10-2022

Field experiments in operations management

Yang GAO
Singapore Management University, ygao@smu.edu.sg

Meng LI

Shujing SUN

Field Experiments in Operations Management

Yang Gao
School of Computing and Information Systems, Singapore Management University, ygao@smu.edu.sg
Meng Li
C.T. Bauer College of Business, University of Houston, mli@bauer.uh.edu
Shujing Sun
Naveen Jindal School of Management, the University of Texas at Dallas, shujing.sun@utdallas.edu

While the field experiment is a powerful and well-established method to investigate causal relationships, operations management (OM) has embraced this methodology only in recent years. This paper provides a comprehensive review of the existing OM literature leveraging field experiments and serves as a one-stop guide for future application of field experiments in the OM area. We start by recapping the characteristics that distinguish field experiments from other common types of experiments and organizing the relevant OM studies by topic. Corresponding to the commonly overlooked issues in field experiment-based OM studies, we then provide a detailed roadmap, ranging from experimental design and implementation to post-experiment analysis. We outline the methodological issues and corresponding solutions when applying field experiments, and conclude by identifying future research directions from an OM perspective.

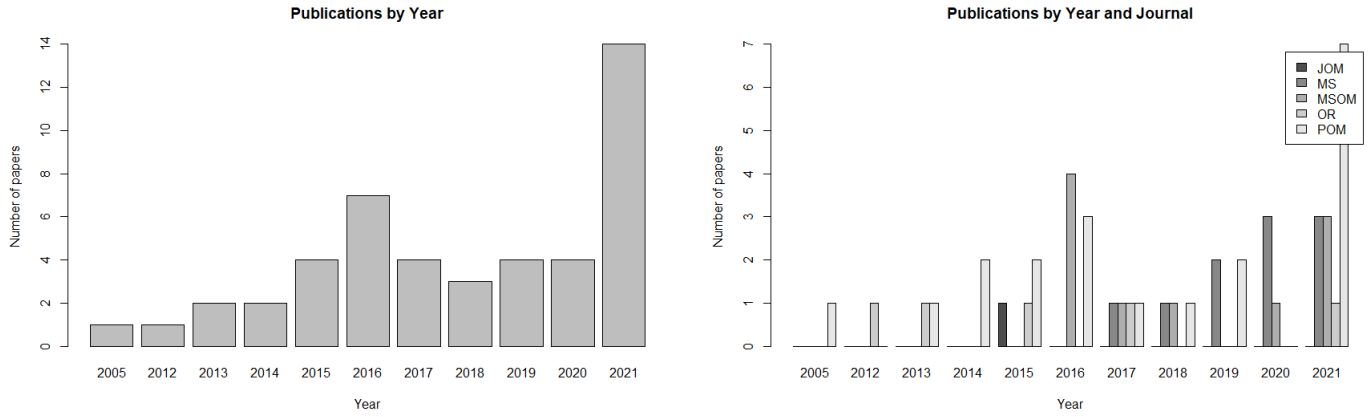
Key words: review, field experiments, methodology, A/B test, IT

1. Introduction

Operations management (OM) is a diverse field that involves a broad assortment of industry sectors, entities, and methodologies. Among the various methodologies, the field experiment has emerged as the most effective solution to address endogeneity concerns and identify causal effects of treatment conditions (Kagel and Roth, 2016). Although the experimental design was developed long ago and has been extensively applied in various areas, the application of experiments in OM is still nascent, with the earliest work dating only as far back as 2005 (Gaur and Fisher, 2005). As evident from Figure 1, fewer than five OM publications utilized the field experiment for most of the years surveyed, although 2021 witnessed a spike in this regard with 14 publications featuring this design.

This recent surge in field experiment-based publications reflects the increased attention paid to the experimental design by OM researchers. With our extensive literary review, this paper provides a comprehensive overview of the current practice and future directions of applying field experiments in the OM area. We first hired a research assistant to read every issue of the top

Figure 1 Distribution of Publications based on Field Experiments



five OM journals between 2005 and 2021: *Journal of Operations Management*, *Management Science*, *Manufacturing & Service Operations Management*, *Operations Research*, and *Production and Operations Management*. This process yielded 192 experimental-based publications, the majority of which are based on non-field experiments, such as lab or quasi-experiments. We filtered out 46 studies that address their research questions with a field experiment design. Our definition of field experiments falls into the *natural field experiment* category developed by Harrison and List (2004), where "*the subjects naturally undertake these tasks and the subjects do not know that they are in an experiment.*" Our reviewed publications cover topics ranging from retail management, service operations, queuing and scheduling to supply chain management, health care, online auction, and transportation management.

There are a few pioneering works reviewing experiment-based studies by OM researchers. Bendoly et al. (2006) review behavioral research that leverages experimental methods. Based on publications between 1985 and 2005, Bendoly et al. (2006) propose a framework (intention, action, and reaction) that identifies the three commonly used behavioral assumptions in analytical OM models and then analyze the literature building on this framework. The majority of publications (approximately 75%) included in this review are laboratory experiments, with only six papers involving field experiments. The following review by Lonati et al. (2018) discusses ways to ensure that experimental research in OM is relevant, realistic, and valid, with a key focus of on the threats that could undermine the internal validity of experimental findings. Based on this, Eckerd et al. (2021) further elaborate on the design choices regarding the effect of demand, incentives, deception, sample construction, and vignettes, as well as the risks and trade-offs surrounding these factors.

The above mentioned review papers touch on the topic of field experiments in one way or another, yet none of them are dedicated to the topic in an OM context, which is uncommon compared to other business fields. For example, there are dedicated review papers on field experiments in

marketing (Gneezy, 2017; Simester, 2017), strategy (Chatterji et al., 2016), accounting and finance (Floyd and List, 2016). Specifically, Gneezy (2017) acknowledges the importance and the potential of field experiments in marketing research. The paper describes the defining features of field experiments using several example studies and proposes three associated considerations: context and experiment type, benefits and costs, design tightness and experimental noises. Simester (2017) reviews the field experiment papers published in top marketing journals between 1995 and 2014, summarizing the requirements to publish a field experiment-based paper and identifying topics that remain relatively understudied in the marketing literature. Chatterji et al. (2016) propose that field experiments have potential for strategy research and categorize two types of experiments, strategy field experiments and process field experiments, using two example studies. Floyd and List (2016) provide a comprehensive guide for conducting field experiments in accounting and finance with a focus on identifying behavioral parameters, generating and analyzing experimental data, summarizing the extant literature, and identifying future opportunities in accounting and finance.

The lack of a systematic review for field experiment-based OM studies is partly due to its gaining popularity only in recent years. Therefore, our paper aims to fill this gap by generating a comprehensive "A-to-Z" instruction of field experiment applications specific to the OM area, as we contend that field experiments have significant potential in OM for several reasons.

First, field experiments provide an alternative data source that empowers OM scholars to consider outcomes of interest long understudied in the OM literature due to data limitations. For example, very few OM studies empirically examine wholesale price despite its importance, because detailed price information is usually confidential and unavailable to OM scholars. Cui et al. (2021), for their part, fill this research gap by conducting a field experiment that manipulates wholesale price quotes. Second, field experiments have great potential to generate actionable insights for OM managers. The outcome variables of interest in OM are often the antecedents of firm performance rather than the direct measure of the performance, and field experiments are suitable for assessing operational processes within a firm. Third, field experiments provide channels for empirical tests of OM theories. Despite the rich history of theory development in OM such as queuing theory, inventory theory, and supply chain coordination, the existing literature lacks empirical evidence of theory application and validation. Owing to new technological capabilities, OM researchers can conduct experiments in settings that were previously difficult to implement otherwise and leverage field experiments as a test bed for theoretical-grounded hypotheses. For instance, Queenan et al. (2019) examine the impact of technology-enabled continuity of care on patients with chronic health conditions, thanks to the availability of a telemonitoring device that allows the sharing and maintenance of data between discharged patients and physicians.

With its thorough and dedicated review of the most up-to-date literature involving field experiments in top OM journals, our paper provides a direct point of reference, including tangible recommendations on improving the current state of practice and the under-explored areas of research to consider. Our paper presents several notable differences and extensions compared to the earlier review works (Bendoly et al., 2010; Lonati et al., 2018; Eckerd et al., 2021). To begin with, we dedicate our review to OM studies that leverage field experiments, with our paper covering the most up-to-date publications from 2005 to the end of 2021, an important distinction as OM researchers have started to embrace field experiments only in recent years. In addition, we discuss the context, design, implementation, analysis, pros and cons of experiments by type, which we believe is a one-stop solution for OM researchers new to the methodology. Then, unlike earlier review works, we distinguish the literature by topic (see Section 3 and the online Appendix), which will inform future research regarding the choice of topic and practice by area of study. Lastly, via a thorough review of the OM literature, we highlight the common methodological issues in the extant practice. Drawing on the economic and psychology literature with its relatively mature practice of implementing field experiments, we propose recommendations for better practice, including experimental design, implementation, and analysis.

2. Background

In this section, we start by recapping several common types of experiments. We then shift gears to field experiments and summarize how field experiments differ from these common types of experiments along several dimensions (see Table 1).

2.1. Common Types of Experiments

A *laboratory (or lab)* experiment is conducted under highly controlled conditions and allows researchers to have precise control over various conditions, including where the experiment takes place, when to implement, whom to include as participants, and what interventions to use. Because of the standardized procedure and accurate measurements, lab experiments have high internal validity and are typically easy to replicate. However, the lab setting is often limited and may lead to unnatural behavior that does not mirror real life, thereby limiting the ecological validity[1] of its findings in their relevance to a wider population in a real-life setting.

A *vignette or scenario-based experiment* is similar to a lab experiment, except that the experimental intervention is done and the outcome measurement is collected via a survey. A randomized stimulus typically presents in the form of text, videos, audio clips, or other stimuli that can fit into

[1] Ecological validity is a subtype of external validity. Specifically, external validity examines whether the study findings can be generalized to other contexts, whereas ecological validity examines whether the study findings can be generalized to real-life settings (Andrade, 2018).

a survey format. Experimenters often administer these randomized manipulations to enumerators using survey questionnaire software (e.g., Qualtrics) or physical survey forms. Due to their ease of implementation, many studies in the OM field have employed vignette experiments (Boyer et al., 2002; Bendoly and Cotteleer, 2008; Venkatesh et al., 2012; Hora and Klassen, 2013; Abbey et al., 2015a; Seifert et al., 2015; Tonin and Vlassopoulos, 2015; Holguín-Veras et al., 2016; Ta et al., 2018; Polyviou et al., 2018). The rise of crowdsourcing platforms, such as Amazon Mechanical Turk, further incentivizes researchers to adopt vignette experiments due to their convenient access, low cost, and flexibility of hiring crowdworkers to perform on-demand tasks. Similar to lab experiments, vignette experiments have high internal validity and replicability but lack ecological validity due to the limitations of the survey setting.

A *natural experiment* happens in a real-life environment and thus more closely reflects real-life situations. A key feature is that natural experiments are not designed or implemented by researchers, instead occurring in real-life environments where individuals are exposed to treatment or control conditions. The treatment is "as-if random" because there is no explicit randomization process designed by researchers. Such a natural setting, however, is a double-edged sword. On the one hand, participants are unlikely to anticipate or be aware of being observed, thereby minimizing the concern of the Hawthorne effect. The real-life environment also engenders high ecological validity of the findings, and reduces participants' likelihood of not complying with the experimental conditions. More importantly, the natural setting is extremely helpful in situations where the randomization of experimental conditions is ethically unacceptable. On the other hand, the absence of researcher control over the experiment poses challenges to causal inference and replication of the study, and it is often a challenging task for researchers to identify a natural shock in the first place. Because of these reasons, there are only a few OM studies that leverage natural experiments (Wang et al., 2019; Adbi et al., 2019; Pan et al., 2020; Cui et al., 2022a).

A *quasi-experiment* largely differs from the above mentioned experiments in that the element of random assignment is missing. As such, internal validity becomes a major concern in the quasi-experiment design because the missing random assignment may lead to discrepancies in its treatment, or in that of control groups. Consequently, additional techniques such as propensity score matching and synthetic control analyses are necessary to alleviate the concern regarding confounding bias (Goldfarb et al., 2022). Nonetheless, a quasi-experiment design has its advantages over other types of experiments, mainly because researchers have direct control over the treatment assignment, and the experiment is typically easier to set up, especially in settings where randomization is impractical or unethical. Since quasi-experiments occur in real-life settings, researchers need not be as concerned about the ecological validity of the findings. For these reasons, several OM studies have leveraged quasi-experiments to address different research questions (DeHoratius and Raman, 2007; Mcafee, 2009; Dhanorkar and Muthulingam, 2020).

