
# Experiments in Supply Chain Management Research: A Systematic Review and Future Directions

| | |
|---|---|
| Journal: | *Journal of Business Logistics* |
| Manuscript ID | JBL-Sep-2023-8651.R2 |
| Wiley - Manuscript type: | Original Article |
| Methodology: | Systematic Literature Review, Experimental Design |
| Keywords: | experimental methods, systematic literature review |
| Abstract: | The supply chain management discipline has seen a tremendous growth in the use of experimental methods. Given the large number of published studies, the time seems opportune to systematically review the use of such approaches. In this note, we consider multiple dimensions of experimental design used in articles published in six of our premier journals. We present these findings and contemplate opportunities for future applications of experimental methods. In particular, we highlight a need to more regularly conduct and report on the results of power analyses and experimental checks, more carefully contemplate the justification and use of WEIRD (Western, educated, industrial, rich, and democratic) participants, develop and test mediated theoretical models, and increase our focus on teams as the unit of analysis when using experimental methods. |
| | |

# Experiments in Supply Chain Management Research: A Systematic Review and Future Directions


## Abstract

The supply chain management discipline has seen a tremendous growth in the use of experimental methods. Given the large number of published studies, the time seems opportune to systematically review the use of such approaches. In this note, we consider multiple dimensions of experimental design used in articles published in six of our premier journals. We present these findings and contemplate opportunities for future applications of experimental methods. In particular, we highlight a need to more regularly conduct and report on the results of power analyses and experimental checks, more carefully contemplate the justification and use of WEIRD (Western, educated, industrial, rich, and democratic) participants, develop and test mediated theoretical models, and increase our focus on teams as the unit of analysis when using experimental methods.



**KEYWORDS**
experimental methods, design of experiments, systematic literature review




***Acknowledgements:*** We thank the Co-Editor-in-Chief, the anonymous Senior Editor, and two anonymous reviewers for their valuable and constructive guidance throughout the review process.

## INTRODUCTION

Often referred to as the gold standard for empirical research (Shadish, Cook, & Campbell, 2002), between-subjects experiments[1] offer several strengths, including high internal validity and an understanding of causality. More recently, we have seen an expansion in the use of experiments in the supply chain management (SCM) discipline. A review of experiments in premier SCM journals, as we will describe below, finds that almost half of all articles employing a between-subjects or mixed design (46%) have been published since 2018.

Despite the strengths and growing use of such experiments, or perhaps as a result of these, our discipline has seen recent debate about best practices. Lonati et al. (2018) criticize the use of vignette-based experiments and deception, express concern about demand effects, and argue for minimum sample sizes and the use of participant performance-based incentives. In response, Eckerd et al. (2021) address these issues and qualify the tradeoffs in making these decisions. Ultimately, there is more agreement than disagreement between these two sets of authors. Indeed, given the growing use of and discussion about experiments in our discipline, it seems opportune to assess the current state of the methodological choices in experiments used in our discipline.

We do so by conducting a methods-focused systematic literature review across six of the premier SCM journals: *Journal of Supply Chain Management*, *Journal of Business Logistics*, *Journal of Operations Management*, *Management Science*, *Manufacturing and Service Operations Management*, and *Production and Operations Management*. This review encompasses articles that use a between-subjects or mixed experimental design, with an objective of assessing research methodology along several dimensions including participant characteristics, data collection venue, use of performance incentives, teams as the unit of analysis, analytical versus conceptual theory, use of manipulation checks, and tests of

[1] We define a between-subjects design as an experiment where different subjects "are studied in different conditions" (Shadish, Cook, & Campbell, 2002, p. 505) and where a subject is assigned to one and only one condition.

statistical power. Our findings suggest several opportunities to improve experimental methods in future research, including reporting the results of power analysis and manipulation checks, more carefully considering the use of WEIRD (Western, educated, industrial rich, and democratic) participant pools, developing mediation models, and studying teams as the unit of analysis. These opportunities also hold the potential to improve the theoretical contributions of future research. Researchers can develop greater confidence in theory testing through improved research methods – for example by appropriately using power analysis and manipulation checks. Supply chain management theory can also be advanced by considering the use of non-WEIRD participant pools, developing theoretical models that account for mediation, and moving to team dynamics and decision making as the unit of analysis.

In the next section of the note, we describe our study's methodology, including data collection, assessment of articles for inclusion, and coding of the articles encompassed in our study. We then describe the analysis of these data and our findings. We conclude by discussing future research opportunities that might help advance the use of experiments in our discipline, by improving methodological rigor, theory development, and insights for practice.

## METHODOLOGY

### Data collection and article identification

A systematic literature review explicitly describes how the review itself was operationalized, providing transparency and replicability (Tranfield, Denyer, & Smart, 2003). For our methods-based review, we followed Carter and Washispack's (2018) Modified AMSTAR criteria, which overlap with Durach et al.'s (2017) guidelines for conducting a systematic literature review (SLR). The modified AMSTAR criteria assess whether an SLR describes the inclusion/exclusion criteria for articles, clearly describes keywords and search criteria, includes two databases in the case of electronic reviews, uses at least two reviewers to determine whether articles are selected for inclusion in the review, and reports

associated inter-coder agreement statistics. The criteria also assure that the SLR states whether articles were purposefully confined to a subset of peer reviewed journals, provides a list of included and excluded articles, lists the characteristics of included studies in an aggregated form such as a table(s) and/or figure(s), and describes the data coding process including inter-coder agreement and definitions of codes.

