On doing relevant and rigorous experiments:

Review and recommendations

Sirio Lonati*
University of Lausanne
Faculty of Business and Economics
Department of Organizational Behavior
Internef 623
CH-1015 Lausanne-Chamberonne, Switzerland
sirio.lonati@unil.ch

Bernardo F. Quiroga
School of Management
Pontificia Universidad Católica de Chile.
Campus San Joaquín,
Av. Vicuña Mackenna 4860
Macul, Santiago RM, Chile
bfquirog@puc.cl

Christian Zehnder
University of Lausanne
Faculty of Business and Economics
Department of Organizational Behavior
Internef 612
CH-1015 Lausanne-Chamberonne, Switzerland
christian.zehnder@unil.ch

John Antonakis
University of Lausanne
Faculty of Business and Economics
Department of Organizational Behavior
Internef 618
CH-1015 Lausanne-Chamberonne, Switzerland
john.antonakis@unil.ch

*Corresponding author

In press Accepted Version (19 October 2018)

Journal of Operations Management

ABSTRACT

Although experiments are the gold standard for establishing causality, several threats can undermine the internal validity of experimental findings. In this article, we first discuss these threats, which include the lack of consequential decisions and outcomes, deception, demand effects and unfair comparisons, as well as issues concerning statistical validity (e.g., minimum sample size per cell, estimating variance correctly). We expose each problem, show potential solutions, and bring to the fore issues of relevance of the findings (i.e., external and ecological validity). Thereafter, we take stock of the state-of-the-science regarding validity threats using a representative sample of 468 recent experiments from 258 articles published in top-tier journals. We compare research practices in three fields of study—management, social psychology, and economics, which regularly use experimental research—to operations management, which has recently begun to use the experimental paradigm. Our results underscore the importance for journals and authors to follow what we identify to be best-practice methodological suggestions (i.e., the "ten commandments" of experimental research). We show that—on average—markers of methodological rigor and generalizability positively and significantly predict the citations received by published articles. Finally, given that experiments are infeasible in some settings, we conclude with a brief review of often overlooked quasi-experimental designs, which are useful for generating strong counterfactuals and hence allow making causal claims in the field.

Keywords: Experimental methods; Research methods; Research impact; Causality; Endogeneity; Quasi-experimentation.

"*If … [a new law] disagrees with experiment … [the law] is wrong. In that simple statement is the key to science. It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is—if … [the law] disagrees with experiment it is wrong. That is all there is to it.*" – Richard Phillips (Feynman, 1965, p. 165), Nobel Prize laureate in Physics in 1965

1. INTRODUCTION

How should experiments be made valid, realistic, and relevant? This question is important to answer given the prominence of experimental research in the applied sciences. Since the end of the 1980s, experimental methods became a common tool to understanding behavioral aspects of decision making in the operations management (OM) discipline (Bendoly, Donohue, & Schultz, 2006). The use of experimental methods in this field has continued to grow and behavioral experiments are currently present in many top tier outlets (e.g., see the most recent special issues on experimental and behavioral OM research, Croson, Schultz, Siemsen, & Yeo, 2013; Gans & Croson, 2008). Are these experiments well done and do they add any useful knowledge beyond other methods?

The answers to the above questions are not evident; nor are there explicit standards that authors can turn to for judging the quality of experimental research in this field. In this article, we aim to provide researchers with guidance on how to conduct useful experimental research in the organizational sciences. This endeavor is not immune to difficulties, because different fields have different viewpoints about experimental methodology (see, e.g., Ariely & Norton, 2007; Hertwig & Ortmann, 2001; Ortmann & Hertwig, 2002). These discrepancies can be especially problematic for organizational scholars, given the multidisciplinary nature of their field and the lack of unified theories, methods, and assumptions (Antonakis, 2017; Gino & Pisano, 2008; Pfeffer, 1993). This

fragmentation is also visible in behavioral OM, which initially covered themes and employed methods coming from the experimental economics tradition (newsvendor game and auctions, see, e.g., Croson & Donohue, 2002; Katok, 2011; Schweitzer & Cachon, 2000); however, with time behavioral approaches to OM have embraced psychological and sociological viewpoints and now includes topics as diverse as groupthink (Riccobono, Bruccoleri, & Größler, 2015), cross-cultural differences (Cannon, Doney, Mullen, & Petersen, 2010; Naor, Linderman, & Schroeder, 2010; Ribbink & Grimm, 2014), trust and trustworthiness (Özer, Zheng, & Chen, 2011; Özer, Zheng, & Ren, 2014), fairness concerns (Katok, Olsen, & Pavlov, 2014; Katok & Pavlov, 2013) and linkages between intrinsic motivation and job design (De Treville & Antonakis, 2006; Rodríguez, Buyens, Landeghem, & Lasio, 2016).

In this article we develop methodological guidelines that strike a balance between diverse methodological traditions and hence speak to a wide range of social scientists. Our contribution is unique vis-à-vis articles that have already discussed experimental and behavioral OM research at length (e.g., Bendoly, Croson, Goncalves, & Schultz, 2010; Bendoly et al., 2006; Gino & Pisano, 2008). We review existing pros and cons for different experimental practices, highlighting trade-offs between different choices, and take a clear position where norms across diverse academic fields diverge. Our end product is a synthesis of best-practices into a uniform methodological paradigm that can help guide future experimental work.

Our methodological guidelines are tailored to the needs of OM scholars; however, they equally apply to organizational fields that were traditionally reluctant to use experimental methods and are now calling for more experimental research (Antonakis, 2017; Chatterji, Findley, Jensen, Meier, & Nielson, 2016; Colquitt, 2008; Eden, 2017; P. M. Podsakoff & Podsakoff, 2018). Our synopsis is relevant too for non-experimentalists, because increasing methodological concerns have raised the bar for identifying causally interpretable results in

studies using observational field data in recent years. The issue of endogeneity, which implies that a modelled independent variable is correlated with the disturbance term in a regression model (Angrist & Pischke, 2008; Wooldridge, 2015), is gaining prominence on the radars of reviewers and editors, both in OM (Guide & Ketokivi, 2015; Ketokivi & McIntosh, 2017), in management, and other social sciences (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Chenhall & Moers, 2007; Hamilton & Nickerson, 2003; Roberts & Whited, 2013; Semadeni, Withers, & Trevis Certo, 2014). Whereas ex-post fixes for endogeneity exist, randomized experiments remain the most convincing solution to the causality issue (Aguinis & Edwards, 2014).

Randomized experiments are, however, not a panacea (Shadish, Cook, & Campbell, 2002). Methodological issues such as deception (Hertwig & Ortmann, 2001; Ortmann & Hertwig, 2002), unwanted demand effects (Zizzo, 2010), small sample sizes (Button et al., 2013; Strube, 1991) and other internal and statistical validity threats can undermine the causal knowledge we can gain from experimental research. Moreover, ecological (Dhami, Hertwig, & Hoffrage, 2004; Highhouse, 2009) and external validity issues (Berkowitz & Donnerstein, 1982; Guala, 2003; Guala & Mittone, 2005; Kessler & Vesterlund, 2015; Ketokivi & McIntosh, 2017) can weaken the generalizability and practical relevance of experiments.

In this review, we will discuss all these issues, show how they are oftentimes tightly interconnected, and identify how experiments should be done to ensure rigor but also relevance. To illustrate where we see room for improvement in OM experimental research, we will benchmark it with experimental research in three close neighbor fields: Management, social psychology, and economics. We document the methodological practices across fields and illustrate empirically that researchers have strong reasons to adopt good practices: As our results

will show, markers of methodological rigor and of potential generalizability predict citations of papers across the academic fields we reviewed.

We structure our article as follows. We first explore the main threats to internal validity, differentiating between ex-ante threats (i.e., faults in the experimental design) and ex-post threats (i.e., incorrect statistical analysis of the data). Then we discuss the most pressing external and ecological validity concerns from an organizational studies viewpoint. Next, we provide a quantified and representative review of the existing experimental research in several social sciences in the past decade. Finally, we propose a toolbox of under-utilized statistical methods that allow researchers to mimic experimental techniques that ensure causal identification with field data when designing an experiment is not possible. We conclude with a summary of our methodological guidelines—the "ten commandments" of experimental research—and thus provide researchers with a portable go-to framework to make experimental research more rigorous and relevant.

2. HOW TO DESIGN AN EXPERIMENT: INTERNAL VALIDITY ISSUES

Experiments can be defined as a type of "study in which an intervention is deliberately introduced to observe its effects" (Shadish et al., 2002, p. 12) on one or multiple dependent variables. That is, the experimenter establishes that an independent variable *x* has a causal impact on a dependent variable *y* by manipulating *x*. In this way *x* is exogenous and varies independently of other causes (i.e., *e*, the disturbance) of *y*.

However, correct causal inference requires clean manipulations and tight control over the experimental environment; otherwise the researchers run the risk of incurring an internal validity error (Campbell, 1957; Shadish et al., 2002). In experimental research, most critical internal validity problems arise if the treatment group and the control group differ not only in one

dimension, that is, the dimension of the intervention, but also in a confounding factor. At one extreme, when the experimental conditions are not randomly assigned, both observable and unobservable variables might be real causes of an observed treatment effect, because the unobserved variables, pooled in *e*, are not equally distributed across treatments. However, in most cases confounds are subtler and are produced by "unfair comparisons" (W. H. Cooper & Richardson, 1986): The constructs employed to contrast two or more theoretical propositions are operationalized with dissimilar care and realism and are, ultimately, not comparable.

2.1. Demand effects

One of the most dangerous types of unfair comparisons in experimental research is unwanted demand effects. Humans are inherently social and their behavior is crucially determined by their interpretation of the social situation they are facing (Hertwig & Ortmann, 2001; Zizzo, 2010). Experimental subjects are no different and construct their beliefs about the researcher's expectations: As a result, the subjects' behaviors might be altered by so-called "demand effects" (also known as demand characteristics). These effects concern the "changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior" (Zizzo, 2010, p. 75). Thus, demand effects are explicit or implicit indications present in the experimental situation—in the instructions of the experiment or in the experimenter's behavior—that can systematically bias participants' response to a treatment. Demand effects represent a central problem for experimentalists as soon as the demand effects correlate with the treatment status. These effects create a confound because treated and untreated subjects differ not only in the treatment status; they differ too in the demand effects they face (Sturm & Antonakis, 2015).

There are several features of experimental settings that facilitate the emergence of demand effects. First, the hierarchical relationship existing between experimenter and

participants might push the subjects to naturally conform to her or his authority (see the pioneering studies on the natural obendience to authority of Milgram, 1963). Second, demand effects lie in the need for social approval and social desirability (Fisher, 1993). This issue is prominent in most experiments, because the behavior of participants is—at least indirectly—observable by the experimenter and possibly by other participants. Third, some elements of the experimental design might be so salient that they might end up revealing the purpose of the experiment to participants, making the hypotheses tested apparent. For instance, if participants are asked either to "maximize their gains" or to "be fair and equitable" in a negotiation simulation, they will surely form a clear opinion of the topic studied. Demand effects may also occur if subjects are prompted as to what is the obvious choice in a particular context. This risk is maximized if subjects are exposed to all or a subset of the manipulations and asked to choose the most appropriate option: For example, asking subjects to choose the most appropriate looking wartime leader from a set of different versions of the same face, ranging from very dominant to very docile will very likely trigger demand effects, because it may become obvious to subjects what the experimenter manipulated (Bøggild & Laustsen, 2016). Lastly, strong demand effects might emerge if experimenters are not blind to treatments and interact directly with the experimental subjects. For instance, a recent replication study of the "power pose" effect (see Carney, Cuddy, & Yap, 2010) has shown that the interaction between experimenter and participants had likely led to an inflation of the original effect sizes, contributing to the replication failure of the apparent effect (Ranehill et al., 2015).

Demand effects generate unfair comparisons because they deliberately hinder one experimental condition in favor of another one. A clear example can be seen in the so-called "recall task" (e.g., Galinsky, Gruenfeld, & Magee, 2003), which is often used to study felt power in psychology. Treated subjects are asked to write an essay about a certain topic (e.g., "write

about a situation where you felt powerful"), whereas participants in the control condition write about a presumably neutral topic (e.g., "write about what you did yesterday"). Apparently, this manipulation should activate a "power prime" in treated subjects making them "feel" power. However, this type of treatment is problematic for two reasons. First, this power manipulation provides treated participants with strong clues about the purpose of the experiment (Sturm & Antonakis, 2015). Moreover, it provides asymmetric clues to the participants: Whereas treated subjects may easily see how the experimenter is trying to study power, the participants in the control condition can hardly form any opinion on the topic studied. As a result, the researchers are not comparing a "high power" group with a "no power" group, but rather a group that is also experiencing a strong demand effect with a group facing no demand at all. Thus, it is very important that researchers not only think very carefully about the implementation of the treatment that they are interested in, but also about the design of their control group. A correct control group should expose experimental subjects to a baseline level of the manipulated variable, so to really observe subjects' behaviors in the absence of the treatment, yet it should *hold constant any demand effect* already present in the treatment condition (e.g., see Study 2 in Bendahan, Zehnder, Pralong, & Antonakis, 2015).

A second mode of unfair comparison hinges on incorrectly specified comparisons between two levels of the same dependent variable. For instance, assume a researcher exposes subjects to a vignette describing either of two possible business scenarios. In the first treatment, a corporation is ostensibly described as environmentally friendly and responsible towards its stakeholder ("high corporate social responsibility" treatment); the second version of the vignette depicts the same company as environmentally damaging and irresponsible ("low corporate social responsibility" treatment). The researchers assess the intentions to buy from each corporation of the experimental subjects as dependent variable. Does contrasting these two treatments identify

the causal effect of corporate social responsibility on subjects' intentions to buy? Not really. The researchers do not know whether any difference in the response of the experimental subjects is due to an aversion towards the irresponsible company, a preference for the responsible one or a mixture of the two explanations. Moreover, the researchers can hardly exclude whether any difference was triggered by two demand effects that are equally blatant, yet are of opposite direction.

The perils of this type of unfair comparison cause a serious internal validity threat, yet are largely overlooked in the social sciences (cf. Schaerer, du Plessis, Yap, & Thau, 2018). For example, Crede (2018) has also recently pointed to the critical flaws of most scholarly articles on the psychological effects of the power pose, which compare expansive and contractive poses, yet do not benchmark them with a neutral pose. Such treatment comparisons are akin to giving to a treated group a medicine (M) that is supposed to cure an illness and to give to the control group a poison (P) that is supposed to aggravate the health status of the patients. That longevity, *y*, is higher in M than P can arise because of a number of reasons, including that M has no effect at all and that P has a detrimental effect, or worse that both treatments reduce longevity with P having a greater effect. Hence, researchers should avoid comparisons that mix up two effects and prefer experimental designs that specify a clear control group representing a neutral level of the manipulated variable and of the related demand characteristics, and/or give different degrees (dosages) of the treatment.

2.2. Non-consequential decision making: Hypothetical studies and missing incentives

Behavioral sciences are "disciplines dealing with the subject of human *actions"* (emphasis added, The Editors of Encyclopædia Britannica, 2016). However, behavioral scientists oftentimes conduct "hypothetical studies" that do not elicit actual behaviors (like vignettes, self-

reports, questionnaires, hypothetical choice scenarios, see, e.g., Baumeister, Vohs, & Funder, 2007; Furr, 2009). Hypothetical studies, similarly to standard questionnaires, suffer from the well-known problems of self-reported measures (Donaldson & Grant-Vallone, 2002; P. M. Podsakoff & Organ, 1986); it is uncertain whether the self-reported intentions will ultimately translate into real actions. Crucially, self-reported assessments can be misrepresented at no cost and are, therefore, particularly prone to demand effects, cheap talk, or socially desirable responses; such hypothetical situations usually do not have real-world tradeoffs or payoffs (Antonakis, 2017). On the contrary, studying actual, consequential behaviors and having tradeoffs and real consequences for the subject (instead of hypothetical self-reports of alleged behaviors) mitigates the impact of demand effects and social desirability bias. We do not, of course, argue that incentivized studies are completely shielded from demand effects. However, in incentivized studies taking an action that is expected to please the researcher or corresponds to the socially desirable choice is at least costly if this action is not the participant's preferred one (e.g., in terms of payoff maximization). We therefore encourage researchers in all social sciences to design experiments in which participants' choices have real consequences.