2.2. Field Experiments: Pros and Cons

A *field experiment* is similar to a lab or vignette experiment in that the experimenter manipulates the treatment variable of interest and randomly assigns participants to different groups. The primary difference is that field experiments are implemented in real-life settings. Given the field context, the participant behavior more closely reflects the real life setting and the experimental findings have higher ecological validity. Compared with natural experiments or quasi-experiments that hinge on an external force, researchers have better control over the design and implementation process of field experiments (Meyer, 1995), thereby achieving high internal validity. Because the study is covert, in that participants do not know that they are the subjects of the study, the field experiment design is less likely to suffer from biases caused by the impact of demand characteristics. In other words, the Hawthorne effect is less of a concern for field experiment-based studies. As such, field experiments represent the most appealing tool for researchers from various fields to examine causal effects, such as information systems (e.g., Looney and Hardin, 2009; Bapna and Umyarov, 2015; Lee and Hosanagar, 2021), marketing (e.g., Tucker and Zhang, 2011; Sahni et al., 2017; Li and Zhu, 2021), and behavioral economics (e.g., Liu et al., 2014; Gee, 2019; Robitaille et al., 2021), as shown in Figure 1.

Apart from these advantages, several caveats should be kept in mind regarding internal validity, external validity, and replicability of field experiments (Roe and Just, 2009). Specifically, the greater the flexibility obtained comes at the cost of less control over extraneous confounding variables, making it challenging for future researchers to replicate the study exactly. Since researchers typically choose a specific setting to conduct a field experiment, one may worry about how well its findings can be generalized to a different context of interest, especially considering issues such as non-compliance with the treatment assignment. Due to the lack of controls for the field, there is also a greater possibility of attrition in the field than in a laboratory. As suggested by James et al. (2017) and Hansen and Tummers (2020), units with and without missing results are systematically different (selective attrition), and such selective attrition could differ by groups of participants (differential attrition). Non-compliance and potential social interactions among participants from different treatment arms may bias treatment effects estimation. Field experiments are typically more expensive and time-consuming than laboratory and natural experiments, and such limitations restrict their broader applicability and occasionally the external validity.

While there is no perfect experiment design, in principle, researchers should consider several important aspects when designing and implementing a field experiment. Besides random assignment, two underlying core assumptions are critical for ensuring the validity of a field experiment (Broockman et al., 2017). The first is *excludability*, which ensures that the only causal channel is by taking the treatment. The second is the *non-interference* assumption, which indicates that the

outcome relies on whether a subject is assigned to one specific treatment status and not to another. Only by satisfying these core assumptions can researchers extrapolate unbiased estimates via a field experiment. Apart from ensuring the above assumptions, another crucial aspect to consider is the insight into the cause of the main effect. In practice, researchers may not be able to design field experiments that allow for evaluating competing explanations. An alternative approach is to evaluate heterogeneous treatment effects by different subgroups, which can be done by incorporating interaction terms rather than merely the primary variable of interest (Simester, 2017).

Table 1 Experiment Comparison

Experiment Type	Assignment	Internal Validity	Threats to Internal Validity	Ecological Validity	Replicability
Lab	Random	High	Attrition, Hawthorne Effect	Low	High
Vignette	Random	High	Attrition, Hawthorne Effect	Medium	High
Natural	As-if Random	Medium	Selection Bias, Social Interaction	High	Low
Quasi	Non-random	Medium	Selection Bias, Social Interaction	High	Low
Field	Random	High	Attrition, Non-compliance, & Social Interaction	High	Low

Notes. "Attrition" refers to the dropout of participants during the experimental period. "Hawthorne effect" refers to a type of reactivity in which individuals modify their behaviors due to the awareness of being observed in an experiment. "Selection Bias" refers to participants' self-selection into the treatment or control groups. "Social Interaction" refers to the scenario where participants infer the treatment assignment or other details about the experiment through interactions with others. "Non-compliance" refers to participants' deviation from the assigned experimental condition.

3. Literature

We focus on the five leading OM journals, including the *Journal of Operations Management*, *Manufacturing & Service Operations Management*, *Production and Operations Management*, *Operations Research*, and *Management Science*. As our primary focus is on field experiment-based studies, we include publications with at least one outcome measure derived from a field experiment. We exclude studies that leverage lab-, quasi-, and natural- experiments from this review due to their different context, design, and analysis. We also restrict our attention to papers with explicit descriptions of the design of the field experiment for adequate evaluation of the experiments involved. After carefully screening the literature, we end up with one paper in *Journal of Operations Management*, ten papers in *Manufacturing & Service Operations Management*, twenty papers in *Production and Operations Management*, five papers in *Operations Research*, and ten papers in *Management Science* (OM department) (see Table 2).

We cluster these studies into five research topics: *workforce management*, *supply chain management*, *retail management*, *service operations*, and *healthcare*. The column labeled "Topic" in Table 2 lists the research topic of each reviewed study.

As shown, nine papers leverage field experiments to explore factors that affect workforce performance in different contexts, including warehouse operations (De Vries et al., 2016; Sun et al., 2022), innovation management (Hutchison-Krupat and Chao, 2014; Wooten and Ulrich, 2017; Cui

Table 2 Literature Review

Study	Journal	Topic	Sample Size	Unit of Analysis	IRB	RL	BC	MC	NC
Abbey et al. (2015a)	JOM	Retail Management	981	Consumer	N	Complete	N	N	N
Cui et al. (2022b)	MSOM	Supply chain management	3,960	Supplier	N	Complete	Y	N	N
Ferreira et al. (2016)	MSOM	Retail Management	6,000	Style	N	Stratified	N	N	N
Zhang et al. (2017)	MSOM	Service operations	30,317	Student	N	Complete	N	N	Y
Cohen et al. (2021)	MSOM	Service operations	537,370	Consumer	N	Stratified	Y	N	N
Bray et al. (2016)	MSOM	Service operations	93,962	Lawsuit case	N	Complete	Y	N	N
Retana et al. (2016)	MSOM	Service operations	2,673	Consumer	N	Complete	N	N	Y
Acimovic et al. (2020)	MSOM	Workforce management	4,771	Agent	N	Complete	Y	N	Y
Craig et al. (2016)	MSOM	Supply chain management	264	Stock keeping unit	N	Complete	N	N	N
Gallino and Moreno (2018)	MSOM	Retail management	435,982 & 2,389,655	Consumer	N	Complete	N	N	Y
Cui et al. (2021)	MSOM	Supply chain management	3,840	Supplier	Y	Complete	Y	N	N
Caro and Gallien (2012)	OR	Retail management	20	Product group	N	Clustered	N	N	N
Feldman et al. (2022)	OR	Retail management	10,421,649	Consumer	N	Complete	Y	N	N
Gallien et al. (2015)	OR	Retail management	34	Article	N	Complete	Y	N	N
Abhishek and Hosanagar (2013)	OR	Service operations	247	Keyword	N	Clustered	N	N	N
Cheung et al. (2017)	OR	Retail management	1,295	Deal	N	N/A	N	N	N
De Vries et al. (2016)	POM	Workforce management	129	Picker	N	Complete	N	N	N
Ta et al. (2021)	POM	Workforce management	466	Crowdsourced participant	N	Complete	N	Y	N
Lu et al. (2021)	POM	Service operations	1,736	Loan	N	Complete	Y	N	N
Haruvy et al. (2014)	POM	Service operations	144 & 96	Auction	N	Complete	N	N	N
Ding et al. (2021)	POM	Retail management	91	Student	N	Complete	Y	Y	N
Riccobono et al. (2016)	POM	Workforce management	71	Student	N	N/A	N	N	N
Wooten and Ulrich (2017)	POM	Workforce management	544	Entry	Y	Stratified	N	N	N
Gaur and Fisher (2005)	POM	Retail management	53	Store	N	Stratified	N	N	N
Zhang et al. (2021)	POM	Supply chain management	2,000	Consumer	N	Complete	N	N	N
Chuang et al. (2016)	POM	Retail management	60	Store	N	Complete	Y	N	N
Queenan et al. (2019)	POM	Healthcare	169	Patient	Y	Complete	Y	N	N
Bichler and Merting (2021)	POM	Service operations	1438 & 1778	Student	N	Complete	N	N	N
Abbey et al. (2015b)	POM	Retail management	1,500	Consumer	N	Complete	N	Y	N
Jung et al. (2021)	POM	Service operations	295	Consumer	N	Complete	Y	N	N
Hardgrave et al. (2013)	POM	Supply chain management	62	Store	N	Complete	N	N	N
Cui et al. (2019b)	POM	Workforce management	423	MTurker	N	Complete	N	N	N
Elmaghraby et al. (2018)	POM	Retail management	508	IT product	N	Stratified	N	N	N
Kistler et al. (2021)	POM	Healthcare	33,057	Surgical case	N	Complete	N	N	N
Tucker and Singer (2015)	POM	Healthcare	196	Work area	N	Stratified	N	N	N
Hutchison-Krupat and Chao (2014)	POM	Workforce management	260 & 280	MTurker	N	Complete	N	N	N
Fisher et al. (2018)	MS	Retail management	15	Product	N	Stratified	N	N	N
Buell et al. (2017)	MS	Service operations	48	Student	Y	Complete	N	N	N
Kesavan and Kushwaha (2020)	MS	Workforce management	541,836	SKU-store	N	Stratified	Y	N	N
Buell and Kalkanci (2021)	MS	Retail management	36,906 & 47,858	Transaction	Y	Complete	N	Y	Y
Cui et al. (2019a)	MS	Retail management	445	Deal	Y	Complete	Y	N	N
Sun et al. (2022)	MS	Workforce Management	782,356	Package	N	Complete	Y	N	N
Cui et al. (2020)	MS	Service operations	598 & 250 & 660 & 293	Request	Y	Complete	N	N	N
Zhang et al. (2020)	MS	Retail management	1,000,000	Consumer	N	Complete	Y	N	Y
Zhang et al. (2019)	MS	Retail management	799,904	Consumer	N	Complete	Y	N	Y
Mejia and Parker (2021)	MS	Service operations	3,200	Ride	Y	Complete	N	N	N

Notes. Y stands for Yes, N stands for No, N/A stands for Not Applicable. IRB stands for institutional review board, RL stands for randomization level, BC stands for balance check, MC stands for manipulation check, and NC stands for non-compliance.

et al., 2019b), agent behaviors (Acimovic et al., 2020; Ta et al., 2021), and more granularly defined contexts (Riccobono et al., 2016; Kesavan and Kushwaha, 2020). Among the five studies in supply chain management, three implement field experiments to examine how to effectively manage supply chains (Hardgrave et al., 2013; Craig et al., 2016; Zhang et al., 2021), and two papers study the wholesale price discrimination from suppliers (Cui et al., 2022b, 2021). There are a total of 17 papers in retail management that employ field experiments to estimate the causal effect. Seven studies examine the impact of pricing strategies through field experiments (Caro and Gallien, 2012; Abbey et al., 2015a; Ferreira et al., 2016; Cheung et al., 2017; Fisher et al., 2018; Elmaghraby et al., 2018; Zhang et al., 2020). The remaining ten studies investigate how various factors affect performance outcome (sales and revenue) in the context of offline retailers (Gaur and Fisher, 2005; Abbey et al., 2015b; Gallien et al., 2015; Chuang et al., 2016; Ding et al., 2021; Buell and Kalkanci, 2021) and online retail management (Gallino and Moreno, 2018; Zhang et al., 2019; Cui et al., 2019a; Feldman et al., 2022). In service operations, researchers have utilized field experiments to causally evaluate the effect of policy changes on performance outcomes, such as consumer outcomes

(Retana et al., 2016; Buell et al., 2017; Zhang et al., 2017; Jung et al., 2021; Lu et al., 2021), auction outcomes (Abhishek and Hosanagar, 2013; Haruvy et al., 2014), scheduling efficiency (Bray et al., 2016; Bichler and Merting, 2021), ride-sharing (Mejia and Parker, 2021; Cohen et al., 2021), and fairness (Cui et al., 2020). OM researchers have also focused on ways to improve various health care outcomes, such as physician performance (Tucker and Singer, 2015), quality and efficiency of care (Queenan et al., 2019; Kistler et al., 2021). For a detailed review of each paper by research topic, please refer to the Appendix Literature Review by Research Topic.