The SCM Journal List (http://www.scmlist.com/) was central in the identification of relevant journals to include in our review. This list is sponsored by many of the universities with top-ranked supply chain management programs and is endorsed by a set of credible scholars in the discipline (Ketchen & Craighead, 2023). Of the eight journals encompassed in the list, we eliminated two: *Operations Research* and *Decision Sciences*. We removed *Operations Research* because of its almost exclusive publication of mathematical modeling articles. *Decision Sciences* publishes articles from a broad array of disciplines, including information systems, organizational behavior, and marketing. This left six journals: *Journal of Business Logistics*, *Journal of Operations Management*, *Journal of Supply Chain Management*, *Management Science*, *Manufacturing and Service Operations Management*, and *Production and Operations Management*. Like *Decision Sciences*, *Management Science* is similarly inter-disciplinary, yet it is nonetheless present as a highly regarded outlet for operations and supply chain management research by a host of top-ranked business schools, compelling us to retain it for inclusion in this review. With this in mind, and to ensure focus, in the case of *Management Science* we followed the method used by the SCM Journal List and only included articles that were handled by the *Operations Management* Department Editor.

In line with recent substantive reviews of the use of experimental methods in the management discipline (Bolinger et al., 2022; Stevenson et al., 2020), we searched for the following keywords in the article abstract, title, or keywords of the six journals: "experiment*", "random*", "policy captur*", "laboratory", or "conjoint". We performed these searches in two databases, ABI/Inform and

SCOPUS, on March 3, 2022. These searches resulted in the identification of 742 unique articles across the six journals (see Table 1 for a summary).

As a key aspect of our inclusion criteria, we focus on identifying studies that employed a between-subjects design (including studies that used both a mixed between-subjects and within-subjects design) where the dependent variable involves human actors who are potentially affected by the study's treatment(s). Two researchers reviewed each article and excluded articles that did not involve human actors and articles that used discriminant analysis, conjoint analysis, policy capturing, or a within-subjects design without a partial or full between-subjects design. The inter-coder agreement rate, calculated as the number of agreements about whether to include/exclude an article divided by the number of possible agreements (Perreault & Leigh, 1989) was 94.51%. Disagreements about article inclusion were discussed between the two researchers and resolved via a consensus approach. This led to the initial exclusion of 507 of the 742 articles. Two researchers then reviewed and coded each of the remaining 240 articles. During the coding process, an additional 38 articles were excluded based on the above exclusion criteria. This resulted in a final list of 202 articles. The 742 initial articles can be found in the on-line supplement, with the 202 included articles noted in bold font.



**TABLE 1** Search results and article inclusion

| Journal | Initial # Identified Articles | Initial # Excluded | Short List | Further Excluded (During Coding) | Final # Articles |
|---|---|---|---|---|---|
| JBL | 45 | 32 | 13 | 3 | 10 |
| JSCM | 26 | 6 | 20 | 3 | 17 |
| JOM | 148 | 92 | 56 | 6 | 50 |
| MS | 45[1] | 18 | 27 | 3 | 24 |
| MSOM | 139 | 113 | 26 | 6 | 20 |
| POM | 339 | 241 | 98 | 17 | 81 |
| Total | 742 | 502 | 240 | 38 | 202 |

[1] Forty-five articles were handled by the Operations Management Department Editor out of 288 articles in *Management Science* that were initially identified through our literature search.

## Data coding

Data were assessed across the codes shown in Table 2. Two of the note's authors independently coded each of the 202 articles. The initial inter-coder agreement rate was 88.99% (range of 85.93% to 91.59% across the six journals). A third author reviewed each of the 202 articles and resolved any disagreements between the initial two reviewers.

**TABLE 2** Data codes used in SLR classification

| Code | Definition |
|---|---|
| Country | Country(ies) location of participants. |
| Study Type | Lab (the study takes place online or in a lab setting) versus Field (the study takes place in a normally occurring situation, such as a company). |
| Team-based Research | The study involved teams (three or more individuals) that worked together toward a common goal (e.g., on a sourcing decision) and could make decisions as a team (unlike, e.g., the Beer Game). |
| Type of Theory | Analytical (mathematical models) versus Conceptual (logic and observation) are used to develop the study's hypotheses/manipulations |
| Data Collection | In-person versus Remote (participants were not physically present at a specific location during data collection – e.g., an online study) |
| Participant Type | Student (who is participating for extra credit, as part of a student pool, as part of a class assignment) versus worker (e.g., consumer, employee, worker participant pool worker such as MTurk) |
| Remote Participants | For Remote participants (where data collection is not in-person), are they Known (participants whose demographics can be reasonably assumed to be known) or Unknown (e.g., MTurk)? |
| Mean Sample Size | Average sample size for a study |
| Power Analysis | A power analysis was conducted |
| Manipulation Check(s) | Manipulation checks were conducted |
| Demand Effects | Demand effects were discussed |
| Use of Deception | The study used deception (e.g., confederates, providing false feedback) |
| SEM to Assess Mediation | Structural equation modeling or a similar approach was used to test for mediation |
| Design | Fractional versus Full Factorial |
| Vignette(s) | A descriptive vignette is used to manipulate the study's factors |
| Performance-based Incentives | The amount of a participant's remuneration is at least partially based on participant performance during the study |

## FINDINGS

We present a summary of our findings in Table 3. Approximately half of the studies in our review set were published since 2018 (the year 2018 or later), and thus 2018 was chosen as a useful point of demarcation for comparison. For associated reference, the final two columns of Table 3 display the findings from 2018 to present (March 2022, when we conducted our searches) and prior to 2018, respectively. All but three of the 202 articles were published after 2000. These findings suggest that between-subjects or mixed experiments involving human actors are a relatively new methodological entrant to empirical SCM research, and that the use of such experiments seems to be rapidly increasing.



