Journal of Supply Chain Management
2021, 57(3), 130–146

EDITORIAL COMMENTARY: ADDRESSING CONFUSION IN THE DIFFUSION OF ARCHIVAL DATA RESEARCH

JASON MILLER
Michigan State University

BETH DAVIS-SRAMEK
Auburn University

BRIAN S. FUGATE
University of Arkansas

MARK PAGELL
UCD Michael Smurfit School of Business

BARBARA B. FLYNN
Indiana University

Supply chain management researchers are increasingly using archival data to push boundaries of existing knowledge. Archival data provide opportunities to address new research questions and offer fresh perspectives on old questions. This editorial seeks to establish a common ground regarding research design, measurement validity, and endogeneity to help both authors and reviewers fully utilize archival data to advance supply chain management knowledge.

Keywords: *archival research; secondary data; research design; measurement validity*

INTRODUCTION

The increased use of archival data represents one of the most pronounced changes taking place in supply chain management (SCM) research (Rabinovich & Cheon, 2011). It seems to be driven by two underlying forces. First, researchers increasingly appreciate how archival data allows them to address previously uninvestigated questions (Calantone & Vickery, 2010). Second, the availability of archival data and tools to examining such data has proliferated, offering unique perspectives on old and new research questions. This evolution represents a strengthening and broadening of the discipline, and we hope to see more rigorous research using archival data. However, along with increased utilization of archival data, we note that its diffusion has been accompanied by some confusion that potentially stymies scientific progress. As such, it is important to proactively address some of the problems that can undermine knowledge development during the diffusion process.

This editorial addresses a handful of pertinent issues regarding the use of archival data for SCM research. This work does not provide a "how-to" guide for archival research in the way Fawcett et al. (2014) do with conceptual, qualitative, and survey-based work. Rather, we aim to establish common ground for using archival data, specifically regarding (1) research design, (2) measurement validity, and (3) endogeneity. As a disclaimer, we remind readers that our thoughts are not the only perspective on these topics. Thus, this work should not be interpreted as the definitive statement on these issues, and we welcome critical feedback and further debate.

The authors would like to thank John Gray for his helpful feedback during the development of this article.

RESEARCH DESIGN CONSIDERATIONS WITH ARCHIVAL DATA

Archival data pose research design considerations that require careful thought. The specific issues addressed include sampling frame, data preparation, and data age/temporality concerns. We also preface the discussion by pointing out that the next three sections do not represent the full set of research design issues for archival data, but rather are the ones that pose the most unique challenges relative to using primary data sources.

Sampling Frame

One key concern with the use of archival data is that the sample in question may not be random. While this editorial assumes the unit of the analysis is the firm for simplicity, these arguments equally apply to other units of analysis (e.g., individuals, teams, networks, industries). For example, archival data on firms' financial performance from the Compustat database are limited to publicly traded entities in the U.S. Similarly, Bloomberg scrapes data from various sources to serve as input into the SPLC database on inter-firm relationships (see Elking et al., (2017) for a discussion), which creates the likelihood that some relationships are missed. Similarly, authors may elect to focus on a segment of the overall population, such as the largest firms. In other cases, authors may be limited to obtaining data from a small number of firms because firms are reluctant to share sensitive performance data and/or provide data in a comparable format (Bartel et al., 2007; Ichniowski & Shaw, 1999; Ichniowski, Shaw & Prennushi, 1997). Other times, scholars may only be able to collect data from only one firm (Bernstein, 2012; Scott, 2018; Wan, 2016; Winter et al., 2012) or a network of firms connected to a single firm (Distelhorst, Hainmueller & Locke, 2017).

To address this concern, we offer two considerations. First, measured constructs can be defined within the specific sample being utilized (Rossiter, 2011). For example, if researchers are examining issues pertaining to corporate social responsibility (CSR) using the Kinder, Lindenberg, and Domini (KLD) database (e.g., Chatterji, Levine & Toffel, 2009; Kang, Germann & Grewal, 2016), the construct can be framed as CSR representing the sample of firms (i.e., large, mostly publicly traded firms) that are rated using KLD's methodology. By identifying characteristics of the firms in the dataset, researchers put themselves in a position to make better generalizations about their work. Using the object (e.g., firm), attribute (e.g., CSR), rater (e.g., KLD) tripartite combination (Rossiter, 2002) is also valuable because it helps clarify how scholars can extend existing work. For example, do results hold using a different approach for rating CSR (e.g., Asset4, FTSE4Good) since these approaches adopt different methodologies (Chatterji, Durand, Levine & Touboul, 2016)?

This brings us to our second consideration, which involves logically defending the scope to which results are generalized (Kane, 2013). For example, publicly traded firms represent a small proportion of the overall population of firms in an industry, and they are generally far larger on average than private entities (Ali, Klasa, & Yeung, 2008; Kull, Kotlar & Spring, 2018). As such, a strong rationale is needed before generalizing findings from large publicly traded firms to smaller, private businesses. The same logic would apply if, for instance, there was a desire to generalize contract manufacturers in one area of the world within one industry sector (e.g., Distelhorst et al., 2017) to a broader set of manufacturing firms.

To make a broader generalization, it would be important to effectively argue that the theoretical mechanisms that bring about the observed relationships in the current sample also likely hold outside the sampling frame. As an example, a recent study by Badorf et al. (2019) used RBV to examine how economies of scale influence the supplier selection decision, contingent on several moderating variables that serve as boundary conditions. Although they used a secondary database exclusive to the automotive industry, they cautiously note the same results could be generalizable to other industries. Given that they drew upon a larger body of research that explained the theoretical mechanisms underlying the observed relationships, this cautious generalization appears reasonable. We note, however, that support for generalization is more likely *not* to exist, in which case researchers can focus more on theoretical contextualization inherent in middle-range theorizing (Stank et al., 2017).

Data Preparation & Reporting

Another research design issue when using archival data that often does not get adequate attention involves data preparation and reporting. It is important to communicate data preparation steps such as addressing missing data, identifying potential errant observations (e.g., nonsensical values), winsorizing extreme values, and other "upfront" issues that should be reconciled prior to further data analysis (see Aguinis, Hill and Bailey, (2019) for a review of best practices). The literature provides many approaches to handle missing data (Enders, 2010) and includes subjective expertise that can be incorporated into analyses (Cudeck & Codd, 2012). As such, our goal is not to suggest a specific set of methodological steps that should be utilized. Rather, the intent is to stress the significance of disclosure about the data and the steps taken to prepare the data for further analysis.

Missing data is a frequent issue that arises in the use of archival data, so it is important to convey the approach for handling this issue. Likewise, it is valuable to explain other means by which data are manipulated. For instance, if researchers choose to winsorize the data, then disclosure about the percentile values would be needed. In some cases, respondents may be removed from the analysis, so it is important to disclose the accompanying logic for their removal. In other cases, nonlinear transformations can be applied to measures in order to reduce the skew of some variables, which should be clearly explained. As a final scenario (and we note that this is not an exhaustive list of examples), rationale should be provided if the number of categories for a continuous measure is reduced by creating ranked scores (e.g., deciles) or through dichotomization. As before, the theme we wish to convey is the importance of disclosing and justifying these "upfront" decisions prior to data analysis.