The literature has also demonstrated that field experiments can generate significant value in OM, such as in workforce management, where Sun et al. (2022) conduct a randomized experiment in four warehouses to causally evaluate the performance of a new "human-centric bin packing algorithm," and find that the non-conformance rate reduces by 5.7% and the average packing time decreases by 4.5%. In supply chain management, Cui et al. (2021) estimate that the economic value of their field experiment with Alibaba ranges from $16.65 million to $17.46 million. In retail management, Caro and Gallien (2012) design and implement an autonomous decision-making process to optimize clearance prices for Zara. They conduct a controlled field experiment in all Belgian and Irish stores in 2008, the results of which indicate that the pricing solution increases clearance revenue by approximately 6%. In service operations, Bray et al. (2016) run a field experiment to measure the effect of switching from a hearing-level first-in-first-out (FIFO) scheduling policy to a case-level FIFO policy in the Roman Labor Court of Appeals. The results indicate that the new scheduling policy decreases the average case duration by 12%, and decreases the probability of a decision being appealed to the Italian Supreme Court by 3.8%.

Despite the great potential of applying field experiments in OM, there are relatively few OM field experiment studies compared to other areas such as marketing or information systems. This fact reflects the challenges of conducting OM field experiments, complicated by two main factors. First, unlike consumer-side field experiments in other management areas that can be done on digital platforms, manager- (or process-) side field experiments in operations and manufacturing settings often have a much smaller sample size. And secondly, unlike the commonly assessed metrics in other fields such as click-through rate, web page views, and conversion rate, performance changes in operations induced by treatment may involve delayed responses and are thus more challenging to track.

4. A Roadmap to Field Experiment

This section describes a roadmap to field experiments, including experimental design, implementation, and analysis. In each subsection, we start by summarizing the standard approaches at different stages of an experiment and then highlight the common issues in field experiment-based

OM studies. We further propose the recommended practices for researchers, which we hope will help guide OM researchers in future applications of field experiments.

4.1. Experimental Design

4.1.1. Ethics Ethical consideration is a critical aspect in designing field experiments with the foremost consideration to ensure that the treatment interventions do not impact any participants in a negative way (Hansen and Tummers, 2020). Before implementing any experiments, it is necessary for researchers to comply with ethics and human subject protocols and work closely with the Institutional Review Board (IRB) or local representatives that oversee the research and experiment design.

Table 2 reveals the surprising finding that among the 46 reviewed studies, only eight explicitly state that they obtained IRB approval before implementing field experiments. This could be due to either not noting the approval in the paper, or to a basic lack of IRB awareness. Considering the importance of ethical consideration, we strongly recommend that future studies in the OM field request approval from the IRB or equivalent institutions before implementation. The objective, comprehensive review from the IRB can help experimenters identify potential risks that may unintentionally harm participants. For studies that have obtained IRB approval, we recommend including the approval details, such as which institutional board the approval is from and what the potential risks and benefits to participants are, in the paper or the online appendix.

Researchers might also consider registering or publishing their research design. For example, on the website `www.socialscienceregistry.org`, researchers can register a random trial by submitting information including trial title, abstract, trial start date, intervention start date, intervention end date, trial end date, and experimental design. For fields such as healthcare or economics, researchers often publish their research design in journals (e.g., *Journal of Development Economics* and *Trials*) before implementation. The journal would carefully review the importance of the field experiment, the power analysis to determine the appropriate sample size, and the planned analyses.

Ultimately, it is the researchers' obligation to address the ethical issues in their field experiments. McDermott and Hatemi (2020) point out that "*a large number of social science field experiments do not reflect compliance with current ethical and legal requirements that govern research with human participants,*" thus initiating a call to establish new standards to protect the public from unwanted manipulation and real harm. By reviewing field experiment-based research in the political science field, Phillips (2021) concurs regarding the ethical issues of field experiments that may escape IRB inspection. Accordingly, Asiedu et al. (2021) propose an appendix of structured ethics for randomized controlled trials to provide details on the following: policy equipoise, researcher role, potential harm to participants and nonparticipants, conflicts of interest, intellectual freedom,

participant feedback, and foreseeable misuse of research results. Although OM studies generally have less potential to harm experiment participants than other social science areas such as political science, we strongly encourage future researchers to follow the aforementioned studies and carefully evaluate the risks to participants before implementing an experiment.

4.1.2. Assignment Mechanism After deciding on the appropriate treatment arms compliant with the relevant ethical considerations, researchers need to select the mechanism for randomly assigning subjects to treatment or control conditions. To achieve proper randomization, researchers must ensure that the assignment mechanism satisfies the following four criteria: (1) the assignment mechanism is independent of covariates and potential outcomes for other units; (2) for every experimental unit, the probabilities of being assigned to the control condition and treatment condition are positive; (3) the assignment mechanism is independent of potential outcomes with given covariates; (4) researchers are aware of and can control the functional form of the assignment mechanism (Imbens and Rubin, 2015).

There are two assignment mechanisms widely adopted by existing studies: *complete randomization* (also known as simple randomization or pure randomization) and *stratified randomization* (also known as block randomization). In a completely randomized experiment, experimenters assign a randomly drawn subset of subjects from the entire population to the treatment condition, and the remaining subjects to the control condition. Researchers may vary the fraction of treatment groups, but the common practice would be 50%, such as when Zhang et al. (2019) randomly assign half of 799,904 users' Alibaba mobile app to the treatment group and the other half to the control group. Stratified randomization can address the need to control and balance the influence of confounding covariates. In a stratified randomized experiment, the entire population is divided into certain strata based on observable covariates, and then researchers randomly assign subjects within each stratum to treatment and control conditions. For example, Gaur and Fisher (2005) generate six clusters based on store-level annual revenue data and then for each cluster randomly assign one store to the control condition and the other two stores to two treatment conditions. An extreme case of stratification is the so-called pair randomization, in which each stratum contains one treated subject and one control subject.

As summarized in Table 2, 74% of reviewed studies (34 out of 46 articles) run their experiments using the complete randomization mechanism, while eight studies adopt stratified randomization. We believe that its ease of implementation is the primary reason for the widespread popularity of complete randomization, although it is worth noting that it may not always be the best practice. In cases with a relatively small sample, the complete randomization mechanism may lead to biased estimation. After carefully screening the selected studies, we noticed that several studies adopted

the complete randomization mechanism even though the sample sizes were relatively small (see Column "Sample Size" of Table 2).

According to the literature, a stratified randomization mechanism is a better choice because it can achieve more balance in observed and unobserved covariates than complete randomization when there are observable factors that strongly correlate with the outcomes (Bruhn and McKenzie, 2009). Building on Athey and Imbens (2017), in their wish list to experimental researchers Czibor et al. (2019) explicitly recommend using stratified randomization in the design phase to increase power and credibility when subjects' characteristics information is available.

Besides the above two common assignment mechanisms, *clustered randomization* is also popular, particularly when interactions between subjects throughout an experiment may damage the causal inference (Athey and Imbens, 2017). Similar to stratified randomization, clustered randomization requires experimenters to begin by dividing subjects into clusters based on a list of covariates with significant differences. A major difference between stratified and clustered randomization is that the latter does not assign treatment or control arms randomly to units within a cluster, instead randomly assigning treatment or control arms to selected clusters such that all subjects in a cluster have the same experimental conditions. For example, Abhishek and Hosanagar (2013) grouped 247 keywords to 29 unique product categories that span frozen meats, seafood, and desserts, which are further assigned to three distinct treatment arms.

4.1.3. Manipulation and Attention Check Given the proper assignment mechanism, researchers must carefully evaluate if participants' reactions to experimental interventions are as desired, done via a manipulation check. Note that it is critical to distinguish a manipulation check from an attention/remembering check, because remembering the manipulation does not necessarily imply that the treatment induces the desired participant reaction (Lonati et al., 2018).

Based on our review, manipulation checks have received limited attention in OM field studies, with only four papers conducting them (see Column "MC" of Table 2). The first reason for the lack of manipulation checks is that the treatment of a field experiment is about changing the environment, rather than changing the participant or their perceptions. Since environmental changes are explicit, and the focus of such studies is on the impact of environmental changes, participant comprehension of an intervention is less of a concern. The second reason is the difficulty in tracking participants and collecting their feedback in the field. In one example, Mejia and Parker (2021) study bias in ridesharing platforms, in which researchers initiate different requests with manipulations of rider characteristics and examine driver bias toward riders. Since researchers have no direct contact with drivers involved in the experiment, it is challenging to ascertain how drivers perceive the treatment (i.e., manipulation of riders' profile pictures). In such cases, a feasible solution is to

collect individual perceptions of the treatment intervention via surveys or interviews for a small subsample of the target population, either post-experiment or during the pilot study (Buell and Kalkanci, 2021; Ding et al., 2021). Alternatively, researchers can perform a manipulation check using a different population, which will reduce the risk of contaminating the field intervention (Abbey et al., 2015b).

In practice, a manipulation check is unnecessary when an intervention is inconsequential and lacks psychological realism (Lonati et al., 2018), yet we nonetheless recommend all researchers implement manipulation checks as they reduce Type I error rates of statistical reference (Abbey and Meloy, 2017) and can minimize the risk of demand effects (Lonati et al., 2018).[2] We summarize several situations where a manipulation check is essential, such as when an intervention attempts to change participants or their perceptions rather than the environment. For instance, Cui et al. (2020) create fictitious guest accounts with either typically African-American or typically white-sounding names to send accommodation requests on Airbnb. One concern that arises is that names signal not only race but also presumed corresponding socioeconomic status. Without a manipulation check on Airbnb host perceptions, it is difficult to assess the extent to which the level of identified discrimination is driven by racial discrimination as opposed to non-race-related socioeconomic factors. Another set of conditions necessary for manipulation checks is when the manipulation is subtle, such as a minor wording change in instruction or messages. In these cases, researchers will need to determine if participants notice the manipulation and how they interpret the change. For instance, with studies that rely on randomized text message interventions with different framing, researchers shall refer to a manipulation check and examine whether participants interpret the message as designed (Cohen et al., 2021; Jung et al., 2021; Lu et al., 2021).

4.2. Experimental Implementation

Researchers often need to collaborate with companies to implement field experiments, which typically involves several key steps.

4.2.1. Preparation Before implementing a designed field experiment, researchers need strong support from the senior level of the collaborating companies. To obtain the buy-in of key decision-makers, researchers can approach the firm via several channels. First, researchers need to interview the managers to learn about opportunities and threats faced by the firm to ensure that the field experiment aligns with the company's strategic goals. Second, researchers must understand the company's internal legal issues, such as consent forms and protections of employee identities.

[2] According to Lonati et al. (2018), demand effects are explicit or implicit indications in the experimental situation that can systematically bias participants' response to treatment.

If possible, researchers should carefully investigate the firm's historical data before implementing an experiment in order to guide them to relevant inferences on what hypotheses to propose, and the likelihood of the proposed hypotheses. It would be also be beneficial to interview front-line managers and IT teams to understand the existing data structure and potential technical challenges associated with experiment implementation and data collection, which will lead researchers to better ways of implementing the experiment. Due to the restrictions and time sensitivity of the field setting, another important issue to consider is the sample size. To decide the proper sample size for analyses, researchers can employ a pilot study to estimate the size of the treatment effect of interest, which will serve as an input of the power calculation (Cui et al., 2021).

4.2.2. Intervention During an intervention, ensuring the ethical treatment of participants is of the utmost importance, and informed consent from experimental subjects is often required. However, adhering to rigid ethical rules by requiring informed consent is sometimes impractical in certain field settings. For example, to investigate whether a buyer's race or gender affects the prices that they pay for used cars, it would be prohibitively difficult to estimate the level of discrimination among used car dealers aware of being in an experiment (List et al., 2008). As a result, List et al. (2008) recommend that "*the benefits and costs of informed consent should be carefully considered in each situation*" and that the informed consent can be relaxed when "*there are minimal benefits of informed consent but large costs.*"

Field interventions are usually implemented by the staffs of the collaborating companies, a fact that decreases the *treatment reliability* because the intervention might not be implemented as originally intended and/or the various people involved might implement the same treatment differently. These conditions increase treatment heterogeneity, which in turn reduces the statistical power of the analysis and biases the estimation results. And so, to ensure treatment reliability, researchers must closely work with the staff who runs the experiment, to which end we summarize three key recommendations. First, researchers should standardize the intervention procedure to minimize staff autonomy during an intervention. Second, in addition to providing written instructions for high-quality and consistent implementation, it is advisable for researchers to conduct a thorough briefing or training of the staff involved in the experiment implementation. Third, it is also helpful to periodically monitor how the treatment is implemented, in order for researchers to resolve problems in its early stages.