**TABLE 3** Summary of findings

| | | All Studies | 2018 - Present | Pre-2018 |
|---|---|---|---|---|
| Number of Articles | | 202 | 94 | 108 |
| Number of Studies | | 272 | 131 | 141 |
| Mean # Studies per Article | | 1.35 | 1.39 | 1.31 |
| Country of Study Participants | US | 68.38% | 64.89% | 71.63% |
| | Germany | 9.93% | 9.92% | 9.93% |
| | China | 5.88% | 7.63% | 4.26% |
| | Unclear | 4.78% | 8.40% | 1.42% |
| | Other* | 11.03% | 9.16% | 12.77% |
| Type of Study | Lab | 88.24% | 86.26% | 90.07% |
| | Field | 11.76% | 13.74% | 9.93% |
| Experimental Design | Fractional | 8.09% | 10.69% | 5.67% |
| | Full Factorial | 91.91% | 89.31% | 94.33% |
| Team-based (vs. Individual or Dyad) | | 2.94% | 3.82% | 2.13% |
| Type of Theory | Analytical | 30.15% | 32.82% | 27.66% |
| | Conceptual | 69.85% | 67.18% | 72.34% |
| Data Collection | In-person | 67.65% | 55.73% | 78.72% |
| | Remote | 32.35% | 44.27% | 21.28% |
| Participants | Students | 58.58% | 51.91% | 63.12% |
| | Non-Students | 40.30% | 48.09% | 34.75% |
| | Both | 1.12% | 0.00% | 2.13% |
| Remote Participants | Known | 41.86% | 55.73% | 90.65% |
| | Unknown[1] | 58.14% | 44.27% | 9.35% |
| Average Sample Size | Field Study | 163,283 | 297,569 | 221 |

| | | | | |
|---|---|---|---|---|
| | Lab Study | 216 | 256 | 181 |
| Power Analysis Conducted | Across Studies | 9.56% | 12.98% | 6.38% |
| | Lab Studies | 9.17% | 11.50% | 7.09% |
| Manipulation Checks | | 30.15% | 27.48% | 32.62% |
| Demand Effects | | 2.57% | 4.58% | 0.71% |
| Use of Deception | | 3.31% | 3.05% | 3.55% |
| Use of SEM to Test Mediation | | 2.57% | 3.05% | 2.13% |
| Use of Vignettes | | 26.84% | 25.95% | 27.66% |
| Vignette Studies Using Realism Checks | | 45.21% | 58.82% | 33.33% |
| Use of Performance-Based Incentives | | 52.57% | 50.38% | 54.61% |

**Note:** [1]MTurk = 77%, Qualtrics = 16%, Other = 7%

Rather than providing a recount and restating of all of the data shown in Table 3, we focus instead on those portions of the findings that are most germane in guiding future research and stimulating discussion about how our discipline might move forward as we continue to use experimental methods. These opportunities can help researchers to develop more insightful theoretical models, identify more managerially useful problems, and enhance methodological rigor. In particular, we describe the findings surrounding the use of power analysis, experimental checks, non-WEIRD participants, mediated theoretical models, and teams as the unit of analysis, in the final section of the paper.[2] Before we do so, we next report the results of a citation analysis using the data shown in Table 3 as independent variables and article citations as the dependent variable.

**Citation analysis**

Citation analysis provides some interesting details regarding the influence that these study factors might have on future research and may provide tell-tale signs of challenges faced by the community as a whole. We consider predictive models of the log number of Google citations and the log of citations per year, since both of these measures are clearly long right-tailed (log transformations

[2] While we are not specifically advocating for a greater proportion of field experiments, their use can complement the key strengths of lab experiments (internal validity and precision of measurement of behavior, McGrath, 1982) by improving realism. For a treatment of an associated issue of participant sampling, see Stevens (2011). We thank the anonymous Senior Editor for raising these issues.

allowed for approximate fits to Normal distributions in accordance with K-S test criteria). Potential independent variables include the variables listed in Table 3, along with the journal in which the article was published, and the article year of publication. Step-wise modeling, driven by Akaike information criteria (AIC), suggested that both the log number of Google citations and the log of Google citations per year were most efficiently predicted by four factors: When the article was published, the sample size, whether analysis occurred at the individual-level, and whether the study appeared in *Management Science* (see Table 4). These four factors yielded an adjusted $R^2$ level of 0.543 in predictions of log citations, and 0.187 for predictions of log citations per year. Each of these factors have significant positive impacts on the predicted variables.

