeneity and is therefore fundamentally flawed. Here, we suggest that the burden of proof be shifted, at least a little, toward those who evaluate the argument, particularly if the authors have made an honest, structured and well-documented effort to address endogeneity. If the authors have completely ignored endogeneity, then they have not accepted the burden of proof in the first place.

We call this *the remediableness consideration*: in raising endogeneity concerns in a review, first establish that endogeneity is indeed a problem. Then, think of ways to convince the authors that it is also remediable, if possible, and recommend feasible solutions. Note that this approach requires reviewers to continually stay abreast of theoretical and methodological developments regarding endogeneity, so that they can make well-informed recommendations on its treatment. Of course, if it becomes evident that only a miracle in terms of an instrumental variable would solve the endogeneity problem in a given application, then reviewers must also act accordingly, perhaps recommending rejection along with advice to take the research problem back to formula, while offering design and analysis strategies that are less vulnerable to endogeneity threats. Sometimes researchers simply run into dead ends and need to rethink their model entirely.

The third pragmatic standard is based on the idea that much like most issues in science and statistical work, endogeneity should not be thought of as a yes/no issue. If we want to turn it into a simple dichotomy, the answer is clear: endogeneity is always a problem. But this is just stating the obvious. Instead, we should seek to examine whether endogeneity is so severe that it plausibly constitutes a problem insofar as the objectives of the inquiry are concerned.

We call this *the reasonableness consideration*: the explicit acknowledgment that the problem of endogeneity is always a matter of degree, much like reliability of measurement and collinearity. Just like there are no methodologically defensible threshold values for reliability and collinearity (e.g., Lance et al., 2006), there are no rules or statistical tests that tell us whether endogeneity is a problem. Instead of rule-following and mechanistic inference, we need careful contextualization and sensitivity analysis to determine what constitutes a reasonable (and an unreasonable) degree of endogeneity in the specific research situation. The notion of *plausible exogeneity* mentioned earlier is a good example of applying the reasonableness consideration: since we know that the exclusion condition neither holds precisely in practice nor is possible to directly test, it simply does not constitute a reasonable standard—it is but an idealization that tells us *what it would take* to eliminate the endogeneity concern altogether. Idealizations serve a useful pedagogical purpose, but as methodological standards for applied research, they are often unreasonable.

5. Conclusion

“If I do x, will my organization's performance improve?” This is the general question underlying the majority of empirical OM research, and it stands at the foundation of just about any form of managerial prescription. But this central question conceals an immediate and fundamental endogeneity trap: if x is a matter of choice, so are a hundred other x's, all of which potentially link to performance. Therefore, many theoretically and practically relevant research questions in OM research suffer from endogeneity by definition. This means our ability to reliably estimate the magnitudes of effects is threatened, which jeopardizes both the theoretical and practical utility of our findings. Indeed, we may not even get the sign right.

Addressing endogeneity starts at acknowledging that it constitutes a dilemma that calls for tradeoffs. In empirical research where statistical methods are applied, the tradeoff involves replacing one set of assumptions with another. An informed tradeoff, in turn, must start at making the assumptions explicit so that they can be evaluated (Table 1 can be useful to this end). A second step is acknowledging that all analyses are based on at least some untestable assumptions. In short, any assumption involving the disturbance term is untestable. Blalock (1991: 332–333) aptly sums up our position: “[W]e must make use of untested assumptions in all research, experimental or not. But this means that we must collectively make such assumptions explicit so that they can be readily challenged, and then rely on a cumulative process through which specific questionable assumptions are challenged, new data are collected to assess such assumptions, and models altered accordingly. No matter how complex the model becomes, however, there must always remain a series of untested assumptions regarding disturbance terms and the unmeasured and unknown variables that affect them.”