The Hawthorne Effect is typically less of a concern in a field experiment, because experimental manipulations are often conducted in a natural enough manner that individuals are unaware of their participation in an experiment (List, 2008). *Reactivity* may still nonetheless occur in field experiments when experimental units learn about the intervention via other channels. Consider

the example of an experiment that manipulates the layout of a web page. If users under different treatment arms happen to know each other or share social interaction, then those experiencing the change could realize that they are in the treatment group. Another example is Sun et al. (2022), which examined order-packing behavior at Alibaba in adjustment to algorithmic prescriptions. In this setting, orders were randomly assigned to either the new algorithm (treatment group) or Alibaba's original algorithm (control group). A worker may then handle packages in both control and treatment groups and become aware of the treatment and control differences in processing orders. In such cases, researchers can alleviate reactivity either by ignoring the informed consent if possible, or by informing the participants that all experimental data collected are anonymous so that experimental participants will be less likely to alter their behavior to cater to anticipated results.

4.3. Experimental Analysis

After carefully designing and implementing an experiment, the next step is analyzing the collected data. While analysis is made relatively straightforward through the convenience of randomization, researchers must be aware of several issues before deriving unbiased statistical estimates.

4.3.1. Randomization (or Balance) Check Appropriate randomization could prevent selection bias by generating comparable treatment and control groups, but it cannot guarantee that these groups will always have identical characteristics. Randomization may yield groups that differ in important observable dimensions, especially in the case of small samples (Czibor et al., 2019). To help the audience assess how similar treatment and control groups are, experimental guidelines—such as CONSORT[3]—recommend presenting the baseline information (mean and standard deviation) of important characteristics between treatment and control groups. Apart from the presentation of baseline difference, researchers usually run a series of t-tests to check if the means of key variables are significantly different across treatment and control groups (Bruhn and McKenzie, 2009). The literature commonly refers to these tests as randomization or balance checks.

As evident from Table 2, 37% of the existing OM literature (17 out of 46 studies) adheres to this practical guideline by conducting a randomization check (i.e., a series of t-tests) before the formal empirical analysis. In practice, with an appropriate random assignment, the necessity of the randomization/balance check remains conceptually unclear. The goal of the significance tests is to assess the probability of chances causing the observed differences (Altman, 1985). Because appropriate randomization already guarantees that any differences are due to chances rather than bias, conducting randomization/balance checks is conceptually unnecessary (Altman, 1985).

[3] `http://www.consort-statement.org/checklists/view/32--consort-2010/510-baseline-data` (accessed by August 13, 2022)

However, when the random assignment is botched, randomization/balance checks become necessary. In such a case, two tests must be conducted to replace the popular t-tests. The first is the omnibus test of joint orthogonality (Hansen and Bowers, 2008), in which researchers run a regression with the treatment assignment as the dependent variable and the baseline characteristics as independent variables, then test the joint hypothesis that the coefficients of the baseline variables are all equal to zero. This omnibus test helps avoid a scenario in which there are significant random differences between variables in the treatment and control groups. The two common choices for the regression models are a linear regression with an F-test and a probit model with a chi-squared test. It is also important that researchers focus on the size of the difference rather than its statistical significance. In particular, Imbens and Rubin (2015) recommend assessing balance in covariate distributions based on the normalized differences, defined as the difference in means between the control and treatment groups divided by the square root of half the sum of the variances of the treatment and control groups. Austin (2009) suggests that a normalized difference of 0.10 or less indicates a good balance of the variable.

When a balance check fails, researchers must conduct thorough follow-up analyses in order to interpret the reasons and outcomes. To begin with, it is critical for researchers to decide which baseline characteristics to control for in the presence of unbalanced baseline characteristics. As suggested by the literature (Altman, 1985; Bruhn and McKenzie, 2009), one should choose control variables "*not on the basis of statistical differences, but on the strength of their relationship to the outcome of interest.*" In other words, a small imbalance in a variable that is highly correlated with the outcome can be more important than a large and significant imbalance for one that is uncorrelated. Re-randomization further provides a means to avoid imbalance due to chance (Morgan and Rubin, 2012), because it checks the balance of baseline characteristics at the time of randomization and re-randomizes if the balance does not satisfy the pre-specified criteria. This process repeats until a balance that satisfies the criteria is reached, and only then is the treatment assignment administered. It is not advisable to include imbalanced variables across treatment groups in the estimation stage, because post hoc adjustments such as variable selections based on failed balance checks can lead to fraudulent conclusions and should thus be avoided (Mutz et al., 2019).

4.3.2. Non-compliance Average treatment effect (ATE), measuring the difference in average outcomes between subjects assigned to the treatment group and subjects assigned to the control group, is the primary metric from field experiments, but a crucial assumption must be satisfied to correctly identify it. In order to accurately derive ATE, every experimental participant must comply with their assignment: subjects in the treatment group take the treatment, and those in

the control group do not. If any participants do not comply with their assignments, researchers will run into the so-called non-compliance issue and produce biased ATE.

Non-compliance is one of the most common complications of field experiments, and occurs when subjects deviate from the assigned experimental condition. It may arise because of random accidents, such as subjects' inability to undergo the assignment or researchers' flawed implementation of the assigned treatment. A much more concerning cause of non-compliance is the systematic differences in behaviors or characteristics between compilers and non-compilers, which may be associated with outcomes of interest. Under such scenarios, even if a researcher perfectly executes random assignment, there is no assurance that the actual receipt of treatment is exogenous because participants might self-select into treatment or control groups, or not accept an assignment at all (Athey and Imbens, 2017). In practice, the receipt of treatment could be a deliberate choice by subjects after taking into account perceptions or expectations of the causal effects of the treatment based on information that researchers may not observe.

Two approaches are commonly used to deal with non-compliance issues (Athey and Imbens, 2017). The first approach is an intention-to-treat analysis (ITT) that focuses on the causal effects of the assignment to treatment rather than the actual receipt of treatment. ITT is the preferred approach for evaluating the overall effect of an intervention like a policy, because the likelihood of non-compliance is relatively high in real-world settings. Three reviewed studies select ITT to evaluate the efficacy of the assignment to treatment, which companies tend to focus on because it is much more controllable than the receipt of treatment (Retana et al., 2016; Gallino and Moreno, 2018; Buell and Kalkanci, 2021). In one related experiment to test the impact of transparency into a company's social responsibility efforts on sales, Buell and Kalkanci (2021) manipulate the content about a company's social responsibility practices in three videos played at a bookstore. Because they cannot directly observe which customers actually watch the videos (the receipt of treatment), the researchers estimate the effect of the treatment assignment (whether the treatment video is playing) rather than the impact of transparency level. One main drawback of the ITT, however, is its poor external validity. Since the assignment mechanism is usually different in new settings, such as when compliance rate is high in a medical trial phase but may be very different when the drug is released to the general public, generalizing the intention-to-treat effects is rather difficult (Imbens and Rubin, 2015).

When external validity is a primary concern, estimating the causal effect of the receipt of treatment is necessary. In this case, an approach preferable to ITT relies on instrumental variables (IV) analysis, with the random assignment to treatment constituting the IV. This approach estimates the compiler average causal effect (CACE), or the local average treatment effect. CACE denotes the effect of the receipt of treatment for the sub-population of subjects who comply with their

assignments, or compilers. Four reviewed studies begin with ITT and then implement IV analysis to estimate the CACE (Zhang et al., 2017, 2019, 2020; Acimovic et al., 2020). In one study to identify the short-term effect of actually viewing price promotions (i.e., receipt of treatment) rather than simply receiving one (i.e., assignment to treatment), Zhang et al. (2020) use the random assignment of customers to treatment and control groups as an IV for viewing promotions and estimate the local average treatment effects of on purchasing behavior using the standard IV setup (i.e., two-stage least square). In the first stage, the authors derive the predicted value of viewing activity by instrument of IV, while in the second stage they regress the outcome variables of interest on the predicted value of viewing activity to derive the CACE.

The justification for random assignment to treatment constituting the IV requires two key assumptions: *unconfoundedness* and *exclusion restriction* (Imbens and Rubin, 2015). The unconfoundedness assumption states that despite the receipt of treatment being confounded by non-compliance, the assignment of treatment is not. In field experiments, appropriate randomization guarantees that this unconfoundedness assumption is satisfied by experiment design. The exclusion restriction assumes that assignment to treatment has no effect on the outcomes of interest other than by indirectly affecting the receipt of treatment. Unlike the unconfoundedness assumption, the validity of the exclusion restriction assumption cannot be guaranteed by randomization. Imbens and Rubin (2015) recommend the double-blind design, in which neither the subjects nor the experimenters are aware of the assignment to treatment, thereby supporting the exclusion restriction. A major drawback of the IV analysis is that the CACE is a local average treatment effect, which may not be generalizable to the entire study population.

Adjusting for non-compliance also depends on whether it is one-sided or two-sided. One-sided non-compliance occurs when subjects in the control group are effectively embargoed from the active treatment, but treated subjects can circumvent the assignment by not taking the treatment. Four reviewed studies specifically discuss potential one-sided non-compliance in their experiment, arsing because they cannot force treated participants to take the treatment unavailable to the control group (Retana et al., 2016; Gallino and Moreno, 2018; Buell and Kalkanci, 2021; Zhang et al., 2020). If non-compliance is symmetric, in that all subjects in the treatment and control groups can choose not to comply with their assigned experimental conditions, we refer to it as two-sided non-compliance. Three reviewed studies encounter this issue because they adopt an encouragement design in which both treatment and control groups can access the random trial, but only the treatment group is randomly assigned to receive encouragement to participate (Zhang et al., 2017, 2019; Acimovic et al., 2020).[4]

[4] The common approaches summarized in this section are based on literature and textbooks, such as Chapters 5 and 6 of Gerber and Green (2012) and Chapters 23, 24, and 25 of Imbens and Rubin (2015). Readers may refer to these materials for more detailed guidance on how to analyze data from randomized experiments with one-sided or two-sided non-compliance.

Despite the common occurrence of non-compliance in field experiments, we find that only a few reviewed OM studies (7 out of 46) explicitly raise concerns about the potential non-compliance issue and implement the above two approaches accordingly (see Column "NC" of Table 2). Several studies did not factor in the non-compliance issue even when there was no guarantee that participants in the treatment group would actually take the treatment. For example, Cui et al. (2022b) study the effect of the buyer's usage of automation by manipulating three recommendation conditions: no recommendation, human recommendation, and AI recommendation. For this randomized experiment, non-compliance might be a concern because suppliers assigned to human or AI recommendation may ignore the recommendation information, in which case Cui et al. (2022b) estimate the effect of the assignment to human or AI recommendation other than the actual receipt. Due to the different implications corresponding to different estimates, we recommend that future researchers carefully evaluate the possibility of non-compliance and make necessary adjustments based on their research interests. If the goal is to estimate the effect of the assignment to treatment, we recommend researchers provide convincing arguments to explain why the assignment to treatment is worth studying in their context and then draw conclusions based on the ITT. If the primary interest is in the effect of the actual receipt to treatment, researchers should begin with the ITT and then use the random assignment as the IV to estimate CACE.

4.3.3. Attrition refers to a scenario where the outcomes of some participants cannot be collected because they drop out of an experiment and is sometimes also considered an extreme case of non-compliance (Czibor et al., 2019). Whether attrition causes a biased estimation depends crucially on its correlation with the treatment. More specifically, if attrition is believed to be independent of the treatment (random attrition), it will not bias the estimation but will reduce its statistical robustness (Duflo et al., 2007). A more concerning scenario, however, is when attrition correlates with treatment (non-random attrition), as it will lead to biased estimates. For example, if participants in the control group have doubts about how the experiment benefits them, they will be more likely to drop out than those in the treatment group, thereby resulting in overestimates of the treatment effects. Non-random attrition is challenging to solve ex-post, especially when respondents and attritors are different among the treatment and control groups, so managing attrition during experimental design or data collection process is essential (Duflo et al., 2007). An effective, yet costly approach to alleviating the attrition concern is to collect participant contact information and then carefully track them even after they leave the program. Glennerster (2017) also provides some helpful insights about minimizing the attrition rate during the data collection process.

To evaluate the impact of non-random attrition, researchers may refer to three types of tests: 1) a *differential attrition rate test* that checks if attrition rates (the rate of participants dropping

out of the experiment) are significantly different across treatment and control groups, 2) a *selective attrition test* that determines if the mean of observable characteristics differs across treatment and control groups conditional on the response status (respondent or attritor), and 3) a *determinants of attrition test* that checks whether available outcomes and covariates are correlated with the response status. In further related research, Ghanem et al. (2021) establish the identifying assumptions of the above tests of treatment effects for both the respondent subpopulation and the study population. If the above tests indicate that non-random attrition remains a problem, researchers may refer to several statistical techniques (e.g., Manski-Lee bounds) to adjust for the attrition bias (Hausman and Wise, 1979; Manski, 1989; Wooldridge, 2002; Grasdal, 2001; Lee, 2009).