**Table 4** Relationship between citations and coded element (step-wise AIC inclusion)

| | Ln(Total Google Cites) | | | | | | Ln(Total Google Cites / Year) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Beta** | **SE(Beta)** | | **Beta** | **SE(Beta)** | | **Beta** | **SE(Beta)** | | **Beta** | **SE(Beta)** | |
| Intercept | 1.734 | (0.318) | *** | 1.731 | (0.315) | *** | 0.683 | (0.289) | *** | 0.684 | (0.285) | ** |
| Published in Management Science | 0.727 | (0.196) | *** | 0.821 | (0.202) | *** | 0.699 | (0.178) | *** | 0.802 | (0.183) | *** |
| Age: (2022 - Publication Year) | 0.199 | (0.014) | *** | 0.187 | (0.015) | *** | 0.056 | (0.013) | *** | 0.043 | (0.014) | *** |
| Ln (Sample Size) | 0.090 | (0.050) | * | 0.095 | (0.050) | * | 0.132 | (0.046) | *** | 0.136 | (0.045) | *** |
| Included Individual-level Analysis | 0.499 | (0.187) | *** | 0.943 | (0.292) | *** | 0.391 | (0.170) | ** | 0.835 | (0.264) | *** |
| Journal of Operations Management | | | | 0.274 | (0.164) | * | | | | 0.333 | (0.148) | ** |
| Described Randomization | | | | -0.557 | (0.282) | ** | | | | -0.512 | (0.255) | ** |
| Included Exp. Checks / Power An. | | | | -0.134 | (0.300) | | | | | -0.248 | (0.271) | |
| Checks x Perform. Incentives | | | | 1.262 | (0.683) | * | | | | 1.153 | (0.618) | * |
| F-value | 56.82 | | | 30.02 | | | 11.86 | | | 7.46 | | |
| d.f. | 185 | | | 181 | | | 185 | | | 181 | | |
| AIC | 126.3 | | | 134.0 | | | 90.4 | | | 95.8 | | |
| Adj-$R^2$ | 0.543 | | | 0.551 | | | 0.187 | | | 0.215 | | |
| R-square | 0.551 | | | 0.570 | | | 0.204 | | | 0.248 | | |

* p<.05
** p<.01
*** p<.001

Additional variables entered, shown to increase the adjusted $R^2$, as an alternative inclusion criterion, were publication in the *Journal of Operations Management*, whether randomization was specified, whether 'any' experimental check or mention of power analysis was included (an aggregated binary term), and the interaction of that binary variable with the inclusion of performance incentives. The binary relating to 'any' check or power analysis was created as an aggregate (if any of these tactics were held, this binary is 1), due to the relatively small incidence of these design approaches across our sample. While this factor in itself did not provide predictive strength, its interaction with the use of performance-based incentives did (i.e., if performance-based incentives are used, a manuscript is likely to receive a bump in citations when proper experimental checks or power analysis is also discussed).

A few basic takeaways are apparent. Large sample sizes and studies involving individual-level analyses see greater citation rates. These two factors are likely intertwined however, and the latter may be more a factor of individual-level studies being simply easier to conduct than critical dyadic or team-level inquiries. Studies that are easier to execute, are also more likely to be conducted, and are in turn more likely to reference similar past studies: A plurality of individual-level studies thus favors citations of prior individual-level studies.

Somewhat concerning is the negative sign on the randomization feature. One would hope that researchers would be more apt to reference well-articulated research designs. However, as per our sample, the community is still woefully behind in the use of a number of fundamental design checking mechanisms, and apparently does not yet appreciate the value of something as basic as sample randomization (even avoids referencing such methods to some extent). There is some sense however that researchers may be starting to appreciate the value of checks in at least a subset of studies (those involving performance-based incentives). Given the popularity of such studies, this may be a sign that such design checks are at least getting attention, if not being fully adopted.

## DISCUSSION AND MOVING FORWARD

Our findings suggest a proliferation of the use of experimental designs in SCM research, with almost half of the reviewed studies being published in the most recent, five-year period. Some of the elements of these studies were relatively constant across time periods. These include the use of WEIRD participants, percentage of lab versus field studies, team-based research, percentage of analytical versus conceptual theory, and the use of manipulation checks, deception, SEM to test mediation, vignettes, and performance-based incentives. However, normatively, some of these percentages, while constant across time, are surprisingly low. These include the use of team-based research (2.94% across studies), manipulation checks (30.15% across studies), and the use of SEM to test mediation (2.57% across studies). While perhaps not surprising, the high percentage of studies using WEIRD participants suggests that much of knowledge derived from experimental methods studies in our discipline is based on participants with relatively homogeneous characteristics and perhaps does not purposefully use non-WEIRD participants.

Other elements of the reviewed studies did change substantially across time periods. These include the use of remote participants (an increase from 21.78% to 44.27%), the use of unknown remote participants (an increase from 9.35% to 44.27%), reporting the results of a power analysis (an increase from 6.38% to 12.98%), and the use of realism checks for vignette studies (an increase from 33.33% to 58.82%). The percentage of studies reporting the results of a power analysis, while more than doubling across time periods, is still surprisingly low.

These findings suggest several opportunities to improve the use of experimental methods in future research. These opportunities include providing more detailed reporting to improve confidence of results in theory testing, broadening the scope of our theorizing through cross cultural comparisons of non-WEIRD participants, developing and testing mediation models, and moving beyond the individual or dyad to investigate teams as a unit of analysis. We outline these opportunities as follows.