Age and Temporality

Because of the rapid changes occurring in the supply chain field, the discipline tends to insist on the use of data that has been "recently" collected. In contrast, other business disciplines routinely publish articles using data that may be decades or even centuries old (e.g., Madsen & Walker, 2017; Silverman & Ingram, 2017). This naturally raises the question about the appropriateness of utilizing "aged" archival data.

In addressing this issue, we offer that a state versus trait distinction (Kenny & Zautra, 2001) provides sound logic for when it is (or not) appropriate to utilize "dated" archival data. The distinction concerns the processes (i.e., mechanisms) (Astbury & Leeuw, 2010; Mahoney, 2001) that are theorized to bring about observed relations. Traits are stable, relatively constant characteristics of the units of analysis, whereas states are less-stable, time-varying characteristics of the units of analysis that depend on multiple spatial–temporal factors (Newsom, 2015). Variables with trait-like properties include local demographical composition (Goodstein, 1994) and organizational composition (Casile & Davis-Blake, 2002), which remain relatively invariant over time. Variables with state-like properties include workers' degree of fatigue (Jin & Lee, 2014) and plants' experience with producing a given product (Levitt et al., 2013).

In addressing the question about data age, there is less concern with using older data when the mechanisms postulated to bring about the empirical relationships are traits that are likely to hold in the future (Astbury & Leeuw, 2010; Hedström & Ylikoski, 2010). For example, Braguinsky et al. (2015) use secondary data from Japanese cotton spinning plants during the 1899–1920 timeframe to shed more light on the implicit assumption that productivity improves after acquisitions due to greater managerial competence. This particular dataset offered a unique level of detail (e.g., providing data about labor and capital "flows" as opposed to "stocks") that facilitated a more nuanced understanding of how acquisitions affected productivity than can typically be found with other more contemporary datasets (e.g., Schoar, 2002).

Merging distinct datasets can lead to questions when time periods of measurement may not perfectly overlap. This issue is less concerning when the data have trait-like properties; that is, the measure(s) of the independent variables are not likely to change over time. A related issue when merging datasets involves a dependent variable temporally preceding the independent variable (e.g., Edelman, 1990; Goodstein, 1994). As an example, there could be rich demographic data insight from the upcoming 2020 U.S. Census. If a study used this data to predict, for instance, hospitals' quality scores for 2019, this is not concerning because local demographics are unlikely to change in 1 year. Further, there is a strong argument that the direction of influence flows from the independent variable to the dependent variable, as it does not seem theoretically plausible that hospital quality shapes local demographics. To mitigate the concern, however, it is important for researchers to theoretically justify such decisions.

MEASUREMENT VALIDITY CONSIDERATIONS WITH ARCHIVAL DATA

SCM scholarship has advocated that different types of measurement validity should be addressed, including face validity, content validity, convergent validity, discriminant validity, criterion validity, and construct validity (Bollen, 1989; Chen & Paulraj, 2004; Mentzer & Flint, 1997). These validity assessments determine whether we can have confidence that a construct is measuring what the researcher intends for it to capture. They incorporate issues beyond how relevant constructs are operationalized by also addressing issues concerning research design, sampling approach, and statistical methodology (Cook & Campbell, 1979; Shadish et al., 2002).

While guilty of it ourselves, the SCM discipline has embraced a "checklist" mentality for validity, whereby researchers run through a list of statistical tests, and, assuming all rules of thumb hold (e.g., strong factor loadings for convergent validity), conclude the measures are valid (Garver, 2019). Our discipline is not alone in this, as the same criticism has also been leveled in marketing (Rigdon et al., 2011; Rossiter, 2011). More problematically, viewing validity as having multiple "types" inadvertently encourages cherry-picking—discussing and emphasizing each facet independently without offering a holistic assessment (Kane, 2013). As the discipline increases its utilization of archival data, more complex validity issues arise that may not be adequately addressed by using the "checklist" approach.

In other fields where archival data research use has matured over time, such as education and psychology, the checklist approach to validity has been supplanted by a more holistic and unitary conceptualization (Kane, 2016; Messick, 1995). Messick (1993, p. 1) defines validity as "an *integrated evaluative judgment* of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment," [emphasis added]. Broadly, it concerns whether researchers are justified to (1) interpret a measured variable or set of measured variables as representing a broader theoretical construct and (2) use these measured variable(s) for a given purpose. More specifically, this concept of validity shifts the focus from statistical rules of thumb (most of which have little justification; Calantone et al., 2017; MacCallum et al., 1999) to validity claims based on logical arguments that are *accompanied by* statistical evidence (Kane, 1992, 2013).

We are not making an argument to move away from the more traditional view of assessing validity (Cook & Campbell, 1979; Shadish et al., 2002). However, a more holistic conceptualization of validity is a complementary approach that is especially helpful with archival data research designs because it moves assessment away from a validity checklist based exclusively on statistical evidence. Researchers using this conceptualization have identified five facets requiring examination to draw an overall validity conclusion (Cook & Beckman, 2006): (1) response process, (2) content evidence, (3) internal structural, (4) relations to other variables, and (5) consequences. Table 1 summarizes these five facets and provides one or more examples of each in relation to establishing the use for archival data. It underscores significant implications for each facet, and the following section highlights several measurement concerns as they apply to the use of archival data in SCM research.

Response Process

Perhaps the most critical facet of validity to consider in utilizing archival data is the response process. In short, it relates to how the data was generated, and it highlights an issue that cannot be ignored when SCM researchers utilize archival data. While not explicitly stated, the response process facet of validity often comes up in the use of survey data. For instance, the use of single-informant, cross-sectional survey data creates the concern that informants may provide inaccurate responses for various reasons (e.g., just marking "7"). As such, results from studies that utilize this kind of data are highly scrutinized (Flynn et al., 2018). We note, however, that response process concerns are not automatically alleviated with the use of archival data. In other words, archival data are not necessarily more "valid" than survey data, nor does it alleviate the possibility of inaccurate responses.

Researchers should note that archival data are oftentimes self-reported and subject to limited audit (e.g., Forbes et al., 2015). In this case, response process concerns about archival data are very similar to those about survey data. For example, companies that self-report their safety performance may fail to report all accidents. As another example, companies that self-report their environmental footprint may not be reporting all their activities. Self-reporting situations such as these are more likely to be the case when firms are not subject to external audits, or when there are not significant repercussions for erroneous reporting. This highlights the significance of the response process because it brings into question how the data were generated, and as a result, the veracity of the responses.

In contrast, there are other situations when researchers can have more confidence in the accuracy of self-reported data. For instance, publicly traded companies must self-report their financials, but these companies are subject to external audit with severe repercussions (e.g., Enron) for misreporting. Thus, the process used to generate company financial data provides greater confidence that the self-reported financial ratios are representative of firms' financial performance. Similarly, primary data may be collected from firms through surveys conducted by government agencies that mandate responses. Examples include the following: (1) European Union law that requires large companies to disclose certain information on the way they operate and manage social and environmental challenges (European Commission, 2019), (2) Australian Bureau of Statistics mandatory survey requiring businesses to report a range of information (Australian Bureau of Statistics, 2019), (3) Brazilian Institute of Geography and Statistics mandatory Consumer Expenditure Survey on the population and economy, and (4) Economic Census conducted by the U.S. Census Bureau every five years requiring companies to disclose business information. While such data are self-reported, concerns about the accuracy of the responses are reduced because individuals who provide knowingly false information are subject to substantial fines and/or other penalties (Ali, Klasa, & Yeung, 2008). In sum, understanding and disclosing the extent to which companies are penalized for misreporting or providing inaccurate responses is critical when utilizing self-reported archival data.