We want to close by returning once more to the third, pragmatic consideration, because we see it as the novel and original argument in our exposition of endogeneity in this paper. Establishing specific standards for reflexiveness, remediableness, and reasonableness—call these the “Three Rs”—is much more challenging than deriving theoretical and empirical standards. Whereas theoretical and empirical treatments have already accumulated a substantial body of literature, the pragmatic approach embodied in the Three Rs is novel, less specific, and very easy to dismiss as lacking methodological rigor. We thus fully realize that adopting a more pragmatic approach will require a significant attitude change among methodological purists who insist on full resolution of the endogeneity problem. We are, however, convinced that if we genuinely wish to address the question “What can we realistically expect empirical scientists to be able to achieve?”, a pragmatic approach is the only option. This is not to say that theoretical and empirical considerations be rejected, rather, we suggest that they be revisited in light of pragmatic considerations. Many existing standards are idealizations, and applying them to research practice translates into unreasonable criteria, which can in turn erect methodological roadblocks to informative substantive inquiry. Instead of pretending that the endogeneity dilemma can be solved, let us carefully derive and understand the tradeoffs it implies and work toward making transparent, informed choices. We are convinced that adoption of a pragmatic attitude could help us get out of what now seems like "an endogeneity impasse" in OM research.

Acknowledgments

We thank John Antonakis, Manos Avgerinos, Mikko Rönkkö, and Brad Staats for their comments on earlier drafts of this paper.

Appendix

Endnotes

The purpose of this appendix is to provide some more details on the empirical approaches to endogeneity and some of the key statistical issues raised in the paper. For more detailed technical treatments, such as proofs and details on the statistical tests, we recommend standard econometric texts such as Baltagi (2013), Cameron and Trivedi (2005), Greene (2011), Kennedy (2008) and Wooldridge (2016). All the topics in this appendix are elaborated in detail in several of these texts.

#1 – The disturbance term

We submit that in most modeling contexts, the label *disturbance term* is more appropriate than *error term* (cf. Lord & Novick, 1968: 38–39), as the latter often connotes the idea of *measurement error*. To be sure, measurement error in a dependent variable is part of its disturbance term, but in general, disturbance terms are in our view best thought of as latent variables that capture the aggregate effect of all the sources of variance aside from the ones modeled by the researcher. In most OM models, the fact that the disturbances account for the majority of variance in the dependent variables is largely due to factors *other than* measurement error.

#2 – How omitted variables and measurement error cause endogeneity

Showing that omitted variables cause an endogeneity problem is straightforward. If the true model is (as in Fig. 2):

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 q + \varepsilon_1$$

but we have not modeled q, then our model is actually (as in Fig. 1):

$$y_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \mu \text{ , where } \mu = \beta_4 q + \varepsilon_1 \text{ is the composite disturbance term.}$$

Unless either all $cov(x_i, q) = 0$ or $\beta_4 = 0$, an endogeneity problem results in that at least some $cov(x_i, \mu)$ are non-zero. Again, in our example, q could represent, among other things, the possible unmodeled organizational factors that correlate with the other regressors in the model and have plausible implications for productivity.

Showing that measurement error leads to an endogeneity problem is similarly straightforward. Suppose that in the model in Fig. 1 we cannot measure x_1 directly but instead observe a proxy x_1^* for it (we assume for the purposes of simplifying this example that we can measure x_2 and x_3 directly, although this will not likely be the case in real applications):

$$x_1^* = x_1 + \nu, \text{or } x_1 = x_1^* - \nu$$

where x_1 and ν are latent variables representing, respectively, the error-free regressor—the *true score*—and the error of measurement. In this case, what we actually estimate is:

$$y_1 = \beta_1 (x_1^* - \nu) + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_1$$

or

$$y_1 = \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + \mu \text{ , where } \mu = \varepsilon_1 - \beta_1 \nu \text{ is the composite disturbance term}$$