**Power analysis**

A critical component of an effective experimental design is that sufficient statistical power is used to evaluate the hypotheses. Power analysis serves as a statistical tool to consider Type I (false positive) and Type II (false negative) error rates and can be done prior to collecting samples to ensure that sufficient sample size is collected given an expected effect size (Cohen, 2013).

Determining statistical power requires specific calculations based on the statistical model that will be used to analyze the data. Several factors influence the power required for a given model, including the estimation method, the number of covariates (if any), number of groups, the use of repeated measures, and whether more complicated models such as mediation or interaction effects will be analyzed; these factors can all change statistical power in meaningful ways. Fortunately, statistical power can be calculated relatively easily by many statistical packages including Stata, SPSS, SAS, R, or tools such as G*Power designed specifically for statistical power calculations. A starting point for each of these tools is included in Appendix A. For more complicated models where standard power calculations cannot be directly applied, simulation approaches can also be used (Fritz & MacKinnon, 2007).

While it is sometimes difficult to estimate effect sizes in a novel experiment, using previously published work as a baseline, estimating based on general effect sizes, or using a pre-test sample to estimate effect sizes are all potential solutions. General rules of thumb to estimate sample sizes, such as the number of observations required per group, are poor approximations for power and should be replaced by power analysis for the statistical model selected in each experiment.

The evidence that less than 10% of studies reported the results of a power analysis is a notable limitation of the current research in our field and a relatively easy methodological concern to address. It is possible that many authors engage in power analysis in their work even when it might not be in the final published manuscript. Yet, it is worth noting that while power analysis is a

relatively small aspect of any single study, the contribution to the entire discipline of including the statistical power is substantial, as noted by previous discussions on the importance of statistical power: "More important than the implications for single studies and individual research projects, statistical power plays a vital role in the development of a consistent cumulative … science." (Abraham & Russel, 2008, p. 286). Doing and reporting power analysis in experimental designs can lay a stronger foundation for future SCM research.

**Appropriate use of experimental checks**

Similar to power analysis, the paucity of manipulation checks is troubling. It can be attributed to any number of factors but most commonly emerges due to a perceived lack of need, or a true lack of knowledge regarding the value it provided in validation. While the latter is difficult to assess, the former can itself be an artifact of the degree to which decision-making contexts are either deliberately selected due to their real-world simplicity, or are highly stylized to remove real-world complexity. While simple decision scenarios are certainly worthy of study, a great many decision making contexts are highly complex, and over stylization can only work against the potential for deriving insights to practice (theoretical contributions are still possible).

Manipulation checks are viewed as generally critical best practices among top contemporary scholars who conduct laboratory experiments. Recent statements to this effect have been made across disciplines from experimental social psychology ("… manipulation checks (MCs) … are critical for the viability of the logical premise of a theoretical hypothesis … Conversely, experiments without MCs suffer from serious deficits of convergent and discriminant validity … because no manipulation can be expected to affect only a single IV", pg. 818, Fiedler et al., 2021), to business and management disciplines such as accounting ("manipulation checks used are crucial for evaluating the validity of an experimental study.", Kotzian et al., 2020, p. 479).

In field experiments, authors can employ non-intrusive verbal protocols or mouse tracking to "monitor some natural mental process," (Harrison & List, 2004, p. 1050). Currently, with digital tracking (IoT) and the availability and rapid advances of AI, it is possible to extract meaningful indicator data from video and audio recorded content. This increases difficultly in imagining scenarios in which (even if only during piloting) assurances of the effectiveness of manipulations cannot be checked on through non-intrusive approaches. Put differently, there is certainly no mandate for manipulation checks to consist of a battery of Likert-type scale questions, cf. Fiedler et al., 2021. Sometimes the onus simply falls upon the researcher to identify clever manipulation checks. With that said, there can certainly be research designs where a manipulation check is desirable, but not feasible. Think for example of a multi-national firm that has agreed to work with researchers to conduct a field experiment where communication messages to warehouse employees are manipulated. The firm may not be willing to allow researchers to communicate with the workers to assess whether the workers noticed the messages.

More broadly, we qualify that our prescription for conducting experimental checks should *not be taken as dogma*. There are instances when manipulation checks may be unnecessary. For example, consider a study that investigates warehouse order picking where participants are given ten tasks (items to pick from inventory). The study's researchers may be interested in investigating factors that can improve productivity in terms of lowering the amount of time it takes to complete each task. If one of the study's manipulated factors is providing participants with more or less time to complete the tasks (say 200 seconds versus 100 seconds to complete the tasks), then this objective difference in time probably does not require a manipulation check. Conversely, if for this same study the researchers want to investigate the effects of time pressure, with the 200 second condition representing low time pressure and the 100 seconds representing high time pressure, then a manipulation check is likely necessary to assess whether participants experienced time pressure. In

other words, the necessity of a manipulation check is contingent in part on the theorizing of the reason for the effect.