How difficult it is to implement consequential choices in an experiment depends on many factors. In field experiments, for instance, the researcher typically does not need to explicitly incentivize any action, because choices are naturally consequential (e.g., the subject has a job to do). In the laboratory, in contrast, experimenters need to design decision-making environments to be consequential (Smith, 1976). In many studies, the easiest way to do so is by directly linking participants' decisions to the monetary compensation that they receive for participating in the experiment. Such financial incentives motivate participants to be attentive and focused on the experiment (Antonakis, 2017; Hertwig & Ortmann, 2001) and clarify the decision situation in which participants find themselves (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001; Smith

& Walker, 1993). In many settings that are of interest to OM researchers, it is clear how to set incentives. For example, Chen-Ritzo, Harrison, Kwasnica, and Thomas (2005) designed an experiment where subjects needed to formulate bids to provide a product. Bids were evaluated in three dimensions: Ask price, delivery lead time, and product quality. Whereas subjects did not know directly how their bids modified the buyer's utility, they knew in advance how much it would cost to provide the auctioned good—given the combination of attributes offered—in the case they would win the auction. At the end of the game, subjects were paid in dollars based on how much profit they made (i.e., bid price minus cost whenever they were the winners of the auction) based on their decisions. In such cases, not making participants' compensation contingent on their performance would seem rather unnatural.

Is our conclusion that un-incentivized experiments should never be run? No, yet we advise researchers to be extremely mindful about their low internal validity and we urge the profession to take the proper countermeasures. On the other hand, it is still appropriate to use vignettes if they are completely purged from demand effects and are based on a clearly quantifiable outcome, linked to a real-world analog (e.g., choosing who among two individuals looks the more competent, where unbeknown to the subjects the individuals depicted competed for a political election, see Todorov, Mandisodza, Goren, & Hall, 2005). Also, because it is difficult to make vignette studies consequential via monetary incentives, we hope to see more researchers employing non-traditional stimulus materials, which can enhance the psychological realism and the immersion in hypothetical experimental environments (video-based vignettes rather than written ones, virtual reality interactions, see, e.g., Aguinis & Bradley, 2014; Antonakis, 2017; Blascovich et al., 2002; DeHoratius, Gürerk, Honhon, & Hyndman, 2015; Harrison, Haruvy, & Rutström, 2011; N. P. Podsakoff, Podsakoff, MacKenzie, & Klinger, 2013).

Moreover, the use of financial incentives is not the only way to render choices consequential in the laboratory. Depending on the research question and the task that participants face in the experiment, other types of extrinsic motivators (e.g., Boyce, Brown, McClelland, Peterson, & Schulze, 1992; Falk & Szech, 2013), intrinsic motivators (i.e., genuine interest in the task), social structures (e.g., a participant's status in the group), psychological rewards (e.g., praise or recognition), and even biological factors (e.g., hormones) can induce consequences as well (Benabou & Tirole, 2003; Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005; Ryan & Deci, 2000). Hence, experimentalists can, and should, create consequential environments by applying all possible sources of motivation, being however mindful about their potential interplays. For instance, when intrinsic motivation is large, additional extrinsic motivators will not necessarily be effective and might even backfire (Ariely, Gneezy, Loewenstein, & Mazar, 2009; Gneezy & Rustichini, 2000; Grant, 2008). As a result, tasks that are perceived as relevant for self-esteem (e.g., score on an IQ test) might require a smaller or perhaps no formal incentivizing mechanism to be consequential.

Finally, it is important to emphasize that the comments above are targeted at experiments that focus on behaviors as the dependent variable: We do not argue that all self-reported measures elicited in experiments are problematic. For instance, if the dependent variable of interest is an emotion or a perception, it makes little sense to make participants' answers consequential. Moreover, if researchers elicit not only participants' behaviors, but also their beliefs about the behavior of other participants, incentivizing beliefs does not only increase belief accuracy, but may—depending on the situation—also distort actual behavior (Blanco, Engelmann, Koch, & Normann, 2010; Gächter & Renner, 2010). Besides, not all types of behavior are equally prone to social desirability bias (e.g., risk preferences can be measured well without an explicit incentive, see, e.g., Dohmen et al., 2011). However, in most cases researchers are well advised to study

strongly consequential behavior in order to minimize the ambiguity about the correct interpretation of the observed data.

2.3. Demand-inducing manipulation checks

When researchers run experimental treatments that affect participants subjectively (e.g., framing or affective reaction), manipulation checks are often required to ensure that an intervention had the desired effect (Perdue & Summers, 1986; Shadish et al., 2002; Sigall & Mills, 1998). Manipulation checks usually consist of one or multiple direct questions, where subjects are asked to report to which extent they were affected by the treatment. Being based on self-reports, however, manipulation checks have the potential to create unwanted demand effects and should be conducted cautiously, despite their importance.

As a general rule, researchers should not run manipulation checks when the experimental treatment consists of objective variations of parameters (e.g., costs or valuations), frames (e.g., gain vs. loss frame) or structures of a game (e.g., order of movers, group size), because such interventions do not need to be actively interpreted by the experimental subjects. If the manipulation is assumed to alter only a psychological state, however, manipulation checks should always be conducted after the main construct of interest (Antonakis, 2017; Kidd, 1976; Perdue & Summers, 1986; Sigall & Mills, 1998); though, ideally, they should be done on a separate sample for the sorts of experiments social scientists normally do. If assessed before the main outcome of interest, manipulation checks provide clear cues on the experimental treatment and indirectly signal how the experimenter hopes participants to behave (Sturm & Antonakis, 2015). They draw the attention of the participants to subtle elements of the experimental situation (i.e., the treatment) that were not so salient prior to the manipulation check and might affect their decision-making.

Performing manipulation checks after the main experiment partially solves this issue, yet such a procedure is not riskless either: If manipulation checks clearly reveal the purpose of the experiment, participants might report biased responses in the manipulation check, because they internalized the demand of the experimenter (Sturm & Antonakis, 2015). Of course, participants may respond in a cognitively consistent matter on the manipulation check following how they responded to the treatment, which is reminiscent of the common-method/source problem (P. M. Podsakoff, MacKenzie, & Podsakoff, 2012). Hence, the optimal solution to this problem is to use an "external manipulation check," that is, test whether the manipulation was effective using a different—yet comparable—sample (Antonakis, d'Adda, Weber, & Zender, 2014; Bendahan et al., 2015). Moreover, unless the manipulation check is objective and not subject to psychological bias, it is also a good idea to elicit manipulation checks for multiple dimensions and/or to also include filler questions (i.e., questions that are not related to the manipulation check), in order not to provide precise clues about the experimental hypothesis. For sake of transparency, scholars should also report all the manipulation checks they conducted in an experiment, and thus disclose all possible sources of demands in their study (Abbey & Meloy, 2017).

Moreover, manipulation checks should not be confused with attention/remembering checks (Abbey & Meloy, 2017). First, a subject correctly recalling the manipulation does not mean that the intervention successfully induced the particular physiological or psychological state of mind hypothesized. Second, if an intervention is inconsequential and lacks psychological realism (i.e., what we usually find in most hypothetical vignette studies), it will induce no specific mental state, nullifying the scope of the manipulation check. In this case, passing the manipulation check will—at best—indicate that the subjects internalized the experimental demand and responded accordingly.

2.4. Deception: Lack of control due to lack of credibility

Deception is a different type of experimental confound that refers to the purposeful act of lying, misleading, or misrepresenting the purpose of the experimental study to the subjects (Antonakis, 2017). Typical examples of deception include lying about the consequence of decisions (e.g., untruly claiming that a particular choice has a monetary consequence for another participant), misrepresenting the identity of other participants (e.g., presenting confederates as normal participants, or claiming that a subject is playing a game with another subject when in reality decisions are taken by a computer), and providing false feedback or distorted information to participants (e.g., misreporting decisions of other participants, or giving false feedback on a subject's performance). Whereas social psychologists and management scholars customarily make use of deception (see, e.g., Ariely & Norton, 2007; Hertwig & Ortmann, 2001), the use of deception is strongly discouraged in economics (Davis & Holt, 1993; Hertwig & Ortmann, 2001; Ortmann & Hertwig, 2002)[1]. Note that failing to specify the treatment conditions or not explaining the purpose of the experiment is not considered deception, but "*obfuscation*" (Antonakis, 2017) and is a broadly accepted practice. In a laboratory experiment, for instance, if a researcher explains to subjects that they will be "matched with another person in the room", s/he is not technically deceiving the participants if subjects can—in reality—be re-matched only within a portion of the people sitting in the room.

We take a very clear position against deception. From our point of view, the main arguments against deception are twofold. Ethically, we find it bad role modeling when scientists institutionalize lying, even if temporarily. Moreover, although most participants do not

[1] Paragraph 8.07 of the "Ethical Principles of Psychologists and Code of Conduct" of the American Psychological Association defines clear standards for the use of deception in psychological research. Among other things, the paragraph states that deception is only to be used if effective nondeceptive alternative procedures are not feasible. Unfortunately, however, the profession seems to have agreed to interpret this rule in a very lax manner, because given its prevalence, it seems that many psychologists use deception out of convenience rather than necessity.

experience much harm by being deceived, and even if psychologists routinely debrief their participants after the study, the use of deception poses a serious threat to one of the main pillars of experimental research: Control (Jamison, Karlan, & Schechter, 2008; Ortmann & Hertwig, 2002). In order to establish a causal effect of an intervention and maintain full control over the experimental information environment, it is of outmost importance that the participants take the information that the researcher provides at face value. On the contrary, a participant who has been deceived before will become increasingly suspicious in subsequent experiments about the instructions provided or the stated purpose of the experiment. In addition, the loss of control may not only affect participants who have been directly deceived. Once the word is out that a particular laboratory uses deceptive methods, this information will spread and may contaminate the whole subject pool (Jamison et al., 2008; Ortmann & Hertwig, 2002). Thus, the simple knowledge that a laboratory, or, at worst, an entire research field, systematically lies to participants might be enough to call the validity of experimental research into question.

However, despite the fact that we strongly advise against the use of deception, we recognize that there are different levels of severity of the problem. For instance, let us assume that some experimental subjects are invited to a bargaining experiment. They are told that they will negotiate with another participant and will be paid according to their performance in the task. In reality, however, they are paired with a computer playing a pre-recorded strategy and are paid with a flat fee. After the experiment, they are debriefed and learn that none of their decisions mattered at all. How much attention or effort will such participants give to the instructions that they receive in the next experiment? We could argue that this experience will affect the level of effort exertion significantly in the future. Conversely, other forms of deception are ethically innocuous, yet could be easily avoided. For example, in several system dynamics papers, subjects are told they will play a Beer Game for X periods, whereas they play the game for $Y < X$ periods

(e.g., Sterman, 1989). There may be good reasons for why subjects should not know when exactly the experiments will end. However, instead of lying to subjects, researches could simply use the "continuation probability" method (see e.g., Dal Bó & Fréchette, 2011 for an example).[2] This method is better suited than deception for at least two reasons. First, subjects have the correct information when they make their decisions (a false belief that there are a finite number of periods may also have an impact on behavior in non-final periods). Second, the method completely avoids the negative spillover effects of deception on future experiment. Alternatively, the number of repetitions of a game could simply remain undisclosed; however, this approach creates uncertainty and—depending on the game—may render predictions and interpretations difficult. The lack of a quantifiable risk of termination in a particular round reduces experimental control, because it generates unpredictable subjective beliefs (i.e., participants might form different expectations about the ending round; moreover, the experimenter has no control over their belief formation process).

The concerns outlined above are most relevant for laboratory research, whereas in many field settings, the reputational problem is less relevant, because most subjects may participate only once in an experiment and communication with potential participants of other field experiments may be unlikely. Moreover, field experimentation often involves additional logistical and budgetary limitations (e.g., management vetoes, logistical and strict ethical constraints, see Eden, 2017), which makes it more complex to avoid deception by design. Whereas we still advise avoiding deception whenever possible, we admit that there are exceptional cases in which the importance of the research may justify the use of a reasonable

[2] There are different versions to implement continuation probabilities. If it is important to have the same number of periods in different sessions (for comparability reasons), the realization of the random process can be determined once for all sessions together.

amount of deception in the field (for an example of obfuscation in the field, see Fehr & Goette, 2007; for an example of deception in the field, see Kröll & Rustagi, 2016).

2.5. Threats to statistical validity

So far, we have focused on design practices that ensure the internal validity of an experimental study before and while data is collected. However, equally important is statistical validity, that is, the right choice of method when data is analyzed (Shadish et al., 2002). A plot of the raw data and a simple comparison of treatment means should be present in all experimental papers and would normally provide substantive answers to the research question. However, after having shown the existence of a treatment effect, many scholars want to dig into the causal mechanism behind it, resorting to regression analysis. In those cases, specific statistical cautions are required. Examining in details statistical validity threats would require extensive discussion; thus, we briefly mention five of the main problems that often plague the analysis of experimental data.

[Table 1 about here]

2.5.1. Incorrect inference

Correct inference requires a correct modeling of the residuals in the regression model. If observations are independent, the residuals can be assumed to be a random variable with mean zero, and to be identically and independently distributed across observations. In many experiments, however, the design of the experiment implies that independence across observations is not necessarily given for all data points. For instance, experimental data regularly entails so-called "clusters" within which observations cannot be treated as independent. In clustered data, residuals are independent across clusters, but not within clusters. Data clusters emerge if there is a common factor that affects all observations within the cluster (e.g., individuals taking decisions in a group setting, where they observe the decisions of group

members). Ignoring the cluster structure leads to biased standard errors and, potentially, false statistical significance (Angrist & Pischke, 2008; Bertrand, Duflo, & Mullainathan, 2004; Cameron & Miller, 2015; Wooldridge, 2015).

A possible solution for the clustering problem is to aggregate the data at the relevant level of analysis (e.g., individual, team, or entire experimental session). With such "collapsed" data, serial correlation between observations in the same group does not pose a problem, because heterogeneity across entities is averaged out. This procedure simplifies the problem of comparing treatment and control groups as a (non-)parametric test for differences-in-means across two or more groups. Multi-level random effects models are also a solution, yet they do not automatically model both the heteroscedasticity and the autocorrelation present in the data (Primo, Jacobsmeier, & Milyo, 2007). Moreover, random effects models assume the random effect (the group specific disturbance, usually referred to as the $u_j$term) to be independent from the covariates; if this assumption fails, the estimator is not consistent and modeling what are called the "fixed-effects" in econometrics is required (see, e.g., Antonakis et al., 2010; Bollen & Brand, 2010; Bou & Satorra, 2018; Ketokivi & McIntosh, 2017)

Correct inference requires the use of cluster-robust estimation of the variance[3] (see the seminal work of K.-Y. Liang & Zeger, 1986), which takes into account that residuals may be arbitrarily correlated within clusters (Bertrand et al., 2004). Typical examples of necessary—though sometimes neglected—clustered standard errors include:

1. Interaction among fixed group. In an auction experiment, two buyers and two sellers interact repeatedly in the same group, either in the same role or switching their roles. The correct cluster level is, hence, the group.

[3] More specifically, the cluster-robust variance estimator is an asymptotically consistent solution for the inference problem (i.e., it converges to the population value of the variance as the sample size increases, Angrist & Pischke, 2008).

2. Repeated interactions. In a Newsvendor experiment, each participant plays several rounds deciding how many units to produce with the objective of maximizing their profit given a demand that is realized after they make their production decision. The individual orders will be autocorrelated over time (e.g., individual characteristics, experience, shocks) and the researcher should cluster the standard errors at the individual level to capture this feature of the data. An alternative to clustering standard errors is to analyze averages per individual across all periods as the unit of analysis.

3. Repeated interactions in changing groups. In a Beer Game, participants interact repeatedly with each other in groups of four, but the groups are randomly recomposed at the beginning of every period. Because participants carry their past experiences from interacting with others from group to group, all choices within the matching group within which people are re-matched are in the same cluster. Thus, if the re-matching takes place within the session, standard errors need to be clustered at the session level (alternatively, the researcher must collapse data at the session level).

In general, it is best practice to cluster standard errors at the highest level possible if it is unreasonable to safely assume the experimental observations are fully independent; however, the researcher should ensure that there are a sufficient number of clusters (i.e., usually between 30 and 50 should do) else inference will be compromised (cf. Cameron, Gelbach, & Miller, 2011). Interested readers can refer to more extensive treatments of this topic to ensure consistency of inference in more complex situations, including where clustering is multiway or disturbances are autoregressive (e.g., Angrist & Pischke, 2008; Bollen & Curran, 2004; Cameron, Gelbach, & Miller, 2011; Cameron & Miller, 2015).

2.5.2. Failed randomization: session effects and small sample sizes

Causal identification of treatment effects requires randomization. If randomization fails, treatment and control groups are not comparable ex-ante and observed differences in the dependent variable cannot be causally attributed to the treatment. As long as the selection process is orthogonal to the treatment, controlling for observable characteristics that are known to influence *y* will partially correct the bias (i.e., these effects are removed from the disturbance and modelled explicitly). Of course, this solution is not perfect because participants who differ in observable characteristics may also differ in unobservable ones. However, if some observable or unobservable variables correlate with both the treatment status and the dependent variable of interest, the treatment effect will be confounded (Angrist & Pischke, 2008; Antonakis et al., 2010; Antonakis, Bendahan, Jacquart, & Lalive, 2014).