TABLE 1

Facets of Validity with Examples

Facet	Definition	Examples
Content evidence	The extent that measures are relevant to the theoretical construct's domain and adequately capture the constructs' scope (Messick, 1995, p. 745).	• Measures from archival data should fall within a construct's domain and should represent the different aspects of a construct to the highest degree possible. • Measures from outside the construct's sampling domain should not be included when measuring the construct.
Response process	The extent the process through which scores are generated helps ensure that scores have veracity (Cook & Beckman, 2006).	• If archival data are self-reported and subjected to limited external auditing, there is the concern that the reported scores may be under- or over-reported. • Individuals reporting scores utilize a different process then intended by those who developed the methodology to generate the data.
Internal structure	The extent scores for multiple measures of the same construct display acceptable internal consistency and do not suffer from issues pertaining to differential item functioning across respondents (Cook & Beckman, 2006).	• Scores across a series of measures utilized to represent a continuous reflective latent variable should display an acceptable degree of internal consistency (e.g., coefficient omega) and have factor loadings in the hypothesized direction. • If factor loadings differ across groups of respondents for whatever reason (e.g., an indicator operates differently within some industries relative to others), it is not possible to say the same construct is measured.
Relations to other variables	The extent scores for a construct display theoretically anticipated relations with other constructs (Downing, 2003).	• If theory predicts that larger firms should have a higher level of environmental sustainability performance, then scores representing firms' environmental sustainability performance should correlate positively with scores representing firm size.
Consequences	The extent to which decision-making based on the measures is justifiable for a certain purpose (Cook & Beckman, 2006).	• If scores are to be used to make decisions that affect respondents (e.g., firms that receive a poor score are publicly singled-out as poor performers), the benefits and costs from such action must be considered. It is possible for scores to be valid for one use (e.g., rank-ordering companies' performance in a domain using the score as representing performance) but not for another (e.g., terminating companies' operating authority if their score falls below a certain threshold).

In other instances, third party "oversight" entities may generate the archival data themselves, which can also relieve response process concerns. For example, the Federal Motor Carrier Safety Administration (FMCSA) generates motor carriers' safety compliance scores by compiling data from different enforcement agencies (Cantor et al., 2016). Similarly, the Health and Safety Executive (HSE) in the U.K. has a public registry of firms that are prosecuted and convicted of safety breaches (Wiengarten et al., 2019). Because of the oversight role of these agencies, researchers can have more confidence that data are accurate.

Our concluding point related to the response process is the acknowledgment that in many cases, archival data started life as primary data collection via surveys. Researchers should consider describing, to the extent possible, how archival data are generated so that response process concerns can be adequately assessed. We add the qualifier "to the extent possible" because there are some cases in which the sources providing archival data may require confidentiality, especially regarding how data are collected. At a minimum, disclosure involves explaining whether the archival data are self-reported, whether it is subject to audit, and if so, how much firms are penalized for inaccurate self-reports. This information alone, however, does not lend to making an adequate assessment of validity. While the response process used to generate secondary data may weaken validity, overall validity judgement is based on the examination and assessment of all the facets of validity (Kane, 1992).

Content Evidence

One frequent issue related to archival data comes up when the dataset lacks measures to represent every aspect of a construct, which raises questions about the content facet of validity (Boyd et al., 1993). For instance, assume a researcher examining buyer–supplier relationships weighs the decision to use data from Bloomberg to measure *firms' dependency on suppliers*. Bloomberg captures the average percentage of a firm's cost of goods sold that each supplier represents. The researcher could look to Elking et al. (2017), who utilized these data to operationalize the construct *firms' financial dependency on suppliers* as the average percentage of a firm's cost of goods sold for the entire set of a firm's suppliers identified by Bloomberg. While this measure aligns with some conceptualizations of dependency (Pfeffer & Salancik, 2003), the buyer–supplier relationship literature also holds that a firm's dependence upon its suppliers is contingent on the ease with which it can replace said suppliers (Heide & John, 1988). Capturing this more perceptual dimension of "ability to switch" would best be addressed through industry concentration data or primary survey data.

As this example demonstrates, archival sources often lack a comprehensive set of measures to fully tap a construct. The question becomes whether it is appropriate to operationalize constructs using archival data that does not fully represent their theoretical underpinnings. We offer that transparency is critical here. For instance, in using the Bloomberg data, the researcher could define buyer dependency, but then explicitly convey the conceptualization is limited to the purchasing concentration dimension of buyer dependency. For further transparency, the construct itself could be labeled "purchasing concentration," as this label would more accurately reflect what is empirically operationalized. It would also be important to note this issue as limitation of the research, given the inability of the data source to operationalize all facets of the construct *firms' dependency on suppliers*. While this may limit the theoretical contribution to some degree, this weakness in content validity should be weighed against the strengths of using this secondary data source. In sum, it is important to address this issue in a holistic manner that weighs both the strengths and weaknesses of a chosen measurement approach against the other facets of validity.

A related content evidence issue that can occur when using archival data involves using scores from only one item to operationalize reflective constructs (Ketchen et al., 2013). This naturally raises the question related to whether it is permissible to use a single item to approximate a construct. One perspective, informed by Little et al. (1999) and Ketchen et al. (2013), is that single items are appropriate when there is a high degree of conceptual overlap between the measure and the construct. Single-item measures are generally deemed acceptable if they tap the center of the construct's domain, or when constructs are objective, unidimensional, and have a clear meaning (Bergkvist & Rossiter, 2007; Rossiter, 2011).

To illustrate, consider firm age and organizational performance. Given that firm age is objective, unidimensional, and has a clear meaning, there is logic to use a single item to capture it. In contrast, organizational performance is complex and multidimensional, consisting of dimensions concerning liquidity, profitability, growth, and stock market performance (Hamann et al., 2013). Consequently, logic would suggest that when using archival data from a source such as Compustat or Financial Analysis Made Easy (FAME), measurement of organizational performance should utilize factor analytic or item response theory techniques to measure each sub-dimensions using multiple indicators to either (1) use the latent variables directly in the analyses or (2) extract factor scores for these sub-dimensions that can be utilized in subsequent analyses (Calantone et al., 2017). In sum, it is important to explain and justify the use of a single indicator to approximate reflective constructs.

Internal Structure

The internal structure facet of validity is relevant when measuring constructs that are theorized as reflective latent variables using multiple indicators. Whether fitting measurement models using factor analytic or item response theory techniques (Wirth & Edwards, 2007), evidence in support of internal structure includes acceptable measurement model fit, theoretically consistent patterns of factor loadings, and acceptable internal consistency (e.g., coefficient omega) (Cook & Beckman, 2006). Because SCM scholars have substantial familiarity with how to execute such statistical approaches, we do not elaborate further on this facet of validity, other than to stress its importance in an overall validity assessment.