Per this example, the line between whether experimental checks are needed is clearly not a matter of whether a treatment consists of an objectively numerical manipulation. There is no guarantee that individuals faced with a numerical parameter are any more likely to observe its relevance (and use it) than a non-numerical one. Further, many studies which consist of numerical parameter manipulations consider multiple parameters simultaneously, and yet anticipate participants to either make appropriate use of a provided formulation or of a mental model for processing such information given stated objectives. Seldom is there a chance to peak into those mental models at each decision instance, thus there can be no guarantee that multiple parameters are each being appreciated in a clean and ordinate fashion prior to incorporation in such models. Yet there is nevertheless often a presumption that appropriate attention to each parameter is taking place (manipulation checks aimed at confirming a relative level of such), and that any decision failure is not a function of crossed signals by such parameters, with one being misinterpreted or ignored due to the level of another (confound checks). There is also often a presumption that a reshaping of the mental model or approach to decision-making as a whole can't occur due to exposure to a specific numerical parameter (Hawthorne checks could prove relevant). In the absence of such checks, one can't disentangle whether participants are consciously under-weighing certain factors, whether they didn't sufficiently observe them, or whether they subconsciously confounded one with another. Since researchers care about claims regarding how participants consciously respond to information in these settings, we should care about whether that response is in fact an attributably conscious one.

We further caution that manipulation checks should not be interpreted as a proxy for data validity. Manipulation checks provide critical evidence that the treatment in question is a key factor

influencing behavior observed. In conjunction with confound checks[3] and other tactics, manipulation checks provide confidence to both researchers and fellow scholars (readers) that a specific treatment in question is having the impact theorized (rather than something else). It is true that these checks are also occasionally used to measure attention to tasks (Kane & Brabas, 2019; Kotzian & Stober, 2020), although that is certainly not their core principle (Fiedler et al., 2021); that is, checks on participant attentiveness to a task are often best managed separately, since they are capturing issues that transcend manipulations (a point that is often confounded in criticisms of manipulation checks).

**Non-WEIRD participants**

The majority of participants across all studies are from WEIRD countries with the U.S. dominating (68.38%), followed by Germany (9.93%). Henrich, Heine, and Norenzayan (2010) who coined the term "WEIRD" encouraged researchers who use WEIRD participants to acknowledge the limitations of their samples. Their key argument was that studies conducted exclusively with WEIRD participants can lead to a narrow understanding of human behavior and limit the generalizability of research findings. The reason for this is that non-WEIRD participants may have different cultural, economic, political, and educational backgrounds. This can have potentially significant implications for experimental research, as non-WEIRD participants may have different behavioral norms, communication styles, attitudes towards authority, or perceptions of justice and fairness. Moreover, their specific backgrounds may affect their motivation, problem-solving skills and styles, and decision-making processes. Therefore, there is a general need to use more non-WEIRD samples in behavioral research.

However, simply sampling non-WEIRD participants would only rely on a broader set of surface-level factors and still risk oversimplifying the complex nature of human behavior. To conduct experimental SCM research effectively, a nuanced approach to data collection and analysis is required.

[3] Of the 272 studies, only 19 (7%) explicitly describe conducting confounding checks.

Such an approach should explore deep-level factors, such as personality differences. By accounting for deep-level differences on the level of the individual, researchers can gain a more comprehensive understanding of human behavior and develop interventions that are more effective and relevant across diverse populations. Hence, we call upon researchers not merely to add non-WEIRD participants but to carefully consider potential deep-level factors in their studies, such as controlling for the Big 5 personality traits.

More broadly, researchers should detail the criteria for participant selection, the rationale behind these criteria, and their alignment with the overall research objectives. This includes explaining why the participants chosen are appropriate for addressing the research questions. The justification of the sample used needs to be closely tied to the existing literature, indicating whether the sample choice aligns with or diverges from previous studies in the field. Such an approach demonstrates that the sample selection is informed by extant research or by identified gaps in the literature. Furthermore, the findings and discussion sections should refer back to the sample choice. This practice aids in contextualizing the results, thereby strengthening the case for their relevance and applicability.

**SEM and mediated theoretical models**

One of the key benefits of experimental designs is the ability to capture causal links through careful manipulation of variables with dependent variables proving a clear testable causal relationship. A properly run experiment using random assignment can effectively limit the potential for confounding effects as the manipulations are orthogonal to the measured outcomes. However, this strength can also be a limitation as manipulating a variable and measuring its outcome on a dependent variable can lack the theoretical richness to fully explain the model. The use of structural equation modeling or a similar approach to assess mediation in a conceptual model was uncommon across the reviewed articles.

Mediated theoretical models can greatly extend the theoretical scope of a research design, but in doing so might sacrifice some of the pure benefits available through experimental design and testing. Authors should carefully consider the inherent tradeoffs between causal tests and correlational designs and several guidelines are listed below for careful consideration (We discuss potentially related issues – common method variance and endogeneity – in the sub-section which follows).