One possible reason for randomization failure is the presence of so called session effects. Ideally, treatments are randomized within session. However, in some laboratory experiments, it is difficult or impossible to allocate different subjects to different treatments within the same session. As a result, sessions are separated by treatment. To minimize the threat that systematic differences across sessions confound the treatment effect, it is crucial that the random assignment of treatments to sessions is done in a balanced way that keeps as many factors as possible constant across treatments (subject pool, laboratory, time of the day; for details, see, e.g., Fréchette, 2011).

Another threat to clean randomization comes from small sample sizes. Small samples may lead to an unbalanced representation of observed and unobserved characteristics subjects in the treatment and control group (Strube, 1991). Moreover, data analyses in smaller samples are very sensitive to outliers (Aguinis, Gottfredson, & Joo, 2013), are heavily unstable with respect to the analytical strategy chosen (Button et al., 2013), and have low statistical power (Ioannidis, 2005);

at the research field level, small sample sizes tend to inflate the effect size of the treatment, given that only significant results are usually published (Antonakis, 2017; Button et al., 2013; Ioannidis, 2008). But how small is small? Determining an exact threshold that applies in all situations is clearly complex: The magnitude of the treatment effect investigated might vary dramatically across situations and only an a priori power calculation can provide accurate answers. However, recent methodological suggestions call for at least 20 (Simmons, Nelson, & Simonsohn, 2011) or more recently 50 observations per cell (Simmons, Nelson, & Simonsohn, 2013). In a series of simple Monte Carlo simulations, we also identify as 50 the minimal sample size requirement per experimental cell in simple empirical settings. Clearly, larger samples are preferable, yet smaller samples might still be enough: Our results—despite being by no means complete and based on certain assumptions—suggest that a sample size of 25 per cell has roughly a 26% failed randomization rate with one unobservable binary covariate (see Appendix A.3), which might be acceptable in some situations. Even though studies with small sample sizes should be avoided whenever possible, if ethical or logistical issues prevent the collection of large samples, researchers should alleviate this problem by: (a) showing the ex-ante comparability of the treatment and the control groups via summary statistics of the main control variables (i.e., age, gender, university major, etc. should be equally represented in the two conditions) or (b) including control variables in regression model, and hence show that the treatment effect is not directly driven by their inclusion or exclusion (for a brief overview, see Eden, 2017, p. 96).

To deal with small samples, researchers can also turn to fractional designs to reduce the treatment dimensionality without losing identification. This method consists of selecting only a subset of the treatments that would be run in a full factorial experiment, by constructing a set of conditions that allow the attribution of the statistical effect uniquely to a factor (Gunst & Mason, 2009; Nair et al., 2008). Thus, fractional designs can be useful in operations research: When

testing some manufacturing conditions, for instance, the number of factors and their potential levels can easily become unmanageable (see, e.g., Adenso-Diaz & Laguna, 2006). With fractional experiments, however, researchers can only make cross cell comparisons and cannot estimate interaction effects. Also, average marginal effects (i.e., main effects) can only be estimated if the interaction term can be safely assumed to be zero. For most applications in the social sciences, the researchers should not only justify the choice of a fractional design based on a pure logistical need (e.g., lack of funding, participants' sample is small); they should also use clear theoretical grounds to justify which conditions to run.

2.5.3. Endogenous mediators and causal non-identification

Experiments are a key way to solve the endogeneity problem typical of cross-sectional studies. However, endogeneity bias can also be hidden in experimental data, and mediation models are especially prone to this issue. OM, as well as management and applied and social psychology scholars, are often interested in the mechanism behind a certain treatment effect and engage in mediation analyses (see Baron & Kenny, 1986). Usually, those researchers include a measured mediator variable when analyzing the effect through which the experimental treatment apparently impacts a dependent variable. However, this procedure is potentially problematic: If the mediator is not randomly assigned and an omitted cause correlates both with the mediator and with the dependent variable, an endogeneity problem will emerge. The possibility of simultaneity may also be present, as will, in most cases, measurement error (thus rending the mediator endogenous).

Endogeneity leads to inconsistent and biased estimates, posing serious problems both in large and small samples. In order to correct for this bias, the authors should examine whether the mediator is actually endogenous (via an Hausman test, Hausman, 1978) and use an instrumental variable estimator (Antonakis et al., 2010; Emsley, Dunn, & White, 2010) instead of the standard

OLS or ANOVA. Finally, researchers should also mind the problem of specious mediation (modeling the wrong mediator that correlates with the true mediator), as well as causally unidentified mediation chains (Fischer, Dietz, & Antonakis, 2016): A causally identified mediation model needs to have at least a unique exogenous predictor per endogenous mediator, the unique exogenous variable/s must be excluded from the *y* equation, and the system of equations must be estimated using an instrumental variable estimator (Antonakis et al., 2010; Antonakis, Bendahan, et al., 2014).

The nature of this problem can be better understood with an example. Let us assume that a researcher wants to study the effects of stockout on customer satisfaction. The experimenter presents experimental subjects with several versions of a vignette where participants are told that a hypothetical company has either incurred a stockout due to a technical failure or that a company has incurred in a stockout because of the negligence of a supplier. The researcher assesses three perceptual measures: (a) how much the experimental subjects attribute responsibility to the company for the stockout, (b) the anger of potential customers toward the company and (c) their hypothetical intention to buy from the described firm again (cf. Hartmann & Moeller, 2014). Anger, predicted by treatment status, is theorized to mediate the relationship between responsibility attribution and future intention-to-buy. Estimating this model using a Structural Equation Model and not "instrumenting" the potentially endogenous mediator can be problematic, because several omitted causes might predict the endogenous variables (e.g., personal characteristics of an individual, demand effects, transient effects). An instrumental variable estimator is, thus, needed. However, the model has two endogenous variables (i.e., attribution of responsibility and anger), making it even more complex to achieve causal identification: The researchers would need to find at least two instruments to assess the causal

claims of their theory, meaning that they should have manipulated two variables that would strongly predict the mediators.

2.5.4. Lagged dependent variables

In several OM applications, researchers are interested in estimating models with lagged dependent variables used as covariate (e.g., demand at time t depends on demand at time $t-1$). The logic is clear: The level of an outcome today might be dependent on yesterday's level and the researcher might be interested in explicitly modeling the adjustment dynamics. However, the lagged variable might be endogenous itself (Bollen & Curran, 2006; Fischer et al., 2016; Wooldridge, 2010). Where is the issue? The lagged variable is correlated with the individual-specific error term, which is correlated with the dependent variable and with all its lags. As soon as the individual-specific error term is correlated with even one of the other independent variables, the whole equation will suffer from endogeneity bias regarding that variable, whose effect cannot be interpreted; of course, the effects of exogenous variables in the equation, which will be orthogonal to the lagged variable, can be interpreted. If the interpretation of the lag is important, the solution to this problem is slightly involved (T. W. Anderson & Hsiao, 1982; Arellano & Bond, 1991; Blundell & Bond, 1998). Intuitively, however, it reduces to using further lags of the dependent variable as instruments for the first difference of the lagged variable, and the implemented solution is available in many standard statistical packages such as Stata and R. Models that fail to use this correction present endogeneity bias, regardless whether the panel effects are specified as random or fixed, and estimate parameters inconsistently.

2.5.5. Issues of non-compliance

As in medical experiments where some participants fail to take the treatment as prescribed, participants in social sciences experiments might not comply with instructions in an experiment (e.g., L. H. Liang et al., 2018, where participants had to do a task that they did not

find appropriate and some did not comply). Note, this issue is different from failing a manipulation check, wherein a participant may still have been treated (though the below remedy will also correct for possible biased estimates in this case too). Similarly, in the field, subjects may not follow all the rules set by an experimenter and—ultimately—fail to be treated[4]. All these issues are known as "non-compliance" problems and pose a challenge to randomized experiments when the non-compliance status is correlated with the treatment status (Glennerster & Takavarasha, 2013; Gupta, 2011). Internal validity is compromised if non-compliers are removed: The experimenter will not be able to correctly recover the Average Treatment Effect (ATE), because an unobservable variable (e.g., an individual characteristic of the subject) affects both the likelihood of complying to the treatment and the outcome of the experiment.

Excluding subjects who were assigned to a treatment but did not receive it (i.e., the non-compliers) is not a viable solution, because randomization ensures balance on observable and unobservable characteristics only ex-ante (Montgomery, Nyhan, & Torres, 2016; Reis & Judd, 2000). However, statistical solutions allow the recovery of a causally interpretable estimate of the treatment effect, even though they alter the interpretation and the generalizability of the experimental results. A first possibility is to carry out the analysis on all subjects, both the ones who received the treatment and the ones who did not (the so-called "intention to treat analysis", Gupta, 2011). This analysis requires no assumptions, yet does not estimate the treatment effect per se, but rather the effect of being randomly allocated to the treatment group (i.e., being eligible for treatment). Whether or not this method is appropriate depends on the research question. For example, if the researcher is interested in the overall effect of a field intervention (e.g., the impact

[4] Such scenario is different from non-random attrition, that is, the systematic loss of some subjects during the experiment (which can be solved with selection correction or other methods, see, e.g., Duflo, Glennerster, & Kremer, 2007; Heckman, 1979)

of the introduction of a virtual collaboration tool on the productivity of the teams in an organization), the intention to treat analysis yields the desired result.

A second solution involves using instrumental variables and exploits the fact that the random allocation into a treatment is the perfect instrument, being—by definition—independent from any experimental outcome or pre-existing individual characteristics. The experimenter can run a two stage model, whereby the allocation into a condition predicts whether the treatment was actually received (e.g., if the manipulation was successful for that specific individual), which, in turn, further explains the outcome of interest (Duflo et al., 2007; Gerber & Green, 2008). The obtained estimate does not represent the ATE anymore, but the LATE (Local Average Treatment Effect), which measures the treatment effect in the compliers subpopulation (Angrist, Imbens, & Rubin, 1996). Returning to the above example, this method would therefore reveal the impact of the virtual collaboration tool on the employees who actually use it. Thus, it is crucial that researchers carefully evaluate which method fits their research question to avoid misinterpretation of observed effects.

3. EXTERNAL VALIDITY: GENERALIZABILITY TO ENVIRONMENTS OUTSIDE THE LABORATORY?

So far, our discussion has focused on methodological elements that ensure the internal validity of experimental studies. Yet, rigor alone—although the first step toward understanding the nature of a phenomenon and a necessary condition for knowledge application—is not enough (cf. Vermeulen, 2005). Rigorous studies only have scientific value in applied domains if they are also relevant; for basic research, the relevance criterion is not immediately obvious nor should it necessarily be (Antonakis, 2017).

What relevance means heavily depends on the question in which the researcher is interested. In many studies, the purpose of an experiment is to test a particular behavioral pattern predicted by a theory or to disentangle different psychological mechanisms behind an observed effect (see Roth, 1987). For studies of this type, relevance is determined by the degree to which the decision environment allows for clean and powerful tests of the theory or mechanism in question. The laboratory offers great advantages for this purpose, because the researcher has full control over the decision environment in which participants operate (Rubinstein, 2001). In the laboratory, institutions and complex systems can be exogenously manipulated at relatively low costs and the researcher can easily create counterfactual states of interest that might not be possible to set up in reality. These features allow researchers to pin down relevant causal effects, which are notoriously hard to identify outside of the laboratory.

However, OM research and, even more generally, the management discipline as a whole is also practice-oriented and aims at making policy suggestions to a professional audience (Holmström, Ketokivi, & Hameri, 2009; Meredith, 2001; Vermeulen, 2005). If experiments aim at informing policy or management, their relevance is largely determined by the generalizability of results to real-life environments outside of the laboratory. The issue of generalizability is, however, not specific to data generated in the laboratory. Many field studies identify a phenomenon in a particular setting and whether or not the findings can be extrapolated to other environments remains to be determined (see Falk & Heckman, 2009 for a detailed discussion).

Nevertheless, many researchers are especially skeptical when it comes to the generalization of laboratory results. The most common concerns are that laboratory settings are artificial and unrealistic (see, e.g., Berkowitz & Donnerstein, 1982; Colquitt, 2008) and that student participants are not representative of the populations of interest (Levitt & List, 2007a, 2007b). We understand the prima facie appeal of these arguments, but we believe that much of

the skepticism regarding the generalizability of laboratory evidence is based on misunderstandings about existing evidence on the topic (Camerer, 2015; Falk & Heckman, 2009; Kessler & Vesterlund, 2015). Still, we agree that there are limits to what the laboratory studies can accomplish. For instance, what we find particularly troublesome is the use of vignette-type experimental studies (especially using students), where organizational-level operations management phenomena are studied, and wherein claims are made to provide policy implications. In the following, we discuss our view in more detail and summarize existing evidence on this issue.

3.1. Generalization: Can we extrapolate findings from the lab to environments outside of the lab?

In order to discuss whether laboratory findings are generalizable, we first must define what exactly we mean by external validity. Some authors argue that external validity requires that the quantitative effect identified in one set-up generalizes to other ones (Levitt & List, 2007a, 2007b). It is clear that most laboratory experiments do not satisfy this strict standard (Ketokivi & McIntosh, 2017). Effect sizes typically depend on parameter choices and it is often difficult to tell what constitutes a "realistic" set of parameters. However, in line with Kessler and Vesterlund (2015), we argue that the quantitative definition of external validity makes little sense in the context of laboratory studies. The reason is that most experimental studies aim at identifying the direction of an effect rather than its precise quantitative size. Thus, the criterion for evaluating the external validity of these findings needs to be whether the sign of the effect remains constant across different environments. In this qualitative sense, external validity is ensured if the environmental differences between the laboratory and the field do not change the sign of the average marginal effect of a treatment. Even though the empirical assessments of external

validity in OM are still relatively rare, examples of laboratory experiments that have passed the test of qualitative external validity exist also in OM-related topics (Bolton & Ockenfels, 2014; Bolton, Ockenfels, & Thonemann, 2012; Olivares, Terwiesch, & Cassorla, 2008).

However, how useful are laboratory experiments for policy makers and managers if these studies "only" get the sign of an effect right, but cannot determine the magnitude of the effect? Schram (2005, p. 232) answers this question as follows: "[a]fter a theoretical design, a test [of a new airplane] in a wind tunnel is the stage of laboratory experimentation. If it does not 'crash' in this experiment, the plane is not immediately used for the transport of passengers, however. One will typically conduct further tests in the wind tunnel under extreme circumstances. In addition, further testing including 'real' flights without passengers will be conducted." In the same way, laboratory experiments typically constitute the first step in a series of studies. Given that many laboratory experiments are often tests of the behavioral assumptions or implications of a theory rather than tools for assessing the size of a statistical effect (Highhouse, 2009; Rubinstein, 2001), they are useful for capturing directional effects of theoretically informed treatments, yet await field tests for a more precise assessment of the practical effect.

One important advantage of laboratory experiments is that they are very easy to replicate (Camerer, 2015). Significant results will typically be investigated repeatedly and researchers will expose the effects to various tests to examine their robustness. Once an effect has survived these robustness checks in the laboratory and boundary conditions have been identified, researchers can move to the field to explore the effect size in the environment of interest. Thus, instead of seeing laboratory and field experiments at loggerheads and arguing about their relative superiority in particular contexts, researchers should start seeing these approaches as complements that reinforce each other (Camerer, 2015; Falk & Heckman, 2009; Kessler & Vesterlund, 2015). Overall, we see the balance between laboratory and field experiments as a matter of

"methodological fit" (Edmondson & McManus, 2007) that should be answered in light of the maturity of a research program.

Finally, it is also important to discuss what researchers should do if qualitative external validity fails. It is important to understand that failure to replicate a treatment effect in the field does not imply that the laboratory result is necessarily "wrong" (Kessler and Versterlund 2015). The replication may have failed because the laboratory environment misses an important element that is present in the field. A good laboratory experiment is similar to a good theory (Bacharach, 1989; De Treville, Edelson, Kharkar, & Avanzi, 2008; Wacker, 1998; Whetten, 1989): It strives to parsimony, simplifies the question of interest to its essence and abstracts from all irrelevant side-aspects that may hinder the clean identification of a causal effect (Plott, 1991). Sometimes, however, simplification goes too far and the design of an experiment leaves out a crucial element that characterizes the relevant field environment. Still, finding the crucial elements and real-world analogs in the laboratory requires more and not fewer laboratory experiments, because laboratory experiments are a simple and effective tool to identify the boundary conditions under which an effect is present or not (Falk and Heckman 2009).