Relations to Other Variables

Consistent with the Mentzer and Flint (1997) explanation of nomological validity, a variable's pattern of relations with other variables provides evidence of validity to the extent that the pattern of relationships conforms to existing theoretical expectations (Downing, 2003). One example of using a measure's relations to other variables to provide evidence for validity is Bloom et al.'s (2019) study, which utilizes archival survey data from the U.S. Census Bureau to measure "structured" management practices. The authors provide evidence that the aggregate management practice score captures establishments' use of structured management practices by empirically documenting that establishments with higher management practice scores have higher labor productivity. Establishing this relationship strengthened validity because it aligns with economic theory arguments that firms' use of structured management practices should positively affect productivity (Syverson, 2011). In contrast, failing to find relationships consistent with existing theory indicates the need to look at how constructs have been operationalized.

Consequences

The consequences facet of validity concerns how decision-makers utilize measures to take actions that affect the units of analysis. The critical issue is that a measure may be valid for one use but not another due to these actions having different impacts on the units of analysis. The consequences facet of validity is of central importance in fields such as education and psychology where decisions are made as a result of archival measures (e.g., test scores), that strongly affect individuals (e.g., college admissions) (Cook & Beckman, 2006; Kane, 2013; Messick, 1993). To illustrate a SCM setting where consequence issues play a role in validity claims, consider the Federal Motor Carriers Safety Administration's practice of publicizing safety data for carriers operating in the U.S. Taking the integrative validity approach, these measures offer a strong rationale for shippers, brokers, and insurance companies to use these data as an input into carrier selection decisions. In contrast, more evidence would be needed for regulators to use these measures for terminating a carrier's operating authority. In sum, if measures developed and/or utilized in SCM research become used for decisions that have consequential implications, then this facet of validity will become critical to address.

Other Relevant Validity Issues

In addition to the validity implications in the previous sections, we have observed two other validity considerations relevant to utilizing archival data. The first issue relates to the consistency of interpretations. Validity claims center, in part, on whether one or more measures can be interpreted to represent a construct (Kane, 2013). An issue often overlooked with archival data occurs when a theoretical construct using an observed measure or set of observed measures is used by different research teams to represent different theoretical constructs. For example, previous research has used firms' R&D intensity (i.e., the ratio of R&D expenditures over sales) to represent constructs including R&D investments, managerial discretion, and asset specificity (see Ketchen et al., 2013 for a detailed discussion). In this case, while using R&D intensity to represent R&D investments seems well-founded, there is reason to challenge whether R&D intensity represents managerial discretion or asset specificity.

Firm size, as a proxy for an unobservable theoretical concept or as a control for omitted variables bias, is another variable that has been used to represent a wide range of theoretical constructs, such as firm financial slack, resource availability, power/dependence, bureaucracy, managerial discretion, proprietary costs, political costs, information output, organizational complexity, competitive advantages (Bonaccorsi, 1992; Bujaki & Richardson, 1997; Diamantopoulos et al., 2014; Lapointe-Antunes et al., 2006; Yu et al., 2015). While firm size may be justifiable in each instance for a corresponding theoretical construct, the cumulative effect of the multiple interpretations of firm size results in conflicting predictions and unexplainable relationships. Kimberly (1976, p. 586) noted the following over four decades ago, which indicates little advancement in addressing this issue:

> …the conclusion is that the concept of size as it has been generally used by organizational researchers is too global….A more differentiated approach

> is needed. It might be desirable ultimately to expunge the word size from the lexicon of organizational research and to develop a new vocabulary which captures the variety formerly encompassed under that general rubric. It is perfectly possible that the important variable is not really size at all, and that this has simply been a heading under which researchers lacking any sharper theoretical perspective have lumped many variables together. Even if it is felt that jettisoning the concept would be premature, at the very least a more differentiated approach would lead to the identification of several aspects of the global construct.

Although it has been given little attention in SCM research apart from Ketchen et al. (2013), this represents a critical issue. In particular, using the same observed measures to represent different theoretical constructs is particularly troublesome when utilizing the same set of firms over the same time horizon. As a hypothetical example, assume that there are eight published papers that utilize a dataset of publicly traded firms from 2001–2018, and each one includes R&D intensity as an observed measure. However, in three papers R&D intensity is used as a single-item measure to capture the construct *R&D investment*; in two papers, R&D intensity is used as a single-item measure to capture the construct *managerial discretion*; and in the remaining three papers, R&D intensity is used as a single-item measure to capture the construct *asset specificity*. Because the same observed measure has been used to represent three distinct theoretical constructs, it becomes impossible to evaluate the accumulated findings across the eight papers—even though all eight used the same sample of firms over the same period of time. This creates a "theoretical under-identification" issue because it becomes challenging to determine which construct was actually represented. Not only does this hamper a larger theoretical contribution, but it potentially creates inaccurate or misleading managerial insights.

A second validity-related issue occurs when archival data are utilized to develop new constructs to incorporate into existing theory. The ideal situation in utilizing archival data is to link measures to existing constructs, but in some cases, this may be difficult due to the uniqueness of the data or the context. For example, Miller, Golicic, and Fugate (2017) adopted definitions regulators had given to various sub-dimensions of motor carrier safety to serve as the theoretical definitions for these constructs. Similarly, Bartel, Ichniowski, and Shaw (2007) incorporate guidance from managers at valve production plants to create a construct of computer numeric control (CNC) machine quality, which they measure by calculating the negative logarithm of the number of CNC machines needed to produce a product. Thus, rather than starting with existing constructs and trying to operationalize these constructs using archival data, these authors directly incorporated definitions created by the data provider or industry experts as a means to extend existing theory (Kane, 2001).

New measures derived from archival data do not necessarily lack validity because they have not been previously established in the literature. New construct development offers a logical step forward in SCM theory development, but due diligence in addressing validity concerns includes several key elements. First, strong rationale is necessary to explain how the new construct fits into existing theory. Second, to the extent possible, it is important to disclose the response process utilized to generate the data. Third, if the information is available, the most direct evidence of validity in new construct development includes an explanation for if and how industry participants utilize the archival data to make decisions. Finally, scores for new constructs should correlate with the scores from other constructs in a manner consistent with theory. This element of new construct validity is not always feasible, however, because there may not be an alternate index available to correlate the scores. In sum, nontraditional approaches are likely to become more common over the coming years (Cortina, Aguinis, & Deshon, 2017), especially as the SCM field increases the utilization of archival data (Garver, 2019)

ENDOGENEITY CONSIDERATIONS IN ARCHIVAL DATA

The SCM field has become increasingly concerned with the threat of endogeneity when empirically driven research utilizes regression analysis. The threat of endogeneity occurs when a predictor of interest is correlated with the error term in a structural equation. The reason for the correlation is that the estimated model is incorrectly specified by (1) excluding relevant predictors that are uniquely correlated with the focal predictor and the dependent variable, (2) predictors being measured imperfectly (i.e., measurement error), or (3) not accounting for simultaneity (Wooldridge, 2009). In primary data collection, researchers have greater ability to address endogeneity concerns through research design and by collecting data on key theoretical confounding variables and potential instruments for their focal measures. When utilizing archival data, however, researchers generally have no control over data collection, so addressing endogeneity can be arduous. We add to previous SCM discussions on endogeneity (e.g., Ketokivi & McIntosh, 2017; Lu, Ding, Peng & Chuang, 2018) by focusing on more specific concerns that can result from the use of archival data.