1. Design experiments and statistical tests considering the research goals. If the goal of the research is a pure test of a theory, causal relationships that are carefully controlled can yield more rigorous tests. If the experimental design is more exploratory to establish new theoretical models, a more open experimental design might be desired to capture a larger potential scope.
2. Recognize whether the experimental designs and statistical models are causal or correlational, as multiple simultaneously collected DVs are not necessarily causal. In an experimental design manipulating A, and measuring B, and C afterwards, the causal links that could be identified are between A → B, and A → C (with a potentially confounding effect on C introduced by the measurement of B). Because B and C are not randomly assigned, they are no longer orthogonal variables and have the same limitation as a correlational model. This means that any argument for mediation between A → B → C or A → C → B relies on a theoretical argument for the link, and is not a pure test of causality. Further, such a mediation model requires that B and C demonstrate sufficient discriminant validity.
3. There are experimental designs that are explicit statistical tests of mediated models in experiments that could be considered (see Pirlott and Mackinnon (2016) which highlights the limitations and strengths of several experimental mediation designs).
4. The value of a research design should be considered in toto, and the limitation of one experimental design can be balanced by another research design explicitly designed to address certain limitations. Multiple-staged experiments, multi-study experimental designs, and multi-method designs can increase the validity of any experimental design.
5. In correlational models such as mediated SEM when several variables are measured at the same time, theoretical justification for the independent nature of the models and the causal direction posited by the model should be rigorously done, using factor analysis and theoretical justification for the connections. Two highly similar constructs in a mediated model (i.e., A → B → C when B and C are analogous to one another) has a greater potential for misspecification of the model.

*Common method variance and endogeneity*

One advantage of experimental designs is that random assignment to conditions creates a research design that is robust to concerns regarding common method variance (CMV) and endogeneity (Cooper et al., 2020). CMV, or "systematic error variance shared among variables measured with and introduced as a function of the same method and/or source" (p. 763, Richardson, Simmering & Sturman, 2009) and endogeneity, when an explanatory variable is

correlated with the error term (Lu, Ding, Peng & Chaung, 2018) can lead to validity concerns regarding regression results. Because random assignment is exogenously manipulated it cannot correlate with the error term or omitted variables. Because of this, controlled experiments are often a recommended solution to reduce endogeneity (Lu, Ding, Peng & Chuang, 2018).

It is worth noting that while most controlled experiments using randomization are immune to concerns regarding CMV and endogeneity due to random assignment, there are some cases where CMV and endogeneity can still be a concern:

1. Randomization does not apply to the variable of interest (e.g., randomizing the number of rounds in a laboratory game when the variable of interest is price which varies by some other factor).
2. Randomization does not match the unit of analysis (e.g., the random assignment occurs for a specific lab session involving multiple participants and the statistical analysis is at an individual level).
3. Randomization only applies to a part of the tested model (e.g., the authors are testing X->Y->Z and manipulate X with random assignment, and then measure Y and Z. Because Y and Z are not orthogonal through random assignment, any analysis of the relationship between Y and Z would be exposed to potential CMV and endogeneity concerns).
4. Randomization fails (e.g., one randomized group has a higher drop-out rate that is related to the variables of interest).

Whenever possible, we recommend researchers take advantage of the strength random assignment offers methodologically, and if not possible then researchers appropriately employ methods to address CMV and endogeneity.

**Teams**

Team-based research was defined as a study of three or more individuals who interact and work together toward a common goal, who may have different roles and responsibilities and who can make decisions as a team (Mathieu et al., 2017). This differs from studies where individuals might act as suppliers who compete in a reverse auction, or the beer-game, where individuals are not able to directly interact and make decisions with individuals in other organizations across the supply chain. Team-based research that employs the team as the unit of analysis is uncommon, accounting for a small percent (2.94%) of studies, even in the more recent, 2018-present time period (3.82%).

Yet in many companies, employees spend 80 percent or more of their time working collaboratively (Cross, 2021). From production work teams to sourcing decisions, supplier development, new product development, strategic negotiations, supply chain disruptions, product recalls, and creating strategic alliances, much of the work and decision making in SCM occurs in teams. It thus seems surprising that only about three percent of studies investigated the team as the unit of analysis. It may be that authors are cognizant of the challenges these settings present, and are aware that large-sample, single respondent studies (albeit imperfect) are generally well received by the present review process. Regardless of the reason, team-based experimental designs can lead to valuable insights by allowing researchers to investigate the dynamics of multiple team members as they interact.

One example from the literature that we reviewed is the work of Cantor and Jin (2019), who examine how perceptions of social loafing can affect the decisions of higher-performing production line team members to help poorly performing team members. They find that team perception of social loafing decreases helping behavior, even among altruistic team members. A second example is Franke, Eckerd, and Foerstl's (2022, p. 965) investigation of 136 team-based sourcing decisions. Here, the authors find that while functional goal misalignment results in status conflict – as predicted – this effect can be lessened when team members have differing levels of design "to control and influence" others. They further find that this decrease in status conflict improves sourcing team performance. Both of these studies represent real world contexts in terms of how SCM work occurs, and both studies required a team-level investigation to yield their findings.

**Concluding comments**

The SCM discipline has seen tremendous growth in the use of experimental methods. Our systematic literature review allows us to assess the current state and use of experimental methods. Many of the design elements that we examined do not necessarily suggest an imbalance or otherwise improper use of experimental methods. In other cases, there appear to be opportunities to expand

our use of experiments to test and develop SCM theory. All of these opportunities also improve insights for practice by enhancing methodological rigor (power analysis and manipulation checks) and increasing realism and generalizability (mediated models, teams as the unit of analysis, and non-WEIRD participants).

## References

Abraham, W. T., & Russell, D. W. (2008). Statistical power analysis in psychological research. *Social and Personality Psychology Compass*, *2*(1), 283-301.