Among 46 reviewed studies, we identify instances of attrition in eight studies directly referencing cases in which a proportion of the participants cannot continue the field experiment. For instance, Cui et al. (2020) mention that the fictitious accounts created to serve as experimental participants were suspended by Airbnb, the platform in their experiment. Only half of the eight studies provide empirical evidence to demonstrate that attrition does not hurt the robustness of their findings. Kistler et al. (2021), for example, removed thousands of surgical cases from their sample because of incomplete information, but did provided a footnote to alleviate the attrition concern: "*Empirical analysis of surgical cases excluded due to missing data indicates no evidence of systematic bias*". For the rest four studies, no empirical evidence is provided to address the attrition issue, which could significantly damage the validity of empirical analyses. In De Vries et al. (2016), for example, 14 out of 143 participants were removed from the sample "*because of missing data for one or more of the relevant testing variables*", but no test was conducted to check if the attrition is random.

4.3.4. Heterogeneous Treatment Effect Besides estimating the average treatment effects for the entire sample, researchers are also interested in exploring their heterogeneity. Heterogeneous treatment effects are also commonly referred to as conditional average treatment effects, which is an average treatment effect specific to a subgroup of subjects defined by subjects' pre-treatment characteristics (Athey and Imbens, 2017). Uncovering heterogeneity helps in the design of new policies and to better understand underlying mechanisms, while also an ideal way to enhance the external validity of a field experiment-based study. Half of the reviewed studies (23 out of 46) conduct relevant analyses to uncover the heterogeneity of treatment effects.

The most commonly adopted approach to estimating the heterogeneous treatment effect is the subsample analysis, in which analyses are separately conducted separately on subgroups divided according to certain observed characteristics. Alternatively, one can add interaction terms between the treatment variable and the observed characteristics to the empirical model. Researchers should be wary that using pre-treatment variables as moderators is preferable and that interacting the

treatment with post-treatment covariates should be avoided because it will lead to a biased estimation (Montgomery et al., 2018). The subsample analysis requires researchers to use ad hoc discretion to select and test interactions, which becomes an almost infeasible task when the data set contains a large number of covariates. Even if researchers manage to select the subgroups, one may be concerned about the multiple testing/comparison issue and then question the validity of the p-values. For example, when numerous subsample analyses are conducted, the probability that at least one result looks statistically significant at the five percent level may be considerably greater than even when the treatment has no effect on anyone (Athey and Imbens, 2017). Several approaches, such as pre-analysis plans (Casey et al., 2012; Olken, 2015), are proposed to alleviate the multiple testing/comparison concerns.

Another increasingly popular approach for identifying heterogeneous treatment effects is to apply machine learning methods, which are attractive because they automate the search for subgroups, making ex-post selection by researchers no longer necessary. The widely adopted machine learning approaches for heterogeneous treatment effects include LASSO (Imai and Ratkovic, 2013), regression tree (Athey and Imbens, 2016), random forests (e.g., Foster et al., 2011; Wager and Athey, 2018; Lechner, 2018; Athey et al., 2019), "Metalearners" (Künzel et al., 2019), and Bayesian machine learning methods (Imai and Strauss, 2011; Green and Kern, 2012; Taddy et al., 2016). Based on our review, we find that machine learning approaches for estimating treatment effect heterogeneity have not drawn much attention from OM researchers, as evidenced by the fact that none of the 46 reviewed studies applied them. As such, we encourage OM researchers to consider machine learning approaches for testing the heterogeneous treatment effects, especially when a large number of covariates is available.

5. Practice and Practicality

Field experiments are generally difficult to implement, with collaborative field experiments presenting even further challenges. Because published papers seldom describe implementation details, researchers thus find it hard to gauge the level of practical obstacles involved in field experiments (Lopez Mateos et al., 2022). To provide a better understanding of the practical issues surrounding experiment implementation, we discuss the common practices of firms conducting field experiments and provide guidance for researchers aiming to improve their practicality.

One particular type of field experiments, the "A/B" test that randomly assigns a user to one of two or more treatment arms, has gained popularity driven by the recent advance in digital transformation and data analytics (Lopez Mateos et al., 2022). Experimenting with product features on online platforms like e-commerce is much less costly than in traditional industry processes like manufacturing, and technological advancements enable companies to run numerous tests simultaneously and reliably interpret experimental results.

We summarize below the general procedures used by firms when implementing field experiments.

1. **Identify the goal.** The goal can be engagement, clicks, page views, or revenue–anything from clicking a button or link to product purchases.

2. **Decide the places to optimize.** Depending on their goal, firms decide where to implement the experiment. For example, web pages with high traffic allow the experimenters to quickly collect data, whereas those with low conversion rates or high drop-off rates might have more room for improvement.

3. **Generate hypotheses and design the experiment.** The experimenter might propose a theoretical hypothesis of why one variation like changing the color of a button is better than another in reaching the identified goal.

4. **Run experiment.** Companies often use A/B testing software such as Optimizely and Google Optimize to randomly assign users to one of the variations and monitor their actions.

5. **Analyze result.** The A/B testing software designates one of the monitored variations as the baseline and compares it with the remaining variations. The software then reports the difference across variations and the corresponding measures of statistical significance such as p-values and confidence intervals. Researchers should be careful about the statistical inference from field experiments based on basic A/B tests, because as suggested by Berman and Van den Bulte (2021), 70% of false discoveries are actually due to null effects rather than low power. For studies based on A/B tests, researchers should refer to different test designs to reduce the false discovery rate and use two-stage designs with multiple variations rather than basic A/B tests. It is also helpful that the researchers leverage econometric tools such as difference-in-differences to account for heterogeneity in performance baseline and trend over time, or apply causal forest analysis to generate richer empirical results regarding the heterogeneous treatment effects.

To assess whether a partnership with a particular company is viable, researchers may consider several general rules, such as whether a collaborating company has sufficient IT expertise, sufficient scale, reputation, and low staff turnover (see Glennerster (2017) for more details). One issue important for researchers is data availability, as companies tend to have increasingly strict data-sharing policies. Accordingly, we suggest that researchers first consider conducting either a preliminary analysis based on existing data sets, or pilot A/B tests to evaluate the collaboration's potential and facilitate better experiment design.

Another issue for researchers to keep in mind is that their interests might diverge from those of the company. For instance, researchers often emphasize the importance of randomization and study questions with potential theoretical contributions rather than immediate practical implications. As a result, researchers typically need to employ a much more sophisticated design for testing the mechanism than the A/B tests now widely adopted by companies. Before proposing

any field experiments to the collaborating company, researchers must ensure that the experiment design aligns with its interests. Then, during implementation, researchers might alter the design to accommodate the practice. During a collaboration with a leading platform company, for example, one of the authors proposed different pricing algorithms for the firm and planned to test their performance. It was still impossible, however, for the platform to completely follow the proposed algorithms because the machine learning algorithms they had been employing were already rather complex. As a result, the company decided to incorporate our idea into the existing algorithm and compare the modified algorithm with its original version.

6. Challenges and Future Directions

Although the recent advancement of IT has empowered firms with a wide range of levers and enabled OM researchers to expand the application of field experiments, utilizing them in OM is still much more challenging than in other fields for three reasons. First, it is often difficult to generate a large sample with sufficient experimental variations in OM. Compared to the business-to-consumer market, business-to-business markets like supply chains are generally more difficult for researchers to find enough subjects to participate in field experiments. Second, the field experiment settings on the manager (or supply) side are often more transparent than other experiment settings on the customer (or demand) side, because subjects in the former settings are often more closely related to each other and might become aware of other treatment arms through social interactions. Third, there could be much greater effort required to implement a field experiment in operations settings, given that operation parameters such as inventory, capacity, and staffing, are costly and take time to adjust, thereby causing a delay in the treatment effects.

Based on our literature review, we summarize potential directions for field experiments from an OM perspective.

6.1. Theory Testing

The OM literature has traditionally focused on specific problems, such as inventory control and queuing, to develop theories. For example, while the newsvendor model is a foundational OM concept, with laboratory studies documenting its famous "pull-to-center effect" (Schweitzer and Cachon, 2000), there is so far no experiment to verify the existence of this effect in field-based settings.

Another possible theory to test through field experiments is the supply chain contract. In the supply chain literature, an important research area is the performance of wholesale price contracts, revenue sharing contracts, and buyback contracts (Cachon, 2003). The literature has compared the performance of these contracts in both theory and in the lab, and documented the discrepancy between the theoretical prediction and laboratory results. For example, one important theoretical

prediction is that the supply chain can achieve coordination through a revenue sharing contract and a buyback contract, and in fact, the two contracts are equivalent (Cachon, 2003). Laboratory studies, meanwhile, find that these two contracts cannot achieve system coordination and are not equivalent. Given this impasse, field studies may help validate theoretical assumptions and predictions, and explain any discrepancies in the existing findings.

Manufacturing flexibility has been another important topic in the OM literature, in which researchers have analyzed the performance of different flexibility structures (Jordan and Graves, 1995). One important finding is that the "chaining" configuration can almost achieve the performance of full flexibility, while possessing only a limited degree of flexibility (Jordan and Graves, 1995; Hopp et al., 2004). Yet, there is no empirical evidence validating this well-known structure, so leveraging field study may help advance this literature.

While OM also has a long history of theory development in other areas such as queuing and inventory management, very few field experiments have be leveraged to test these theories (Yu et al., 2020). Considering the often simplified assumptions in the theoretical works, it would be meaningful to verify the practicability of the assumptions through field experiments. If the assumptions hold, it is natural to further validate the theoretical predictions. If the assumptions deviate from the empirical settings, it is crucial to identify the conditions under which the theoretical predictions would hold in practice, which will help further advance the theory.

6.2. Optimal Experimental Design

As a best practice, companies want to search across possible designs for one with the highest profit or the lowest cost, which is essentially an optimization, or operations problem. While this is meaningful in practice, the OM field lacks systematic studies investigating general rules regarding optimal experimental design. An exception is Li et al. (2015), in which the authors study how the number of pricing experiments changes when the size of the category grows. They find that the number of experiments does not grow exponentially with respect to the number of products. Similarly, Bhat et al. (2020) explore optimal experiment design in a setting in which the randomization of field experiments can be inefficient.

Since experiment design, or testing hypotheses using data, is not a new practice, the recent advance in popularity of A/B tests brings fresh perspectives. For example, online experimentation allows companies to continuously monitor the field experiment and decide whether to stop and continue the test at any time (Johari et al., 2022). In addition, when firms outsource the online experimentation platform or software, it is often a "black box" with limited or incomplete information available to the company implementing A/B tests. It would be interesting to incorporate these new features of online experimentation and develop applicable design strategies.

6.3. OM Measure

Existing field experiments mainly focus on consumer side measures, with limited attention to operational processes and related measures. While the supply chain literature has studied wholesale prices and lead time decisions, field experiments on these metrics are scarce, due to the difficulty in observing and collecting data. To overcome this, OM researchers should work more closely with practitioners to learn the up-to-date industry practices, which would help researchers obtain new and meaningful metrics that leverage the new approach to collecting OM-related data.

6.4. Research Prioritization

Because existing field experiment-based OM studies tend to answer context-specific questions rather than more general questions, it is difficult for future researchers to replicate the field experiment and thus deepen the understanding of a specific topic. Given how other areas tend to benefit from a focus on research topics, such as the study of newsvendor experiments in the behavioral operations literature, we believe that an agreement regarding more general research directions needs to be prioritized to facilitate future applications of field experiments.

7. Conclusion

The selection problem has long been a concern in empirical studies, thereby preventing researchers from correctly identifying causal effects. One of the best approaches to solving this problem is through randomized experiments, by which researchers can estimate the causal impact of treatment variable(s) on the outcome variable(s) of interest. The random assignment mechanism ensures that any differences between the treatment and control arms are caused by the intervention rather than pre-existing differences between subjects.

However, unlike other areas such as marketing, information system, or finance and accounting, where large samples are available to conduct field study, it is more challenging to acquire empirical data and implement field experiments in OM. Given the increasing trend and potential regarding the application of field experiments to OM (Cohen et al., 2022), we review the existing OM literature leveraging field experiments. We then summarize the common practice, the often overlooked methodological issues, and the corresponding solutions when applying field experiments. We also identify future research directions from an OM perspective, and we hope this paper can provide guidelines for OM researchers in future applications of field experiments.

References

Abbey JD, Blackburn JD, Guide Jr VDR (2015a) Optimal pricing for new and remanufactured products. *Journal of Operations Management* 36(1):130–146.

Abbey JD, Meloy MG (2017) Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management* 53(1):63–70.

Abbey JD, Meloy MG, Guide VDR, Atalay S (2015b) Remanufactured products in closed-loop supply chains for consumer goods. *Production and Operations Management* 24(3):488–503.

Abhishek V, Hosanagar K (2013) Optimal bidding in multi-Item multislot sponsored search auctions. *Operations Research* 61(4):855–873.

Acimovic J, Parker C, F Drake D, Balasubramanian K (2020) Show or tell? Improving inventory support for agent-based businesses at the base of the pyramid. *Manufacturing & Service Operations Management* 24(1):664–681.

Adbi A, Chatterjee C, Drev M, Mishra A (2019) When the big one came: A natural experiment on demand shock and market structure in India's influenza vaccine markets. *Production and Operations Management* 28(4):810–832.

Altman DG (1985) Comparability of randomised groups. *Journal of the Royal Statistical Society: Series D (The Statistician)* 34(1):125–136.

Andrade C (2018) Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian Journal of Psychological Medicine* 40(5):498–499.

Asiedu E, Karlan D, Lambon-Quayefio M, Udry C (2021) A call for structured ethics appendices in social science papers. *Proceedings of the National Academy of Sciences* 118(29).

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey S, Imbens GW (2017) The econometrics of randomized experiments. *Handbook of Economic Field Experiments*, volume 1, 73–140 (Elsevier).

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. *The Annals of Statistics* 47(2):1148–1178.

Austin PC (2009) Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation* 38(6):1228–1234.

Bapna R, Umyarov A (2015) Do your online friends make you pay? A randomized field experiment on peer influence in online social networks. *Management Science* 61(8):1902–1920.

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

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

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

Berman R, Van den Bulte C (2021) False discovery in A/B testing. Forthcoming at Management Science.

Bhat N, Farias VF, Moallemi CC, Sinha D (2020) Near-optimal ab testing. *Management Science* 66(10):4477–4495.

Bichler M, Merting S (2021) Randomized scheduling mechanisms: Assigning course seats in a fair and efficient way. *Production and Operations Management* 30(10):3540–3559.

Boyer KK, Olson JR, Calantone RJ, Jackson EC (2002) Print versus electronic surveys: A comparison of two data collection methodologies. *Journal of Operations Management* 20(4):357–373.

Bray RL, Coviello D, Ichino A, Persico N (2016) Multitasking, multiarmed bandits, and the italian judiciary. *Manufacturing & Service Operations Management* 18(4):545–558.

Broockman DE, Kalla JL, Sekhon JS (2017) The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis* 25(4):435–464.

Bruhn M, McKenzie D (2009) In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics* 1(4):200–232.

Buell RW, Kalkanci B (2021) How transparency into internal and external responsibility initiatives influences consumer choice. *Management Science* 67(2):932–950.

Buell RW, Kim T, Tsay CJ (2017) Creating reciprocal value through operational transparency. *Management Science* 63(6):1673–1695.

Cachon GP (2003) Supply chain coordination with contracts. *Handbooks in operations research and management science* 11:227–339.

Caro F, Gallien J (2012) Clearance pricing optimization for a fast-fashion retailer. *Operations Research* 60(6):1404–1422.

Casey K, Glennerster R, Miguel E (2012) Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *The Quarterly Journal of Economics* 127(4):1755–1812.

Chatterji AK, Findley M, Jensen NM, Meier S, Nielson D (2016) Field experiments in strategy research. *Strategic Management Journal* 37(1):116–132.

Cheung WC, Simchi-Levi D, Wang H (2017) Technical note—dynamic pricing and demand learning with limited price experimentation. *Operations Research* 65(6):1722–1731.

Chuang HHC, Oliva R, Liu S (2016) On-shelf availability, retail performance, and external audits: A field experiment. *Production and Operations Management* 25(5):935–951.

Cohen MC, Fiszer MD, Kim BJ (2022) Frustration-based promotions: Field experiments in ride-sharing. *Management Science* 68(4):2432–2464.

Cohen MC, Fiszer MD, Ratzon A, Sasson R (2021) Incentivizing commuters to carpool: A large field experiment with Waze. Forthcoming at Manufacturing & Service Operations Management.

Craig N, DeHoratius N, Raman A (2016) The impact of supplier inventory service level on retailer demand. *Manufacturing & Service Operations Management* 18(4):461–474.

Cui R, Ding H, Zhu F (2022a) Gender inequality in research productivity during the COVID-19 pandemic. *Manufacturing & Service Operations Management* 24(2):707–726.

Cui R, Li J, Li M, Yu L (2021) Wholesale price discrimination in global sourcing. *Manufacturing & Service Operations Management* 23(5):1096–1117.

Cui R, Li J, Zhang DJ (2020) Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Science* 66(3):1071–1094.

Cui R, Li M, Zhang S (2022b) AI and procurement. *Manufacturing & Service Operations Management* 24(2):691–706.

Cui R, Zhang DJ, Bassamboo A (2019a) Learning from inventory availability information: Evidence from field experiments on Amazon. *Management Science* 65(3):1216–1235.

Cui Z, Kumar PM S, Gonçalves D (2019b) Scoring vs. ranking: An experimental study of idea evaluation processes. *Production and Operations Management* 28(1):176–188.

Czibor E, Jimenez-Gomez D, List JA (2019) The dozen things experimental economists should do (more of). *Southern Economic Journal* 86(2):371–432.

De Vries J, De Koster R, Stam D (2016) Aligning order picking methods, incentive systems, and regulatory focus to increase performance. *Production and Operations Management* 25(8):1363–1376.

DeHoratius N, Raman A (2007) Store manager incentive design and retail performance: An exploratory investigation. *Manufacturing & Service Operations Management* 9(4):518–534.

Dhanorkar S, Muthulingam S (2020) Do E-Waste laws create behavioral spillovers? Quasi-Experimental evidence from California. *Production and Operations Management* 29(7):1738–1766.

Ding Y, Tu Y, Pu J, Qiu L (2021) Environmental factors in operations management: The impact of air quality on product demand. *Production and Operations Management* 30(9):2910–2924.

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

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

Elmaghraby WJ, Gopal A, Pilehvar A (2018) Starting prices in liquidation auctions for IT equipment: Evidence from field experiments. *Production and Operations Management* 27(11):1906–1927.

Feldman J, Zhang D, Liu X, Zhang N (2022) Customer choice models versus machine learning: Finding optimal product displays on Alibaba. *Operations Research* 70(1):309–328.

Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* 18(1):69–88.

Fisher M, Gallino S, Li J (2018) Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management Science* 64(6):2496–2514.

Floyd E, List JA (2016) Using field experiments in accounting and finance. *Journal of Accounting Research* 54(2):437–475.

Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30(24):2867–2880.

Gallien J, Mersereau AJ, Garro A, Mora AD, Vidal MN (2015) Initial shipment decisions for new products at Zara. *Operations Research* 63(2):269–286.

Gallino S, Moreno A (2018) The value of fit information in online retail: Evidence from a randomized field experiment. *Manufacturing & Service Operations Management* 20(4):767–787.

Gaur V, Fisher ML (2005) In-store experiments to determine the impact of price on sales. *Production and Operations Management* 14(4):377–387.

Gee LK (2019) The more you know: Information effects on job application rates in a large field experiment. *Management Science* 65(5):2077–2094.

Gerber AS, Green DP (2012) *Field Experiments: Design, Analysis, and Interpretation* (W. W. Norton), ISBN 978-0-393-97995-4.

Ghanem D, Hirshleifer S, Ortiz-Becerra K (2021) Testing attrition bias in field experiments .

Glennerster R (2017) The practicalities of running randomized evaluations: Partnerships, measurement, ethics, and transparency. *Handbook of Economic Field Experiments*, volume 1, 175–243 (Elsevier).

Gneezy A (2017) Field experimentation in marketing research. *Journal of Marketing Research* 54(1):140–143.

Goldfarb A, Tucker C, Wang Y (2022) Conducting research in marketing with quasi-experiments. *Journal of Marketing* 86(3):1–20.

Grasdal A (2001) The performance of sample selection estimators to control for attrition bias. *Health Economics* 10(5):385–398.

Green DP, Kern HL (2012) Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly* 76(3):491–511.

Hansen BB, Bowers J (2008) Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 23(2):219–236.

Hansen JA, Tummers L (2020) A systematic review of field experiments in public administration. *Public Administration Review* 80(6):921–931.

Hardgrave BC, Aloysius JA, Goyal S (2013) RFID-enabled visibility and retail inventory record inaccuracy: Experiments in the field. *Production and Operations Management* 22(4):843–856.

Harrison GW, List JA (2004) Field experiments. *Journal of Economic Literature* 42(4):1009–1055.

Haruvy E, Popkowski Leszczyc PTL, Ma Y (2014) Does higher transparency lead to more search in online auctions? *Production and Operations Management* 23(2):197–209.

Hausman JA, Wise DA (1979) Attrition bias in experimental and panel data: The gary income maintenance experiment. *Econometrica: Journal of the Econometric Society* 47(2):455–473.

Holguín-Veras J, Amaya-Leal J, Cantillo V, Van Wassenhove LN, Aros-Vera F, Jaller M (2016) Econometric estimation of deprivation cost functions: A contingent valuation experiment. *Journal of Operations Management* 45(1):44–56.

Hopp WJ, Tekin E, Van Oyen MP (2004) Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science* 50(1):83–98.

Hora M, Klassen RD (2013) Learning from others' misfortune: Factors influencing knowledge acquisition to reduce operational risk. *Journal of Operations Management* 1(31):52–61.

Hutchison-Krupat J, Chao RO (2014) Tolerance for failure and incentives for collaborative innovation. *Production and Operations Management* 23(8):1265–1285.

Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.

Imai K, Strauss A (2011) Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* 19(1):1–19.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press).

James O, John P, Moseley A, van Ryzin G, Jilke S (2017) Field experiments in public management. *Experiments in public management research: Challenges and contributions* 89–116.

Johari R, Koomen P, Pekelis L, Walsh D (2022) Always valid inference: Continuous monitoring of a/b tests. *Operations Research* 70(3):1806–1821.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Science* 41(4):577–594.

Jung M, Cho D, Shin E (2021) Repairing a cracked mirror: The heterogeneous effect of personalized digital nudges driven by misperception. *Production and Operations Management* 30(8):2586–2607.

Kagel JH, Roth AE (2016) *The Handbook of Experimental Economics*, volume 2 (Princeton university press).

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182–5190.

Kistler JT, Janakiraman R, Kumar S, Tiwari V (2021) The effect of operational process changes on preoperative patient flow: Evidence from field research. *Production and Operations Management* 30(6):1647–1667.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of The National Academy of Sciences* 116(10):4156–4165.

Lechner M (2018) Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487* .

Lee D, Hosanagar K (2021) How do product attributes and reviews moderate the impact of recommender systems through purchase stages? *Management Science* 67(1):524–546.

Lee DS (2009) Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies* 76(3):1071–1102.

Li H, Zhu F (2021) Information transparency, multihoming, and platform competition: A natural experiment in the daily deals market. *Management Science* 67(7):4384–4407.

Li JQ, Rusmevichientong P, Simester D, Tsitsiklis JN, Zoumpoulis SI (2015) The value of field experiments. *Management Science* 61(7):1722–1740.

List JA (2008) Homo experimentalis evolves. *Science* 321(5886):207–208.

List JA, et al. (2008) Informed consent in social science. *Science* 322(5902):672.

Liu TX, Yang J, Adamic LA, Chen Y (2014) Crowdsourcing with all-pay auctions: A field experiment on taskcn. *Management Science* 60(8):2020–2037.

Lonati S, Quiroga BF, Zehnder C, Antonakis J (2018) On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management* 64(1):19–40.

Looney CA, Hardin AM (2009) Decision support for retirement portfolio management: Overcoming myopic loss aversion via technology design. *Management Science* 55(10):1688–1703.

Lopez Mateos D, Cohen MC, Pyron N (2022) Field experiments for testing revenue strategies in the hospitality industry. *Cornell Hospitality Quarterly* 63(2):247–256.

Lu X, Lu T, Wang CA, Wu R (2021) Can social notifications help to mitigate payment delinquency in online Peer-to-Peer lending? *Production and Operations Management* 30(8):2564–2585.

Manski CF (1989) Schooling as experimentation: a reappraisal of the postsecondary dropout phenomenon. *Economics of Education review* 8(4):305–312.