When theoretical models hypothesize that change in one construct affects another construct in some manner (e.g., X has a linear effect on Y), endogeneity from omitted predictors becomes relevant. This necessitates reporting baseline econometric results by including control variables that serve to isolate the unique effects of focal predictor(s) (Wooldridge, 2009). Control variables that reduce concerns about omitted variable bias are those that have significant unique relationships with both the dependent variable *and* the focal predictor(s) (Mauro, 1990) and reside causally upstream from the focal predictor(s) (Keele, Stevenson, & Elwert, 2019).

Due to omitted variable bias concerns, one practice that seems especially prevalent with archival data is to indiscriminately add control variables. There is reason to *discourage* this practice, as it clashes with methodological best practice (Carlson & Wu, 2012; Spector & Brannick, 2011). Rather, a reasoned and thoughtful approach can be undertaken in assessing the need for and number of control variables. A logical explanation is needed to offer why particular control variables should be included *over and above other control variables included in the model*, keeping in mind two considerations.

First, arbitrary inclusion of control variables can unduly affect the meaning of the estimated regression parameter of interest (Keele, Stevenson, & Elwert, 2019; Xu et al., 1994). Breaugh (2006) provides a didactic example in the context of basketball players' height and rebounding. As he explains, including players' weight as a control in a regression model alters the meaning of height such that height represents lankiness because height's regression weight captures the effect of changing height by one unit while holding weight constant. Carlson and Wu (2012) and Spector and Brannick (2011) echo these concerns, with Carlson and Wu (2012, p. 431) stating, "Adding more CVs [control variables] does not make a study more rigorous. Unless there is a very sound reason that including a specific CV accomplishes an unambiguous and meaningfully statistical control objective, studies adding CVs may confound, rather than enhance, the interpretations of findings. When in doubt, leave them out." Second, Allison (1995) notes that excluding a nonsignificant control variable does not significantly affect the parameter estimate(s) for the focal predictor(s). Appendix A illustrates this in detail by considering a scenario where a researcher has one dependent variable, one focal predictor, and three control variables.

Another prescribed way of addressing omitted variable concerns is to utilize instrumental variable estimation (see Wooldridge, 2009; Angrist & Pischke, 2009; see Semadeni, Withers & Certo, 2014) for best practice when using instruments). This, however, can be challenging when using archival data, given that (1) effect sizes in archival data are often small, which raises concerns regarding weak instruments (Semadeni et al., 2014), and (2) researchers are often limited by the available data, which raises concerns that no valid instruments can be found (Lu et al., 2018). Weak instruments that explain little unique variance in the focal predictor and especially invalid instruments such as instruments that uniquely predict the dependent variable in the structural equation are both more problematic than using no instruments. Hence, in these specific cases, there is rationale for avoiding the use of instruments (Rossi, 2014). It is also important to present baseline results that do not use instruments as a basis for comparing estimates (Semadeni et al., 2014). Likewise, it is informative to evaluate the differences in parameter precision (i.e., size of standard errors) with and without instruments.

In instances when instrumental variables cannot be identified, two additional factors are worth considering. First, it is important to evaluate whether hypotheses are grounded in theory and practical relevance. Second, inclusion of theoretically grounded control variables becomes more important. If both of these factors have been satisfactorily addressed, there is rationale to have a reasonable degree of confidence that the hypothesized relationships are not being unduly affected by omitted variable bias. Likewise, transparency about this issue involves the acknowledgment of the potential for endogeneity and how it could affect the results.

A related issue that has been largely unaddressed in endogeneity discussions is when regression models hypothesize moderation or curvilinear effects. Examining these kinds of nonlinear effects can create a situation where finding enough valid instruments in an archival dataset becomes highly problematic. This is because it is inappropriate to generate predicted values for the linear terms of potentially endogenous independent variable(s) and *then* form the product terms (Angrist & Pischke, 2009). For example, if authors are interested in the interaction between variables X and Z and are concerned both are endogenous, they cannot estimate $\hat{X}$ and $\hat{Z}$ and then form the product term as $\hat{X}\hat{Z}$ even if they have valid instruments for X and Z. Rather, valid instruments are needed for all three associated terms (i.e., X, Z, and XZ). Likewise, testing the quadratic effect of X on Y requires instruments for both X and X^2, as it is incorrect to estimate $\hat{X}$ and then square this predicted value to generate $\hat{X}^2$ that is included in the regression model (Angrist & Pischke, 2009). This issue becomes more problematic as effects become more complex; for example, *seven* sets of valid instruments would be needed to identify all terms involved in a three-way interaction if all constituent variables are viewed as endogenous.

Given the challenge of finding appropriate instruments for curvilinear or moderation effects, endogeneity due to omitted variables is less pressing when theoretically driven moderation and/or curvilinear effects are hypothesized, and results are consistent with those predictions. Ketokivi and McIntosh (2017) note that the extent of the endogeneity concern must first be addressed theoretically. That is, finding evidence of complex interactions and curvilinear effects that are consistent with theory provide one input for scholars to construct an "inference to the best explanation" case (Lipton, 2004). Further, the philosophy of science literature (Mayo, 1991; Meehl, 1990) infers that specific contingencies must hold to find evidence consistent with moderation or curvilinear effects. Specifically, Leavitt et al. (2010, p. 660) states, "Additionally, greater confidence can be given to studies that specify moderated relationships. Although Meehl's paradox (1967) states that a high-power test yields a nearly 50% chance of significant results falling in the direction of a hypothesis, specifying the direction of a two-way interaction reduces this likelihood to 25%; specifying the direction of a three-way interaction reduces this to only 12.5%." Consequently, when interaction and/or curvilinear hypotheses are soundly grounded in the literature and the results offer corroboration, the risk of omitted variables biasing the parameter estimates is substantively reduced. In this situation, two considerations are paramount: (1) sensitivity to the difficulty of finding the additional number of instrument variables required, and (2) the significance of including theoretically relevant control variables to assuage endogeneity concerns.

Finally, in some cases, it is important to recognize that endogeneity may not be a concern because of the nature of the research question. A particular benefit that can come from the use of archival data is the ability to characterize and evaluate a change across time. For instance, a research question could be, "how has retailers' inventory performance changed since the end of the last recession?" Answering this kind of question involves utilization of a structured latent curve model (Blozis, 2004; Browne, 1993) or mixed effects model (Cudeck, 1996) that estimates a unique change trajectory of each retailers' inventory performance over the given time horizon. In this case, theoretical relationships between constructs are not being tested; rather, it involves identifying an algebraic function that adequately describes change over time and quantifying the degree of heterogeneity in retailers' change processes. Likewise, there may be interest in asking "how stable is retailers' rank ordering of inventory performance since the end of the last recession?" This research question could be answered by estimating autoregressive panel data structural equation models (Little, 2013) or patterned correlation matrices (Browne, 1977; Yung, Browne & Zhang, 2015). As these examples illustrate, endogeneity does not exist when scholars are interested in estimating unconditional effects.

In sum, endogeneity concerns are contingent on the nature of the research question being tested and the purpose of the estimated statistical model. In some settings, endogeneity due to omitted variables will be particularly concerning—as noted by Ketokivi and McIntosh (2017)—but the magnitude of concern is specific to each investigation. Consequently, it is advisable for researchers to be transparent about endogeneity issues and to clearly articulate their rationales as to the choices that they have made. As before, transparency is most critical.

RECAPITULATION & CONCLUSION

Archival data allow SCM researchers to push the boundaries of existing knowledge, especially since these sources often provide authors access to panel and time series data that expand the range of research questions that can be examined. Furthermore, many of the entities that generate archival data, such as the Bureau of Labor Statistics, can devote far more resources to data collection than small research teams. For example, the U.S. Census Bureau's M3 survey obtains monthly shipments, inventories, and orders from "most manufacturing companies with $500 million or more in annual shipments" (Census Bureau, 2019ureau, 2019). Consequently, archival data may provide the ability to answer heretofore unexamined questions by allowing researchers to leverage data that they themselves could not collect.

However, archival data are not a magical "silver bullet" and should not be perceived as inherently better than other data sources. We have presented a number of considerations that we hope will be helpful for both authors and reviewers (summarized in Table 2). The discipline will continue to advance as researchers establish logically grounded validity claims when they utilize archival data to measure theoretical constructs (Messick, 1995); as they clearly explain their sampling frame and scope of generalizability (Kane, 2013); and as they emphasize disclosure of the response process and the data preparation process. Furthermore, there is a need for transparency about potential endogeneity concerns due to omitted variables (Ketokivi & McIntosh, 2017) while simultaneously being cognizant that use of very weak or invalid instrumental variables can be more problematic than using no instruments (Rossi, 2014). We hope this editorial, and the references herein, help to lay the foundation for future SCM studies that fully utilize archival data to advance knowledge.

TABLE 2

Summary of Issues and Considerations

Issue	Question	Consideration Summary
Research design		
Sampling frame	What if the sample of firms in the dataset is not random?	Many data sources do not randomly draw from the overall population of firms, limiting the generalizability of the results. If the sample is nonrandom, define constructs within the specific sample from the data source and logically defend the scope of generalization of the results.
Data preparation & reporting	How should "upfront" issues before data analysis be addressed?	Data often have issues related to missing data, errant observations, or skewness. Transparently report data preparation steps and the approach used to address such issues.
Age of data	Is it appropriate to utilize "aged" data?	Age of data is less problematic when it has trait-like characteristics that remain relatively stable and consistent over time. Justify the use of aged data by highlighting the underlying theoretical mechanisms that are likely to be stable across time and/or providing empirical support.
Temporality	What if variables from one dataset temporally precedes variables from another dataset?	When merging datasets, the time periods may not always overlap. Justify temporality effects by demonstrating that the measures in question are time-invariant and/or providing a strong theoretical argument for their relationship.
Measurement validity		
Response process	How was the data generated?	Archival data can be self-reported and subject to limited audit. Disclose whether the data are self-reported, if it is subject to audit, and if firms are penalized for inaccurate self-reports.
Content evidence: construct domain	What if measures do not fully represent a construct's theoretical underpinnings?	Data sources may not include every theoretical dimension of a construct. Define the theoretical construct by conceptualizing it as it is represented by the measures in the dataset, and highlighting it as a limitation.
Content evidence: single items	Is it permissible to use one measure to operationalize a construct?	Single items are acceptable if they are objective, unidimensional, and have a clear meaning. Clearly explain and justify the use of single items.
Internal structure		Items must have reflective psychometric properties in order to

(continued)

TABLE 2 (continued)

Issue	Question	Consideration Summary
	Do a set of items to measure a reflective construct have acceptable psychometric properties?	measure a construct theorized to be a reflective latent variable. Explain the nature of the estimated measurement model, providing evidence that the measurement model displays acceptable fit, summarizing key parameters from the measurement model (e.g., factor loadings), and reporting coefficient omega as a measure of reliability.
Relations to Other Variables	Does a measure display theoretically consistent relations with other measures that are used to represent other constructs?	Finding theoretically consistent patterns of relations strengthens authors' validity claim for the focal measure. Draw on existing theoretical arguments to explain why a given measure should relate to other measures, and provide empirical evidence documenting these relationships.
Consequences	How are measures used in decision-making?	Measures may be justified to use for certain decision-making purposes, but not for others. In offering managerial insights, be explicit in explaining how data can be used for decisions, and be cognizant of unintended negative consequences that may come as a result.

Issue	Question	Transparency Recommendation
Measurement validity (cont.)		
Consistency of interpretations	What if the measures have been used to measure different constructs?	Using the same measures for different constructs can hamper the possibility of evaluating accumulated findings across studies. Provide clear logic that explains how the operationalized measures are consistent with the construct's theoretical underpinning.
New construct development	Can data be utilized to develop new constructs?	Constructs that have not been previously established can be newly developed from the data as a means to extend theory or build new theory. When developing new measures, provide a strong rationale for how the new construct fits into existing theory if the purpose is to extend theory, clearly address response process concerns, and if possible, explain how industry participants utilize the data for decision-making.

(continued)

TABLE 2 (continued)

Issue	Question	Transparency Recommendation
Endogeneity		
Linear effects	How should endogeneity be addressed in models with linear effects?	There is greater concern when the explanatory power of a linear model is low, and utilizing valid instruments can be challenging. When possible utilize valid instruments, but if instruments are very weak or invalid, include theoretically relevant control variables, noting the inability to identify a valid instrument as a limitation.
Complex effects	How should endogeneity be addressed in models with complex effects?	It is difficult to utilize instrumental variable estimation when testing for moderation or curvilinear effects. When testing complex effects, highlight that these effects are less susceptible to endogeneity concerns and further minimize the threat with relevant control variables.
Not applicable	When is endogeneity not an issue?	Some research questions render endogeneity as a moot point, specifically when the goal is to describe what has happened over time versus testing relationships between constructs. When this is the case, explicitly note that the research question and subsequent analysis are not threatened by endogeneity because authors are estimating unconditional effects.

REFERENCES

Aguinis, H., Hill, N. S., & Bailey, J. R. (2019). Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*, 1094428119836 485.

Ali, A., Klasa, S., & Yeung, E. (2008). The limitations of industry concentration measures constructed with Compustat data: Implications for finance research. *Review of Financial Studies*, 22(10), 3839–3871.

Allison, P. D. (1995). The impact of random predictors on comparisons of coefficients between models: Comment on Clogg, Petkova, and Haritou. *American Journal of Sociology*, 100(5), 1294–1305.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.

Astbury, B., & Leeuw, F. L. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation*, 31(3), 363–381. https://doi.org/10.1177/1098214010371972.

Australian Bureau of Labor Statistics (2019). *Quarterly business indicators survey*. Available at: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/business+indicators.