3.2. Subject Pool Effects: Using Students, Representative Samples or Professionals as Participants

Many laboratory experiments use students as participants. This practice is frequently criticized, because students are neither representative of the general population nor do they correspond to the populations that are of interest in most studies (Levitt & List, 2007a, 2007b). So, why do experimental researchers use students so frequently? Students are very convenient participants. Most laboratories are located at universities where students are readily available. Moreover, students tend to have low incomes and are thus likely to be motivated by the monetary

incentives offered in experiments. They also tend to be rather intelligent and understand abstract and complex instructions; moreover, they are used to interacting in computer-mediated environments.

In terms of qualitative external validity, using students is only problematic if there are reasons to believe that the sign of an identified effect would be different in another subject pool. However, existing evidence suggests that this possibility is rather unlikely. Students are often similar to the general population and to professionals (see, e.g., C. A. Anderson, Lindsay, & Bushman, 1999; Fréchette, 2014; Herbst & Mas, 2015; Hosoda, Sone-Romero, & Coats, 2003; Mitchell, 2012). Clearly, it is true that students are a highly selected sample, which may imply that their behavior does not correspond perfectly to the behavior of other, potentially more relevant, groups. For instance, several studies have found that students differ from the general population in the importance given to social preferences: They seem to be consistently less pro-social than the general population, with economics students being less pro-social than other comparable students (see, e.g., Belot, Duch, & Miller, 2015; Exadaktylos, Espín, & Branas-Garza, 2013; Falk, Meier, & Zehnder, 2013). However, as discussed above, there are many reasons for why findings of laboratory experiments can rarely be interpreted quantitatively in any case, so that these differences in magnitudes of effects should not be seen as a major issue.

Moreover, existing evidence in OM and operations research seems to point at a qualitative similarity between the basic decision-making strategy of students and experienced managers. An early experimental study contrasted the performance of students and professionals in bargaining games (Siegel & Harnett, 1964); these researchers found that both sets of participants maximized joint profits and engaged in 50-50 splits. More recently, Bolton et al. (2012) studied a newsvendor task, finding evidence that managers behave similarly to students. More specifically, managers also display pull-to-center bias, that is, they put orders that are biased toward the center

point of the demand distribution. According to the Bolton et al. (2012) data, students are better at extracting analytic information from existing demand data, whereas managers are better at the beginning of the experiment, because they seem to use effectively information coming from their experience when the historical demand data are lacking. Interestingly, students end up over-performing managers.

Clearly, if researchers are interested in the quantitative magnitude of a treatment effect, using the population of interest as subjects in a laboratory experiment may be appealing (Stevens, 2011). However, researchers should keep in mind that the effect size is not only determined by the subject pool. A quantitatively interpretable result also requires that the decision environment (i.e., the stake size, time frame, group size, etc.) is identical or at least highly similar to the relevant field setting. Moreover, close resemblance to the real world increases the familiarity of the participants with the experimental scenario (D. J. Cooper, Kagel, Lo, & Gu, 1999; Montmarquette, Rullière, Villeval, & Zeiliger, 2004). However, tight affinity to reality can come at a cost, confounding the experimental result (Smith, 1976): For instance, whereas framing effects might enhance mundane realism (e.g., cover stories), several researchers have shown that they might bias the behavioral outcomes measured in an experimental study (Andreoni, 1995; Kühberger, 1998; Tversky & Kahneman, 1981).

To conclude, the aim of research in the social sciences is to identify general behavioral patterns that are valid beyond the setting in which they are identified (Kessler & Vesterlund, 2015). Laboratory experiments are characterized by a high level of control so that they can identify behavioral principles in a clean and systematic manner. Experimentalists are aware of the fact that the magnitude of the effects detected in a particular study heavily depends on the details of the setting; however, there is large agreement that the sign of the effect generalizes to other environments. It is also important to remember that generalizability alone is not a signal of high

quality. If the same endogeneity prone correlation generalizes to various environments, this spurious relationship tells us nothing about the causal impact of one variable on another. Thus, as stated earlier, it makes no sense to see a trade-off between rigor and relevance, because relevance ultimately requires rigor.

4. REVIEW OF EXPERIMENTAL PRACTICES

In this section, we examine whether experimental research complies with methodological best practices. We analyze the last 10 years of experimental research in OM. We limit the scope to 2006-2016, because other authors have already analyzed OM experimental research until 2006 (Bendoly et al., 2006); thus, our review casts light on current research practice. We focus on four journals, widely considered as the top-tier outlets in OM: (a) *Management Science*[5]; (b) *Production and Operations Management*; (c) *Manufacturing & Service Operations Management*; (d) *Journal of Operations Management*. These journals are the only journals in OM that are part of the Financial Times 50 list and the UT Dallas Research list that also publish experimental work. On top of reviewing the current experimental research in OM, we also examined its relative standing compared to a random samples of experimental research published in top outlets in management (i.e., *Academy of Management Journal, Journal of Applied Psychology, Journal of Management*, *Strategic Management Journal*, *Personnel Psychology*), economics (i.e., *Quarterly Journal of Economics*, *American Economic Review*, *Econometrica*, *Review of Economic Studies*, *Journal of Political Economy*) and psychology (i.e., *Journal of Personality and Social Psychology, Journal of Experimental Social Psychology, Organizational Behavior and Human Processes, Psychological Science*; note, although the latter is a general psychology journal it is a common outlet for social-psychology articles).

[5] We only focus on the Operations Management division of Management Science. Management Science also publishes experimental research covering other fields and their inclusion would have prevented a clear comparison between Management Science and other OM journals.

4.1. Method

To include all possible published papers in the chosen OM journals, we first reviewed the existing publications using a search on Web of Science (i.e., search term in the "topic" field: "experiment" or "experiments") and manually browsed through the abstract of all the issues published in the period 2006-2016. We excluded articles using natural experiments, re-analyses of already published experimental data, and articles in advance of print. Concerning the experimental research in the other fields, we selected a random sample of articles of non-OM disciplines: A random selection of articles should allow to have an unbiased representation of a discipline's average experimental study and maintain a feasible workload level for the coders (cf. Antonakis et al., 2010; Bergh & Perry, 2006; Fischer et al., 2016; Kacmar & Whitfield, 2000). Using a search term in the "topic" field on Web of Science, we downloaded the list of all articles explicitly mentioning the words "experiment" or "experiments" either in the title, in the abstract or in the keywords, randomly ordered them and manually selected the first 50 articles that were relevant for our analysis.

We then coded information from the articles sampled. Research assistants collected bibliographical information, whereas the first and the second author of this paper manually coded for the main markers of internal, statistical and external/ecological validity presented in the first part of the article. We took several steps to ensure that the coding procedures were as objective as possible. We first developed an initial version of the coding manual and adjusted it after two training phases on articles not included in the final sample. Afterwards, the two coders separately coded 20 articles from the reviewed sample. The two coders discussed their disagreements and calibrated their coding policies after consultation with the two other authors; the coders were then randomly assigned to independently code half of the sampled articles each (for details on the entire procedure, see Appendix A.1).

The two coders finally coded ten articles already coded by the other coder: We used data from these 20 articles to check coding agreement using the κ statistic (Viera & Garrett, 2005). Given that not all the coded dimensions were finally included in the data analysis, we limited our attention only to the categories reported in the remainder of the paper. On this subsample, the agreement between coders was *78.1*%, which was clearly better than the *9.7*% expected agreement due to chance, leading to a $\kappa = .76$ ($z = 53.82$, $p < .01$, see Appendix A.2 for details on this procedure). Such a level of agreement can be classified as "substantial" (Landis & Koch, 1977). Finally, following the computation of the inter-rater agreement statistics, the coders re-checked their coding choices and made some corrections in three categories, because of systematic differences in the coding decisions (see Appendix A.2 for details[6]). As a result, the reported κ statistics should be regarded as a conservative (i.e., lower-bound) proxy of the actual inter-coder agreement. Thus, although the agreement is very high, it is not perfect, which of course is reflected in some degree of judgment to extract the coding categories.

For sake of transparency, all data and code necessary to replicate the κ computation, as well as the results reported in the rest of the article are publicly available at http://dx.doi.org/10.17632/5wc48wpmpk.1.

4.2. Analysis and results

We located 258 articles (111 from OM, 47 from economics, 50 from social psychology and 50 from general management), for a total of 468 experimental studies[7] (i.e., several papers

[6]It came to our attention during the revision of this article that we did not code for non-compliance issues in the reviewed articles. Given time constraints, we took a random sample of 20 articles from OM journals and of 20 articles from non-OM disciplines to check the baseline rate. We found only two studies presenting a non-compliance issue. In both cases, however, the problem was correctly dealt with: The first article used an instrumental variable strategy, whereas the second article used a simple Intention-to-treat analysis, yet justified this choice from a theory viewpoint (results available upon request). Thus, due to its low occurrence rate, we excluded this category from the remaining of the analysis.

[7] The dataset of articles initially allocated to coders contained 267 articles. During the coding procedures, we excluded 9 articles from the analysis, because we verified that they did not meet all inclusion criteria (i.e., they were initially included either because of a manual mistake or because their abstract incorrectly suggested they were an experimental study; see Appendix A.1 for details on the procedure).

reported more than a single study). The definition of all coding categories can be found in Appendix A.2; complete descriptive statistics can be found in Table A.3 of the Appendix. Before moving to the data analysis, it is worth mentioning that our data clearly underline how the reviewed experimental research in OM appears not to be a uniform field. Whereas a subset of OM experimental research has a clear experimental economics/decision science undergirding (55%), a second subset of studies is closer to the applied psychology, management or organizational behavior area (44%); the pure engineering/operations research disciplines have a minimal representation (1%). Thus, when illustrative, throughout our analysis we differentiate between an "economics-based OM" and "non-economics-based OM" clusters, so to have a more fine-grained understanding of relevant patterns within the field.

4.2.1. Internal validity differences

We first map the variance in internal and statistical validity across different fields, with the objective of summarizing relevant descriptive trends at the academic field level. In Table 2, we regress the likelihood of adopting the internal and statistical validity suggestions on five academic field dummy variables (with economics being the baseline/omitted category), as well as a series of bibliographical controls (e.g., average number of studies, impact factor of publications, etc.), journal, publication-year and coder fixed effects. We use dichotomous dependent variables, indicating whether an experimental study followed or did not follow each of our methodological suggestions. We employ a linear probability model, rather than a limited dependent variable model for ease of interpretation with respect to the marginal effects (see Angrist & Pischke, 2008 for a technical argument in favor of this choice)[8]. For increasing the interpretability of our results, we mean-center all covariates (but not the dummy coefficients of interest): As a result, the

[8] Given that some academic fields always have the same value for specific dependent variables (i.e., we coded as 0 or as 1 all experimental studies coming from the same discipline), logit or probit models would have been unable to estimate the relevant dummy variable coefficient (Caudill, 1988).

constant term can be interpreted as the predicted probability for economics to engage in either of our methodological suggestions (at the means of all covariates), whereas the beta coefficients of the academic field dummies can be interpreted as marginal effects compared to economics.

The results show that both OM subfields seem to better address the quest for rigor than management and social psychology do; yet the OM subfields usually lag behind economics in this dimension. Both OM subfields are significantly more likely than economics to be non-incentivized and to report no form of compensation to the subjects (column (1), $\beta = .17$, $p < .01$ for economics-based OM; $\beta = .42$, $p < .01$ for non-economics OM); however, economics-based OM papers are much more likely than non-economics ones to be incentivized. Only studies from the non-economics OM subfield have a significant probability of reporting demand-driven treatment comparisons (column (3), $\beta = .26$, $p < .05$) and to be non-consequential (column (5), $\beta = .25$, $p < .05$), whereas none of the OM subfields seems to use deception or manipulation checks before the dependent variable or to deceive participants often (see columns (2) and (4)).

In almost all dimensions, management and social psychology seem to be less likely to adopt our best-practice suggestions than does OM; however, management experiments normally show a lower probability of engaging in internal validity issues as compared to social psychology (aside from the probability of engaging in manipulation checks before the dependent variable, see column (2)). Clearly, some of the differences we observe might be explained by different research foci proper to each discipline. For instance—as we have argued before—designing consequential decision environments or including incentives in an experiment might not be needed in fields that study emotions or perceptual dependent variables (e.g., social psychology). However, incentives of this sort are easier to implement in disciplines that are more focused on financial metrics or other objective dependent variables (e.g., economics or OM).

[Insert Table 2 about here]

Even though economics-based OM scores better than non-economics OM studies in internal validity, the former's situation is somewhat reversed for the statistical validity dimensions; moreover, OM as a whole seems to lag behind our suggestions in this area. Both subfields are more likely than economics to estimate incorrectly specified models (column (6), β = .12, p = .11 for economics-based OM; β = .16, p < .05). Moreover, studies sampled from the economics-based OM field are more likely to fail to report correctly clustered standard errors than non-economics-based OM (column (7), β = .17, p < .05 and β = -.05, p = .51). Finally, both OM subfields are more likely than economics to have a small sample (column (8), β = .36, p < .01 for economics-based OM and β = .20, p < .10). We also report F tests for the differences in coefficients between the two OM subfields directly in Table (2).

Concerning the other disciplines, economics appears to be the most rigorous field when it comes to statistical validity. Surprisingly, studies sampled from social psychology are less likely than the ones from economics to estimate incorrect standard errors (β = -.10, p < .10). There is, however, a caveat to the interpretation of these results. Almost 20% of studies in social psychology employ simple "differences-in-means" as the sole statistical test in their analysis (e.g., *t*-test, non-parametric test for equality of means or medians). Whereas these techniques are correct per se, if they do not violate the assumptions of the estimator and are done at the correct level of analysis, not reporting a regression or structural equation modeling analysis limits what can be tested in terms of multivariate effects. As a result, the statistical practice in this field shields these studies from problems emerging from incorrect estimation and clustering of standard errors.

4.2.2. Potential generalizability to organizational environments

In the second part of the analysis, we focus on proxies of generalizability to real organizations and to the working population, which usually represent the focus of the analysis in

OM and other organizational sciences. In Table 3, we regress three main proxies of potential generalizability on the academic field dummies and control for the same set of bibliographic control variables and fixed effects used in Table 2. The three proxies we chose are (a) use of EMBA students or (b) working adults as experimental subjects (i.e., sample generalization to real organizations) and (c) proportion of field experiments in each discipline (i.e., operational generalization). EMBA and working adults are "closer" than undergraduate students to the ideal participant in many studies of OM and management, namely, a working professional in a real organization; however, this point is not necessarily true for social psychology and economics, because these fields generalize beyond organizational phenomena. Moreover, field experiments represent the apex of realism in experimental research for all disciplines, to the extent that they measure behavior in its naturalistic environment. Hence, these proxies are measures of generalizability *to*—and not *across*—real professionals and organizations. That is, they do not quantify the actual external validity of an experimental finding, which would require replication studies and meta-analyses; yet, they indicate their potential generalizability to environments and populations that represent the ultimate interest for organizational sciences.

Overall, we do not find clear evidence for a trade-off between internal validity and potential generalizability. The fields that were previously found to lag behind regarding standards of internal validity (i.e., non-economics OM, management and social psychology) are not systematically more generalizable. However, our data show that economics-based OM could improve significantly in the generalizability domains. In particular, our results underline that economics-based OM makes less widespread use of working adults as experimental subjects than does non-economics-based OM (column (1), $\beta = .00$, $p = .98$ and $\beta = .18$, $p < .05$). Furthermore, economics OM is less likely to employ graduate or EMBA students (column (2), $\beta = .04$, $p = .18$ and $\beta = .15$, $p < .01$). However, as introduced above, both OM subfields employ significantly

less field experimentation compared to economics (column (3), β = -.39, p < .01 and β= -.35, p < .01). We report formal F tests for the differences in coefficients between the two OM subfields directly in Table (3).

Interestingly, we did not identify any article in social psychology reporting a field experiment; the management discipline hardly relies on field experimentation either (β = -.35, p < .01); the academic discipline that most frequently employs field experimentation is economics. Contrary to economics and economics-based OM, management and social psychology make significant use of EMBA students (β = .10, p < .05 and β = .05, p < .05 respectively); however, they do not appear to make use of working adults as experimental samples (both β coefficients are non-significant). To conclude, it is important to note that these results should be interpreted with caution: Many studies in social psychology and economics are interested in research questions that do not directly relate to organizational environments; thus, it is easy to see why they might not need to use EMBA or working adult samples.

[Insert Table 3 about here]

4.2.3. Citation analysis: Let the market decide

The heterogeneity of experimental practices we observe might be partly driven by the fact that the reviewed papers belong to different disciplines with distinct research interests and specializations. Part of the differences in observed experimental practices might therefore be justifiable by the specificities of the different fields. To examine this conjecture, we conducted a citation analysis (see, e.g., Antonakis, Bastardoz, Liu, & Schriesheim, 2014), because citations are a measure of academic relevance.

We thus regressed the number of citations received by each paper in our sample (retrieved manually from Scopus in September 2017) on all the markers of rigor and generalizability. Specifically, we used five indicator variables capturing the following aspects of internal validity:

(i) no incentive/compensation for the experimental subject, (ii) non-consequential study, (iii) deception, (iv) demand-driven comparisons of treatments, and (v) manipulation check assessed before the measurement of the main dependent variable. We also used three indicator variables that coded for the generalizability potential of the study: (i) sample composed of working adults, (ii) sample composed of EMBA students, and (iii) field experiment. As proxies of statistical validity, we employed three measures already reported (i.e., incorrect estimation, incorrect inference, small-sample) and added a fourth indicator variable measuring whether the study employed simple "differences-in-means" as statistical test rather than more complex regression-based analyses. This variable is included in our analysis for interpretation issues: The 16% of studies reporting only "differences-in-means" in our sample never engage in incorrect estimation or in incorrect inference simply because they do not report a regression-based model. Additionally, we also included a large set of bibliographical controls (e.g., article age, article age square, citations received by the author's team, etc., see Table 4 for the complete list), coders fixed effects and journal fixed effects or—in alternative specifications—academic field fixed effects, in order to reduce the likelihood of omitted variable bias. Given that the number of citations received are counts, we used negative binomial regressions (Gardner, Mulvey, & Shaw, 1995), which correctly models the "overdispersion" of the data present in our dataset (see Figure 1 for a graphical overview of this pattern).

[Insert Figure 1 about here]

Our results can only be interpreted from a prediction standpoint and not causally (i.e., the design characteristics of a study are clearly not randomly assigned)[9]. The main results (see

[9]Antonakis et al. (2014) use an instrumental variable estimator to address the endogeneity problem of one of their independent variables. More specifically, they instrument the presence of endogeneity in a study with several bibliometric variables (e.g., author previous citations). In our setting, however, this possibility is prevented: Each internal, statistical and external/ecological validity proxy would need to be instrumented with a presumably exogenous variable and our dataset does not contain such a large number of potential instruments.

column (1), Table 4) underline that the baseline set of bibliographical controls and fixed effects have a strong joint explanatory power (Wald test: $\chi^2(27) = 837.59$, $p < .01$). Most covariates seem to follow a similar pattern compared to the results reported by Antonakis, Bastardoz, et al. (2014), who engage in a similar citation analysis. However, some indicators for internal validity and of potential generalizability have significant explanatory power vis-à-vis academic relevance. Experiments that do not make use of any type of incentive in the laboratory (e.g., credit- or financial-based, both related or unrelated to actual performance obtained in the experimental task) are cited less ($\beta = -.42$, $p < .01$), as too are experiments that trigger demand effects ($\beta = -.23$, $p < .10$). If combined, the inclusion of all internal validity issues leads to an increase in the explanatory power of the model (Wald test: $\chi^2(5) = 11.05$, $p = .05$). The use of field experimentation has a positive and significant effect on total citations received ($\beta = .34$, $p < .10$). Overall, however, the generalizability proxies do not further increase the fit of the model significantly (Wald test: $\chi^2(3) = 5.40$, $p = .15$). Similarly, the inclusion of statistical validity proxies leads to no improvement to the fit of the model (Wald test: $\chi^2(4) = 3.19$, $p = .53$).

To obtain an intuitive representation of some of these effects, we provide a graphical representation of our estimates in Figure 2, where we depict the average marginal effect of the main markers of internal, statistical and external/ecological validity on the citations received by an article (based on the regression specification of column (1), Table 4). It is immediate to see that the effect of most predictors, despite being not always individually significant, suggests that rigor and generalizability increase the impact of an article. Also, in Table 4 we report the linear combination of the effect of not employing our validity markers' in experimental studies. This estimate ($-.10$, $p < .05$ for the model in column (1)) pools together the effect of all our

suggestions and underlines their overall relevance: Following the suggestions we propose should provide a citation advantage to journals.

Finally, the second regression specification (see column (2), Table 4) reinforces almost all other reported findings, which are also robust across additional specifications and estimation methods (see Appendix A.4). To ensure that these effects are not driven by some individual academic fields or journals, in columns (3)-(6) of Table 4 we also report the results of a series of regressions estimated within each field. This analysis suggests that differences across fields are evident, yet most internal and external validity markers previously identified as important determinants of citations remain relevant predictors, even though they might drop below traditional significance levels (because of the lower statistical power caused by the limited sample size).

[Insert Figure 2 about here]

However, some within-field results are noteworthy. First, and contrary to the trend in other fields, the analysis suggests that OM rewards incorrectly estimated models in the "market for citations" ($\beta = .51$, $p < .10$, column (3)). This finding is somewhat worrisome, given that we have also documented how OM is among the fields lagging behind our statistical validity suggestions: Perhaps because the field has started to consider some of these statistical issues only recently (see Ketokivi & McIntosh, 2017), the profession still needs to discount the effect of incorrectly estimating models on academic impact. Second, non-consequential studies are rewarded in social psychology ($\beta = .39$, $p < .10$, column (6)), but not in other fields. As mentioned in the previous subsections, this result could be driven by the object of study of social psychology, which tends to focus more on perceptions rather than objective actions, making it more complex to design consequential experiments. Note also that social psychology experiments most frequently report self-reported/non-behavioral outcomes (almost 63%). Third—and

surprisingly—economics seems to reward small sample studies ($\beta = .87$, $p < .01$, column (4)), contrary to OM, which weakly devalues them ($\beta = -.33$, $p < .10$, column (3)). However, it should be noted that economics is the field that reports by far the lowest frequency of small samples (23% against more than 40% for all other fields). Although it is not clear as to what drives this effect, it may be that the few small sample studies in economics treat especially relevant topics or present other design features that make them exceptionally noteworthy. Lastly, only management seems to put a premium on experimental studies using working adults ($\beta = .25$, $p < .10$, column (5)), whereas OM does not significantly reward any external/ecological validity proxy. This result is somewhat puzzling in light of the attention towards practical relevance of OM: Intuitively, we would have expected the profession to strongly reward practices like field experimentation. However, field experiments represent a small percentage of OM reviewed studies (only 8%): Once more, unobservable characteristics of these few field experiments might make them less appealing for the citation market.

[Insert Table 4 about here]

Finally, we investigate whether experiments are more or less cited than other types of articles within each academic field: In other words, does publishing experimental research "pay off" for researchers? To answer this question, we first computed the number of citations per year received by an average article predicted by our main model (on the basis of the same specification as in Table 4, column (1), and using citations received per year rather than total citations as dependent variable). We then computed both the predicted total citations received by the average 2-year-old and by the average 5-year-old article we reviewed and contrasted them with two measures: (a) the median 2-year impact factor of each field and (b) the 5-year impact factor of the journals we included in our review (both retrieved from Scopus in September 2017).

Our results are displayed in Table 5 and suggest that the average experimental study we reviewed receives more citations than the median article in its field. If we restrict our attention only to the journals we reviewed in our article (i.e., which represent top publications in each domain), experimental articles are cited more than the average article published in top tier outlets in OM (*+44%, p = .10*); this effect is not significant, similarly to what we observe for economics (*+11%*) and social psychology (*+28%*). Only experiments in management are cited significantly less than average articles in top tier outlets (*-55%, p < .01*). This effect is strong, yet ambiguous to interpret. Even though the management field has shown some adversity toward experimentation (see Colquitt, 2008; Highhouse, 2009), our review has also underlined that experiments in management studies are generally neither especially rigorous, nor especially generalizable. Thus, it remains unclear whether the skepticism towards experimentation is a cause or an effect of this occurrence.

[Insert Table 5 about here]

5. QUASI-EXPERIMENTAL TECHNIQUES: A QUEST FOR RELEVANCE AND RIGOR WHEN EXPERIMENTAL TOOLS ARE UNAVAILABLE

Our review and results call for a strong complementarity (and not substitutability) between field evidence and laboratory experiments. However, applying the best practices of experimentation we have outlined in our article might be impracticable in some naturalistic OM-relevant settings (e.g., real organizations). Business problems are often ill-defined and practical, financial and/or ethical issues make it hard to randomly allocate entities, whether people, groups, or organizations to different conditions and managers might be reluctant to apply experimental methods in their own teams (e.g., Eden, 2017; Grant & Wall, 2009). So, how should researchers

proceed if they encounter an interesting research environment, but randomized field experiments are out of the question?

In this section, we outline several quantitative methods, which are relatively unused in OM and which are equipped to produce pragmatic and rigorous research in real business settings. These approaches do not require randomization and, under some stringent assumptions, allow making strong causal claims even in small sample conditions: (a) the Difference-in-Differences design, (b) the Regression Discontinuity design, and (c) the Synthetic Control Method. Our intent is not to provide an extensive exposé of those methods; rather, we showcase the intuition behind them, discuss when their identification assumptions are likely to be met, and point readers toward more in-depth resources. We finally link these approaches to methodological frameworks that are more common in OM and management research, showing their possible complementarities.

5.1. Difference-in-Differences

In situations where randomization is not feasible, we can exploit both within- and between-entity variability through the Difference-in-Differences (DD) estimator. The intuition of this method mimics the logic of randomized experiments. The researcher compares the historical outcome of two similar entities (e.g., two divisions of a single firm) in two distinct periods. In the first period, no unit is exposed to a treatment, whereas in the second period only one entity is treated. Given that the two entities are not randomly allocated, a simple comparison of their outcomes in the second period might lead to a spurious treatment effect, because other differences between the units may be driving the effect.

The DD estimate circumvents this problem by measuring any pre-existing difference between the two untreated entities assessed in the first time period, as well as the variation of the outcome over time that would have been observed even without the intervention (see the classic

example of Card & Krueger, 1994). If the time differences between groups are stable, then any observed difference in the time trends after the intervention is most likely due to the treatment (for a brief introduction to DD, see section 4.4 of Antonakis et al., 2010). In addition, two assumptions are needed for the DD estimator to be consistent and for the inference to be correct: (a) exogenous treatment, (b) cluster-robust inference (for details on the DD estimator, see Angrist & Pischke, 2008; Bertrand et al., 2004).

For an example of how OM scholars can apply the DD framework, let us assume a researcher would like to estimate the causal effect of a change in some processes of a company. Supposing the researcher is able to find a company with at least two similar plants, she would need to convince the firm's management to adopt the treatment in one plant and to leave the second plant untouched. If the time trends in performance across plants are relatively constant prior to the intervention, the researcher can measure the performance before and after the intervention, both for the treated plant and for the control one to estimate the treatment effect. If the researcher can assume that any difference between the treatment and the control plan would remain stable over time without the intervention, partialling out the first difference will allow the researcher to isolate all time-related confounds and to focus only on the treatment effect.

Clearly, the larger the between (as well as the within) sample size, the more credible is the DD causal effect. However, DD requires—at minimum—two entities: One control and one treated group, and of course, observations over time. Still, the validity of DD hinges on the constant time-trend assumption, which is credible if treated and untreated entities are similar (e.g., two plants or two sub-units of the same firm) and face comparable environments (e.g., same market conditions). The same presumption might not hold if the treated and non-treated entities are more heterogeneous (e.g., units located in different continents might not follow the same time trend). Also, researchers should carefully consider the possibility of spillover effects across

studied entities, which is especially likely if treated and non-treated units are somehow connected (e.g., communication between managers or workers in geographically close plants).

5.2. Regression Discontinuity Design

Let us assume a researcher wants to study the relationship between socially responsible procurement and financial performance of corporations. This question is especially interesting for OM, where scholars have started to integrate sustainability in their framework (Kleindorfer, Singhal, & Wassenhove, 2005; Linton, Klassen, & Jayaraman, 2007). Ideally, one would like to run field experiments that randomly allocate many corporations to different procurement policies, but such studies are most likely infeasible for logistical reasons. In such situations, a Regression Discontinuity Design (RDD) provides an attractive alternative. In RDD setups, the treatment is not assigned randomly. Instead, an assignment variable deterministically allocates some units to a natural treatment group according to a pre-specified threshold, either in time, space, or some descriptive variable (e.g., Imbens & Lemieux, 2008; Lee & Lemieux, 2010). For instance, Flammer (2015) uses this logic and contrasts the performance of firms in which a proposal related to Corporate Social Responsibility (CSR) was either passed or failed with a small margin in shareholder meetings. The intuition is that treated and untreated units (i.e., companies that either adopted or opposed the CSR proposal by a handful of votes) can be considered comparable and *as if* randomized, because the board members have—individually—little to no control over the assignment variable. As a result, the researcher does not need to model the selection into treatment and "close call" votes can generate random variation around the cutoff, that is, at 50% vote share (for a recent example and the types of robustness checks to conduct see also Arvate, Galilea, & Todescat, 2018). Using close-call elections in one way to model a RD. There are, however, many ways in which RDD can be implemented by using premeasured assignment

variables in time and space, or even truly exogenous shocks (for ideas, see Bastardoz, Jacquart, & Antonakis, 2017).

5.3. Synthetic control method

Consider the case where a researcher is interested in studying the effect of a specific quality-control program on the output quality of a given product. A plant in a large corporation (plant A) has implemented a new quality-control program (i.e., our "treatment"), whereas all other plants of the same group still employ the old quality-control policy. However, the researcher is unable to employ a DD design, because the data referring to plant A before the implementation of the new quality-control program are unavailable. RDD cannot be used, either, because there is no assignment variable that allocated different plants to different programs. In other words, the researcher has just no chance of observing an actual counterfactual.

A last resort for estimating causal effects in small *n*-size contexts is the so-called Synthetic Control Method (Abadie, Diamond, & Hainmueller, 2010, 2011, 2015). Without using instrumental variable techniques (cf. Angrist & Pischke, 2008), the work of Abadie and colleagues addresses the counterfactual problem by allowing researchers to build their own "synthetic" (i.e., simulated) counterfactual. Applied to our example, the intuition of the Synthetic Control Method is to use plant A as the treated unit and employ a matching algorithm to create a mock plant A that is constructed as a weighted average of all plants that were not treated. More specifically, the Synthetic Control Method will approximate the output quality of plant A, had it not experienced the treatment (i.e., the quality-control program). Confronting the dynamics of the outcome for the treated and the synthetic entity allows estimating the causal effect of the adoption of the quality-control program. To further check the validity of the method, the researcher can then execute some "placebo tests," simulating the introduction of the treatment either before or

after the actual adoption of the quality-control program or assigning the treatment to plants where it was never actually introduced. If the synthetic counterfactual is well-constructed, these placebo simulations should not identify a statistical effect of the simulated treatments. On the contrary, a large placebo effect would suggest that important differences between actual and synthetic units can be triggered by noise: Thus, any observed treatment effect is likely unreliable.

5.4. From explanation to exploration: Design Science and case studies

The aforementioned quasi-experimental methods, which are useful for providing causal explanations, are rather rare in the OM literature: A search on Web Of Knowledge in autumn 2018 revealed that only 7 articles in the reviewed OM journals have ever reported either in their abstracts or in their keywords the terms "Regression Discontinuity" (0 occurrences), "Difference in differences" or "Diff-in-diff" (4 occurrences), "quasi experiment*" (3 occurrences) or "Synthetic control method" (0 occurrences)[10]. We hope these methods will receive more attention in the near future, because they allow for solid causal claims on the basis of a rigorous quantitative framework using the counterfactual logic (Rubin, 1974). These methods can also provide fruitful avenues to complement other frameworks already present in OM: Design Science and case study research.

The Design Science (DS) paradigm is an approach with roots in engineering; here, technically complex problems are studied, where engineers have complete control over the environment and can alter one element of the process at will (note, in the engineering context a pre-post statistical test can estimate the causal effect of the intervention on the process). The DS paradigm has also been applied to OM and other disciplines to tackle the theory-practice gap (Holmström et al., 2009; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007; Van Aken, 2005;

[10]For Management Science, we only considered the Operations Management division, as well as articles reporting no specific division. We further excluded articles that mentioned any of the keywords, yet did not engage in a quasi-experimental empirical analysis.

Van Aken, Chandrasekaran, & Halman, 2016; Von Alan, March, Park, & Ram, 2004). However, using a DS logic in a complex microeconomic system that is dynamic and includes human players brings complications and internal validity confounds that typically render causal interpretations impossible. In contrast to traditional experimental approaches, which develop interrelated hypotheses and test them via some statistically rigorous method with the objective of predicting and/or explaining a phenomenon, the DS paradigm is therefore rather suitable to uncover hypotheses about effects of interventions on organizational performance in a particular environment (Von Alan et al., 2004). DS focuses on the practical usefulness of a treatment and tends to favor proximity to reality (Van Aken et al., 2016; Von Alan et al., 2004). The logic behind this approach is to study a process, intervene, and then observe the effect of the intervention (Hevner, March, Park, & Ram, 2004; Holmström et al., 2009).

DS may be appropriate for ill-defined research questions, where context-relevant knowledge must be collected and analyzed to explore novel solutions to newly-formalized business problems. Thus, DS research and experiments serve two different purposes; yet they can be complementary: A DS approach can help to formalize the problem and provide an initial policy suggestion; experimentation aims at formally testing the causal impact of the policy.

To simply demonstrate the point, suppose the business objective of a publisher is to increase the subscriptions to a journal. A DS study might reveal that a change in editorship (i.e., the appointment of a more qualified and renowned person) is followed by an increase in the journal's impact factor a few years later. This observation gives rise to the hypothesis that the quality of the editor affects the impact factor. To empirically establish a causal link, however, counterfactual thinking becomes key, because alternative explanations need to be excluded. For example, the increase in the impact factor could also be a field specific effect (e.g., more journals in that field were indexed in the Web of Science that year) or it could be a general effect (e.g.,

citation inflation). In this particular case, a direct experiment is difficult to implement, but the quasi-experimental approaches discussed above (e.g. a "diff-in-diff" design that uses comparable journals as a control group) could be applied. That is, we can compare the difference in outcomes in an entity from before and after a DS intervention relative to an entity receiving no intervention; doing so would result in a difference-in-differences estimator and thus allow for causal claims to be made.

Quasi-experimentation also retains some of the virtues of case study research with respect to focusing on a fine-grained and in-depth understanding of the dynamics at play in a specific situation (Meredith, 1998). Being especially "close to the action", the case study logic is helpful to formalize new research questions and hypotheses. However, when it comes to causality, experimentation is superior to "snapshot" case studies, because focusing on a single case might lead to sampling on the dependent variable (i.e., to employ a criterion—like exceptional performance—to select a case, and then study the antecedents and consequences of the criterion, failing to study all entities that do not meet the criterion itself). Quasi-experimental techniques are comparable to existing case study frameworks for generating counterfactuals (see, e.g., "polar opposites" and "experimental template for case study research", Eisenhardt & Graebner, 2007; Gerring & McDermott, 2007), which makes them closer also to the experience of qualitative researchers. Yet again, with observations over time, quantification, and a strong counterfactual, one can easily extend this paradigm to make causal claims.

Overall, we are confident that the quasi-experimental methods we introduced earlier can be very useful complements to the current toolkit of OM and other social sciences too, given that these methods are applicable in small *n*–size design conditions. Quasi-experimentation represents a causally solid approach for evaluating artifacts' performance (Mettler, Eurich, & Winter, 2014), yet ensures a logistically simpler alternative to field-experimentation. Obviously, these methods

still present some logistical challenges compared to less obtrusive—though less internally valid—observational research methods. Still, researchers should consider choosing from a more complete menu of identification strategies that can exploit either non-randomized interventions or naturally-occurring variations.

6. DISCUSSION

Our review and our empirical results underline significant margins of improvement regarding experimental methods in several social sciences. In particular, management disciplines seem to lag behind best practice in statistical analysis of experimental data and in some dimensions of internal validity (e.g., demand effects, non-consequential studies and incentivization of decisions) and potential generalizability (e.g., paucity of field experimentation). After having presented a series of recommendations on how to make an experimental study rigorous, more generalizable, and relevant for practice, our empirical analysis has revealed that the advice we promulgated is rather wide-ranging and helps, not hurts, an article in the academic marketplace for citations. That is, across different fields, the markers of internal validity and external/ecological validity we recommended positively and significantly predict citations received by experimental articles. Publishing well-done experimental research benefits journals and rewards authors.

Some of our recommendations are similar to the ones already presented in manuals and reviews of experimental method both in OM (e.g., Katok, 2011) and other disciplines (e.g., Duflo et al., 2007; Gerber & Green, 2008; Hertwig & Ortmann, 2001). However, our unique contribution is to cover the most disparate aspects about experimental design, analysis, and interpretation and to find a balance between multiple approaches and needs coming from the different "methodological histories" of experimental OM (mainly psychology and economics).

Another distinctive characteristic of our article is the measurement the state of different academic fields vis-à-vis the methodological guidelines we propose.

Overall, our arguments suggest that methodological rigor and research relevance are complementary rather than substitutes. Rigor is an obvious antecedent of generalizability and practical relevance (Highhouse, 2009): Research with low internal validity misrepresents reality and is neither generalizable, nor useful (Antonakis, 2017; Gulati, 2007; Vermeulen, 2005). Next, some of the most important markers of internal validity that we have reviewed in this article are beneficial—and not antithetic—to experimental realism and generalizability. For instance, consequential decision environments are as important for achieving control in experimental settings as they are crucial for increasing the psychological realism and the subjects' immersion in the experimental environment.

However, the results of our study bring to the fore an important question: Given that, on average, good methodological practice is generally rewarded in the citation market, why do we sometimes observe a gap between current experimental work and our methodological suggestions? We can only speculate, yet it is likely that experimental research that does not comply with our guidelines is logistically simpler to conduct, less time-consuming and, ultimately, economically convenient. For instance, running non-consequential vignette experiments directly in the class where the experimenter teaches is faster and cheaper than organizing an incentivized experiment with real participant interaction. To this extent, researchers might find more profitable to produce more but lower-quality studies. Similar, using deception out of convenience makes running experiments quite easy. However, such a state of affairs is unfortunate and we urge the publication system gatekeepers to adopt stricter evaluation criteria in this regard because in doing so both relevance and rigor will improve. Also, reorganization of

doctoral training in experimental methods is required (cf. Aiken, West, & Millsap, 2008; Antonakis et al., 2010).

6.1. The ten commandments of experimental research

We synthetize our recommendations below and trust that they can serve as a useful guide for authors; the below is summarized in Table 6:

I. **Rigor is the sine qua non of experimental research**: Research that is not rigorous cannot be relevant, not even in vocationally oriented academic fields.

II. **When deciding about the appropriate type of experiment, use laboratory experiments for tightly controlled theory testing and detailed analysis of causal mechanisms; prefer field experiments if effect sizes within a particular population or context are of central interest.** Laboratory experiments provide tight control; however, to be internally valid, they must be consequential and avoid demand effects. Moreover, they are mostly suitable for individual- or group-level phenomena (having real-world behavioral, attitudinal, and decisional analogs). Field experiments, contrarily, allow establishing the magnitude and the generalizability of an effect size in naturalistic conditions and in the population of interest. As a general rule, empirical settings that can be simplified to a single type of decision (i.e., newsvendor quantities, auction-pricing) can be most appropriately first explored with laboratory experiments, particularly with regards to directionality and biases, whereas more context-specific decisions are best studied in the field.

III. **Experimental demand effects must not correlate with the treatment (i.e., they must be constant across treatments); specify an appropriate baseline to cleanly estimate the treatment effect.** Researchers should always construct a control group (i.e., baseline condition) representing a neutral level of the manipulated variable and of the related demand

characteristics; if possible they should also manipulate levels of the focal variable. Comparing only two treatments with opposite expected effects (i.e., a "medicine" and a "poison") does not identify which condition causes the observed effect and generates unfair comparisons.

IV. **Avoid hypothetical experiments (e.g., vignettes, conditional-choice type experiments) that are not consequential and are prone to demand effects.** Non-consequential, or self-reported assessments may be appropriate if the dependent variable of interest is an emotion or a perception. However, when it comes to studying behavior, such experiments are on the lowest rung of experimental evidence; they may represent a preliminary indication of the validity of a theory only if purged from demand effects.

V. **Use manipulation checks carefully and parsimoniously to minimize the risk of demand effects**. Manipulation checks should be always conducted after the measurement of the main experimental outcome and, ideally, be run on a separate—yet comparable—sample. Also, manipulation checks are meaningless—they are just "remembering checks"—when manipulations are non-consequential or solely demand-driven.

VI. **Avoid deceiving participants, particularly in laboratories.** Researchers should never misrepresent an experimental situation; deception should be used very sparingly and only as a last resort. Obfuscation or concealing the real purpose of the experiment, as well as some details of the design is permitted. In field experiments the use of deception might be appropriate if subject pool contamination is very unlikely and practical restrictions prohibit the use of non-deceptive solutions.

VII. **Guard against endogeneity by eliminating confounds via design or estimation where possible.** Simple comparisons of treatment means and plots of the row data are often sufficient to analyze experimental results. However, the required countermeasures should be

taken when modeling endogenous mediators, lagged dependent variables, or when there is serial correlation, failed randomization, or imperfect compliance to treatment.

VIII. **Ensure a sufficient sample size per cell for proper randomization of treatments to participants.** Use of within-session randomization is recommended whenever possible. If sessions have to be separated by treatment, it is crucial to minimize the risk of session effects by balanced randomization. Small sample sizes (i.e., $n < 50$) need to be avoided if possible. If external factors impose a small sample, it is important to establish ex-ante comparability of treatment groups and to use control variables in regressions.

IX. **The sample used must match the research question that will be answered.** Student participants are readily available and have the intellectual capacity to understand complex instructions. These characteristics make them ideal for initial theory tests and the establishment of general behavioral, attitudinal, and decisional mechanisms. EMBA students or employees can be a useful first step to test the practical relevance of a finding without giving up to the advantages of the laboratory.

X. **Use an appropriate quasi-experimental technique if field-experimentation is impossible.** Regression Discontinuity Design, Difference-in-Differences, and the Synthetic Control Method could be used more frequently to exploit premeasured assignment variables (either naturally occurring or implemented with the collaboration of a firm). They are less obtrusive than field experiments, yet mimic experimental control when key assumptions are met.

Each of our guidelines is self-standing, yet our "ten commandments" are also deeply interrelated. Together, they sketch a methodological protocol that can help to standardize the empirical practice in especially multidisciplinary fields and foster communication among experimentalists who come from different fields. Also, a common empirical framework can

support a fruitful scientific cycle, whereby laboratory and field experimentation complement each other and, together with quasi-experiments, represent a powerful tool to test and amend existing theories.

[Insert Table 6 about here]

6.2. Limitations

Finally, there are some limitations that should temper some of our results. Given space constraints, we have only treated a fraction of all possible topics pertaining to experimental methods. For instance, we did not discuss the novel opportunities and risks related to the emergence of electronic marketplaces (e.g., Amazon mTurk), which are now a common tool for many experimental researchers (Horton, Rand, & Zeckhauser, 2011). Similarly, we have not discussed some of the limits related to field experimentation (Eden, 2017), or advanced design and statistical issues (e.g., multiple hypotheses testing corrections, List, Shaikh, & Xu, 2016); discussing pre-analysis plans would have been very useful too (e.g., Olken, 2015).

As for our results, first, the information we extracted from the reviewed articles was manually coded and is based on a limited subset of articles; it also required several judgement calls. Next, our results are descriptive; what causes the differences we observe across disciplines as well as what causally drives citations is not clear. We leave to future research the objective of enlarging the scope of our quantitative review and of exploring more thoroughly the causality matter. Similarly, we hope future inquiries will build on our work and will review more advanced methodological topics related to experimentation.

7. CONCLUSION

The epigraph at the beginning of our article nicely captures the need for experimental observation. Of course, experiments must be well done and they are not a panacea. The infamous

experiments that suggested neutrinos travelled faster than the speed of light were even replicated; yet, so contrary were these results to theory that most physicists did not believe them. Faulty cabling and measurement finally explained the mishap (Reich, 2012). Similarly, Jacques Benveniste destroyed his career by publishing an article in *Nature* that initially looked well done and experimentally demonstrated that homeopathic preparations had effects. However, these results violated the laws of chemistry and a tightly controlled experiment showed Benveniste's team fell prey to experimenter expectation effects because of a lax experimenter blinding protocol (Maddox, Randi, & Stewart, 1988). Given the respect scientists have for the experimental method, it is easy to see how specious experimental results might lead researchers down the garden path. Thus, we cannot restate enough times that rigor is the *sine qua non* of all research including experimental research.

Experiments are becoming mainstream. Several academic fields that had been historically averse to experimental methods like finance and financial accounting (Haigh & List, 2005; Libby, Bloomfield, & Nelson, 2002), political science (McDermott, 2002), and even archeology (Outram, 2008) are now embracing them. Experiments represent a great opportunity for advancing scientific knowledge, as well as lay down an important methodological challenge: Because experimental evidence should be the final judge for the validity of our theories, we have an intellectual obligation to employ the best possible experimental methods.

In this article, we have discussed how to make experimental research in operations management—and in the social sciences more in general—valid, realistic, and relevant for practice. Our theories must rest on solid empirical grounds; the practice and the science of management demand cutting-edge methods to solve the challenges facing our rapidly changing society. We are confident that experimentation—if rigorously conducted—represents an

important method for discovering new knowledge and for ultimately testing and contributing to the developing of new theory.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association, 105*(490), 493-505.

Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software, 42*(13), 1-17.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science, 59*(2), 495-510.

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.

Adenso-Diaz, B., & Laguna, M. (2006). Fine-tuning of algorithms using fractional experimental designs and local search. *Operations Research, 54*(1), 99-114.

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., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270-301.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*(1), 32.

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8*(1), 3-9.

Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics, 18*(1), 47-82.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics, 110*(1), 1-21.

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

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*: Princeton university press.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly, 28*(1), 5-21.

Antonakis, J., Bastardoz, N., Liu, Y., & Schriesheim, C. A. (2014). What makes articles highly cited? *The Leadership Quarterly, 25*(1), 152-179.

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

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions *The Oxford Handbook of Leadership and Organizations* (pp. 93-117).

Antonakis, J., d'Adda, G., Weber, R., & Zender, C. (2014). Just words? Just speeches? On the economic value of charismatic leadership. *NBER Reporter, 4*.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies, 58*(2), 277-297.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies, 76*(2), 451-469.

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

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

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14*(4), 496-515.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Bastardoz, N., Jacquart, P., & Antonakis, J. (2017). *When ANOVA gets it wrong: A re-introduction to the Regression Discontinuity design.* Paper presented at the Academy of Management Proceedings.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*(4), 396-403.

Belot, M., Duch, R., & Miller, L. (2015). A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior & Organization, 113*, 26-33.

Benabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies, 70*(3), 489-520.

Bendahan, S., Zehnder, C., Pralong, F. P., & Antonakis, J. (2015). Leader corruption depends on power and testosterone. *The Leadership Quarterly, 26*(2), 101-122.

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.

Bendoly, E., Donohue, K., & Schultz, K. L. (2006). Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management, 24*(6), 737-752.

Bergh, D. D., & Perry, J. (2006). Some predictors of SMJ article impact. *Strategic Management Journal, 27*(1), 81-100.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37*(3), 245-257.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics, 119*(1), 249-275.

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H.-T. (2010). Belief elicitation in experiments: is there a hedging problem? *Experimental Economics, 13*(4), 412-438.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry, 13*(2), 103-124.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics, 87*(1), 115-143.

Bøggild, T., & Laustsen, L. (2016). An intra-group perspective on leader preferences: Different risks of exploitation shape preferences for leader facial dominance. *The Leadership Quarterly, 27*(6), 820-837.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 89*(1), 1-34.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467): John Wiley & Sons.

Bolton, G. E., & Ockenfels, A. (2014). Does laboratory trading mirror behavior in real world markets? Fair bargaining and competitive bidding on eBay. *Journal of Economic Behavior & Organization, 97*, 143-154.

Bolton, G. E., Ockenfels, A., & Thonemann, U. W. (2012). Managers and students as newsvendors. *Management Science, 58*(12), 2225-2233.

Bou, J. C., & Satorra, A. (2018). Univariate Versus Multivariate Modeling of Panel Data: Model Specification and Goodness-of-Fit Testing. *Organizational Research Methods, 21*(1), 150-196.

Boyce, R. R., Brown, T. C., McClelland, G. H., Peterson, G. L., & Schulze, W. D. (1992). An experimental examination of intrinsic values as a source of the WTA-WTP disparity. *American Economic Review, 82*(5), 1366-1373.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376.

Camerer, C. F. (2015). The promise and success of lab-field generalizability in experimental economics: A critical reply to Levitt and List. In G. Frechette & A. Schotter (Eds.), *Handbook of Experimental Economic Methodology ed* Oxford: Oxford University Press.

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

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics, 29*(2), 238-249.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317-372.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297.

Cannon, J. P., Doney, P. M., Mullen, M. R., & Petersen, K. J. (2010). Building long-term orientation in buyer–supplier relationships: The moderating role of culture. *Journal of Operations Management, 28*(6), 506-521.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review, 84*(4), 772-793.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science, 21*(10), 1363-1368.

Caudill, S. B. (1988). Practitioners corner: An advantage of the linear probability model over probit or logit. *Oxford Bulletin of Economics and Statistics, 50*(4), 425-427.

Chatterji, A. K., Findley, M., Jensen, N. M., Meier, S., & Nielson, D. (2016). Field experiments in strategy research. *Strategic Management Journal, 37*(1), 116-132.

Chen-Ritzo, C.-H., Harrison, T. P., Kwasnica, A. M., & Thomas, D. J. (2005). Better, faster, cheaper: An experimental analysis of a multiattribute reverse auction mechanism with restricted information feedback. *Management Science, 51*(12), 1753-1762.

Chenhall, R. H., & Moers, F. (2007). The issue of endogeneity within theory-based, quantitative management accounting research. *European Accounting Review, 16*(1), 173-196.

Colquitt, J. A. (2008). From the editors publishing laboratory research in AMJ: A question of when, not if. *Academy of Management Journal, 51*(4), 616-620.

Cooper, D. J., Kagel, J. H., Lo, W., & Gu, Q. L. (1999). Gaming against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers. *American Economic Review, 89*(4), 781-804.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71*(2), 179-184.

Croson, R., & Donohue, K. (2002). Experimental economics and supply-chain management. *Interfaces, 32*(5), 74-82.

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

Dal Bó, P., & Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review, 101*(1), 411-429.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*: Princeton university press.

De Treville, S., & Antonakis, J. (2006). Could lean production job design be intrinsically motivating? Contextual, configurational, and levels-of-analysis issues. *Journal of Operations Management, 24*(2), 99-123.

De Treville, S., Edelson, N. M., Kharkar, A. N., & Avanzi, B. (2008). Constructing useful theory: The case of Six Sigma. *Operations Management Research, 1*(1), 15-23.

DeHoratius, N., Gürerk, Ö., Honhon, D., & Hyndman, K. B. (2015). Understanding the behavioral drivers of execution failures in retail supply chains: An experimental study using virtual reality. *Unpublished working paper*.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*(6), 959-988.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association, 9*(3), 522-550.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2), 245-260.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics, 4*, 3895-3962.

Eden, D. (2017). Field Experiments in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 4*(1), 91-122. doi:10.1146/annurev-orgpsych-041015-062400

Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *Academy of Management Review, 32*(4), 1246-1264.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *Academy of Management Journal, 50*(1), 25-32.

Emsley, R., Dunn, G., & White, I. R. (2010). Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research, 19*(3), 237-270.

Exadaktylos, F., Espín, A. M., & Branas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports, 3*(1213), 1-6.

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535-538.

Falk, A., Meier, S., & Zehnder, C. (2013). Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples. *Journal of the European Economic Association, 11*(4), 839-852.

Falk, A., & Szech, N. (2013). Morals and markets. *Science, 340*(6133), 707-711.

Fehr, E., & Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review, 97*(1), 298-317.

Feynman, R. (1965). *The character of physical law*. Cambridge: MIT Press.

Fischer, T., Dietz, J., & Antonakis, J. (2016). Leadership Process Models: A Review and Synthesis. *Journal of Management, 43*(6), 1726-1753.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*(2), 303-315.

Flammer, C. (2015). Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach. *Management Science, 61*(11), 2549-2568.

Fréchette, G. R. (2011). Session-effects in the laboratory. *Experimental Economics, 15*(3), 485-498.

Fréchette, G. R. (2014). Experimental economics across subject populations. The handbook of experimental economics, vol. 2. Eds. Kagel J., Roth A: Princeton, NJ: Princeton University Press.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality, 23*(5), 369-401.

Gächter, S., & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics, 13*(3), 364-377.

Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology, 85*(3), 453.

Gans, N., & Croson, R. (2008). Introduction to the special issue on behavioral operations. *Manufacturing & Service Operations Management, 10*(4), 563-565.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118*(3), 392-404.

Gerber, A. S., & Green, D. P. (2008). Field experiments and natural experiments *The Oxford Handbook of Political Methodology*.

Gerring, J., & McDermott, R. (2007). An experimental template for case study research. *American Journal of Political Science, 51*(3), 688-701.

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

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*: Princeton University Press.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics, 115*(3), 791-810.

Grant, A. M. (2008). The significance of task significance: Job performance effects, relational mechanisms, and boundary conditions. *Journal of Applied Psychology, 93*(1), 108-124.

Grant, A. M., & Wall, T. D. (2009). The neglected science and art of quasi-experimentation: Why-to, when-to, and how-to advice for organizational researchers. *Organizational Research Methods, 12*(4), 653-686.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science, 70*(5), 1195-1205.

Guala, F., & Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology, 12*(4), 495-515.

Guide, V. D. R., & Ketokivi, M. (2015). Notes from the editors: Redefining some methodological criteria for the journal. *Journal of Operations Management*(37), v-viii.

Gulati, R. (2007). Tent poles, tribalism, and boundary spanning: The rigor-relevance debate in management research. *Academy of Management Journal, 50*(4), 775-782.

Gunst, R. F., & Mason, R. L. (2009). Fractional factorial design. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*(2), 234-244.

Gupta, S. K. (2011). Intention-to-treat concept: a review. *Perspectives in Clinical Research, 2*(3), 109-112.

Haigh, M. S., & List, J. A. (2005). Do professional traders exhibit myopic loss aversion? An experimental analysis. *The Journal of Finance, 60*(1), 523-534.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization, 1*(1), 51-78.

Harrison, G. W., Haruvy, E., & Rutström, E. E. (2011). Remarks on virtual world and virtual reality experiments. *Southern Economic Journal, 78*(1), 87-94.

Hartmann, J., & Moeller, S. (2014). Chain liability in multitier supply chains? Responsibility attributions for unsustainable supplier behavior. *Journal of Operations Management, 32*(5), 281-294.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251-1271.

Heckman, J. J. (1979). Sample Selection Bias As A Specification Error. *Econometrica, 47*(1), 153-161.

Herbst, D., & Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science, 350*(6260), 545-549.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(03), 383-403.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*(3), 554-566.

Holmström, J., Ketokivi, M., & Hameri, A. P. (2009). Bridging practice and theory: A design science approach. *Decision Sciences, 40*(1), 65-87.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399-425.

Hosoda, M., Sone-Romero, E. F., & Coats, G. (2003). The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology, 56*(2), 431-462.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med, 2*(8), e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648.

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), 477-488.

Kacmar, K. M., & Whitfield, J. M. (2000). An additional rating method for journal articles in the field of management. *Organizational Research Methods, 3*(4), 392-406.

Katok, E. (2011). Using laboratory experiments to build better operations management models. *Foundations and Trends® in Technology, Information and Operations Management, 5*(1), 1-86.

Katok, E., Olsen, T., & Pavlov, V. (2014). Wholesale pricing under mild and privately known concerns for fairness. *Production and Operations Management, 23*(2), 285-302.

Katok, E., & Pavlov, V. (2013). Fairness in supply chain contracts: A laboratory study. *Journal of Operations Management, 31*(3), 129-137.

Kessler, J., & Vesterlund, L. (2015). The external validity of laboratory experiments: The misleading emphasis on quantitative effects. *In Handbook of Experimental Economic Methodology, Oxford University Press, Oxford, UK, 20*.

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

Kidd, R. F. (1976). Manipulation checks: advantage or disadvantage? *Representative Research in Social Psychology, 7*(2), 160-165.

Kleindorfer, P. R., Singhal, K., & Wassenhove, L. N. (2005). Sustainable operations management. *Production and Operations Management*, *14*(4), 482-492.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673-676.

Kröll, M., & Rustagi, D. (2016). Shades of dishonesty and cheating in informal milk markets in India. *Unpublished working paper*.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes, 75*(1), 23-55.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281-355.

Levitt, S. D., & List, J. A. (2007a). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d'économique, 40*(2), 347-370.

Levitt, S. D., & List, J. A. (2007b). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives, 21*(2), 153-174.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

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

Libby, R., Bloomfield, R., & Nelson, M. W. (2002). Experimental research in financial accounting. *Accounting, Organizations and Society, 27*(8), 775-810.

Linton, J. D., Klassen, R., & Jayaraman, V. (2007). Sustainable supply chains: An introduction. *Journal of Operations Management, 25*(6), 1075-1082.

List, J. A., Shaikh, A. M., & Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *NBER Paper 21875*.

Maddox, J., Randi, J., & Stewart, W. W. (1988). " High-dilution" experiments a delusion. *Nature, 334*(6180), 287.

McDermott, R. (2002). Experimental methods in political science. *Annual Review of Political Science, 5*(1), 31-61.

Meredith, J. R. (1998). Building operations management theory through case and field research. *Journal of Operations Management, 16*(4), 441-454.

Meredith, J. R. (2001). Hopes for the future of operations management. *Journal of Operations Management, 4*(19), 397-402.

Mettler, T., Eurich, M., & Winter, R. (2014). On the Use of Experiments in Design Science Research: A Proposition of an Evaluation Framework. *CAIS, 34*, 10.

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371-378.

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*(2), 109-117.

Montgomery, J. M., Nyhan, B., & Torres, M. (2016). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62*(3), 760-775.

Montmarquette, C., Rullière, J.-L., Villeval, M.-C., & Zeiliger, R. (2004). Redesigning teams and incentives in a merger: An experiment with managers and students. *Management Science, 50*(10), 1379-1389.

Nair, V., Strecher, V., Fagerlin, A., Ubel, P., Resnicow, K., Murphy, S., . . . Zhang, A. (2008). Screening experiments and the use of fractional factorial designs in behavioral intervention research. *American Journal of Public Health, 98*(8), 1354-1359.

Naor, M., Linderman, K., & Schroeder, R. (2010). The globalization of operations in Eastern and Western countries: Unpacking the relationship between national and organizational culture and its impact on manufacturing performance. *Journal of Operations Management, 28*(3), 194-205.

Olivares, M., Terwiesch, C., & Cassorla, L. (2008). Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science, 54*(1), 41-55.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives, 29*(3), 61-80.

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

Outram, A. K. (2008). Introduction to experimental archaeology. *World Archaeology, 40*(1), 1-6.

Özer, Ö., Zheng, Y., & Chen, K.-Y. (2011). Trust in forecast information sharing. *Management Science, 57*(6), 1111-1137.

Özer, Ö., Zheng, Y., & Ren, Y. (2014). Trust, trustworthiness, and information sharing in supply chains bridging China and the United States. *Management Science, 60*(10), 2435-2460.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Perdue, B. C., & Summers, J. O. (1986). Checking the success of manipulations in marketing experiments. *Journal of Marketing Research, 23*(4), 317-326.

Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review, 18*(4), 599-620.

Plott, C. R. (1991). Will economics become an experimental science? *Southern Economic Journal, 57*(4), 901-919.

Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. L. (2013). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *Journal of Applied Psychology, 98*(1), 99-113.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*(1), 539-569.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management, 12*(4), 531-544.

Podsakoff, P. M., & Podsakoff, N. P. (2018). Experimental Designs in Management and Leadership Research: Strengths, Limitations, and Recommendations for Improving Publishability. *The Leadership Quarterly*.

Primo, D. M., Jacobsmeier, M. L., & Milyo, J. (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly, 7*(4), 446-459.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. *Psychological Science, 26*(5), 653-656.

Reich, E. S. (2012). Embattled neutrino project leaders step down. *Nature News*. doi:10.1038/nature.2012.10371

Reis, H. T., & Judd, C. M. (2000). *Handbook of research methods in social and personality psychology*: Cambridge University Press.

Ribbink, D., & Grimm, C. M. (2014). The impact of cultural differences on buyer–supplier negotiations: An experimental study. *Journal of Operations Management, 32*(3), 114-126.

Riccobono, F., Bruccoleri, M., & Größler, A. (2015). Groupthink and Project Performance: The Influence of Personal Traits and Interpersonal Ties. *Production and Operations Management, 25*(4), 609-629.

Roberts, M. R., & Whited, T. M. (2013). Chapter 7 - Endogeneity in Empirical Corporate Finance1. In G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.), *Handbook of the Economics of Finance* (Vol. 2, pp. 493-572): Elsevier.

Rodríguez, D., Buyens, D., Landeghem, H., & Lasio, V. (2016). Impact of lean production on perceived job autonomy and job satisfaction: An experimental study. *Human Factors and Ergonomics in Manufacturing & Service Industries, 26*(2), 159-176.

Roth, A. (1987). *Laboratory experimentation in economics, and its relation to economic theory*: Lanham, University Press of America.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubinstein, A. (2001). A theorist's view of experiments. *European Economic Review, 45*(4), 615-628.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54-67.

Schaerer, M., du Plessis, C., Yap, A. J., & Thau, S. (2018). Low power individuals in social power research: A quantitative review, theoretical framework, and empirical test. *Organizational Behavior and Human Decision Processes, 149*, 73-96. doi:https://doi.org/10.1016/j.obhdp.2018.08.004

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology, 12*(2), 225-237.

Schweitzer, M. E., & Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science, 46*(3), 404-420.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*: Wadsworth Cengage learning.

Siegel, S., & Harnett, D. L. (1964). Bargaining behavior: A comparison between mature industrial personnel and college students. *Operations Research, 12*(2), 334-343.

Sigall, H., & Mills, J. (1998). Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review, 2*(3), 218-226.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. *Paper presented at the Meeting of the Society for Personality and Social Psychology, New Orleans, LA*.

Smith, V. L. (1976). Experimental economics: Induced value theory. *American Economic Review, 66*(2), 274-279.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry, 31*(2), 245-261.

Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science, 35*(3), 321-339.

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

Strube, M. J. (1991). Small sample failure of random assignment: a further examination. *Journal of Consulting and Clinical Psychology, 59*(2), 346-350.

Sturm, R. E., & Antonakis, J. (2015). Interpersonal Power. *Journal of Management, 41*(1), 136-163.

The Editors of Encyclopædia Britannica. (2016). Behavioral science. Retrieved from https://www.britannica.com/science/behavioral-science

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Van Aken, J. E. (2005). Management research as a design science: Articulating the research products of mode 2 knowledge production in management. *British Journal of Management, 16*(1), 19-36.

Van Aken, J. E., Chandrasekaran, A., & Halman, J. (2016). Conducting and publishing design science research: Inaugural essay of the design science department of the Journal of Operations Management. *Journal of Operations Management, 47-48*, 1-8.

Vermeulen, F. (2005). On rigor and relevance: Fostering dialectic progress in management research. *Academy of Management Journal, 48*(6), 978-982.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine, 37*(5), 360-363.

Von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75-105.

Wacker, J. G. (1998). A definition of theory: research guidelines for different theory-building research methods in operations management. *Journal of Operations Management, 16*(4), 361-385.

Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14*(4), 490-495.

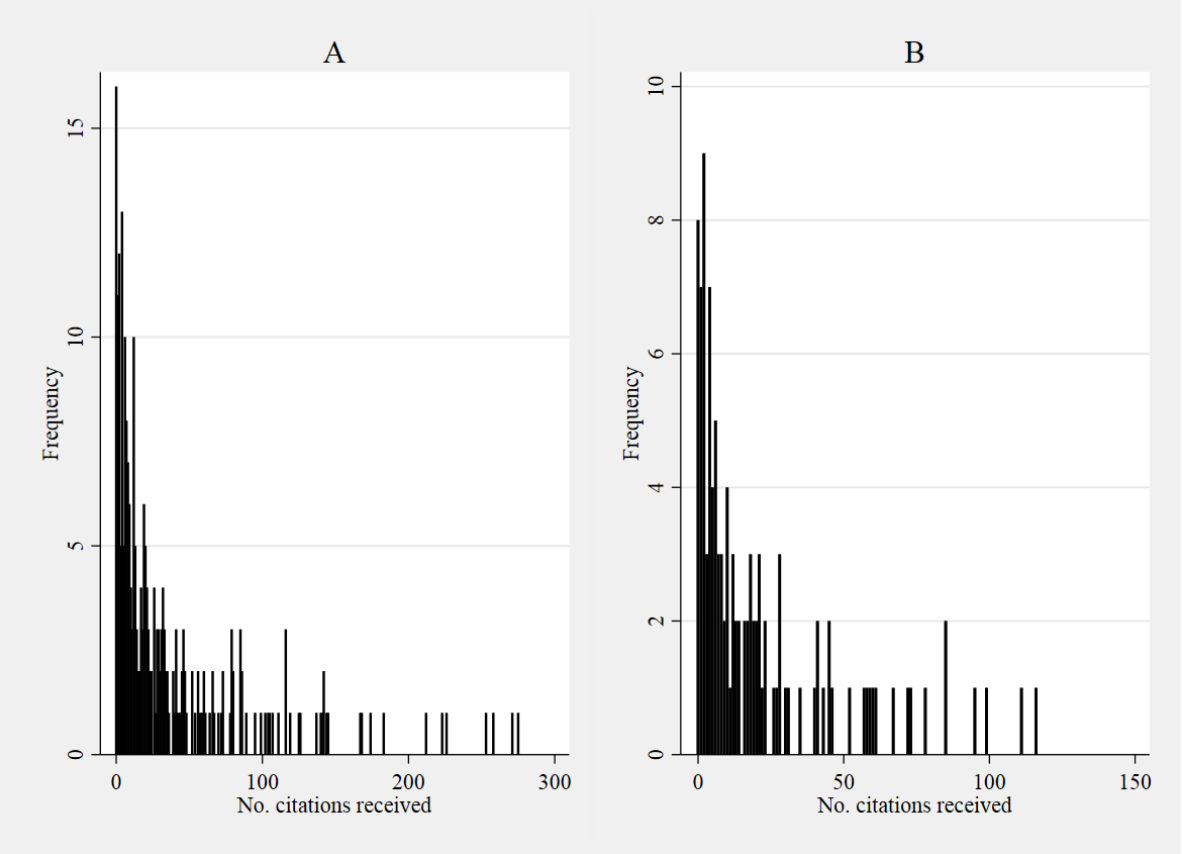
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*: MIT press.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*: Nelson Education.