The question then becomes: Under what conditions does the estimable model produce trustworthy estimates of the true model parameters (which is what the researcher obviously wants to obtain)? Put differently, can we proxy x_1 with x_1^* with without jeopardizing our inferences about the β_i? The typical assumption in OM measurement in particular and most measurement contexts in general is that the measurement error $\nu = x_1^* - x_1$ is uncorrelated with the true score x_1 (Lord and Novick, 1968: 36). However, this means that the observed x_1^* unavoidably correlates with ν, as we can show using expected values:

$$cov(x_1^*, \nu) = E(x_1^* \nu) - E(x_1^*)E(\nu)$$

Given the typical assumption that $E(\nu) = 0$, we have:

$$cov(x_1^*, \nu) = E(x_1^* \nu) = E((x_1 + \nu)(\nu)) = E\left(x_1 \nu + \nu^2\right) = E(x_1 \nu) + E\left(\nu^2\right) = E\left(\nu^2\right) \neq 0.$$

If x_1 correlates with ν, then it also correlates with the composite disturbance term μ. As a result, using the observed x_1^* to proxy x_1 leads to an endogeneity problem. This is the well-known drawback of proxy variables, commonly used in management research (Ketchen et al., 2012). An obvious remedy is to give up the use of proxies and to model measurement error explicitly by modeling common factors, an approach which has been frequently used in OM research since the early 1990s. Modeling the regressors as common factors would of course require multiple indicators to measure each regressor, in order to be able to partition them into measurement error and true, construct-driven variance.

In sum, measurement error in the regressors is more or less guaranteed to result in an endogeneity problem. But again, what is relevant is the magnitude of the problem, which is determined largely by the amount of measurement error, in this example, $var(\nu)$. In examining whether measurement error could be causing problems, researchers should always examine the reliability of the measures used for the regressors. If the regressors are empirically salient and can be measured with high precision, they can be considered *plausibly exogenous* from a measurement error point of view.

#3 – Endogeneity is a judgment call

We openly acknowledge that we the authors of this manuscript, too, disagreed on whether the x's would correlate with ε_2 (we agreed on the two other endogeneity concerns). We finally arrived at a compromise position of *plausible exogeneity* of the x's with respect to ε_2, that is, $cov(x_i, \varepsilon_2) \cong 0$. Although the correlation between the x's and ε_2 is definitely non-zero, it constitutes in our view a lesser concern for endogeneity than the correlation between ε_1 and ε_2 or the correlation between the x's and ε_1. In the context of Fig. 1, the notion of plausible exogeneity is closely connected with the idea that it is unrealistic to expect any full mediation hypothesis to be exactly true (MacKinnon, 2008: 69). This acknowledgment should demonstrate that endogeneity is always a judgment call, and that research teams should engage in discussion and debate on its potential causes and severity. The question of how serious an endogeneity problem is also depends fundamentally on how accurately the researcher wishes to estimate model parameters.

#4 — Two-stage least squares, Sargan-Hansen J-test, and Hausman test

The most popular statistical solution to the endogeneity problem is *two-stage least squares (2SLS)* regression, which involves applying a set of instrumental variables to the troublesome regressors before conducting the regression of theoretical interest (although in reality, current statistical software packages execute both stages in one step). If the conditions for instrumental variables are satisfied, this approach will allow for consistent estimation of the parameters of interest, despite endogeneity due to omitted variables or measurement error.

Let us consider the model in Fig. 3, where x_1- x_3 can be thought of as instruments of y_1. In the first stage, one would regress y_1 on x_1- x_3 to obtain $\widehat{y}_1$, a predicted value of y_1. A useful predicted value obviously hinges on the covariance of y_1 and x_1- x_3 being non-zero, formally known as *the relevance condition*. In the second stage, y_2 is regressed not on y_1 but the expected value $\widehat{y}_1$ obtained in the first stage. The second-stage regression yields a consistent estimate of β_4 as long as the instruments x_1-x_3 are uncorrelated with ε_2, which is formally known as *the exclusion condition*. Seeing how the exclusion condition is critical is straightforward: the utility of the approach relies on the assumption that the variance in $\widehat{y}_1$ is uncorrelated with the disturbance term ε_2. But because $\widehat{y}_1$ is a linear combination of x_1-x_3, it is uncorrelated with the disturbance term only if x_1-x_3 satisfy the exclusion condition. Note therefore that in applying an instrumental variable, the OLS assumption that the regressor does not correlate with the disturbance term is simply replaced by the IV assumption that the instruments do not correlate with the disturbance term.

If there are more instruments than troublesome (endogenous) regressors (as in Fig. 3), a statistical test known as the *Sargan-Hansen J-test* is available to examine whether the instrumental variables are uncorrelated with the disturbance term. This involves an examination of the correlation between the second-stage residuals and the instruments. Under the condition of no association between the instruments and the second-stage residuals, the *J*-statistic is asymptotically distributed as a chi-square variable with p - k degrees of freedom, where p is the number of instruments and k the number of endogenous regressors. Returning to the model in Fig. 3, the *J*-test would thus involve the examination of the correlation between x_1-x_3 and $\widehat{\varepsilon}_2$. The *J*-test in this case has two degrees of freedom ($p = 3$, $k = 1$). One could run the test by first estimating the model in Fig. 3, saving the values of $\widehat{\varepsilon}_2$ and then regressing them on x_1-x_3 . The coefficients in this regression should not be significant. We should note that the number of degrees of freedom in the *J*-test equals the degrees of freedom of the structural equation model in Fig. 3 (see Antonakis et al., 2014).

But again, in applying the *J*-test, one must make an untestable assumption: at least one of the instruments satisfies the exclusion restriction. A further problem is that the overidentifying restrictions that allow the *J*-test to be performed can still be statistically correct even if the instruments and the disturbances in the explanatory equation are actually correlated (see Parente and Santos Silva, 2012, for details). For these reasons, the test is in practice problematic, and Kennedy (2008: 144) called it not a test but “a test of sorts.” The technical details of the Sargan test are discussed in Sargan (1958) as well as the works cited in the beginning of this appendix.

If one has instrumental variables available, one can check the need for applying an instrument by using the *Hausman test* (a.k.a. Durbin-Wu-Hausman test), which tests whether the OLS and 2SLS estimates indeed differ from one another. If they do differ, the plausible reason is the bias in the OLS estimate due to endogeneity, which is of course why the instrument would be needed in the first place. Returning to Figs. 1 and 3, the Hausman test is testing whether the values of β_4 obtained in the two models are the same. The estimates of β_4 obtained from estimating models in Figs. 1 and 3 correspond, respectively, to the OLS and the 2SLS estimates. We conjecture that most OM researchers would, however, estimate the models in Figs. 1 and 3 using the full-information maximum-likelihood estimator available in most SEM software packages.

It might be useful to note that there are other, equivalent implementations of the Hausman test, which OM researchers might end up using without realizing that they are actually conducting Hausman tests. For instance, in many SEM applications researchers add correlations between disturbance terms to examine whether model fit is significantly improved; this approach is justified, because one is rarely able to account for all possible sources of covariation between two dependent variables. With respect to Figs. 1 and 3, a one degree of freedom, chi-square difference test between two models—one allowing a correlation between ε_1 and ε_2 and the other not—is in fact comparable to the Hausman endogeneity test (Antonakis et al., 2010). Because many OM researchers are likely more familiar with SEM instead of econometrics, making the connection may help OM researchers to better understand how endogeneity problems and their assessment relate to more conventional model evaluation procedures.

Another important connection between an SEM and econometrics approach can be made by comparing the fixed-effects (FE) and random-effects (RE) models in Figs. 4 and 6. If we remove the time-invariant z_1 from the RE model, the RE model becomes nested within the FE model. This enables a chi-square difference test to compare the two models, and a significant value on the chi-square means the less restricted FE model is more suitable (Bollen and Brand, 2010). This is effectively the Hausman test for testing FE versus RE.

A benefit of taking an SEM approach is that the graphical representation of the models helps us see the assumptions on which the statistical tests are erected. For instance, in order for the Hausman endogeneity test (i.e., comparing models in Figs. 1 and 3) to be trustworthy, the instruments must satisfy the untestable exclusion condition. If the model in Fig. 3 is misspecified, the Hausman test cannot be trusted. In order for the Hausman test of FE versus RE to be trustworthy (Figs. 4 and 6), in turn, the regressor x must satisfy the exclusion condition and the disturbance terms, conditional on x and the unobserved effect μ, must be uncorrelated (Wooldridge, 2002: 266). This is yet another important reminder that every inferential test is based on untestable assumptions. The actual test and its technical details are discussed in Hausman (1978) as well as the works cited in the beginning of this appendix.

#5 — Variance over time vs. change over time

Figs. 4 and 6 are instructive in showing how exactly panel-data models—FE or RE—relate to time. To say these models “explain variance over time” is easily misinterpreted. What models in Figs. 4 and 6 explain is cross-sectional variance in y at four different time points—nothing is actually being examined *over time* per se. To clarify, if we indeed believe the RE model in Fig. 6 is the correct model, in order to obtain a consistent estimate of β_1, all we need to measure is x_1, z_1, and y_1. In other words, in order to consistently estimate the parameters of the conventional RE panel regression model, one does not actually need panel data. The RE can simply be thought of as being part of the composite disturbance term $\varepsilon_1 + \mu$, and as long as this disturbance does not correlate with the regressors, the estimate of β_1 will be consistent (e.g., Wooldridge, 2002: 492). Of course, having multiple waves of measurement improves the estimate of β_1, but this is an additional benefit, not a necessary condition. This should establish that the parameters estimated in panel data regression are neither inherently longitudinal nor do they have anything to do with what happens over time.

In stark contrast, a latent change model requires, by definition, multiple waves of data, because the focus is specifically on what happens *over time*. If we had just one wave of data, the model in Fig. 5 could not be estimated. With two waves of data, one could estimate both the intercept and the linear change component (although no degrees of freedom would be left to examine whether a linear change trajectory fits the data). Multiple waves of data would improve the estimate of both the intercept and the change trajectory, and would also enable testing of whether change is linear, and modeling non-linear change trajectories. We see a lot of potential for latent change models in OM research, where their application is currently non-existent.

#6 — Fixed versus random effects

The distinction of FE versus RE is an endless source of confusion. Gelman (2005) counted five different, incommensurate ways of making the distinction, noting that problems arise from thinking that one distinction implies or subsumes another. In this paper, the distinction in the context of panel data modeling is conveyed by Figs. 4 and 6. Both effects can be thought of as latent variables, so both are random effects, statistically speaking (Wooldridge, 2002: 496). The two are different in that the FE can correlate with the regressors, but the RE cannot. The two are similar in that neither can correlate with the disturbance term. This distinction is consistent with the treatment of panel models most econometric texts.

In latent change models, such as that shown in Fig. 5, the distinction is markedly different: the individual-level intercepts (β_0) and slopes (β_1) are sometimes called the random effects and the overall mean intercepts and slopes the fixed effects (Curran and Hussong, 2002: 69). The random intercepts and random slopes are thus simply individual-level deviations from the overall means. In latent change models, the question is not whether effects are modeled as random or fixed, they are simultaneously modeled as *both*. In general, framing longitudinal model specification as being a forced choice between FE and RE is strange (e.g., Mundlak, 1978). In most OM modeling situations, two kinds of effects are relevant: (1) the effect observed over time at the level of the specific observational unit (this is the *within effect*, WE); and (2) the effect observed cross-sectionally across observational units (this is the *between effect*, BE). Given the interest in these effects, the FE versus RE formulation is awkward precisely because FE corresponds to WE, but RE corresponds to *the average* of WE and BE. In what modeling situation is the selection between examining either WE or the average of WE and BE pertinent (e.g., Certo et al., 2016)? More fundamentally, what is the substantive relevance of the average of BE and WE in the first place? What if BE is of interest to the researcher? Neither FE nor RE will be desirable, unless of course WE equals BE (which is what the Hausman test is effectively testing). But it is easy to see how BE and WE could both be relevant but their values unequal in the population. Established procedures exist to incorporate both BE and WE explicitly in a *hybrid model*, which avoids the contrived choice of FE or RE (Bell and Jones, 2015; Mundlak, 1978). Again, STATA procedures are readily available (Schunk, 2013).

References

Adler, P.S., Goldoftas, B., Levine, D.I., 1999. Flexibility versus efficiency? A case study of model changeovers in the Toyota production system. Organ. Sci. 10, 43–68.
Antonakis, J., Bendahan, S., Jacquart, P., Lalive, R., 2010. On making causal claims: a review and recommendations. Leadersh. Q. 21, 1086–1120.
Antonakis, J., Bendahan, S., Jacquart, P., Lalive, R., 2014. Causality and endogeneity: problems and solutions. In: Day, D.V. (Ed.), The Oxford Handbook of Leadership and Organizations. Oxford University Press, New York, pp. 93–117.
Ashley, R.A., Parmeter, C.F., 2015. When is it justifiable to ignore explanatory variable endogeneity in a regression model? Econ. Lett. 137, 70–74.
Baltagi, B.H., 2005. Econometric Analysis of Panel Data, third ed. Wiley, New York.
Baltagi, B.H., 2013. Econometric Analysis of Panel Data, fifth ed. Wiley, New York.
Bell, A., Jones, K., 2015. Explaining fixed effects: random effects modeling of time-series cross-sectional and panel data. Political Sci. Res. Methods 3, 133–153.
Bendoly, E., van Wezel, W., Bachrach, D.G. (Eds.), 2015. The Handbook of Behavioral Operations Management: Social and Psychological Dynamics in Production and Service Settings. Oxford University Press, New York.
Bettis, R., Gambardella, A., Helfat, C., Mitchell, W., 2014. Quantitative empirical analysis in strategic management. Strat. Manag. J. 35, 949–953.
Blalock Jr., H.M., 1991. Are there really any constructive alternatives to causal modeling? Sociol. Methodol. 21, 325–335.
Bollen, K.A., 2000. Modeling strategies: in search of the holy Grail. Struct. Equ. Model. 7, 74–81.
Bollen, K.A., 2012. Instrumental variables in sociology and the social sciences. Annu. Rev. Sociol. 38, 37–72.
Bollen, K.A., Brand, J.E., 2010. A general panel model with random or fixed effects: a structural equations approach. Soc. Forces 89, 1–34.
Bollen, K.A., Curran, P.J., 2006. Latent Curve Models: a Structural Equation Perspective. Wiley, New York.
Bolton, G.E., Ockenfels, A., Thonemann, U.W., 2012. Managers and students as newsvendors. Manag. Sci. 58, 2225–2233.
Bromiley, P., Rau, D., 2016. Operations management and the resource based view: another view. J. Oper. Manag. 41, 95–106.
Bun, M.J.G., Windmeijer, F., 2011. A comparison of bias approximations for the two-stage least squares (2SLS) estimator. Econ. Lett. 113, 76–79.
Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: Methods and Applications. Cambridge University Press, New York.
Certo, S.T., Withers, M.C., Semadeni, M., 2016. A tale of two effects: using longitudinal data to compare within- and between-firm effects. Strat. Manag. J. (in press).
Clougherty, J.A., Duso, T., Muck, J., 2016. Correcting for self-selection based endogeneity in management research: review, recommendations, and simulations. Organ. Res. Methods 19, 286–347.
Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. Rev. Econ. Stat. 94, 260–272.
Cook, T.D., Campbell, D.T., 1979. Quasi-experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin Company, Boston.
Curran, P.J., Hussong, A.M., 2002. Structural equation modeling of repeated measures data: latent curve analysis. In: Moskowitz, D.S., Hershberger, S.L. (Eds.), Modeling Intraindividual Variability with Repeated Measures Data: Methods and Applications. Lawrence Erlbaum Associates, Mahwah, N.J., pp. 59–85
Davis, J.A., 1985. The Logic of Causal Order. Sage Publications, Beverly Hills, CA.
Duncan, O.D., 1972. Unmeasured variables in linear panel models. Sociol. Methodol. 4, 36–82.
Dyer, J.H., 1996. Specialized supplier networks as a source of competitive advantage: evidence from the auto industry. Strat. Manag. J. 17, 271–291.
Freedman, D.A., 1991. Statistical models and shoe leather. Sociol. Methodol. 21, 291–313.
Gelman, A., 2005. Analysis of variance—Why it is more important than ever. Ann. Stat. 33, 1–33.
Greene, W.H., 2011. Econometric Analysis, seventh ed. Prentice Hall, Upper Saddle River, N.J.
Guide, V.D.R., Ketokivi, M., 2015. Notes from the editors: redefining some methodological criteria for the journal. J. Oper. Manag. 37, v–viii.
Hamilton, B.H., Nickerson, J.A., 2003. Correcting for endogeneity in strategic management research. Strateg. Organ. 1, 51–78.
Hausman, J.A., 1978. Specification tests in econometrics. Econometrica 46, 1251–1271.
Hitt, M.A., Xu, K., Carnes, C.M., 2016. Resource based theory in operations management research. J. Oper. Manag. 41, 77–94.
Hossain, T., List, J.A., 2012. The behavioralist visits the factory: increasing productivity using simple framing manipulations. Manag. Sci. 58, 2151–2167.
Imai, K., Keele, L., Tingley, D., 2010. A general approach to causal mediation analysis. Psychol. Methods 15, 309–334.
Imai, K., Yamamoto, T., 2013. Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. Polit. Anal. 21, 141–171.
Jeong, J., Yoon, B.H., 2010. The effect of pseudo-exogenous instrumental variables on Hausman Test. Commun. Stat. - Simul. Comput. 39, 315–321.
Jokisaari, M., Nurmi, J.-E., 2009. Change in newcomers' supervisor support and socialization outcomes after organizational entry. Acad. Manag. J. 52, 527–544.
Keele, L., Tingley, D., Yamamoto, T., 2015. Identifying mechanisms behind policy interventions via causal mediation analysis. J. Policy Anal. Manag. 34, 937–963.

Kennedy, P., 2008. A Guide to Econometrics, sixth ed. Blackwell Publishing, Malden, MA.

Ketchen, D.J., Ireland, R.D., Baker, L.T., 2012. The use of archival proxies in strategic management studies: castles made of sand? Organ. Res. Methods 16, 32–42.

Kiviet, J.F., 2016. When is it really justifiable to ignore explanatory variable endogeneity in a regression model? Econ. Lett. 145, 192–195.

Kontopantelis, E., Doran, T., Springate, D.A., Buchan, I., Reeves, D., 2015. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 350, h2750.

Kraay, A., 2012. Instrumental variables regressions with uncertain exclusion restrictions: a Bayesian approach. J. Appl. Econ. 27, 108–128.

Krause, M.S., 2013. The incompatibility of achieving a fully specified linear model with assuming that residual dependent-variable variance is random. Qual. Quant. 47, 3201–3204.

Lance, C.E., Butts, M.M., Michels, L.C., 2006. The sources of four commonly reported cutoff criteria: what did they really say? Organ. Res. Methods 9, 202–220.

Lance, C.E., Vandenberg, R.J. (Eds.), 2008. Statistical and Methodological Myths and Urban Legends: Received Doctrine, Verity, and Fable in the Organizational and Social Sciences. Routledge, New York.

List, J.A., 2011. Does market experience eliminate market anomalies? The case of exogenous market experience. Am. Econ. Rev. 101, 313–317.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

MacCallum, R.C., 1998. Commentary on quantitative methods in I/O research. Indus.-Organ. Psychol. 35 online article, no page numbers given.

MacDuffie, J.P., 1995. Human resource bundles and manufacturing performance: organizational logic and flexible production systems in the world auto industry. Indus. Labor Relat. Rev. 48, 197–221.

MacKinnon, D.P., 2008. Introduction to Statistical Mediation Analysis. Erlbaum, Mahwah, N.J.

Maner, J.K., 2014. Let's put our money where our mouth is: if authors are to change their ways, reviewers (and editors) must change with them. Perspect. Psychol. Sci. 9, 343–351.

McArdle, J.J., 2009. Latent variable modeling of differences and changes with longitudinal data. Annu. Rev. Psychol. 60, 577–605.

Merton, R.K., 1968. The Matthew effect in science. Science 159, 56–63.

Monteverde, K., Teece, D.J., 1982. Supplier switching costs and vertical integration in the automobile industry. Bell J. Econ. 13, 206–213.

Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, Cambridge.

Mundlak, Y., 1978. On the pooling of time series and cross section data. Econometrica 46, 69–85.

Murray, M.P., 2006. Avoiding invalid instruments and coping with weak instruments. J. Econ. Perspect. 20, 111–132.

Muthén, B.O., Asparouhov, T., 2015. Causal effects in mediation modeling: an introduction with applications to latent variables. Struct. Equ. Model. 22, 12–23.

Parente, P.M.D.C., Santos Silva, J.M.C., 2012. A cautionary note on tests of overidentifying restrictions. Econ. Lett. 115, 314–317.

Pearl, J., 2009. Causality: Models, Reasoning and Inference, second ed. Cambridge University Press, Cambridge, MA.

Preacher, K.J., Wichman, A.L., MacCallum, R.C., Briggs, N.E., 2008. Latent Growth Curve Modeling. Sage, Thousand Oaks, CA.

Reeb, D., Sakakibara, M., Mahmood, I.P., 2012. From the Editors: endogeneity in international business research. J. Int. Bus. Stud. 43, 211–218.

Roberts, M.R., Whited, T.M., 2013. Endogeneity in empirical corporate finance. In: Constantinides, G.M., Harris, M., Stulz, R.M. (Eds.), Handbook of the Economics of Finance, 2A. Elsevier, Amsterdam, pp. 493–572.

Rumelt, R.P., Schendel, D., Teece, D.J., 1991. Strategic management and economics. Strat. Manag. J. 12, 5–29.

Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. Econometrica 26, 393–415.

Schroeder, R.G., Flynn, B.B. (Eds.), 2001. High Performance Manufacturing: Global Perspectives. Wiley, New York.

Schunk, R., 2013. Within and between estimates in random-effects models: advantages and drawbacks of correlated random effects and hybrid models. Stata J. 13, 65–76.

Semadeni, M., Withers, M.C., Certo, S.T., 2014. The perils of endogeneity and instrumental variables in strategy research: understanding through simulations. Strat. Manag. J. 35, 1070–1079.

Skinner, W., 1969. Manufacturing - missing link in corporate strategy. Harv. Bus. Rev. 50 (3), 136–145.

Skinner, W., 1974. The focused factory. Harv. Bus. Rev. 52 (3), 113–121.

Stock, J.H., Trebbi, F., 2003. Retrospectives: who invented instrumental variable regression? J. Econ. Perspect. 17, 177–194.

Tan, T.F., Staats, B.R., 2016. Behavioral drivers of routing decisions: evidence from restaurant table assignment. In: Proceedings of the 2016 Manufacturing and Service Operations Management Conference (M&SOM), Auckland, New Zealand.

Tchetgen Tchetgen, E.J., Shpitser, I., 2012. Semiparametric theory for causal mediation analysis: efficiency bounds, Multiple robustness, and sensitivity analysis. Ann. Stat. 40, 1816–1845.

VanderWeele, T.J., 2010. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology 21, 540–551.

VanderWeele, T.J., 2015. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press, New York.

VanderWeele, T.J., Chiba, Y., 2014. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. Epidemiol. Biostat. Publis Health 11, e9027.

Whetten, D.A., 1989. What constitutes a theoretical contribution? Acad. Manag. Rev. 14, 490–495.

White, H., Chalak, K., 2010. Testing a conditional form of exogeneity. Econ. Lett. 109, 88–90.

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

Wooldridge, J.M., 2016. Introductory Econometrics: a Modern Approach, sixth ed. Cengage Learning, Boston, MA.

Ziliak, S., McCloskey, D., 2008. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor, MI.