Badorf, F., Wagner, S. M., Hoberg, K., & Papier, F. (2019). How supplier economies of scale drive supplier selection decisions. *Journal of Supply Chain Management*, 55(3), 45–67.

Bartel, A., Ichniowski, C., & Shaw, K. (2007). How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. *Quarterly Journal of Economics*, 122(4), 1721–1758.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44(2), 175–184.

Bernstein, E. S. (2012). The transparency paradox: A role for privacy in organizational learning and operational control. *Administrative Science Quarterly*, 57(2), 181–216.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I., & Van Reenen, J. (2019). What drives differences in management practices? *American Economic Review*, 109(5), 1648–1683.

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological Methods*, 9(3), 334–353.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley.

Bonaccorsi, A. (1992). On the relationship between firm size and export intensity. *Journal of International Business Studies*, 23(4), 605–635.

Boyd, B. K., Dess, G. G., & Rasheed, A. M. (1993). Divergence between archival and perceptual measures of the environment: Causes and consequences. *Academy of Management Review*, 18(2), 204–226.

Braguinsky, S., Ohyama, A., Okazaki, T., & Syverson, C. (2015). Acquisitions, productivity, and profitability: Evidence from the Japanese cotton spinning industry. *American Economic Review*, 105(7), 2086–2119.

Breaugh, J. A. (2006). Rethinking the control of nuisance variables in theory testing. *Journal of Business and Psychology*, 20(3), 429–443.

Breaugh, J. A. (2008). Important considerations in using statistical procedures to control for nuisance variables in non-experimental studies. *Human Resource Management Review*, 18(4), 282–293.

Browne, M. W. (1977). The analysis of patterned correlation matrices by generalized least squares. *British Journal of Mathematical and Statistical Psychology*, 30(1), 113–124.

Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions* (Vol. 2, pp. 171–197). New York, NY: Elsevier.

Bujaki, M. L., & Richardson, A. J. (1997). A citation trail review of the uses of firm size in accounting research. *Journal of Accounting Literature*, 16, 1–27.

Calantone, R. J., & Vickery, S. K. (2010). Introduction to the special topic forum: Using archival and secondary data sources in supply chain management research. *Journal of Supply Chain Management*, 46 (4), 3.

Calantone, R., Whipple, J. M., Wang, J., Sardashti, H., & Miller, J. W. (2017). A primer on moderated mediation analysis: exploring logistics involvement in new product development. *Journal of Business Logistics*, 38(3), 151–169.

Cantor, D. E., Corsi, T. M., Grimm, C. M., & Singh, P. (2016). Technology, firm size, and safety: Theory and empirical evidence from the US motor-carrier industry. *Transportation Journal*, 55(2), 149–167.

Carlson, K. D., & Wu, J. (2012). The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, 15 (3), 413–435.

Casile, M., & Davis-Blake, A. (2002). When accreditation standards change: Factors affecting differential responsiveness of public and private organizations. *Academy of Management Journal*, 45 (1), 180–195.

Census Bureau (2019). *Manufacturers' shipments, inventories, & orders: How the data are collected*. Available at: https://www.census.gov/manufacturing/m3/how_the_data_are_collected/index.html.

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal*, 37(8), 1597–1614.

Chatterji, A. K., Levine, D. I., & Toffel, M. W. (2009). How well do social ratings actually measure corporate social responsibility? *Journal of Economics & Management Strategy*, 18(1), 125–169.

Chen, I. J., & Paulraj, A. (2004). Towards a theory of supply chain management: The constructs and measurements. *Journal of Operations Management*, 22(2), 119–150.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Earlbaum Associates Inc.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, 119(2), 166.e7–166.e16.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cortina, J. M., Aguinis, H., & DeShon, R. P. (2017). Twilight of dawn or of evening? A century of research methods in the Journal of Applied Psychology. *Journal of Applied Psychology*, 102(3), 274–290.

Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research*, 31(3), 371–403.

Cudeck, R., & Codd, C. L. (2012). A template for describing individual differences in longitudinal data, with application to the connection between learning and ability. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 3–24). Charlotte, NC: Information Age Publishing Inc.

Diamantopoulos, A., Ring, A., Schlegelmilch, B., & Doberer, E. (2014). Drivers of export segmentation effectiveness and their impact on export performance. *Journal of International Marketing*, 22(1), 39–61.

Distelhorst, G., Hainmueller, J., & Locke, R. M. (2017). Does lean improve labor standards? Management and social performance in the Nike supply chain. *Management Science*, 63(3), 707–728.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830–837.

Edelman, L. B. (1990). Legal environments and organizational governance: The expansion of due process in the American workplace. *American Journal of Sociology*, 95(6), 1401–1440.

Elking, I., Paraskevas, J. P., Grimm, C., Corsi, T., & Steven, A. (2017). Financial dependence, lean inventory strategy, and firm performance. *Journal of Supply Chain Management*, 53(2), 22–38.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

European Commission (2019). *Directive 2014/95/EU*. Available at: https://ec.europa.eu/info/business-economy-euro/company-reporting-and-auditing/company-reporting/non-financial-reporting_en.

Fawcett, S. E., Waller, M. A., Miller, J. W., Schwieterman, M. A., Hazen, B. T., & Overstreet, R. E. (2014). A trail guide to publishing success: tips on writing influential conceptual, qualitative, and survey research. *Journal of Business Logistics*, 35(1), 1–16.

Flynn, B., Pagell, M., & Fugate, B. (2018). Survey research design in supply chain management: The need for evolution in our expectations. *Journal of Supply Chain Management*, 54(1), 1–15.

Forbes, S. J., Lederman, M., & Tombe, T. (2015). Quality disclosure programs and internal organizational practices: Evidence from airline flight delays. *American Economic Journal: Microeconomics*, 7(2), 1–26.

Garver, M. S. (2019). Threats to the validity of logistics and supply chain management research. *Journal of Business Logistics*, 40(1), 30–43.

Goodstein, J. D. (1994). Institutional pressures and strategic responsiveness: Employer involvement in work-family issues. *Academy of Management Journal*, 37(2), 350–382.

Hamann, P. M., Schiemann, F., Bellora, L., & Guenther, T. W. (2013). Exploring the dimensions of organizational performance: A construct validity study. *Organizational Research Methods*, 16(1), 67–87.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36(1), 49–67.

Heide, J. B., & John, G. (1988). The role of dependence balancing in safeguarding transaction-specific assets in conventional channels. *Journal of Marketing*, 52(1), 20–35.

Ichniowski, C., & Shaw, K. (1999). The effects of human resource management systems on economic performance: An international comparison of US and Japanese plants. *Management Science*, 45(5), 704–721.

Ichniowski, C., Shaw, K., & Prennushi, G. (1997). The effects of human resource management practices on productivity: A study of steel finishing lines. *American Economic Review*, 87(3), 291–313.

Jin, G. Z., & Lee, J. (2014). Inspection technology, detection, and compliance: evidence from Florida restaurant inspections. *RAND Journal of Economics*, 45(4), 885–917.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211.

Kang, C., Germann, F., & Grewal, R. (2016). Washing away your sins? Corporate social responsibility, corporate social irresponsibility, and firm performance. *Journal of Marketing*, 80(2), 59–79.

Keele, L., Stevenson, R. T., & Elwert, F. (2019). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, 8, 1–13.

Kenny, D. A., & Zautra, A. (2001). Trait–state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *Decade of behavior. New methods for the analysis of change* (pp. 243–263). Washington, DC: American Psychological Association.

Ketchen, Jr, D. J., Ireland, R. D., & Baker, L. T. (2013). The use of archival proxies in strategic management studies: castles made of sand? *Organizational Research Methods*, 16(1), 32–42.

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

Kimberly, J. R. (1976). Organizational size and structuralist perspective: A review, critique, and proposal. *Administrative Science Quarterly*, 21, 571–597.

Kull, T. J., Kotlar, J., & Spring, M. (2018). Small and medium enterprise research in supply chain management: The case for single-respondent research designs. *Journal of Supply Chain Management*, 54 (1), 23–34.

Lapointe-Antunes, P., Cormier, D., Magnan, M., & Gay-Angers, S. (2006). On the relationship between voluntary disclosure, earnings smoothing and the value relevance of earnings: The case of Switzerland. *European Accounting Review*, 15(4), 465–505.

Leavitt, K., Mitchell, T. R., & Peterson, J. (2010). Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, 13(4), 644–667.

Levitt, S. D., List, J. A., & Syverson, C. (2013). Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *Journal of Political Economy*, 121(4), 643–681.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). New York, NY: Routledge.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Little, T., Lindenberger, U., & Nesselroade, J. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, 4(2), 192–211.

Lu, G., Ding, X. D., Peng, D. X., & Chuang, H. H. C. (2018). Addressing endogeneity in operations management research: Recent developments, common problems, and directions for future research. *Journal of Operations Management*, 64 (1), 53–64.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84–99.

Madsen, T. L., & Walker, G. (2017). Competitive heterogeneity, cohorts, and persistent advantage. *Strategic Management Journal*, 38(2), 184–202.

Mahoney, J. (2001). Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum*, 16(3), 575–593.

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108(2), 314–329.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, 58(4), 523–552.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1 (2), 108–141.

Mentzer, J. T., & Flint, D. J. (1997). Validity in logistics research. *Journal of Business Logistics*, 18 (1), 199.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, 1993(2), i–18.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.

Miller, J. W., Golicic, S. L., & Fugate, B. S. (2017). Developing and testing a dynamic theory of motor carrier safety. *Journal of Business Logistics*, 38(2), 96–114.

Newsom, J. T. (2015). *Longitudinal structural equation modeling*. Ney York, NY: Routledge.

Pfeffer, J., & Salancik, G. R. (2003). *The external control of organizations: A resource dependence perspective*. Stanford, CA: Stanford University Press.

Rabinovich, E., & Cheon, S. (2011). Expanding horizons and deepening understanding via the use of secondary data sources. *Journal of Business Logistics*, 32(4), 303–316.

Rigdon, E. E., Preacher, K. J., Lee, N., Howell, R. D., Franke, G. R., & Borsboom, D. (2011). Avoiding measurement dogma: a response to Rossiter. *European Journal of Marketing*, 45(11/12), 1589–1600.

Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science*, 33 (5), 655–672.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, 19(4), 305–335.

Rossiter, J. R. (2011). Marketing measurement revolution: The C-OAR-SE method and why it must replace psychometrics. *European Journal of Marketing*, 45(11/12), 1561–1588.

Schoar, A. (2002). Effects of corporate diversification on productivity. *Journal of Finance*, 57(6), 2379–2403.

Scott, A. (2018). Carrier bidding behavior in truckload spot auctions. *Journal of Business Logistics*, 39(4), 267–281.

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

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

Silverman, B. S., & Ingram, P. (2017). Asset ownership and incentives in early shareholder capitalism: Liverpool shipping in the eighteenth century. *Strategic Management Journal*, 38(4), 854–875.

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2), 287–305.

Stank, T. P., Pellathy, D. A., In, J., Mollenkopf, D. A., & Bell, J. E. (2017). New frontiers in logistics research: theorizing at the middle range. *Journal of Business Logistics*, 38(1), 6–17.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature*, 49(2), 326–65.

Wan, X. (2016). Timing of the effects: A dynamic analysis of pack-size variety, demand, and cost. *Journal of Business Logistics*, 37(3), 271–283.

Wiengarten, F., Fan, D., Pagell, M., & Lo, C. K. (2019). Deviations from aspirational target levels and environmental and safety performance: Implications for operations managers acting irresponsibly. *Journal of Operations Management*, 65 (6), 490–516.

Winter, S. G., Szulanski, G., Ringov, D., & Jensen, R. J. (2012). Reproducing knowledge: Inaccurate replication and failure in franchise organizations. *Organization Science*, 23(3), 672–685.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological methods*, 12(1), 58–79.

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

Xu, K., Windle, R., Grimm, C., & Corsi, T. (1994). Re-evaluating returns to scale in transport. *Journal of Transport Economics and Policy*, 28, 275–286.

Yu, S., Mishra, A. N., Gopal, A., Slaughter, S., & Mukhopadhyay, T. (2015). E-procurement infusion and operational process impacts in MRO procurement: Complementary or substitutive effects? *Production and Operations Management*, 24 (7), 1054–1070.

Yung, Y. F., Browne, M. W., & Zhang, W. (2015). Fitting direct covariance structures by the MSTRUCT modeling language of the CALIS procedure. *British Journal of Mathematical and Statistical Psychology*, 68(1), 178–193.

APPENDIX

EXCLUSION OF A NONSIGNIFICANT CONTROL VARIABLE

Per Allison (1995) excluding a nonsignificant control variable does not significantly affect the parameter estimate(s) for the focal predictor(s). The following scenario illustrates this logic:

Assume there is one dependent variable (Y), one focal predictor (X), and three control variables (C_1, C_2, and C_3). One way to conceptualize omitted variable bias is in terms of partial correlations (Cohen et al., 2003). Consider the case of a single omitted variable (Z). Omitting Z will only bias the estimate of X if two conditions hold (Mauro, 1990; Wooldridge, 2009). The partial correlation between X and Z must differ significantly from zero holding constant C_1, C_2, and C_3; in other words $r_{X,Z|C_1,C_2,C_3} \neq 0$. Additionally, the partial correlation between Y and Z must differ significantly from zero holding constant X, C_1, C_2, and C_3; in other words $r_{Y,Z|X,C_1,C_2,C_3} \neq 0$. *Unless both conditions exist*, the parameter estimate of X will be similar regardless of whether Z is included or excluded from the model (Allison, 1995). An additional point can be made based on this example per Mauro (1990, p. 316): (1) If Z is strongly correlated with C_1, C_2, and C_3, then there is less concern that omitting Z will affect the regression weight for X, and (ii) if C_1, C_2, and C_3 are strongly correlated with X, then there is also less concern that omitting Z will result in the effect of X being unduly influenced. With this being said, if C_1, C_2, and C_3 are very strongly correlated with X, there is the concern that the unique variation left over in X may poorly represent the underlying theoretical construct X is designed to measure (Breaugh, 2008).