Bolinger, M. T., Josefy, M. A., Stevenson, R., & Hitt, M. A. (2022). Experiments in strategy research: A critical review and future research opportunities. *Journal of Management*, *48*(1), 77-113.

Cantor, D. E., & Jin, Y. (2019). Theoretical and empirical evidence of behavioral and production line factors that influence helping behavior. *Journal of Operations Management*, *65*(4), 312-332.

Carter, C. R., & Washispack, S. (2018). Mapping the path forward for sustainable supply chain management: A review of reviews. *Journal of Business Logistics*, *39*(4), 242-247.

Chun, Y., Harris, S. L., Chandrasekaran, A., & Hill, K. (2022). Improving care transitions with standardized peer mentoring: Evidence from intervention based research using randomized control trial. *Journal of Operations Management*, *68*(2), 185-214.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cooper, B., Eva, N., Fazlelahi, F. Z., Newman, A., Lee, A., & Obschonka, M. (2020). Addressing common method variance and endogeneity in vocational behavior research: A review of the literature and suggestions for future research. *Journal of Vocational Behavior*, *121*, 103472.

Cross, R. L. (2021). *Beyond Collaboration Overload: How to Work Smarter, Get Ahead, and Restore Your Well-Being*. Harvard Business Review Press.

Durach, C. F., Kembro, J., & Wieland, A. (2017). A new paradigm for systematic literature reviews in supply chain management. *Journal of Supply Chain Management*, *53*(4), 67-85.

Eckerd, S., DuHadway, S., Bendoly, E., Carter, C. R., & Kaufmann, L. (2021). On making experimental design choices: Discussions on the use and challenges of demand effects, incentives, deception, samples, and vignettes. *Journal of Operations Management*, *67*(2), 261-275.

Fiedler, K., McCaughey, L., Prager, J. 2021. Quo Vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*(4), 816–826.

Franke, H., Eckerd, S., & Foerstl, K. (2022). Rising to the Top: Motivational Forces Influencing Status Conflict in Sourcing Teams. *Production and Operations Management*, *31*(3), 963-983.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*(3), 233-239.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009-1055.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29.

Kane, J.V., Barabas, J. 2019. No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science*, *63*(1), 234–249.

Ketchen Jr, D. J., & Craighead, C. W. (2023). What constitutes an excellent literature review? Summarize, synthesize, conceptualize, and energize. *Journal of Business Logistics*, *44*(2), 164-169.

Kotzian, P., Stober, T. 2020. To be or not to be in the sample? On using manipulations checks in experimental accounting research. *Accounting Research Journal*, *33*(3), 469-482.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*, 19-40.

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*, 53-64.

Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 452.

McGrath, J. E. (1982). Dilemmatics: The study of research choices and dilemmas. In McGrath, J. E., Martin, J., & Kulka, R. A. (1982). Judgment calls in research. Sage Publications, Thousand Oaks, CA.

Mir, S., Aloysius, J. A., & Eckerd, S. (2017). Understanding supplier switching behavior: The role of psychological contracts in a competitive setting. *Journal of Supply Chain Management*, *53*(3), 3-18.

Nicks, S. D., Korn, J. H., & Mainieri, T. (1997). The rise and fall of deception in social psychology and personality research, 1921 to 1994. *Ethics & Behavior*, *7*(1), 69-77.

Perreault Jr, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, *26*(2), 135-148.

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*, 29-38.

Pulles, N. J., & Loohuis, R. P. (2020). Managing Buyer-Supplier Conflicts: The Effect of Buyer Openness And Directness On A Supplier's Willingness to Adapt. *Journal of Supply Chain Management*, *56*(4), 65-81.

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods*, *12*(4), 762-800.

Rong, K., Zhou, D., Shi, X., & Huang, W. (2022). Social Information Disclosure of Friends in Common in an E-commerce Platform Ecosystem: An Online Experiment. *Production and Operations Management*, *31*(3), 984-1005.

Rungtusanatham, M., Wallin, C., & Eckerd, S. (2011). The vignette in a scenario-based role-playing experiment. *Journal of Supply Chain Management*, *47*(3), 9-16.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207-222.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Wadsworth, Cenage Learning, Belmont, CA.

Stevens, C. K. (2011). Questions to consider when selecting student samples. *Journal of Supply Chain Management*, *47*(3), 19-21.

Stevenson, R., Josefy, M., McMullen, J. S., & Shepherd, D. (2020). Organizational and management theorizing using experiment-based entrepreneurship research: Covered terrain and new frontiers. *Academy of Management Annals*, *14*(2), 759-796.

## APPENDIX A: Tools for Calculating Power

**Calculating Power with Reference Tables**
Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* Academic Press.

**Calculating Power in STATA**
Power, precision, and sample size | Stata
https://www.stata.com/features/overview/power-and-sample-size/

**Calculating Power in SAS**
Introduction to Power and Sample Size Analysis (sas.com)
https://support.sas.com/documentation/onlinedoc/stat/131/intropss.pdf

**Calculating Power in R**
A Practical Guide to Statistical Power and Sample Size Calculations in R (r-project.org)
https://cran.r-project.org/web/packages/pwrss/vignettes/examples.html

**Calculating Power in SPSS**
Power Analysis - IBM Documentation
https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=features-power-analysis

**Calculating Power in G*Power**
Universität Düsseldorf: G*Power (hhu.de)
https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower