Received: 17 May 2019 | Revised: 23 October 2020 | Accepted: 27 October 2020

DOI: 10.1002/joom.1128

TECHNICAL NOTE

WILEY

# On making experimental design choices: Discussions on the use and challenges of demand effects, incentives, deception, samples, and vignettes

**Stephanie Eckerd**[1] | **Scott DuHadway**[2] | **Elliot Bendoly**[3] | **Craig Carter**[4] | **Lutz Kaufmann**[5]

[1]Haslam College of Business, University of Tennessee, Knoxville, Tennessee

[2]Portland State University, Portland, Oregon

[3]Fisher College of Business, The Ohio State University, Columbus, Ohio

[4]W.P. Carey School of Business, Arizona State University, Tempe, Arizona

[5]WHU – Otto Beisheim School of Management, Vallendar, Germany

**Correspondence**
Stephanie Eckerd, Haslam College of Business, University of Tennessee, Knoxville, TN.
Email: seckerd@utk.edu

**Handling Editor:** Tyson Browning.

**Abstract**

In late 2018, the *Journal of Operations Management* published an invited methods article by Lonati et al. (2018) to provide guidance to authors on how to design behavioral experiments to achieve the rigor required for consideration in the journal. That article was written as a response to a number of behavioral research submissions to *JOM*, each dealing with interesting topics but viewed by the editors to possess poor design choices at inception. While the Lonati et al. (2018) piece provides experimental guidance fitting to certain research agendas, questions have arisen concerning whether and how exactly to implement some of the points that it makes, and how to best address trade-offs in the design of behavioral experiments. Questions have also arisen concerning how to apply these concepts in operations management research. This technical note seeks to address these questions, by diving into the details of research risks and trade-offs regarding demand effects, incentives, deception, sample selection, and context-rich vignettes. The authors would like to recognize the input of a large number of senior scholars in the *JOM* community who have provided support and feedback as we have sought to help authors tease out what can reasonably be done in designing strong behavioral experiments that fit various research agendas.

**KEYWORDS**

deception, demand effects, Experimental methods, incentives, sample, vignettes

## 1 | INTRODUCTION

Behavioral research in the *Journal of Operations Management (JOM)* and the Operations Management (OM) field draws from a wide variety of disciplines and methodological approaches. As the scope of the field broadens (Croson et al., 2013), it requires researchers to be more knowledgeable in multiple reference disciplines (Bendoly et al., 2010). The benefits in expanding OM's foundational focus are, of course, numerous; foremost is that this allows researchers to ask interesting research questions and apply the most appropriate methodologies for a specific research question. Yet, the combination of such diverse methodologies also requires trade-offs in research design, and can be

Stephanie Eckerd and Scott DuHadway authors contributed equally to this work.

  

challenging for authors and reviewers trying to successfully navigate such a wide range of methods using the highest standards.

The recent invited article by Lonati et al. (2018) presents "the 'ten commandments' of experimental research" (p. 20) with the end goal of providing "a synthesis of best practices into a uniform methodological paradigm that can help guide future experimental work" (p. 20). These authors raise critical concerns about many of the challenges associated with experimental work, identifying several pitfalls that should be avoided. However, given the way their recommendations are presented in the article (particularly Table 1), there is a risk of OM researchers potentially overlooking the nuance implicit to their methodological recommendations and interpreting them as more restrictive than perhaps intended. While Lonati et al. (2018) acknowledge some of these nuances, and provide reasons to employ different experimental design choices beyond exclusive adherence to the "ten commandments," there is a need to further flesh out guidance and compensate for misunderstandings that have been observed by many experienced authors in the *JOM* community. In this technical note, we therefore extend the discussion started by Lonati et al. (2018), offering further perspectives on ways that researchers and reviewers can carefully navigate research design choices. The primary design decisions that have raised questions in the OM community concern demand effects, incentive alignment, deception, sample issues, and context-rich vignette experiments.

In this note, we provide a roadmap, departing from the Lonati et al. (2018) discussion, that guides OM researchers as they explore a breadth of important research questions, rigorously drawing from a variety of theoretical foundations (Campbell & Stanley, 2015; Cook et al., 2002). We explore trade-offs arising in designing and administering experimental research in OM (Bickman & Rog, 2008; McGrath, 1982; Scandura & Williams, 2000). Consideration of trade-offs does not imply a lack of standards or rigor; rather it enforces explicitness relating to potentially incompatible goals, the "strengths and weaknesses of alternative means for pursuing goals,", and justification of research design choices (Collier et al., 2004, p. 223). We provide insights regarding these experimental trade-offs in consideration of evidence in Section 2. In Section 3, we consider the interdependence of these design decisions, and provide examples of how these trade-offs may be made across the entire array of design decisions with published papers from the OM field.

## 2 | CONSIDERATIONS OF DIFFERENT RESEARCH DESIGN CHOICES THAT MAXIMIZE RIGOR

All research methods have limitations, and as such no one method is the "best" or "correct" or "true" method (McGrath, 1982). Trade-offs are necessary in all experimental work. In this section, we develop the discussion around five issues raised by Lonati et al. (2018), exploring the nuances that are likely to arise in the OM context when designing and administering experiments: demand effects, incentive alignment, deception, sample issues, and context-rich vignette experiments. The guidance offered in this section is summarized in Table 1.

### 2.1 | What are the risks of experimental demand effects and how should they be addressed?

One of the many criticisms levied against experimental research involving human subjects concerns the threat of experimental demand effects (also referred to as demand characteristics). These effects represent "changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior (behavior 'demanded' from them)" (Zizzo, 2010, p. 75). For demand effects to pose a risk to experimental research, participants need to change their behavior based on their belief of the researcher's desired outcomes, and this change needs to be correlated with the objectives of the research (see Zizzo, 2010). Lonati et al. (2018) address demand effects in their third "commandment," focusing on describing situations where demand effects pose a particularly strong risk. We elaborate on the perceived risks of demand effects and provide several ways to proactively address and test for them.

Several recent substantive empirical investigations provide guidance regarding the risk of demand effects. de Quidt et al. (2018) explore demand effects in economic games, and Mummolo and Peterson (2019) explore demand effects in vignette experiments. de Quidt et al. (2018) deliberately induced demand effects in opposite directions and measured changes in targeted behavior to measure their risk. Using approximately 19,000 participants across 11 canonical economic games (e.g., the dictator game and trust game), they concluded the following (p. 3268, emphasis added):

> Our first finding is that responses to the weak treatments are modest, averaging around 0.13 standard deviations, varying from close to 0 for unincentivized real effort

**TABLE 1** Research design choices for rigorous experiments

| | | | |
|---|---|---|---|
| *Demand effects* | *Greater concern when*:<br>• The research question involves a sensitive topic (e.g., ethical behavior) where participants might guess what the researcher wants.<br>• The manipulations are salient to participants.<br>• Participants and researchers interact directly. | *Lesser concern when*:<br>• Using a between-subjects design (participants do not see more than treatment).<br>• The research question is not obvious to participants.<br>• Participants are unlikely to guess the research hypotheses.<br>• The experiment is double-blinded. | *Suggestions for OM experimental research*:<br>• Test for demand effects if it pertains to the research question or context.<br>• Use a pilot study to test manipulation effectiveness.<br>• Use between-subject designs for research where demand effects are viewed as a concern, or appropriate mitigation measures for within-subject designs. |
| *Incentives* | *Financial incentives beneficial when*:<br>The research question and desired behaviors are related to economic motivations and incentives.<br>The external research context being studied has similar incentives.<br>Behaviors can be connected to decision-based incentives in meaningful ways. | *Financial incentives problematic when*:<br>Decision-based incentives would bias respondents, increase demand effect risks, or reduce external validity.<br>The behaviors in the experiment are not well-defined choices between different economic outcomes or are otherwise unrelated to economic choices.<br>The experiment involves deception or withholding of information that could be perceived as influencing compensation. | *Suggestions for OM experimental research*:<br>Make the experiment as close as possible to the research context being analyzed such as rewarding respondents like real-world reward scenarios.<br>Use decision-based incentives when they are connected to the research question.<br>Carefully consider the research question and the behaviors being observed to decide between decision-based incentives or general incentives. |
| *Deception* | *Deception may be warranted when ALL of the following apply*:<br>• Deception is necessary to answer the research question.<br>• The benefits outweigh the potential harms.<br>• Respondents are fully debriefed and participant harm is minimized.<br>• The research complies with IRB or appropriate ethical guidelines regarding the use of deception. | *Avoid deception when*:<br>• Nondeceptive alternatives could be used instead of deception.<br>• The potential harms outweigh the benefits.<br>• The sample population is frequently used for experiments such as a laboratory.<br>• The deception would influence or could be perceived as influencing respondent compensation. | *Suggestions for OM experimental research*:<br>• Use deception only when necessary for the research question and the benefits outweigh the potential harm.<br>• Consider respondent harm, sampling norms of laboratories, available alternatives, and the fit between the research question and experimental design before using deception.<br>• Be transparent about the usage of deception, and follow IRB guidelines (in the U.S.) and appropriate ethical guidelines. |
| *Samples* | *General samples allowable when*:<br>• The research question emphasizes individual behaviors that are generalizable to many contexts.<br>The experiment context does not require in-depth knowledge of any particular context. | *Targeted samples preferred when*:<br>• The research question is dependent on context.<br>• The level of analysis is at a higher level such as organizational or interfirm contexts.<br>The samples should have adequate experience to fully understand the context. | *Suggestions for OM experimental research*:<br>• Match the unit of analysis and context to the sample population.<br>• More general sampling approaches such as student populations or general online sampling methods are appropriate when the behavior in the research question is universal and generalizable to many contexts.<br>• Use more specific sampling approaches when the context of the experiment is important and |

(Continues)

**TABLE 1** (Continued)

| | | | |
|---|---|---|---|
| | | | when the behaviors observed are meant to be generalizable to more specific samples.<br>Use multiple sample populations when possible. |
| *Vignettes* | *Vignettes more appropriate when*:<br>The context of the experiment is necessary for the research question.<br>The topic is sensitive.<br>Manipulations on different scenarios are realistic and respondents and researchers are blind to the manipulations.<br>Respondents would have an appropriate grasp of the experimental context. | *Vignettes less appropriate when*:<br>The manipulation lacks meaningful realism to the participants.<br>Participants lack context to make informed and realistic decisions.<br>Determining effect sizes. | *Suggestions for OM experimental research*:<br>Match the research design to the research question and embrace the appropriate epistemological foundation for the research question.<br>Be transparent in the type of research and the foundations used in the research.<br>When using manipulations in all experimental designs, make such interventions as consequential and realistic as possible. |

> to 0.29 standard deviations for trust game second movers. *In most tasks, our estimates are not significantly different from zero. Overall, we interpret these results as suggesting that demand effects in typical experiments are likely to be small*. Responses to our strong demand treatments are much larger, with bounds averaging 0.6 standard deviations and ranging from 0.23 to 1.06 standard deviations. While these bounds are likely more conservative than required in most applications, they illustrate that participants can respond substantially to strong signals about the researcher's objective, and thus researchers are right to pay close attention to potential demand effects in their studies.

It is important to qualify the difference between these findings. The "weak" manipulation informs participants of the experimental hypothesis and the expected direction of their findings, with these instructions varying by treatment. As de Quidt et al. (2018) noted regarding the weak effect condition, "We believe that these treatments are likely to be more informative than implicit signals about demand in typical studies, so in our view these bounds will be sufficient for most applications" (p. 3267). The "strong" manipulation included asking participants to specifically answer questions in one way ("You will do us a favor if...") and demonstrated that it is possible to create demand effects by asking participants to answer in that specific way. Overall, these results suggest that similar, well-designed experiments are likely not to be impacted by demand effects, but that participants will respond in certain ways if asked.

Mummolo and Peterson (2019) investigated demand effects using five different hypothetical vignettes grounded in various political science contexts. The manipulations used in Mummolo and Peterson (2019) were highly similar to the "weak" manipulation of de Quidt et al. (2018), as they included only a hint regarding the research hypotheses in their instructions.[1] Mummolo and Peterson concluded that "across five surveys that involve more than 12,000 respondents and over 28,000 responses to these experiments, we fail to find evidence for the existence of [experimental demand effects] EDEs in online survey experiments" (Mummolo and Peterson (2019), p. 518). Similar to de Quidt et al. (2018), Mummolo and Peterson (2019) were able to create a demand effect in some extreme conditions, such as when respondents were given financial incentives to follow explicit demand effects: "When this added incentive is present, we are sometimes able to detect differences in observed treatment effects that are consistent with the presence of [experimental demand effects] EDEs. But on average, pooling across all our experiments, we still see no detectable differences in treatment effects even when financial incentives are offered" (2019, p. 518).

Comparing findings across these studies, the evidence suggests demand effects were smaller in the vignette manipulations in Mummolo and Peterson (2019) than in the economic games in de Quidt et al. (2018), although for both groups the total impact was low. It is possible that the difference is due to the different demand effect manipulations between the two studies, the specific

experiments selected, or the different contexts. For example, economic games tend to be more abstract in comparison to vignettes, which typically include a richer context (Camerer, 1997). When considering choices in an abstract thought experiment (i.e., the dictator game or trust game), it may be that participants are likely to rely more on situational cues (i.e., hints from the researcher) in selecting their behaviors, leading to heightened demand effects relative to vignette experiments where the contextual information provided is used to make behavioral decisions. Furthermore, although the vignettes studies in Mummolo and Peterson (2019) were conducted in political science, and not OM, their "inability to uncover evidence of hypothesis-confirming behavior across multiple samples, survey platforms, research questions and experimental designs suggests that longstanding concerns over demand effects in survey experiments may be largely exaggerated" (2019, p. 528). Some of the issues addressed in their vignettes in fact directly correlate to questions studied in an OM context, for example the implications of race on resumé treatment is remarkably similar to the context of consumer attitudes of crowdsourced delivery drivers based on ethnicity (Ta et al., 2018). Moreover, some of the underlying theories explaining behavior in these studies, for example framing effects, are the same as those used in OM research (e.g., Abbey et al., 2019; Wuttke et al., 2018). Taken together, these results suggest that experimental research can be carried out in the OM context in a way that minimizes demand effects. Further research into the demand effects of strong manipulations is warranted.

Although the empirical evidence that is available suggests that the risk posed by demand effects is expected to be limited in the OM context, it is worth considering, as Lonati et al. (2018) did, the kinds of strong manipulations that might occur in this context, and whether such strong manipulations would yield demand effects that would be sufficiently large to cause concern. Even in these cases, examples of remedies can be found from the literature. For example, Dhar et al. (2018), concerned that socially desirable responses might interact with their treatment in a field experiment, incorporated a short-form Marlow-Crowne module to measure respondents' propensity to offer socially desirable baselines.[2] Similarly, researchers concerned about demand effects (for example, when the manipulations might be particularly salient to participants or the researcher interacts with the participants directly in some way) could benefit through direct measurement of potential demand effects, and correlating those biases to the treatments. de Quidt et al. (2018), pp. 3288–3289 describe a within-subject approach that can be added at the end of a study to construct bounds for demand effects for each participant. Tsutsui and Zizzo (2014) measure respondents' "[experimental demand effect] EDE Sensitivity," and find evidence that this measure was "uncorrelated with behavior" in all regressions and robustness tests (p. 238), concluding that demand effects were not a concern for their study.

Researchers could also proactively use designs that are more robust, although not immune, to demand effects, such as between-subjects designs rather than within-subject designs (de Quidt et al., 2019). Researchers should make this choice only after carefully considering the trade-offs in these methods with regards to the context of the research question being studied and in light of the practical implementation of the research study (Charness et al., 2012). Other approaches are well covered in econometric textbooks covering experimental designs (see de Quidt et al., 2019) and include double-blind experiments, identical and neutral instructions across experimental manipulations, minimizing experimenter-participant interaction, not informing participants regarding study hypotheses, specific conditions, and expected outcomes, and efforts to mask researcher intent. Most of these remedies are relatively low-cost and, while they do not guarantee complete immunity from demand effects, they do reduce these risks and should be used whenever possible. Effectively managing demand effects through research design is a good example of how to off-set potential trade-offs in conducting rigorous and relevant experimental research.

## 2.2 | Critical considerations regarding the use of incentives

Another key consideration in experimental design involves if and how participants are to be financially rewarded for their participation. Lonati et al. (2018) address the use of incentives in their fourth "commandment," recommending that researchers "avoid hypothetical experiments with hypothetical choices" (p. 20). The situation is more nuanced, however, as it is critical to align the research question, the research context, and experimental incentives. For some research questions, decision-based financial incentives are appropriate. For example, Liu et al. (2015) examined whether participants chose shared transportation with shared costs or private transportation with individual costs. Incentives are core to their research question and coincide with targeted behaviors, and so are clearly an appropriate design choice. However, for other research questions, decision-based incentives are inappropriate, and their introduction could create a confound. For example, Ried et al. (n.d.) explored information leakages in buyer–supplier–supplier triads, and how a supplier's observation of a

buyer leaking the information of an unrelated supplier causes the focal supplier to lose trust in that buyer and reduces wilingness to share information with the buyer. The research question, context, and absence of financial incentives are appropriately matched. Inappropriate introduction of decision-based incentives could lead participants to guess at the "right" answers to earn greater monetary rewards rather than simply answering honestly (Slater, 1980).

In OM research, it is not atypical to encounter phenomena with no explicit extrinsic motivation, such as employees enacting organization-level decisions, where there is no expectation of direct, explicitly related rewards. For example, research investigating the decision to initiate a recall (Ball et al., 2018) does not depend on behavioral incentives because incentives are unrelated to the research question or the context in which the question is embedded. There are also good examples of mixed-method approaches. Cantor and Jin (2019) examined how production line factors influence intrinsically motivated helping behaviors, and do not incorporate use of incentives in their main study. The absence of incentives in this instance represents an appropriate match of research question, context, and design choice. In post hoc tests, Cantor and Jin (2019) introduced team- and individual-level performance incentives, and found that while individual incentives were not related to helping, team-based incentives were. Again, the research question (i.e., do incentives influence helping?) coincided with the design choice.

Even in contexts in which incentives are appropriate, there is ample empirical evidence that financial rewards meaningfully change behavior, and this can happen in unintended, nontrivial ways. Not surprisingly, these effects depend on a number of factors: the nature of the task (Eckartz et al., 2012), the amount or type of incentive (Holt, 1986; Holt & Laury, 2005), the value placed on incentives by participants (Ariely et al., 2009), as well as a host of additional framing issues. For example, incentive alignment fails to improve performance when participants are risk averse, yet risk averse participants improve their performance under flat-rate schemes (Cadsby et al., 2016). There are also important cultural differences in responses to incentive framing (Lee, Ribbink, & Eckerd, 2018). Financial incentives can even be harmful to decision-making efforts. In Camerer and Hogarth's (1999) evaluation of 74 studies comparing the use of incentives, they identified overlearning effects, overexertion effects, and self-conscious behaviors (i.e., "choking") as potentially adverse consequences of incentivization. Meloy et al. (2006) found that incentives can impact participants' mood, leading to biased information processing and overconfidence.

Our point throughout this section is not to suggest that research should or should not use decision-based incentives. Rather, we illustrate that the use of financial incentives in experiments carries its own risks. Researchers and reviewers should carefully consider how incentives match the research question and the research context to maximize the internal validity of the experiment.

## 2.3 | Deception should be used carefully and sparingly in experimental research

A fast reading of Lonati et al.'s (2018) sixth commandment ("VI. Deception. Do not deceive participants; obfuscation is allowed.", Table 1, p. 20) can lead a reader to understand a universal requirement for perfect truth in the administration of behavioral experiments. It is important to note that Lonati et al. (2018) are forgiving of obfuscation (i.e., the withholding of critical information), as well as a "reasonable amount of deception" applied in field experiments when the importance of the research question might justify it (p. 23). The debate surrounding the use of deception in experiment research is not a new one, with many point – counterpoint arguments available in the literature (e.g., Barrera & Simpson, 2012; Krasnow et al., 2020). Much of this debate is centered on how economists and social psychologists differ concerning what degree of deception is considered acceptable and what constitutes an appropriate exception. The debate illustrates that the degree of deception can be described on a continuum. For example, consider the use of a computer confederate in an experimental setting. Should it be considered deception to instruct participants, "You will be playing with a partner" when the partner is a computer? Consider further if the instructions said, "You will be playing against another player in the room" when the computer player was locally hosted on the machine in the room. While most might agree that "an explicit misstatement of fact" (Nicks, Korn, & Mainieri, 1997) qualifies as deception, a recent survey of experimental economists found there is considerable variability as to what constitutes deception and obfuscation (Samek, 2019). Thus, the problem is more difficult than asking when one can—or must—use deception; fundamentally, the concept is too ill-defined to answer such a query. Clearly, it is inconsistent with a scientific approach to draw an arbitrary line on deception; researchers need to carefully consider the benefits and risks of operating anywhere along the continuum.

A key issue regarding deception is its potential to generate confounding effects, that is, that the act of deceiving a participant will cause the participant to alter their

behavior. To date, while there are very few empirical evaluations of the methodological costs of deception, there are a few from which we can draw some conclusions. Ortmann and Hertwig (Hertwig & Ortmann, 2008a, 2008b; Ortmann & Hertwig, 2002) conducted several evaluations across psychology studies using deception. They reported that when participants have direct and specific knowledge that deception was used, this impacted their behavior; yet, if that knowledge was indirect or general, targeted behaviors were not affected. Another set of studies (also reported on by Hertwig & Ortmann, 2008a) finds mixed results regarding whether deceived participants experience resentment (ascertained after debriefing), but that telling participants they will be deceived beforehand can alter performance, and that deception that arouses suspicion in participants may reduce conformity behaviors. The take-aways are that researchers using deception would be wise to wait until all data for the research effort has been conducted before debriefing participants, and that it may be worthwhile to post-test participants for suspicion.

A second issue concerns the use of repeated participant pools, as is the case with many undergraduate student laboratories. The potential risk posed is that a participant, once deceived in a study, will in the future be more suspicious of experiments, and that ambient distrust will lead to confounding behavior. There is indeed some evidence that deception can lead to less willingness to participate in future studies, particularly when the participant is both deceived and receives smaller compensation (Jamison et al., 2008). Note this affects recruiting efforts, *not* confound effects. Yet, for labs using repeat sampling pools it is a risk worth considering. Jamison et al. (2008) suggested that, rather than banning deception altogether, participant pools could be maintained separately. Avoidance of deception may most profitably be thought of as a *lab rule*, as opposed to a *research rule*.

A third issue involves the ethical considerations underlying deception and obfuscation. We would like to add to the discussion the role of review boards. Researchers in the U.S. must have their research designs vetted by Institutional Review Boards, all of which have special sections dedicated to the issues of deception or withholding of information. Similar Ethical Review Boards exist in many research universities in Europe, as well. These guidelines and gatekeepers help to uphold the standards of ethical research practice, although a systematic understanding of their breadth and applicability to issues of deception and obfuscation could be beneficial for the OM community.

Use of deception requires an evaluation of trade-offs and consideration of the research question. For some research questions, deception works in opposition to the goals of the experiment, thus detracting from internal validity. For other research questions, the trade-off involves the ability to model important contextual cues in a controlled way, versus not representing those factors in the experiment at all (Ariely & Norton, 2007). For example, in studies of unethical behavior, confederates may be used to provide an initial modeling of the unethical behavior observed by participants so that ensuing contagion effects can be assessed (e.g., Gino et al., 2009). Although few experiments using deception have been conducted in OM (Lonati et al., 2018), those that do tend also to make use of confederates in order to control the treatments participants are exposed to or to provide important contextual knowledge (e.g., Eckerd et al., 2013; Sommer et al., 2020 and described in Section 3 below). Per Jamison et al. (2008), researchers should carefully consider its role, and use deception or information withholding only when the research is not otherwise practicable, risks are minimal, and potential benefits outweigh potential risks; as a reviewer, consider the trade-offs presented.

## 2.4 | Issues in experimental sampling

Two issues often are relevant when it comes to experiment samples: sample size and the sample population. The issue of sample size is addressed by Lonati et al. (2018) in their eighth "commandment"; the advice provided in the table itself is that to "ensure an appropriate sample size per experimental cell for covariate balance ($n > 50$ per cell)" (p. 20). One must reference the footnote for the caveat that advanced planning of an experimental study should, if possible, include a test of power which would indicate an appropriate sample size. We agree with this well-established advice (Verma & Goodale, 1995), and would add to it that it is important to clearly differentiate between the number of *participants* versus the number of *observations*. Often in experiment research, multiple observations are collected from a single participant (see the streams of literature evaluating the bullwhip effect and newsvendor decision-making, as examples). Where this experimental design is used, it is the number of independent observations that typically matters. For a more in-depth review of sample size considerations, we refer readers to Lenth (2001), who pointed out additional practical and ethical concerns relating to sample size, beyond statistical considerations.

The issue of sample population tends to be a trickier concern in our field. Laboratory-based studies often use students as participants. This is advocated for in Lonati et al. (2018) in their ninth "commandment," and we agree that students are convenient participants because it

is easy to get them into university labs, and they are relatively cost-effective participants. However, theoretically, an argument needs to be made for the appropriateness of a sample, which means matching the sample to the focal research question (Thomas, 2011). For research evaluating universalistic theories, student research participants are generally "safe" (Stevens, 2011). However, even when the arguments for using students are sound, it should still be acknowledged in all studies employing student samples that their use represents a convenience sample, and the convenience of student samples comes at a cost. Research has described student participant pools as "WEIRD," meaning from cultures that are Western, educated, industrialized, rich, and democratic (Henrich et al., 2010a, 2010b). These restrictions can be quite meaningful when interpreting problems in the context of OM, where focal research questions orbit global problems. We are not suggesting this de-legitimizes the use of student participant pools, only that researchers need to be transparent regarding limitations and trade-offs inherent in their use. One effective mechanism to help reduce the concerns regarding the bias of a specific sample is to replicate the results using additional samples that are drawn from a different pool (McGrath, 1982). Replication of results across samples strengthens the triangulation and robustness of any research.

Increasingly, studies in the OM space are adopting online platforms (e.g., Amazon Mechanical Turk [MTurk], Qualtrics panels) to administer experiments (see Lee, Seo, & Siemsen, 2018). These online platforms greatly expand potential participant pools, but also are subject to trade-offs (Aguinis & Lawal, 2012, 2013). A key initial issue is whether the sample is appropriate for the research question; for example, online panels can be particularly useful when there are confidentiality concerns, such as when asking about abusive supervision (Porter et al., 2019) or supply chain fraud (DuHadway et al., 2020). A thorough exposition of the strengths and weaknesses of crowdsourcing platforms is offered by Goodman and Paolacci (2017). Among the strengths are reduced costs, participant diversity, and a well-referenced section purporting the strong data quality achieved through MTurk. Hauser et al. (2018) provided evidence and solutions for some of the more common concerns relating to the use of MTurk samples, including insufficient effort, language comprehension issues, and misrepresentativeness. While some studies have reported increased attentiveness of MTurk samples as compared to student samples (Hauser & Schwarz, 2016; Klein et al., 2014), the use of attention checks with any sample is an important experimental protocol (Abbey & Meloy, 2017; Kane & Barabas, 2019). Finally, it is possible through certain online platforms (e.g., MTurk) to build a personal panel of participants who can be properly screened and tracked over time (Litman et al., 2017; Peer et al., 2012; Sharpe Wessling et al., 2017).

While much of this guidance is specific to MTurk, by and large the lessons conveyed are applicable to many of the other commonly used platforms (Google Surveys, Prolific, etc.). Importantly, the concerns regarding online participant pools are not unique to OM. The guidance here originates from a variety of disciplines, in particular management and marketing, which have learned to successfully navigate this potentially rich resource.

## 2.5 | Vignette (and other nonconsequential) studies

Behavioral OM has tended to maintain a focus on actual behaviors, or "in-task" behaviors (Bachrach & Bendoly, 2011; Croson et al., 2013). However, while actions represent one plausible dependent variable, intentions, attitudes, and affect also are important outcomes amenable to experimental study. These "out-of-task" perceptual measurements often are the underlying drivers of observed behaviors (Ajzen, 1991; Gino & Pisano, 2008), and can be evaluated using vignette study designs (or nonconsequential decision making, in Lonati et al., 2018, and addressed in their fourth "commandment"). Ultimately, the appropriateness of such design decisions comes down to the purpose of the research; for research intended to develop theory, it is reasonable to use "more artificial, stylized scenarios," but where the research purpose involves testing theory to better understand real behavior, then enhanced realism undeniably aids in that effort (Morales et al., 2017, p. 474). A quick perusal of Lonati et al. (2018) may leave the reader with the impression that they advocate a categorical ban on vignette studies; while the authors do offer some nuance on the topic, given the increasing prevalence of vignette studies in OM and the value they stand to offer, we find it useful to expand on this discussion. As with any other research method, there are best practices to be adhered to when conducting a vignette study (see Aguinis & Bradley, 2014; Rungtusanatham et al., 2011; Weber, 1992).

Correctly designed vignette experiments situate participants in an operational scenario, or a storyline, that is carefully crafted to realistically depict the problem setting. The vignette consists of baseline information about the setting that is consistent across all treatments (a common module), and manipulations of the independent variable conveyed by different versions of the scenario (experimental cues modules) that are randomly distributed to participants by treatment (Rungtusanatham et al., 2011).[3]

Vignette studies have a strong history of use in other disciplines, where they are used for their ability to balance the challenges regarding internal versus external validity (Aguinis & Bradley, 2014). Benefits of vignette studies are that they overcome weak external validity of traditional experiments, cover a broad range of relevant factors, and increase "the generalizability of context specific results" (Atzmüller & Steiner, 2010, p. 137).

Despite these benefits, there are also trade-offs to consider. First, while one benefit is in introduction of context comparable to what one might experience in "real life," there may be situations in which the context is too dense or broad in scope for a vignette study to adequately capture (Lohrke et al., 2010). A vignette needs to contain the information essential to understanding the context, or it may lead to a situation where the participant projects their own experiences or knowledge to fill in the gaps (Wason et al., 2002). A second challenge involves the effective manipulation of variables. Manipulations must be salient to the participant, yet in their exposition it is important to protect against framing effects (Wason et al., 2002). Different treatments should be as similar as possible, to avoid potential confounds in the experiment (Rungtusanatham et al., 2011).

In some fields, such as marketing, there is a current and concerted effort to improve the realism of vignette studies where it is a fit to the research question (Morales et al., 2017). Efforts to enhance realism can be particularly useful when testing theory, for example, but may be less pertinent when research is developing new theory (Morales et al., 2017). Mechanisms for increasing realism can involve the use of video or audio clips in vignette studies, as briefly pointed to in Lonati et al. (2018). These techniques can help increase participant engagement (Caro et al., 2012) and the ecological validity of vignette studies (Bateson & Hui, 1992). For example, Victorino et al. (2013) enhanced the realism of the scenarios used in their OM research examining service scripting by creating videos of service encounters (see also Seawright & Sampson, 2007, for an example of a video vignette methodology applied to waiting lines). Technological advancements are even facilitating the use of virtual reality studies for full-immersion, multi-sensory experiences (Aguinis & Edwards, 2014). Another design approach for enhancing participants' engagement involves opportunities to seek out or probe for additional information, such as conducting internet searches (Caro et al., 2012). Yet, even these design choices are not free of trade-offs; as studies become more immersive, there are more opportunities for confounding effects, increased costs, and logistical challenges. Victorino and Dixon (2016) provide excellent methodological guidance for the development of video experiments.

As Croson et al. (2013, p. 4) pointed out: "Using [out-of-task psychometric measures] as correlates in the context of experiments or surveys is a technique that will enrich the field of behavioral operations". There is much to be learned from studies more deeply exploring the processes of judgment and decision-making, and one of the prominent techniques for doing so is through vignette experiments. Through the guidance offered in this section, we hope to continue to see high-quality, well-designed vignette studies in the OM space.

## 3 | CONTRIBUTIONS OF OM RESEARCH FROM MULTIPLE PERSPECTIVES

While in Section 2 we focused on each discussion point (demand effects, incentive alignment, deception, sample issues, and context-rich vignette experiments) independent of one another, it is also useful to think about how trade-offs occur across designs. Because OM is an applied field, our research questions tend to depend heavily on context, which implies that participants in experiments may need to have work experience. It makes little sense to incentivize managers with money, as the quantities are generally not sufficiently motivating to get them to leave work and travel to a lab. Such participants are more likely to be motivated with access to study results through a white paper. They may also be willing to invest in the experiment even if they perceive it as nonconsequential if they are convinced that it will result in the generation of useful knowledge. Vignette studies may thus yield important insights in the OM context even in the absence of perfectly aligned incentives.

As a further, more detailed example of how these research design considerations trade off in practice, we consider two experiments in our field that on the surface appear to be highly related, yet rely on fundamentally different design choices to explore their research question. Each of the following experiments explore individuals' perceptions of value for products. However, a more detailed exploration of these experiments shows that they rely on different foundations in their analysis to appropriately match their research design to their research question.

Agrawal et al. (2015) investigated the typical customer's perceived value of original manufactured goods when refurbished versions are available from either the original manufacturer or a third party. The research context is a purchasing decision where a general consumer pays for the product. Agrawal et al. (2015) matched the context to the research question by using a sampling of data from MTurk, presenting simplified choices, and

paying each respondent $1 and a chance to receive a decision-based incentive of $200 based on their specific choices in the experiment. This is an appropriate research sample, as their question focuses on general population purchasing behaviors, and MTurk is shown to be reasonably representative of the general population (Goodman et al., 2013). Had they used a more targeted sample (e.g., CEOs or students), their findings would be weaker due to decreased generalizability.

The Agrawal et al. (2015) experiment presents limited context, describing primarily the choice between two products. Again, this is the correct approach because the research question is not highly dependent on the context and the choice is easy for participants to understand. Adding additional superfluous details in the experiment would detract from the strength of their manipulation. The use of a decision-based incentive is correctly matched in the experiment to the incentives in reality, where consumers benefit directly from purchases in the form of the goods they select. Had they selected a flat-rate incentive for participation, the findings would be weaker. The authors' selection of sample, context, incentive, and overall experimental design is clearly based on their research question of interest: selecting a general consumer sample, matching a simplified consumer purchasing environment, using incentives since they are connected to the research question of valuations, and using conjoint analysis with limited context outside of the choices presented.

Bendoly and Cotteleer (2008) investigated managerial perceptions of value for different characteristics of enterprise resource planning systems. The research context is a manager's purchasing decision where a firm pays for the product. Bendoly and Cotteleer (2008) matched the context of the research question by using a targeted sample of managers asked to review written cases, and then to provide their own evaluations of the relative usefulness of communication capabilities of enterprise resource planning systems specifically in those contexts. This is the correct research sample, because the population of interest is highly focused. Had Bendoly and Cotteleer (2008) instead used student participants or a general MTurk sample, unfamiliar with the real-world complexity related to the task, their results would be weaker. Specifically, the sampling strategy/population of interest increases the generalizability of their results. Incentives are not discussed, as they are not relevant for the choice. Unlike the consumer's purchasing decision, managers typically do not receive pay—or receive direct incentives for—for these types of organizational purchases; any such rewards are typically indirect. If Bendoly and Cotteleer (2008) had used decision-based incentives, it would have weakened their findings as connecting incentives to the outcome variables would likely increase demand effects by leading the participants to simply guess what answers would generate greater rewards (e.g., choosing the most/least expensive without consideration for their organizational needs).

Despite superficially similar research questions, these efforts require different approaches regarding sample selection, use of hypothetical scenarios, depth of context provided, and decision-based incentives. If the authors had applied different design choices, then their experiments would decrease in both rigor and relevance. For example, if Bendoly and Cotteleer (2008) had used a generalized online sample or students as opposed to managers, abandoned the rich context of a case study in favor of simple context-free choices, or rewarded participants financial compensation based on the decisions they

**TABLE 2** Effective use of research design decisions

| | Agrawal et al. (2015) | Bendoly and Cotteleer (2008) |
|---|---|---|
| Research question | How do *consumers* value product characteristics for *personal purchases*? | How do *managers* value product characteristics in *organizational purchases*? |
| Incentives | Decision-based incentive, matching the personal purchase scenario | No decision-based incentive, matching the organizational purchase scenario |
| Context | Limited context is presented to focus on the primary choice of refurbished products | Detailed context is provided to ground the experiment in reality |
| Sample | MTurk, representative of the general population | Managers, representative of the focused target population |
| Deception | No deception was applied; it was not necessary to suggest to participants details that were not valid in the context of the experiment | No deception was applied; it was not necessary to suggest to participants details that were not valid in the context of the experiment |

made, both the internal and external validity of the research, and thus its potential contribution, would be substantially lessened. Similarly, had Agrawal et al. (2015) used a narrow target population rather than a general sample, added the irrelevant context of a case study, or did not provide decision-based incentives these choices would have diminished the value of their work. Table 2 compares how these two experiments make research design decisions based on the framework presented in Table 1.

Other well-designed research experiments might take a combination of multiple research designs and methods. For example, Sommer et al. (2020) investigated how individuals and teams explore complex problems. The approach is inherently multi-disciplinary as it employed a combination of normative modeling and experiments, and used a variety of research design choices bridging both aspects of the key issues identified in Tables 1 and 2. The research context was intentionally simple, as the research question was on the search process, not the actual decisions being made. To enhance the realism of the search process, the authors provided live feedback from a market analyst. Rather than being an individual providing feedback as it is presented in the simulation, the feedback was done using artificial intelligence (i.e., a confederate) that provided market analysis based on actual experimental performance and decisions made. This design choice was well justified in the research article. Even though the experimental task was a highly simplified context, the authors also incorporated the use of vignettes as part of the pretraining exercise to verify the efficacy of the student sample. The student sample was also justified, given that for the research context, "students are expected to be fairly homogeneous with respect to task-specific competencies and knowledge" (Sommer et al., 2020, p. 10). The ability to make research design choices across the multiple epistemological backgrounds of operations research enhances experimental validity.

## 4 | CONCLUSIONS

The OM field draws on multiple foundations, and those foundations do not always agree on methodological "best practices." Lonati et al. (2018) started an important discussion in our field by provocatively presenting a set of "ten commandments" for experimental research (Table 1, p. 20). While the Lonati et al. (2018) piece provides experimental guidance fitting to certain research agendas, questions have arisen concerning whether and how exactly to implement some of the points that it makes, and how to best address trade-offs in the design of behavioral experiments. Questions have also arisen concerning how to apply these concepts in OM. In this note, we have elaborated on the design choices surrounding demand effects, incentives, deception, sample issues, and the use of vignettes. We have carefully depicted the trade-offs inherent in making design decisions in experimental research. Consistent with the long-standing perspective of McGrath (1982), there is no single "best" method; we agree that it is very difficult to lay down a set of guidelines that fairly and adequately addresses different research foundations. We suggest that the path forward is an acknowledgement and appreciation of these differences. As eloquently stated by Croson (2005, p. 145) "Each researcher needs to make their own methodological decisions based on the objectives of the experiment, the methods currently used in their field, and the audience they wish to address." We can all likely agree that OM is a cross-disciplinary field; we build on the foundations of other disciplines to inform our own work (Bendoly et al., 2010; Gino & Pisano, 2008). Weaving the foundations of these disciplines into rigorous research requires that we as a community invest together to explore and understand the nuances so that we can produce actionable research efficiently, and so that we can provide guidance to JOM authors and review teams that maximizes the value of their investments.

## ACKNOWLEDGMENT

The authors would like to give thanks to the following individuals who provided their time, feedback, and support to this effort: Herman Aguinis, John Aloysius, Daniel Bachrach, Kenneth Boyer, David Cantor, Aravind Chandrasekaran, Thomas Choi, Rachel Croson, Jonathan de Quidt, Jan Fransoo, John Gray, Paul Green, Robert Handfield, Susan Helper, Manpreet Hora, Tim Kraft, Xenophon Koufteros, Kevin Linderman, Robert Lount, Anant Mishra, Anand Nair, Erik Peterson, Rebecca Walker Reczek, Kenneth Schultz, Tobias Schoenherr, Enno Siemsen, Brad Staats, John Sterman, and Morgan Swink. The authors also would like to thank the members of our anonymous review team, who devoted much time and effort helping us to structure and refine this note. Everyone's contributions certainly led to a stronger and more useful contribution.

## ORCID

*Stephanie Eckerd* https://orcid.org/0000-0002-9996-4752

*Scott DuHadway* https://orcid.org/0000-0002-3718-7871

*Elliot Bendoly* https://orcid.org/0000-0002-0158-8403

## ENDNOTES

[1] For example, in the Partisan News Experiment, participants in the "weak" condition receive the following instructions: "You will now be asked to consider some hypothetical (not real) online news items and to indicate which news item you would most prefer to read. The purpose of this exercise is so we can measure whether the news outlet offering an article influences how likely people are to read the article." In the "strong" (explicit) condition, the last sentence is modified to: "The purpose of this exercise is so we can measure whether people are more likely to choose a news item if it is offered by a news outlet with a reputation of being friendly toward their preferred political party." (Mummolo and Peterson, Table 2, p. 522).

[2] It is worth clarifying here that social desirability bias and demand effects are separate risks to experimental validity. Social desirability bias is to answer surveys in a way that is viewed favorably by others, while demand effects represent changes in experiment behavior associated with treatment effects.

[3] An example vignette from Joshi and Arnold (1997) is provided in Appendix A to clearly show baseline and experimental cue modules. Although having originated in marketing, this baseline scenario and context has been replicated across numerous disciplines, including OM (Ganesan et al., 2010; Ro et al., 2016; Su et al., 2017).

## REFERENCES


Abbey, J. D., Kleber, R., Souza, G. C., & Voigt, G. (2019). Remanufacturing and consumers' risky choices: Behavioral modeling and the role of ambiguity aversion. *Journal of Operations Management*, *65*(1), 4–21.

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53*, 63–70.

Agrawal, V. V., Atasu, A., & Van Ittersum, K. (2015). Remanufacturing, third-party competition, and consumers' perceived value of new products. *Management Science*, *61*(1), 60–72.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351–371.

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, *51*(1), 143–174.

Aguinis, H., & Lawal, S. O. (2012). Conducting field experiments using eLancing's natural environment. *Journal of Business Venturing*, *27*(4), 493–505.

Aguinis, H., & Lawal, S. O. (2013). eLancing: A review and research agenda for bridging the science–practice gap. *Human Resource Management Review*, *23*(1), 6–17.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–555.

Ariely, D., & Norton, M. I. (2007). Psychology and experimental economics: A gap in abstraction. *Current Directions in Psychological Science*, *16*(6), 336–339.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology*, *6*, 128–138.

Bachrach, D. G., & Bendoly, E. (2011). Rigor in behavioral experiments: A basic primer for supply chain management researchers. *Journal of Supply Chain Management*, *47*(3), 5–8.

Ball, G. P., Shah, R., & Donohue, K. (2018). The decision to recall: A behavioral investigation in the medical device industry. *Journal of Operations Management*, *62*, 1–15.

Barrera, D., & Simpson, B. (2012). Much ado about deception: consequences of deceiving research participants in the social sciences. *Sociological Methods and Research*, *41*(3), 383–413 Published: AUG 2012.

Bateson, J. E., & Hui, M. K. (1992). The ecological validity of photographic slides and videotapes in simulating the service setting. *Journal of Consumer Research*, *19*(2), 271–281.

Bendoly, E., & Cotteleer, M. J. (2008). Understanding behavioral sources of process variation following enterprise system deployment. *Journal of Operations Management*, *26*(1), 23–44.

Bendoly, E., Croson, R., Goncalves, P., & Schultz, K. (2010). Bodies of knowledge for research in behavioral operations. *Production and Operations Management*, *19*(4), 434–452.

Bickman, L., & Rog, D. J. (Eds.). (2008). *The SAGE handbook of applied social research methods*, Thousand Oaks, CA: Sage publications.

Cadsby, C. B., Song, F., & Tapon, F. (2016). The impact of risk-aversion and stress on the incentive effect of performance-pay. *Experiments in Organizational Economics*, *19*, 189–227.

Camerer, C. (1997). Rules for experimenting in psychology and economics, and why they differ. In W. Albers, W. Güth, P. Hammerstein, B. Moldovanu & E. van Damme (Eds.), *Understanding Strategic Interaction* (pp. 313–327). Heidelberg, Berlin: Springer. https://doi.org/10.1007/978-3-642-60495-9_25

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1–3), 7–42.

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*, Boston, MA: Ravenio Books.

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.

Caro, F. G., Ho, T., McFadden, D., Gottlieb, A. S., Yee, C., Chan, T., & Winter, J. (2012). Using the internet to administer more realistic vignette experiments. *Social Science Computer Review*, *30*(2), 184–201.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8.

Collier, D., Brady, H.E. and Seawright, J., 2004. Critiques, responses, and trade-offs: Drawing together the debate. *Rethinking social inquiry: diverse tools, shared standards*, 195–228.

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

Croson, R. (2005). The method of experimental economics. *International Negotiation*, *10*(1), 131–148.

Croson, R., Schultz, K., Siemsen, E., & Yeo, M. L. (2013). Behavioral operations: the state of the field. *Journal of Operations Management*, *31*(1–2), 1–5.

de Quidt, J., Haushofer, J., & Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, *108*(11), 3266–3302.

de Quidt, J., Vesterlund, L., & Wilson, A. J. (2019). Experimenter demand effects. In *Handbook of research methods and applications in experimental economics*, Northampton, MA: Edward Elgar Publishing.

Dhar, D., Jain, T. and Jayachandran, S., 2018. Reshaping adolescents' gender attitudes: Evidence from a School-Based Experiment in India (No. w25331). *National Bureau of Economic Research*, 1–47.

DuHadway, S., Talluri, S., Ho, W., & Buckoff, T. (2020). Light in dark places: The hidden world of supply chain fraud. *IEEE Transactions on Engineering Management*, 1–14. https://doi.org/10.1109/TEM.2019.2957439.

Eckartz, K., Kirchkamp, O. and Schunk, D., 2012. *How do incentives affect creativity?* CESifo Working Paper, No. 4049, Center for Economic Studies and Ifo Institute, Munich.

Eckerd, S., Hill, J. A., Boyer, K. K., Donohue, K., & Ward, P. T. (2013). The relative impact of attribute, severity, and timing of psychological contract breach on behavioral and attitudinal outcomes. *Journal of Operations Management*, *31*(7/8), 567–578.

Ganesan, S., Brown, S. P., Mariadoss, B. J., & Ho, H. (2010). Buffering and amplifying effects of relationship commitment in business-to-business relationships. *Journal of Marketing Research*, *47*(2), 361–373.

Gino, F., Gu, J., & Zhong, C. B. (2009). Contagion or restitution? When bad apples can motivate ethical behavior. *Journal of Experimental Social Psychology*, *45*(6), 1299–1302.

Gino, F., & Pisano, G. (2008). Toward a theory of behavioral operations. *Manufacturing & Service Operations Management*, *10*(4), 676–691.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, *44*(1), 196–210.

Hauser, D., Paolacci, G. and Chandler, J.J., 2018. Common concerns with MTurk as a participant pool: Evidence and solutions. https://doi.org/10.31234/osf.io/uq45c

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2–3), 111–135.

Hertwig, R., & Ortmann, A. (2008a). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior*, *18*(1), 59–92.

Hertwig, R., & Ortmann, A. (2008b). Deception in social psychological experiments: Two misconceptions and a research agenda. *Social Psychology Quarterly*, *71*(3), 222–227.

Holt, C. A. (1986). Preference reversals and the independence axiom. *American Economic Review*, *76*(3), 508–515.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, *95*(3), 902–912.

Jamison, J., Karlan, D., & Schechter, L. (2008). To deceive or not to deceive: The effect of deception on behavior in future laboratory experiments. *Journal of Economic Behavior & Organization*, *68*(3–4), 477–488.

Joshi, A. W., & Arnold, S. J. (1997). The impact of buyer dependence on buyer opportunism in buyer-supplier relationships: The moderating role of relational norms. *Psychology & Marketing*, *14*(8), 823–845.

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.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, S., Bernstein, M. J., ... Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142–152.

Krasnow, M. M., Howard, R. M., & Eisenbruch, A. B. (2020). The importance of being honest? Evidence that deception may not pollute social science subject pools after all. *Behavioral Research Methods*, *52*(3), 1175–1188.

Lee, Y. S., Ribbink, D., & Eckerd, S. (2018). Effectiveness of bonus and penalty incentive contracts in supply chain exchanges: Does national culture matter? *Journal of Operations Management*, *62*, 59–74.

Lee, Y. S., Seo, Y. W., & Siemsen, E. (2018). Running behavioral operations experiments using Amazon's Mechanical Turk. *Production and Operations Management*, *27*(5), 973–989.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences. *Behavior Research Methods*, *49*(2), 433–442.

Liu, C., Mak, V., & Rapoport, A. (2015). Cost-sharing in directed networks: Experimental study of equilibrium choice and system dynamics. *Journal of Operations Management*, *39*, 31–47.

Lohrke, F. T., Holloway, B. B., & Woolley, T. W. (2010). Conjoint analysis in entrepreneurship research a review and research agenda. *Organizational Research Methods*, *13*, 16–30.

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.

McGrath, J. E. (1982). Dilemmatics: The study of research choices and dilemmas. In J. E. McGrath, J. Martin, & R. A. Kulka (Eds.), *Judgment calls in research*, Beverly Hills, CA: Sage publication.

Meloy, M. G., Russo, J. E., & Miller, E. G. (2006). Monetary incentives and mood. *Journal of Marketing Research*, *43*(2), 267–275.

Morales, A. C., Amir, O., & Lee, L. (2017). Keeping it real in experimental research—Understanding when, where, and how to enhance realism and measure consumer behavior. *Journal of Consumer Research*, *44*(2), 465–476.

Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2), 517–529.

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.

Ortmann, A., & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, *5*(2), 111–131.

Peer, E., Paolacci, G., Chandler, J. and Mueller, P., 2012. Selectively Recruiting Participants from Amazon Mechanical Turk Using Qualtrics. SSRN (May), 2100631.

Porter, C. O., Outlaw, R., Gale, J. P., & Cho, T. S. (2019). The use of online panel data in management research: A review and recommendations. *Journal of Management*, *45*(1), 319–344.

Ried, L., Eckerd, S., Kaufmann, L., & Carter, C.. n.d.. Spillover effects of information leakages in buyer-supplier-supplier triads. Forthcoming in the *Journal of Operations Management*.

Ro, Y. K., Su, H. C., & Chen, Y. S. (2016). A tale of two perspectives on an impending supply disruption. *Journal of Supply Chain Management*, *52*(1), 3–20.

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.

Samek, A., 2019. *To deceive or not to deceive: The debate about deception in economics. Evidence base*. Retrieved from https://healthpolicy.usc.edu/evidence-base/to-deceive-or-not-to-deceive-the-debate-about-deception-in-economics/

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*(6), 1248–1264.

Seawright, K. K., & Sampson, S. E. (2007). A video method for empirically studying wait-perception bias. *Journal of Operations Management*, *25*(5), 1055–1066.

Sharpe Wessling, K., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, *44*(1), 211–230.

Slater, P. E. (1980). *Wealth addiction*, New York, NY: Dutton publication.

Sommer, S. C., Bendoly, E., & Kavadias, S. (2020). How do you search for the best alternative? Experimental evidence on search strategies to solve complex problems. *Management Science*, *66*(3), 1395–1420.

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

Su, H. C., Chen, Y. S., & Ro, Y. K. (2017). Perception differences between buyer and supplier: The effect of agent negotiation styles. *International Journal of Production Research*, *55*(20), 6067–6083.

Ta, H., Esper, T. L., & Hofer, A. R. (2018). Designing crowdsourced delivery systems: The effect of driver disclosure and ethnic similarity. *Journal of Operations Management*, *60*, 19–33.

Thomas, R. W. (2011). When student samples make sense in logistics research. *Journal of Business Logistics*, *32*(3), 287–290.

Tsutsui, K., & Zizzo, D. J. (2014). Group status, minorities and trust. *Experimental Economics*, *17*(2), 215–244.

Verma, R., & Goodale, J. C. (1995). Statistical power in operations management research. *Journal of Operations Management*, *13*(2), 139–152.

Victorino, L., & Dixon, M. J. (2016). Testing service innovation: A methodological review of video experiments. *Service Science*, *8*(2), 234–246.

Victorino, L., Verma, R., & Wardell, D. G. (2013). Script usage in standardized and customized service encounters: Implications for perceived service quality. *Production and Operations Management*, *22*(3), 518–534.

Wason, K. D., Polonsky, M. J., & Hyman, M. R. (2002). Designing vignette studies in marketing. *Australasian Marketing Journal*, *10*(3), 41–58.

Weber, J. (1992). Scenarios in business ethics research: Review, critical assessment, and recommendations. *Business Ethics Quarterly*, *2*(2), 137–160.

Wuttke, D. A., Donohue, K., & Siemsen, E. (2018). Initiating supplier new product development projects: A behavioral investigation. *Production and Operations Management*, *27*(1), 80–99.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*(1), 75–98.

**How to cite this article:** Eckerd S, DuHadway S, Bendoly E, Carter C, Kaufmann L. On making experimental design choices: Discussions on the use and challenges of demand effects, incentives, deception, samples, and vignettes. *J Oper Manag*. 2020;1–15. https://doi.org/10.1002/joom.1128

## APPENDIX

The following vignette is reproduced from Joshi and Arnold (1997). The headings are not visible to the participant. We added the notations in brackets differentiating baseline modules and experimental cues modules. The vignette begins and concludes with baseline (i.e., common) modules. Each participant randomly received one of the two dependence experimental cues modules (low or high), and one of the two relational norms experimental cues modules (low or high).

### Introduction (baseline module)

You are a purchasing manager responsible for the purchase of microchips for a midsize electronic equipment manufacturer. Microchips are an important component for the equipment that you manufacture; therefore, they need to be purchased on a regular basis. You have one existing supplier for this component.

### Low dependence (experimental cues module)

As purchasing manager responsible for microchips, you find yourself in a situation wherein it is not difficult for you to find a suitable replacement for the existing supplier. If you decide to stop purchasing from this supplier,

you could easily replace their volume with purchases from alternative suppliers. There are many competitive suppliers for microchips and you can switch to them without incurring any search costs. Switching suppliers is not going to have any negative effects on the quality or design of the equipment that you manufacture. Your production system can be easily adapted to use components from a new supplier. The procedures and routines that you have developed are standard and they are equally applicable with any supplier of this component. The skills that your people have acquired in the process of working with the supplier can easily be changed to fit another supplier's situation. You can therefore terminate your relationship with your present supplier without incurring any costs.

**High dependence (experimental cues module)**

As purchasing manager responsible for microchips, you find yourself in a situation wherein it is difficult for you to find a suitable replacement for the existing supplier. If you decide to stop purchasing from this supplier, you could not easily replace their volume with purchases from alternative suppliers. There are very few, if any, competitive suppliers for microchips and you cannot switch to them without incurring significant search and verification costs. Switching suppliers is also going to have negative effects on the quality or design of the equipment that you manufacture. Your production system cannot be easily adapted to use components from a new supplier. The procedures and routines that you have developed are unique and hence they are not applicable with any other supplier of this component. The skills that your people have acquired in the process of working with the supplier cannot easily be changed to fit another supplier's situation. You cannot therefore terminate your relationship with your present supplier without incurring significant costs.

**Low relational norms (experimental cues module)**

Both you and your supplier bring a formal and contract-governed orientation to this relationship. Exchange of information in this relation- ship takes place infrequently, formally, and in accordance to the terms of a prespecified agreement. Even if you do know of an event or change that might affect the other party, you do not divulge this information to them. Strict adherence to the terms of the original agreement characterizes your relationship with this supplier. Even in the face of unexpected situations, rather than modifying the contract, you adhere to the original terms. You have an arm's-length relationship with your supplier. You do not think that the supplier is committed to your organization—in fact, you think that if you did not carefully monitor this supplier's performance, they would slack off from the original terms. Above all, you see your supplier as an external economic agent with whom you have to bargain in order to get the best deal for yourself.

**High relational norms (experimental cues module)**

Both you and your supplier bring an open and frank orientation to the relationship. Exchange of information in this relationship takes place frequently, informally, and not always according to a prespecified agreement. You keep each other informed of any event or change that might affect the other party. Flexibility is a key characteristic of this relationship. Both sides make ongoing adjustments to cope with the changing circumstances. When some unexpected situation arises, the parties would rather work out a new deal than hold each other responsible to the original terms. You tend to help each other out in case of unexpected crises. If your supplier is unable to fulfill an order, they recommend an alternative source of supply for the same. Above all, you have a sense that your supplier is committed to your organization and that they work with you keeping your best interests in mind. You see each other as partners, not rivals.

**Conclusion (baseline module)**

Recently, the supplier informed you that they are involved in a labor dispute. Consequently, they are temporarily unable to guarantee on-schedule delivery. This creates some uncertainty for your organization. Delayed delivery of microchips, may, for example, cause problems for your organization in meeting delivery schedules to customers. The supplier has called to get your regular order. Drawing from experience, how would you be most likely to react in this situation? Please rate each of these statements to the extent that they match with your expectation of your reaction.