Mcafee A (2009) The impact of enterprise information technology adoption on operational performance: An empirical investigation. *Production and Operations Management* 11(1):33–53.

McDermott R, Hatemi PK (2020) Ethics in field experimentation: A call to establish new standards to protect the public from unwanted manipulation and real harms. *Proceedings of the National Academy of Sciences* 117(48):30014–30021.

Mejia J, Parker C (2021) When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* 67(1):166–184.

Meyer BD (1995) Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics* 13(2):151–161.

Montgomery JM, Nyhan B, Torres M (2018) How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science* 62(3):760–775.

Morgan KL, Rubin DB (2012) Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40(2):1263–1282.

Mutz DC, Pemantle R, Pham P (2019) The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician* 73(1):32–42.

Olken BA (2015) Promises and perils of pre-analysis plans. *Journal of Economic Perspectives* 29(3):61–80.

Pan X, Dresner M, Mantin B, Zhang JA (2020) Pre-Hurricane consumer stockpiling and Post-Hurricane product availability: Empirical evidence from natural experiments. *Production and Operations Management* 29(10):2350–2380.

Phillips T (2021) Ethics of field experiments. *Annual Review of Political Science* 24:277–300.

Polyviou M, Rungtusanatham MJ, Reczek RW, Knemeyer AM (2018) Supplier non-retention post disruption: What role does anger play? *Journal of Operations Management* 61:1–14.

Queenan C, Cameron K, Snell A, Smalley J, Joglekar N (2019) Patient heal thyself: Reducing hospital readmissions with technology-enabled continuity of care and patient activation. *Production and Operations Management* 28(11):2841–2853.

Retana GF, Forman C, Wu DJ (2016) Proactive customer education, customer retention, and demand for technology support: Evidence from a field experiment. *Manufacturing & Service Operations Management* 18(1):34–50.

Riccobono F, Bruccoleri M, Größler A (2016) Groupthink and project performance: The influence of personal traits and interpersonal ties. *Production and Operations Management* 25(4):609–629.

Robitaille N, House J, Mazar N (2021) Effectiveness of planning prompts on organizations' likelihood to file their overdue taxes: A multi-wave field experiment. *Management Science* 67(7):4327–4340.

Roe BE, Just DR (2009) Internal and external validity in economics research: Tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics* 91(5):1266–1271.

Sahni NS, Zou D, Chintagunta PK (2017) Do targeted discount offers serve as advertising? Evidence from 70 field experiments. *Management Science* 63(8):2688–2705.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46(3):404–420.

Seifert M, Siemsen E, Hadida AL, Eisingerich AB (2015) Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management* 36(1):33–45.

Simester D (2017) Field experiments in marketing. *Handbook of Economic Field Experiments*, volume 1, 465–497 (Elsevier).

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* 68(2):846–865.

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

Ta H, Esper TL, Tokar T (2021) Appealing to the crowd: Motivation message framing and crowdsourcing performance in retail operations. *Production and Operations Management* 30(9):3192–3212.

Taddy M, Gardner M, Chen L, Draper D (2016) A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics* 34(4):661–672.

Tonin M, Vlassopoulos M (2015) Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science* 61(8):1795–1811.

Tucker AL, Singer SJ (2015) The effectiveness of management-by-walking-around: A randomized field study. *Production and Operations Management* 24(2):253–271.

Tucker C, Zhang J (2011) How does popularity information affect choices? A field experiment. *Management Science* 57(5):828–842.

Venkatesh V, Chan FK, Thong JY (2012) Designing e-government services: Key service attributes and citizens' preference structures. *Journal of Operations Management* 30(1-2):116–133.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wang Y, Goes P, Wei Z, Zeng D (2019) Production of online Word-of-Mouth: Peer effects and the moderation of user characteristics. *Production and Operations Management* 28(7):1621–1640.

Wooldridge JM (2002) Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese economic journal* 1(2):117–139.

Wooten JO, Ulrich KT (2017) Idea generation and the role of feedback: Evidence from field experiments with innovation tournaments. *Production and Operations Management* 26(1):80–99.

Yu Q, Zhang Y, Zhou YP (2020) Delay information in virtual queues: A large-scale field experiment on a ride-sharing platform. *Available at SSRN 3687302* .

Zhang D, Allon G, Van Mieghem J (2017) Does social interaction improve learning outcomes? Field evidence from massive open online education. *Manufacturing & Service Operations Management* 19(3):347–367.

Zhang DJ, Dai H, Dong L, Qi F, Zhang N, Liu X, Liu Z, Yang J (2020) The long-term and spillover effects of price promotions on retailing platforms: Evidence from a large randomized experiment on Alibaba. *Management Science* 66(6):2589–2609.

Zhang DJ, Dai H, Dong L, Wu Q, Guo L, Liu X (2019) The value of pop-up stores on retailing platforms: Evidence from a field experiment with Alibaba. *Management Science* 65(11):5142–5151.

Zhang N, Kannan K, Shanthikumar G (2021) Nudging a Slow-Moving High-Margin product in a supply chain with constrained capacity. *Production and Operations Management* 30(1):11–27.

Appendix. Literature Review by Research Topic

A. Workforce Management

A handful of studies leverage field experiments to explore the factors affecting workforce performance and, thus, generate valuable insights for business managers. Two papers focus on warehouse operations. De Vries et al. (2016) investigate the extent to which the incentive system of the tasks (competition-based vs. cooperation-based) and regulatory focus of pickers (prevention-focus vs. promotion-focus) influence picking performance with three manual picker-to-parts order picking methods (parallel, zone, and dynamic zone picking). Through a controlled field experiment conducted in a warehouse environment, they find that for promotion-focused pickers, when using a parallel picking method, a competition-based incentive system leads to a higher level of productivity than a cooperation-based incentive system; moreover, when using a zone picking method, it is more productive to use a cooperation-based incentive system. Moreover, prevention-focused pickers become more productive in zone picking with a cooperation-based incentive system than a competition-based incentive system but deliver a similar productivity performance with different incentive systems in the other two picking methods. Sun et al. (2022) propose a "human-centric bin packing algorithm" to predict packing workers' non-conforming behavior and improve workers' packing efficiency in a warehouse. Collaborating with the Alibaba Group for two weeks, they conduct a randomized experiment in four warehouses to causally evaluate the performance of the algorithm design: the non-conformance rate is lowered by 5.7% and the average packing time is decreased by 4.5%.

Furthermore, three papers examine the role of financial incentive, feedback design, and evaluation on innovation management. Hutchison-Krupat and Chao (2014) study how rewards and penalties impact the decisions of individuals engaged in a collaborative innovation initiative. The experimental results demonstrate that financial rewards have a positive effect on resource allocation, whereas financial penalties have a negative effect on resource allocation. Wooten and Ulrich (2017) conduct a set of field experiments using two online contest websites for a logo design to compare the impact of three different feedback treatments (no feedback, random feedback, and directed feedback[5]) on idea generation. The results suggest that directed feedback results in the highest level of participation and random feedback induces more participation than no feedback. With respect to outcome, while directed feedback improves the average quality of entries submitted, no feedback or random feedback may produce better top-end entry quality. Further, the authors find that, under directed feedback, the variance in quality declines as the contest progresses. Cui et al. (2019b) examine the efficacy of two idea evaluation processes, scoring vs. ranking. They find that the scoring process strictly outperforms the ranking process in terms of the likelihood of selecting the highest-quality ideas. This finding remains robust when there is a tie in the ranking process. In addition, when they reduce the number of ideas to be evaluated from eight to three, the efficacy of the two idea evaluation processes becomes similar. Moreover, the efficacy of the ranking process is improved when additional information is provided, but the efficacy of the scoring process does not change.

[5] Directed feedback is the in-process feedback that is highly correlated with the final quality rating of the entry.

The remaining papers focus on more granularly-defined contexts, but all of them aim to improve worker performance. Riccobono et al. (2016) investigate how "groupthink concurrence-seeking behavior" (GTB) interacts with group member personal traits and interpersonal ties and, in turn, influences the project performance in the context of business process reengineering (BPR) projects. The findings demonstrate that the negative effect of GTB on group performance in BPR projects is indeed moderated by personal traits and interpersonal ties. More specifically, while perceived control, conscientiousness, and interpersonal evaluation mitigate this negative impact, confidence and previous relationships amplify the negative impact of GTB on group performance. By manipulating the merchants' discretionary power to override the data-driven decision-making (DDD) tool in a field experiment, Kesavan and Kushwaha (2020) find that a merchant overriding the DDD tool reduces profitability by 5.77%. Moreover, the analysis of the product lifecycle reveals that merchants increase profitability for growth-stage products but decrease profitability for products in mature and declining stages.

Two papers study agent behaviors. In the context of mobile money, Acimovic et al. (2020) examine how different types of guidance and the provision of in-person training help agents improve the service quality. The results from a field experiment with 4,771 agents in Tanzania suggest that in-person training paired with explicit recommendations leads to an improvement in agent performance. In contrast, agents in other experimental conditions show no statistically significant change in their performance. Further, they find that the performance improvements concentrate on agents who never replenished their money at a bank and maintained sub-optimal replenishment levels in the pre-treatment period. Ta et al. (2021) draw on the foundations of the self-determination theory and the heuristic-systematic model to examine the manners in which variations in messages presented to crowdsourced agents can serve as a mechanism to enhance participation and associated performance outcomes. Data from a field experiment involving a retail inventory audit task reveal that messages appealing to the crowd's consumer identity—as opposed to crowdsourcing platform identification or firm identification—generally lead to superior performance outcomes, particularly shorter reservation time, higher task quality approval, and post-task satisfaction. However, these effects are contingent on the valence of the message frame and the nature of the task.

B. Supply Chain Management

Studies in supply chain management are increasingly implementing field experiments to causally evaluate means to effectively manage supply chains. Hardgrave et al. (2013) investigate the effectiveness of the visibility into the movement of inventories in a supply chain, which is enabled by Radio Frequency Identification (RFID), in reducing retail store inventory record inaccuracy (IRI). The results from two different field experiments reveal that the effectiveness of RFID in reducing IRI varies by category and RFID ameliorates the effects of known determinants of IRI. The study also finds that the technology is most effective for product categories characterized by these determinants. Based on a field experiment at a major supplier of branded apparel (i.e., Hugo Boss), Craig et al. (2016) study the impact of a supplier's inventory service level on demand from its retailer customers. Their findings demonstrate a positive relationship between historical fill rate and current retailer demand, and this relationship is stronger for retailers that order more frequently. Further, they show that the increase in demand is primarily the result of retailer customers placing larger

orders, rather than the changes in retailer assortment or order cycles. Moreover, Zhang et al. (2021) develop a support vector machine (SVM) approach to identifying customers who could be nudged to purchase a slow-moving high-margin product (i.e., long-tail product) with constrained capacity in a supply chain. They run a field experiment to evaluate the performance of approaches in nudging the identified customers and find that their SVM-based approach outperforms other approaches.

Two recent papers focus on wholesale price discrimination from suppliers in a supply chain. Cui et al. (2022b) explore how a buyer's AI strategy affects the wholesale price received from suppliers. In collaboration with a trading company, they design and conduct a randomized field experiment in which suppliers' wholesale price quotes are compared across female, male, and chatbot buyer types under AI and no recommendation conditions. The results indicate that suppliers price-discriminate against a not-so-smart chatbot buyer by providing a higher price quote. However, introducing a smart control and signaling that the supplier is recommended by a smart system can effectively reduce the price quoted for chatbot buyers. Furthermore, they show that when automation and smartness are jointly adopted by buyers, AI will deliver the most value in procurement. To investigate the wholesale price discrimination in business-to-business markets, Cui et al. (2021) collaborate with a global trading company that runs a field experiment. Their study provides evidence supporting the wholesale price discrimination against white buyers in global sourcing. But there is no evidence indicating price discrimination based on country: buyers from the U.S. and South African markets receive the same price quote from suppliers. They further show that price discrimination disappears when buyers present market information (i.e., market price) to suppliers, whereas price discrimination remains when social information (i.e., a referral from a previous customer) is presented.

C. Retail Management

One stream of literature in retail management aims to examine the impact of pricing strategies through field experiments. In collaboration with Zara, Caro and Gallien (2012) design and implement an autonomous decision-making process to optimize clearance prices. To evaluate the effectiveness of the pricing solution, they conduct a controlled field experiment in all Belgian and Irish stores in 2008, the results of which indicate that the pricing solution increases the clearance revenue by approximately 6%. Further, Abbey et al. (2015a) run an experiment with U.S. consumers to investigate the optimal pricing of new and remanufactured products. The investigation reveals two distinct segments of consumers: one segment shows indifference between new and remanufactured products and is highly sensitive to price discounts, whereas the other segment prefers new products over remanufactured products and is relatively insensitive to price discounts. Ferreira et al. (2016) develop an algorithm to predict future demand for new products and translate the demand forecasts into a pricing policy by collaborating with an online retailer, Rue La La. Through a field experiment, they find that the pricing decision algorithm increases the revenue from first exposure styles by approximately 9.7%, while barely influencing aggregate sales. Cheung et al. (2017) propose a pricing algorithm to address a dynamic pricing problem and utilize a field experiment conducted by Groupon to demonstrate that implementing their algorithm increases daily bookings by 116% and daily revenue by 21.7%. Fisher et al. (2018) propose a best-response pricing strategy for retailers that follow a competition-based dynamic-pricing strategy. The results from a field experiment at a leading Chinese online retailer validate the proposed strategy: adopting

the strategy increases the daily categorical revenue by 11%. Collaborating with a large liquidation company for IT equipment, Elmaghraby et al. (2018) run a field experiment by manipulating auction starting prices. The paper provides insight into the effect of starting prices on the final auction prices of returned IT products and finds evidence of cross-product dependencies. Zhang et al. (2020) show the effect of dynamic pricing through price promotion on consumer behavior. In the short term, the shopping cart promotion significantly boosts consumers' purchasing probability and expenditure. In the long run, price promotion may cause customers to behave more strategically in the sense that more products are added to the shopping cart, and the price paid for a product without promotions decreases. Moreover, they observe that the long-term effects spill over to sellers without the shopping cart promotion.

The other stream of literature in retail management focuses on how various factors affect the performance outcome, such as sales and revenue, and propose strategies for business practitioners accordingly. Several studies have been conducted in the context of offline retailers. Gaur and Fisher (2005) present an experimental methodology to help retail managers measure how demand varies with price. The application of the experimental methodology at a toy retailer generates an unexpected finding: demand increases with price in certain cases. To understand consumer perception of remanufactured products in closed-loop supply chains, Abbey et al. (2015b) utilize an experiment and conclude that the perceived attractiveness of remanufactured products can be affected by factors such as price discount, brand equity, product quality perceptions, and negative attribute perceptions (e.g., disgust). Further, they find that green consumers and consumers who consider remanufactured products green usually perceive remanufactured products as more attractive. In collaboration with Zara, Gallien et al. (2015) develop and test a decision support system for allocating limited stock by location over time. Based on a worldwide field experiment with 34 articles in 2012, they estimate the system to increase season sales by approximately 2.2%, to reduce the number of unsold units at the end of the regular selling season by 4.3%, to increase the proportion of shipments sold by 1.4%, and to increase the proportion of demand converted into sales by 1.5%. Chuang et al. (2016) conduct a field experiment in a national retailer's store set to explore the operational and financial feasibility of adopting external shelf-audits. For treated stores, they use transactional data to detect abnormal operations and respond to possible shelf out-of-stocks (OOS)[6] by sending auditors to correct empty shelves and incorrect inventory records. They find that Stock Keeping Units (SKU) in the treatment group are less likely to have shelf-OOS and inventory record inaccuracy, and experience a significant increase in sales. More importantly, they find the external shelf-audits to be economically feasible because the required auditing efforts after a transitional period are low. Furthermore, Ding et al. (2021) analyze how air quality affects the demand for different product color options. The results from an observational study, a field experiment, and two lab experiments consistently suggest evidence of greater demand for blue-color product options on air-polluted days than that on clear days. Through two field experiments and complementary online experiments, Buell and Kalkanci (2021) find that transparency into both internal and external responsibility initiatives tends to dominate generic brand marketing in motivating consumer purchases, and transparency into a company's

[6] Shelf OOS refers to the scenario where the item is in the store but customers cannot find it.

internal responsibility practices can be at least as motivating for consumer sales as transparency into its external responsibility initiatives.

The remaining studies are in the context of online retail management. Gallino and Moreno (2018) examine the value of virtual fit information in online retail with a randomized field experiment in which the availability of virtual fit information is manipulated. They find that providing virtual fit information increases conversion rates and order value, and reduces fulfillment costs from returns and home try-on behaviors (i.e., customers ordering multiple sizes of the same product). Zhang et al. (2019) study the value of the omnichannel retail strategy, particularly short-lived and experientially oriented pop-up stores, in a large-scale field experiment. They randomly assign approximately 800,000 customers to either receive a message about a pop-up store event or not receive any message related to the event. The results reveal that receiving the message increases foot traffic to the pop-up store and, in turn, boosts expenditure at participating retailers' online stores after the event ended. From the platform's perspective, they find that pop-up store visits have a spillover effect: retailers that sell related products but do not participate in the pop-up store event experience an increase in consumers' purchases as well. Cui et al. (2019a) run two field experiments on Amazon to study whether and how consumer purchase behavior is impacted by the inventory availability information. By creating exogenous shocks on the availability information for a random subset of Amazon lightning deals, the paper demonstrates that a decrease in product availability causally attracts more consumers to purchase in the future. In terms of magnitude, a 10% increase in past claims leads to a 2.08% increase in cart add-ins in the next hour. To identify the optimal set of products to display online, Feldman et al. (2022) run a field experiment to compare two approaches: traditional customer choice models and a featured multinomial logit (MNL) model (driven by machine learning). They conclude that the MNL-based approach can significantly improve the revenue generated per consumer visit, which is due to the closer integration of MNL with the downstream optimization problem.

D. Service Operations

Field experiments have become a popular tool for studies focusing on service operations to causally evaluate the effect of a policy on performance outcomes, such as consumer outcomes (Retana et al., 2016; Buell et al., 2017; Zhang et al., 2017; Jung et al., 2021), efficiency (Bray et al., 2016; Bichler and Merting, 2021), and fairness (Cui et al., 2020), etc.

Numerous studies have investigated whether the provision of additional information improves individuals' service experience and service outcomes. Retana et al. (2016) study the impact of service providers' efforts to educate customers on consumer outcomes. By analyzing a field experiment executed by a major public cloud infrastructure services provider in 2011, they find that proactive education reduces by half the number of customers who churn from the service during the first week. Further, educated customers ask 19.55% fewer questions during the first week of their tenure and increase their accumulated usage of the service by 46.57% in the eight months after sign-up. Finally, they provide evidence that the treatment effects are strongest among customers who have less experience with the provider. Further, Buell et al. (2017) exploit two field and two laboratory experiments to demonstrate that the introduction of visual transparency between consumers and producers not only improves customer perceptions but also increases service quality

and efficiency. Zhang et al. (2017) analyze whether encouraging social interaction among students improves learning outcomes in massive open online courses by randomly assigning the opportunity to visit the course discussion board (i.e., social interactions) or one-on-one discussions (i.e., small-group interaction). They estimate that one additional board visit causally increases the probability of a student completing the quiz in the subsequent week by up to 4.3%. Students who followed through and actually conducted one-on-one discussions improved their quiz completion rates and quiz scores by 10% in the subsequent week. Jung et al. (2021) examine how consumers' misconception of their past energy consumption determines their effort provision and daily energy consumption behaviors when people's goal-setting and feedback are provided. Based on a field experiment that exploits the widespread deployment of smart metering services, they find that 1) goal-setting intervention reduces overestimating users' energy consumption while having no impact on underestimating users, and 2) performance feedback makes underestimating users consume less energy. Furthermore, to study the response of borrowers to different types of social notifications, Lu et al. (2021) collaborate with a Chinese P2P platform to implement a field experiment, in which borrowers who failed to repay an installment were randomly assigned to three groups—the control group, the peripheral-circle group and the core-circle group. Their results show that social notifications to either the core-circle or the peripheral-circle decrease the default rate by about 50%. In addition, results from survival analyses show that social notifications targeted at core social contacts affect both the short-term and long-term repayment behavior and their effectiveness increases with repetition, whereas peripheral-circle notifications only have a short-term effect and their effectiveness decreases with repetition.

Two papers examine process decisions in the context of auctions. Through a controlled field experiment, Haruvy et al. (2014) find that higher transparency in terms of auction comparability increases willingness to pay, revenues, time spent on search, price sensitivity, and lower price dispersion between concurrent auctions. However, increasing transparency in terms of the level of detail provided to bidders decreases willingness to pay, revenue, search, and price sensitivity. Abhishek and Hosanagar (2013) propose two bidding policies for multi-item multi-slot sponsored search auctions—myopic policy and forward-looking policy—and evaluate their effectiveness with a field experiment. The results from a DID analysis indicate that the myopic policy can increase the performance of the advertising campaign by 75.38%, and the improvement in the performance led by the forward-looking policy is estimated to be 83.25%.

Two papers focus on scheduling. Considering that it is customers who are being scheduled, researchers study different scheduling policies to reduce waiting time and improve efficiency. Bray et al. (2016) run a field experiment to measure the effect of switching from a hearing-level FIFO scheduling policy to a case-level FIFO policy in the Roman Labor Court of Appeals. The results indicate that the new scheduling policy decreases the average case duration by 12% and the probability of a decision being appealed to the Italian supreme court by 3.8%. Bichler and Merting (2021) run large-scaled field experiments to compare two scheduling mechanisms—Bundled Probabilistic Serial (BPS) and First-Come First-Served (FCFS)—in the contexshoourse assignments. The results of two field experiments show that while the advantages of BPS over FCFS are not large, envy-freeness turns out to be an important advantage of BPS.

There are two studies that address questions related to ride-sharing, a trending topic in transportation management. Mejia and Parker (2021) study how a rider's gender, race, and perception of support for lesbian,

gay, bisexual, and transgender (LGBT) rights impact cancellation rates. By manipulating rider names and profile images on a major ride-sharing platform, they confirm the elimination of any bias at the ride request stage. However, after acceptance, racial and LGBT biases are persistent, while there is no evidence of gender biases. Meanwhile, they find that higher prices due to peak timing can lower cancellation rates for non-Caucasian riders. But this effect does not apply to riders who signal support for the LGBT community. Leveraging the Waze Carpool service to run a digital field experiment, Cohen et al. (2021) examine the impact of two types of incentives—highlighting time-saving and monetary compensation ($10 welcome bonus)—on commuters' intention to carpool. The results suggest that highlighting the high-occupancy vehicle (HOV) lane effectively nudge commuters to carpool. However, highlighting both the HOV lane and the potential time saving does not yield an additional marginal impact compared to only mentioning the HOV lane. As for the monetary incentive, the finding suggests that highlighting the welcome bonus has a rather minimal impact on commuters' intention to carpool.

A recent paper (Cui et al., 2020) examines the social aspects of service operations by exploring ways to reduce racial discrimination. The authors conduct four randomized field experiments on Airbnb by sending accommodation requests from fictitous quest accounts. They find that requests from guests with African-American–sounding names are less likely to be accepted than those with white-sounding names. However, a positive review posted on a guest's page significantly reduces discrimination. They further show the importance of credible peer-generated reviews: a blank review without any content or a nonpositive review can also help attenuate discrimination, yet self-claimed information cannot do so.

E. Health Care

Improvements in healthcare outcomes have been a core component in operations management, where researchers investigate ways to improve physician performance (Tucker and Singer, 2015), care quality (Queenan et al., 2019), and care efficiency (Kistler et al., 2021). Tucker and Singer (2015) randomly select hospitals to implement the 18-month-long, management-by-walking-around (MBWA)-based improvement program, in which senior managers observe frontline employees, solicit ideas regarding improvement opportunities, and work with staff to resolve issues. The paper finds that the program had a negative impact on performance and the impact is moderated by the work area's problem-solving approach: prioritizing easy-to-solve problems was associated with improved performance, while doing that with high-value problems was not successful. They also find that assigning responsibility to senior managers for ensuring that identified problems are resolved results in better performance. Queenan et al. (2019) examine the role of the patient activation measure (PAM)[7] on care quality. Based on a randomized, controlled field experiment with technology-enabled continuity of care intervention, the paper shows a direct effect of health IT, together with its interaction with PAM, reduces readmission over the base case (i.e., without technology-enabled continuity of care). Kistler et al. (2021) examine the impact of operational process changes on the preoperative flow of patients. The treatment is the implementation of centralized decision-making and the introduction of an information technology enabled intraoperative prompt. Accordingly, there are two distinct patient groups:

[7] Patients' skills, knowledge, and motivation to actively engage in their health care are assessed using the patient activation measure (PAM).

the treatment group that is impacted by the implemented operational changes and the control group that was not impacted by the changes. Using the DID framework, the authors show that that information coordination is beneficial by comparing the preoperative patient processing time of the control and treatment groups before and after each process change.