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

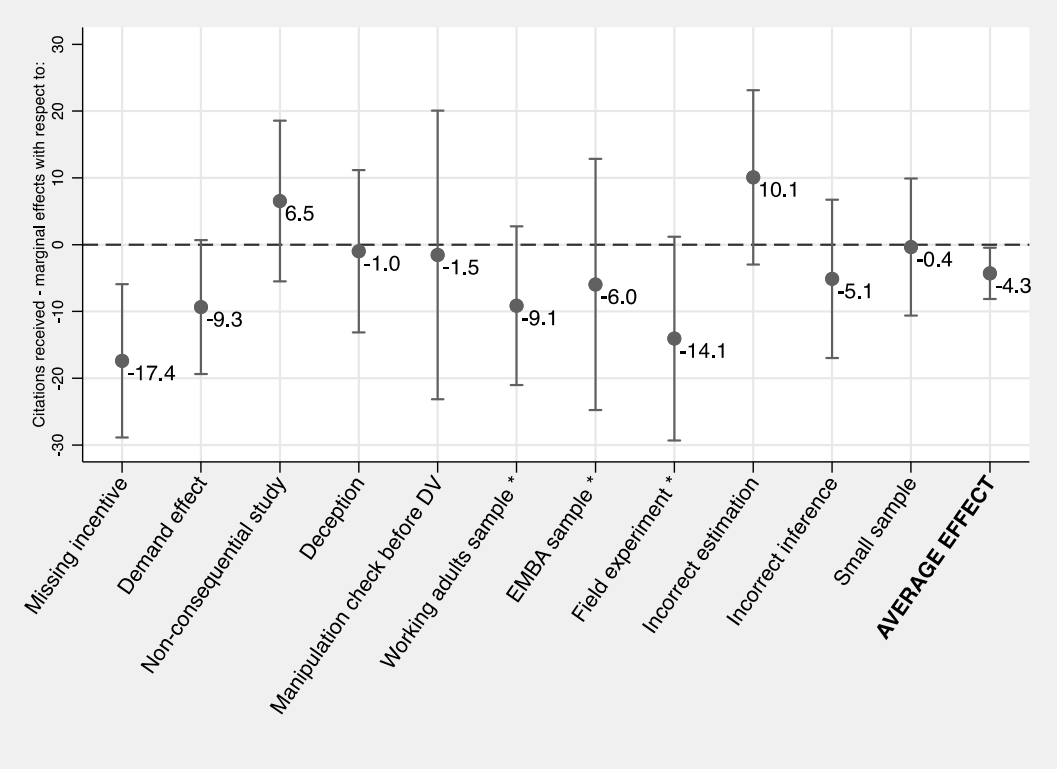
FIGURES AND TABLES

Figure 1. Citations distribution (retrieved from Scopus in September 2017).



The *x*-axis refers to the total number of citations received by an article; the y-axis refers to the number of articles in the sample that exhibit a given number of total citations. Panel A (*n* = 258, *mean* = 38.37, *sd* = 52.16, *min* = 0, *max* = 275, *proportion of 0 citation* = 6.20%) shows the overall distribution for the entire sample, whereas Panel B (*n* = 111, *mean* = 21.68, *sd* = 26.45, *min* = 0, *max* = 116, *proportion of 0 citations* = 7.21%) reports the citation distribution for the OM field.

Figure 2. Effect of validity markers on citations.



Average marginal effect of each internal, statistical and external/ecological validity marker on the citations received by the experimental articles in our sample (based on the regression in column (1), Table 4, with 95% confidence intervals). The graph depicts the average discrete effect of each binary validity indicator on the average total citations received by an article. The last estimate is the overall effect using a linear combination of estimators. External validity indicators (marked by *) are reverse-coded, in order to interpret the average effect as the consequence of *not* employing working adults samples, EMBA samples, or field experiments.

Table 1. The major internal and statistical validity threats in experimental research.

Internal validity threats

Name	*Explanation*
Unfair comparisons and demand effects	• Social desirability and experimenter effects. • Asymmetric demands between treatment and control group (e.g., missing placebo group and "medicine vs poison" treatment comparison).
Non-consequential decision environments	• No linkage between actions made and consequences faced by an experimental subject. • Absence of extrinsic motivators. • In case of absence of formal extrinsic motivators, lack of any form of intrinsic motivator.
Deception	• Misrepresentation of important elements of the experimental setting (especially problematic in laboratory experiments).
Manipulation checks before the assessment of dependent variable	• They induce demand effects, because they make salient the purpose of the experiment in the eyes of participants.

Statistical validity threats

Name	*Explanation*
Incorrect inference	• Lack of cluster-robust standard errors when data are hierarchically clustered.
Failed randomization	• Unbalanced allocation of covariates across treatments due to small sample size (i.e., if $n < 50$ per experimental cell, see the results of our Monte Carlo simulation in Appendix; for similar suggestions, see also Simmons et al., 2013).
Lagged dependent variable, endogenous mediator and local non-identification	• Endogeneity triggered by the autocorrelated error term. • Mediator and outcome variable both depend on an unobserved variable. • Instrumental variable estimator requires at least one instrument, excluded from the *y* equation, per endogenous mediator.

Non-compliance to treatment	• Excluding non-compliers.

Table 2. Internal and statistical validity across fields.

	(1) No compensation/incentive	(2) Manipulation check before DV	(3) Demand effect	(4) Deception	(5) Non-consequential study	(6) Incorrect estimation	(7) Incorrect inference	(8) Small sample ($n < 50$)
OM Econ.	.17***	.01	-.04	.06	-.07	.12	.17**	.36***
	(.06)	(.02)	(.09)	(.05)	(.08)	(.07)	(.09)	(.10)
OM Non-econ.	.42***	.03	.26**	.06	.25**	.16**	-.05	.20*
	(.09)	(.04)	(.12)	(.05)	(.12)	(.08)	(.07)	(.11)
Management	.15*	.14**	.37***	.21***	.50***	.17**	-.06	.36***
	(.08)	(.06)	(.11)	(.07)	(.10)	(.07)	(.07)	(.11)
Psychology	.35***	.04	.45***	.30***	.56***	.15**	-.10*	.32***
	(.09)	(.03)	(.11)	(.08)	(.09)	(.07)	(.05)	(.10)
Constant	-.03	-.01	.17**	-.03	.19***	.02	.15***	.13*
	(.04)	(.01)	(.08)	(.04)	(.07)	(.04)	(.05)	(.08)
R-squared	.18	.12	.38	.31	.51	.11	.14	.36
Coders FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pub. Year Fes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bib. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F statistic	3.78^{***}	.69	15.80^{***}	2.16^{***}	36.40	1.55^{*}	2.10^{***}	10.99^{***}
Difference (a-b)	**	ns	***	ns	***	ns	***	ns

n = 468 experimental studies; Clustered standard errors in parentheses (article-level); ***$p < .01$, **$p < .05$, *$p < .10$, ns: $p > .10$
Omitted category is economics. The distinction between OM Economics-based and OM non-Economics-based was a coded category (observed inter-coder agreement on this variable was 90%). All regressions include bibliographical controls (journal impact factor, number of experimental studies in each article, number of authors, number of references, sum of authors' citation until publication date, sum of authors' articles until publication date, average ranking of the authors' universities), coder fixed effects and publication-year fixed effects. All covariates and fixed-effects are mean-centered. The row "Difference (a-b)" refers to the test for the difference of the coefficients of OM Economics-based and OM Non-Economics-based ($F(1, 257)$).

Table 3. External and ecological validity across fields.

	(1) Working adults	(2) Grad/EMBA students	(3) Field experiment
OM Econ.[a]	.00	.04	-.39***
	(.05)	(.03)	(.09)
OM Non-econ.[b]	.18**	.15***	-.35***
	(.07)	(.06)	(.09)
Management	.05	.10**	-.35***
	(.07)	(.05)	(.09)
Psychology	-.02	.05**	-.42***
	(.04	(.03)	(.09)
Constant	.05	-.03	.43***
	(.04)	(.02)	(.08)
R-squared	.16	.13	.26
Coders Fes	Yes	Yes	Yes
Pub. Year Fes	Yes	Yes	Yes
Bib. Controls	Yes	Yes	Yes
F statistic	1.49*	.85	2.41***
Difference (a-b)	***	**	ns

n = 468 experimental studies; Clustered standard errors in parentheses (article-level); ****p* < .01, ***p* < .05, **p* < .10, ns: *p* > .10.

Omitted category is Economics. The distinction between OM Economics-based and OM non-Economics-based was a coded category (observed inter-coder agreement on this variable was 90%). All regressions include bibliographical controls (journal impact factor, number of experimental studies in each article, number of authors, number of references, sum of authors' citation until publication date, sum of authors' articles until publication date, average ranking of the authors' universities), coder fixed effects and publication-year fixed effects. All covariates and fixed-effects are mean-centered. [a,b]The row "Difference (a-b)" refers to the test for the difference of the coefficients of OM Economics-based and OM Non-Economics-based (*F(1, 257)*).

Table 4. Citation analysis.

	Article citations					
	All fields	All fields	Operations	Econ.	Mngt	Psych.
5-year impact factor		.11***	.04	-.06	.17***	.15**
		(.03)	(.05)	(.08)	(.04)	(.08)
Article age	.79***	.74***	.97***	.79***	.81***	1.03***
	(.08)	(.08)	(.12)	(.16)	(.11)	(.21)
Article age squared	-.04***	-.03***	-.05***	-.04***	-.04***	-.05***
	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)
Number of authors	.08	.08	.10	.28**	-.05	.09
	(.05)	(.05)	(.13)	(.13)	(.08)	(.08)
Number of studies	-.01	.01	.12	-.48***	-.28**	.08
	(.06)	(.06)	(.09)	(.15)	(.12)	(.11)
No. refs	.81***	.62***	.30	.41	1.15***	.91**
	(.19)	(.20)	(.34)	(.42)	(.41)	(.42)
Authors' citations[a]	.01	.01*	-.01	-.00	.01*	.01
	(.00)	(.01)	(.01)	(.01)	(.01)	(.03)
No articles team[a]	-.12	-.13	-.04	-.19	-.35	-.29
	(.10)	(.10)	(.12)	(.91)	(.64)	(1.06)
Team ave. uni. Rank	.05	.07*	.05	.30***	-.13**	.28***
	(.04)	(.04)	(.04)	(.09)	(.06)	(.07)
No incentive[+]	-.42***	-.43***	-.63***		-.54**	-.47*
	(.14)	(.15)	(.21)		(.24)	(.27)
Demand effect[+]	-.23*	-.31**	.01	-.82**	-.19	-.20
	(.12)	(.13)	(.23)	(.40)	(.17)	(.19)
Non-consequential[+]	.16	.15	.09	.39	.30	.39*
	(.15)	(.17)	(.34)	(.46)	(.25)	(.21)
Deception[+]	-.02	-.00	.29	-.05	.00	-.34
	(.15)	(.14)	(.28)	(.31)	(.25)	(.21)
Manip. Check before DV[+]	-.04	.04	-.03		.01	1.09*
	(.27)	(.32)	(.37)		(.15)	(.59)
Working adults[+]	.22	.08	.31	.34	.25*	.04
	(.14)	(.15)	(.26)	(.46)	(.14)	(.30)
Grad/EMBA students[+]	.14	.09	-.16		.47	
	(.23)	(.24)	(.27)		(.29)	
Field experiment[+]	.34*	.32*	.42	.40	.67	
	(.18)	(.19)	(.31)	(.28)	(.41)	
Incorrect estimation[+]	.24	.11	.51*	.18	-.32**	-.48*
	(.16)	(.14)	(.27)	(.57)	(.16)	(.25)
Incorrect inference[+]	-.12	-.15	-.14	.38	-.43	-.47
	(.15)	(.15)	(.18)	(.30)	(.28)	(.42)
Small sample[+] ($n < 50$)	-.01	.02	-.33*	.87***	.45	.34
	(.13)	(.14)	(.19)	(.27)	(.32)	(.22)
Only diff.-in-means	-.03	-.11	.16	-.34	-.17	-.18

	(.12)	(.13)	(.19)	(.42)	(.57)	(.18)
Constant	-.11	-.90**	-.87	-1.10	-1.26	-3.44***
	(.43)	(.46)	(.63)	(1.07)	(.89)	(.96)
Ave. of validity markers	-.10**	-.10*	-.07	.02	-.19***	-.02
Observations (studies)	468	468	162	60	87	159
Coders FE	Yes	Yes	Yes	Yes	Yes	Yes
Publication year FE	Yes	Yes	Yes	Yes	Yes	Yes
Journal FE	Yes	No	No	No	No	No
Field FE	No	Yes	No	No	No	No
Pseudo R2	.15	.13	.14	.14	.23	.16

Clustered standard errors in parentheses (article-level).

*** p<.01, ** p<.05, * p<.10

Negative binomial regression results (over-dispersion parameter ln α always significant at the 1% level). Bibliographical controls include: article age, article age squared, number of authors, number of studies, number of references reported by each article, citations received by the team of authors, articles published by the team of authors, and average ranking of the authors' universities of affiliation. Authors' citations, articles published, ranking of the university and number of references reported by the coded article are divided by 100. The beta coefficients reported represent the difference in the logs of predicted citations, holding constant all other variables. Average effect of validity markers = linear combination of the coefficients of internal, external, and statistical validity markers highlighted with the symbol $^{+}$ (with external validity makers reverse-coded, to interpret the average effect as the consequence of *not* following our methodological suggestions).[a]till publication date;

Table 5. Predicted Vs actual citations.

	1	2	3	4	5	6
Field	**Predicted citations**	**Predicted citations per year (5 years)**	**Median IF of journals in field**	**5-year IF of reviewed journals**	**% diff (3) to (1)**	**% diff (4) to (2)**
OM	2.79	6.43	1.49	4.44	+87%	+44%$^{+}$
Economics	3.06	7.06	.93	6.35	+229%**	+11%
Management	1.79	4.13	1.72	9.17	+4%	-55%***
Psychology	2.85	6.56	2.13^{a}	5.13	+34%	+28%
			1.62^{b}		+76%	

^{a}Psychology; ^{b}Psychology-Social

Asterisk refer to $\chi^2(1)$ test with Bonferroni correction for testing multiple hypotheses

$^{*}p < .10$; $^{**}p < .05$; $^{***}p < .01$, $^{+}p = .10$

Median impact factor data refer to JCR Year 2016; 5-year impact factor is computed as the mean at the academic field level of the impact factors of the journals reviewed in our article.

Table 6. The "ten commandments" of experimental research

I.	**Rigor:** Rigor is the sine qua non of experimental research; relevance follows.
II.	**Laboratory vs field experiments:** Use laboratory experiments for controlled theory testing of real-world analogs; prefer field experiments to estimate the statistical effect for a particular population and context.
III.	**Demand effects and baselines:** Hold experimental demand characteristics constant across treatments and clearly specify a baseline condition.
IV.	**Vignette studies:** Avoid experiments with hypothetical choices.
V.	**Manipulation checks:** Use manipulation checks carefully and parsimoniously.
VI.	**Deception:** Avoid deceiving participants; obfuscation is allowed.
VII.	**Endogeneity:** Guard against omitted variables and noncompliance.
VIII.	**Randomization:** Ensure an appropriate sample size per experimental cell for covariate balance ($n > 50$ per cell).[1]
IX.	**Sample composition:** Match the right sample to the right research question. Undergraduate student samples are legitimate samples to use.
X.	**Quasi-experimentation:** Use appropriate non-experimental identification techniques if field-experimentation is impossible.

Note: [1]We wish to stress that what determines an appropriate sample size is very context specific. The above threshold is purely indicative and based on a series of simulations that consider only a fraction of all possible empirical scenarios (see Appendix A.3). Throughout all simulations, we only studied the distribution of dichotomous covariates (both orthogonal and correlated) randomized into two balanced experimental conditions. Also, the $n > 50$ threshold was constructed only with reference to the issue of failed randomization: Clearly, employing the same rule of thumb will not necessarily ensure sufficient statistical power of an experiment. Thus, researchers should always be wary of any rules of thumb and should take great care to justify their sample size on the basis of field-specific criteria.