Behavioral Empirics and Field Experiments

Maria R. Ibanez¹ and Bradley R. Staats²

¹ Harvard Business School
² University of North Carolina at Chapel Hill

4.1. Going to the field to study behavioral operations

Management researchers seek not only to develop and test rigorous theories, but also to tackle and answer questions that are relevant to the individuals and organizations under study. While this dual purpose requires effort, a focus on building fundamental understanding (rigor) and on considering the application of that understanding (relevance) do not have to be at odds with one another, but instead can be complementary (Stokes, 1997; Tushman & O'Reilly, 2007). As the study of behavioral operations has continued to grow in its understanding of how human aspects alter the idealized state of rational decision making and operational processes more broadly, a growing number of researchers are turning to the field¹ to push deeper in order to find the answers to relevant behavioral questions. Why go to the field? More specifically, how does going to the field extend and enhance what we do in the lab or with analytical models? At its core, it presents the opportunity for researchers to situate their work clearly in the context in which it is actually occurring. This not only creates the potential for additional insight, but it also

¹ When we use the term “the field” we mean real world settings in which participants are conducting their ‘normal work’. In other words, settings that are not contrived to permit a researcher careful study of a given topic, but rather those settings where individuals complete their typical activities. It is important to highlight that the field is not a singular construct; on the contrary, researchers want to think about what contextual factors in the field are most conducive to the behavioral mechanism of interest and then choose a setting appropriately.
helps to overcome methodological challenges (while at the same time adding new challenges). Fieldwork, at its best, permits one to conduct research that is both rigorous and relevant.

More specifically, going to the field offers researchers five advantages as compared to other approaches: (1) the opportunity to establish external validity and identify effect size; (2) an ability to overcome observer bias; (3) valuable context with which to understand phenomena more deeply; (4) the chance to identify time-based effects; and (5) an occasion to investigate more than individual decisions. We now explore each advantage in more detail.

4.1.1. External validity and identification of effect size

First, and most obviously, a methodological strength of field studies is that they permit one to establish external validity. When a new research exploration begins, this may be as simple as using case studies to provide an existence proof – that is, evidence that a phenomenon exists in practice – or to begin to understand the factors that shape the relationships of interest. For example, MacDuffie (1997) identifies differences in process improvement across three automakers and then explores how the problem solving process used in each factory explains this variability.

External validity can also be established through empirical studies that analyze data produced within a company by workers conducting their normal activities. For example, the operations field often traces its roots to the pioneering work of Frederick Taylor (Smiddy & Naum, 1954). Taylor (1911) studied worker’s actions to attempt to find improved tools and methods that could be used to improve overall productivity. By identifying the impact of practices in real-world operational settings, researchers clearly establish the relevance of the work.
Moreover, field work is typically required in order to determine effect size. As Cachon (2012) states, one recipe for interesting research consists of demonstrating that an effect that was thought to be small is large, or that one that was thought to be large is much larger or instead small. While the lab identifies effects, the field brings the context that allows one to estimate the size of the effects. And only the field can bring the full contextual detail, including incentives and high-stakes, to understand complex behaviors such as employee theft (Pierce, Snow, & McAfee, 2014), process compliance (Staats, Dai, Hofmann, & Milkman, 2015), peer effects (Mas & Moretti, 2009), multi-tasking (KC, 2013), front-line workers’ deviance (Ibanez, Clark, Huckman, & Staats, 2016), unethical choices (Pierce & Snyder, 2008), or worker attendance (Duflo, Hanna, & Ryan, 2012). For example, in the context of retail, field studies have been conducted to evaluate how sales are affected by price (Gaur & Fisher, 2005) and external audits of on-shelf inventory positions (Chuang, Oliva, & Liu, 2016).

4.1.2. Overcome observer bias

Second, field studies eliminate (or at least mitigate) observer bias, since data come from subjects’ behaviors in their natural environments and subjects are often unaware of their participation in a study. In addition, even in those cases in which they are notified ahead of time, an observer’s presence grows less salient over time and so participants are likely to forget as time passes and they remain in their familiar environment. As a result, researchers can study behaviors that would otherwise be modified by the subjects, such as discrimination (Bertrand & Mullainathan, 2004; Milkman, Akinola, & Chugh, 2012). For example, Milkman et al. (2012) wished to know if individuals show bias in accepting meeting requests. By randomly assigning email requests from requestors of different gender and ethnicity, they were able to gain insight
into how actual decisions were made, finding that women and minority applicants were significantly less likely to receive a response.

When turning to the field to study human behavior, researchers are increasingly able to rely on technology to remove observer bias. Webb, Campbell, Schwartz, and Sechrest (1966) highlighted the need to go to the field and use trace data to remove the researcher from the picture and avoid demand or Hawthorne effects (Roethlisberger & Dickson, 1934). At the time of Webb et al.’s writing, looking for trace events in practice was quite challenging. For example, to understand how people moved around an office would mean looking for wear patterns in the carpet at night. However, with technological advances, it is now possible to rely on tools such as traffic counters (Perdikaki, Kesavan, & Swaminathan, 2012) or video cameras (Lu, Musalem, Olivares, & Schilkrut, 2013) to study behavior in retail stores, RFID devices to measure process compliance in hand-washing (Dai, Milkman, Hofmann, & Staats, 2015) or email usage to capture interaction patterns (Aral, Brynjolfsson, & Alstyne, 2012).

4.1.3. Context

Third, field research contributes to our understanding of a phenomenon by incorporating the role of context. When our theories are forced to predict what happens in practice then the inevitable gaps that exist in a theory are exposed. Thomas Kuhn (1962) describes the scientific evolutionary process including how the identification of discrepancies leads to modified or even new theories. For science to progress we have to further refine our models, which means identifying boundary conditions, moderators, and other variables of interest – and the field is

---

2 “Hawthorne effects” refer to participants’ change of their behavior simply because they are being watched. In a study seeking to understand the impact of lighting on productivity, this could mean that workers grow more productive when lights are turned up in the production setting, and then later increase their productivity again when the lighting is lowered. The concern would be that the omitted variable driving this effect is not lighting, but rather the researcher’s presence.
helpful here. It is the realistic context that enables identifying important moderators as well as interrelationships with other variables (Clark & Huckman, 2012; Ibanez et al., 2016; KC & Terwiesch, 2011; Nagin, Rebitzer, Sanders, & Taylor, 2002; Narayanan, Balasubramanian, & Swaminathan, 2009; Tucker, 2007).

An example comes from a series of studies on specialization. Building on earlier work in psychology, organizational learning, and operations, Schilling, Vidal, Ployhart, and Marangoni (2003) found in the lab that performance improved more quickly when individuals completed multiple, related tasks, as compared to either just one task (specialization) or unrelated tasks. However, by studying the topic in the field, examining software maintenance tasks, Narayanan et al. (2009) found that neither strategy was always dominant, but rather balance between the two was necessary. Too much specialization was problematic, but so too was too much variety.

4.1.4. Time-based effects

Fourth, the field enables the researcher to capture events over longer time periods, such as interpersonal relationships (e.g., trust) or intrapersonal behavior (e.g., habit formation) Schilling et al. (2003). Creative laboratory studies can find ways to attempt to capture many of these effects – for example by asking subjects to stay for several hours to repeat an exercise or even come back a following week to extend the time of exploration. For example, since Narayanan et al. (2009) wished to study team member familiarity and they knew that they could not proxy it in the lab over an hour, they drew subjects from the same social group for their familiar condition. However, there are times when even the most creative researcher cannot overcome the challenges of the topic under study in order to explore it in the lab. In such cases, by turning to the field, it is possible to gain access to an appropriate setting in order to study the topic of interest.
Scholars have benefited from the long time scales that exist in practice, but not the lab, to study habit formation in gym membership (Charness & Gneezy, 2009; Milkman, Minson, & Volpp, 2014), energy usage (Allcott & Rogers, Forthcoming) or process compliance (Staats, Dai, Hofmann, & Milkman, 2016). These studies have found that interventions to encourage participation work strongly in the short-term, and tend to have persistent yet declining effects over time. For example, Staats et al. (2016) find that process compliance actually declines to pre-intervention levels when electronic monitoring is removed from a hospital unit in the context of hand hygiene. With a time scale of years before the effect is observed, it would be challenging to replicate within the lab.

Building further on the work on specialization described above, Staats and Gino (2012) explored the topic in a Japanese bank and found that time was an important aspect, as specialization was dominant in the short-term, but variety mattered more in the long-term. In both of the latter two papers, the additional insight was gained by looking at data over extended time periods – e.g., month after month of working as opposed to time within an hour. By turning to the field, new interesting and important moderators can be revealed.

As the evolution of work in the study of specialization and variety illustrates, lab and field should be seen as complements rather than substitutes; in particular, although the lab tends to precede the field, researchers can go back to the lab after field experiments (Harrison & List, 2004). For example, Plott and Levine (1978) found in the field that by shaping the agenda for a committee meeting they were able to get the outcome they sought. In particular, a flying club was voting on which aircrafts to include in its fleet and the agenda was altered in order to attempt to have the majority vote for the authors chosen aircraft. After examining this field data, the authors then conducted a laboratory experiment to rule out the possibility that their results were
accidental. Going back to the lab after studying a topic in the field also creates an opportunity to identify mechanisms that underlie results. For example, Cable, Gino, and Staats (2013) conducted a field experiment at an Indian business process outsourcer where they found that an individually-focused onboarding process was related to lower attrition and better operational performance than was an organizationally-focused onboarding process. Although the field experiment suggested that authentic self-expression – being able to act as oneself – was driving the result, the evidence for the mechanism was limited. To address this shortcoming, they then conducted a lab experiment to replicate conceptually the field experiment and to identify the mechanism - which was confirmed to be authentic self-expression. Staats, KC, and Gino (2016) use a similar approach when they analyze a natural experiment that occurred after a warning announcement from the US Food and Drug Administration (FDA). The announcement permits causal identification of how cardiologists behave, with respect to which cardiac stent they choose to use, but additional lab studies aid in mechanism identification for a similar decision-making process.

4.1.5. Beyond individual decision-making

Fifth, while in lab experiments the unit is a person or occasionally a team, going to the field allows for subjects other than individuals. This moving up in levels is important because, although factories, organizations, and countries are made up of individuals, the macro factors that affect them are not always decomposable into individual-level studies. By turning to the field, important theoretical and practical insights can be gained on different levels of analysis. For example, field studies find that lower proximity (resulting from new airline routes) of manufacturing plants to headquarters increases plant-level investment (Giroud, 2013), and that additional capital (resulting from shocks to capital stock generated using randomized in-kind
equipment/inventories or cash grants equivalent to either three or six months of median profits) increases profits by 60% per year for Sri Lankan microenterprises, implying marginal returns above the market interest rates (de Mel, McKenzie, & Woodruff, 2008).

Bloom, Eifert, Mahajan, McKenzie, and Roberts (2013) recruited multi-plant textile manufacturers (woven cotton fabric) in India to conduct an experiment to evaluate the effects of management practices on firm performance, keeping labor and capital inputs constant. Management consulting was used as “a mechanism of convenience” to improve management practices. All plants received initial diagnostic consulting for a month and some light consulting months later to collect data on management and performance. In between, the 14 treated plants received four additional months of implementation consulting that the 6 control plants did not. They measured management practices according to 38 key practices related to factory operations (e.g., recording reasons for machine breakdowns), quality control (e.g., monitoring quality defects’ records), inventory control (e.g., monitoring stock), planning (e.g., sales and operations managers meet regularly), human resources (e.g., performance-based rewards), and sales and orders (e.g., order-wise production planning). The experiment showed that treatment generated a 17% increase in productivity. Reflecting on their findings, the authors argue that the main reason for lack of implementation of managerial practices in the past was that managers did not believe the practices would be profitable, hence suggesting that information constraints explain differences in productivity across firms and countries.

As can be seen by the discussion above, going to the field offers numerous potential advantages for behavioral operations management researchers. In the remainder of this chapter, we review the main empirical methods used to analyze field data, as well as provide examples of
applications. Then, we focus our attention on one area—field experiments—for a deeper exploration.

4.2. Analyzing the Data: Common Empirical Methods

A quick examination of journal articles in operations management reveals numerous empirical methods that can assist researchers in answering questions about human behavior in operations. A number of review articles have focused on many of these research methods (Fisher, 2007; Scudder & Hill, 1998). Given the prevalence of these methods in the field, we quickly review several key methods that have shown or hold promise for behavioral researchers. Depending on the characteristics of the data and the topic studied, five types of empirical methods are commonly used, often in combination.

4.2.1. Reduced Form Analysis of Panel Data

Perhaps the most common empirical approach involves reduced-form estimation using panel data. By using data from multiple units (e.g., individuals, teams, factories) over time, one can use these two dimensions to extract the effects of a particular factor controlling for the heterogeneity across units and/or trends. In other words, the researcher is able to observe how independent variables of interest drive changes in outcomes for focal units over time. One of the oldest streams of literature in this tradition is that of learning curves. The idea that outcomes improve, on average, with increasing cumulative experience has been studied for many years (Lapré & Nembhard, 2010). For example, Argote, Beckman, and Epple (1990) used monthly (annual) data from 13 of the 16 shipyards that built the Liberty Ships during World War II to study learning over time within and across shipyards. Exploiting the two dimensions of the data, the authors included a monthly trend as well as dummy variables for the shipyards, which controlled for time-invariant unmeasured characteristics of the shipyards. A consistent challenge
in panel data studies is missing or incomplete data. A researcher is always concerned with the possibility of omitted variable bias – if a missing variable is correlated with a variable that is included then it is possible that an identified effect could be spurious. Interestingly, Thompson (2007) collected additional data from the National Archives on Liberty Ships and found that when other variables were controlled for (e.g., product mix), the effect of forgetting became much smaller or possibly even zero. Researchers must always seek as detailed data as possible to make sure that models are properly specified – or accept that some questions may not be answered with a given dataset.

Other examples of reduced form analysis of panel data include behavioral drivers of productivity (Bartel, Ichniowski, & Shaw, 2004; Huckman & Staats, 2011; KC & Terwiesch, 2009; Schultz, Juran, & Boudreau, 1999). For example, using project- and individual-level data for 2004-2006 from the Indian software services firm Wipro, Huckman and Staats (2011) studied how team diversity (in terms of customer experience) affected team performance, controlling for project and team characteristics, and found that intrapersonal team diversity (i.e., range of expertise of each individual in the team) helped in handling the difficulties associated with changing tasks, while interpersonal team diversity (i.e., variety in experience across the team) had the opposite effect, interacting with changing tasks to lower performance.

Similarly, this methodology can be used to study individual decision-making (Kim, Olivares, & Weintraub, 2014; van Donselaar, Gaur, van Woensel, Broekmeulen, & Fransoo, 2010). For example, van Donselaar et al. (2010) explored the ordering behavior of retail store managers of a large supermarket chain in Europe. Using data for 19,417 item-store combinations from five stores over the year 2002, the authors documented the drivers of managers’ deviations
from the automated inventory replenishment system, exploiting the panel nature of the data by including store, subgroup, and store—subgroup interaction fixed effects in the regressions.

An important issue to consider when conducting studies using panel data is the data generation process. In other words, how did the observations and the outcome of the observed data come into being. With non-experimental field data, there are often endogeneity concerns arising from the fact that managers and workers make decisions that might be driving the observed results. For example, imagine a researcher had a dataset on 100 retail stores covering their monthly profit and labor hours (perhaps including other control variables). She was interested in understanding whether using more labor hours would permit greater monthly profit and so she ran a regression of profit on labor hours. Would a positive and statistically significant coefficient on labor hours mean that managers should all rush out to put more labor in their stores? Unfortunately, no. Reverse causality is quite plausible in this case as more profitable stores could have more discretionary money for labor and thus the observed relationship would not answer the question at hand. So, what is a researcher to do?

One common technique to address this concern is the instrumental variables (IV) approach (Kesavan, Staats, & Gilland, 2014; Olivares & Cachon, 2009; Randall, Netessine, & Rudi, 2006; Tan & Netessine, 2014). For example, Olivares and Cachon (2009) found that competition increased inventory holdings in the auto industry using data for 235 GM dealerships across the United States over a six-month period (08/15/2006-02/15/2007). In this case, the endogeneity concern arose from the possibility that GM market entry decisions were based on unobserved market characteristics that could themselves drive inventory decisions in the dealerships located in those markets. To control for the endogeneity of market entry decisions, the authors used total population in the (Census) urban area, fringe (peripheral) population,
population density, other measures related to county population growth, and past population and density variables to instrument for competition.

4.2.2. Difference-in-differences

A more general analytical framework for examining policy interventions is known as difference-in-differences (DD). Suppose you have data for a set of units over time and that some of these units are affected by a particular policy change, intervention or event (the treatment group) while the rest are not (the control group). Then, DD estimation can be used to estimate the effect of the intervention on a given outcome. This approach consists of estimating the regression of the outcome variable on an indicator for the treatment group, an indicator for the post-treatment time periods, and the interaction of these two indicators (i.e., an indicator for the treatment group post-treatment) using Ordinary Least Squares (OLS). The key coefficient is the one on the interaction term, which represents the difference between the time changes in the means for the treatment and control group, and provides an unbiased estimate as long as the intervention is not systematically related to other factors that affect the outcome variable. Group-specific and/or time-specific effects are not a problem.

Despite the extensive advantages, this methodology does not come without challenges. Bertrand, Duflo, and Mullainathan (2004) highlight one of the most significant—in the presence of serial correlation, conventional DD standard errors are inconsistent, and significance levels are overestimated. They explored solutions to this problem resulting from serial correlation noting that when the number of groups (in the panel data) is sufficiently large, block bootstrap or clustered standard errors correct for this problem. When the number of groups is relatively small, those corrections won’t work well, and an alternative solution is to collapse the data by taking
the averages for the pre- and post-intervention periods and adjust the t-statistics for the small number of observations in the regression.

DD design is not new, and it is increasingly gaining popularity in operations management (e.g., Gallino & Moreno, 2014; Levine & Toffel, 2010; Pierce, Snow, & McAfee, 2014). For example, Caro and Gallien (2010) use this approach to evaluate the impact of a proposed inventory management policy tested in collaboration with Spanish retailer Zara. Specializing in fast-fashion, Zara only displays articles for which there is inventory for key sizes and colors in the store; when this is not the case, all units of the article are moved to the backroom, which hence can not be sold. This creates dependencies across articles that should be considered when allocating inventory across stores in the network. The authors propose a process for inventory distribution based on an optimization model that considers these dependencies across different sizes and colors of each article, and test this process by running a pilot with Zara. The results proved the process valuable and Zara decided to expand it broadly.

When using a DD approach, researchers should be mindful the “parallel trends” condition to ensure the validity of the approach. The treatment and control groups should have parallel slopes prior to the treatment – in other words, they are behaving similarly and so the control group provides a valid counterfactual for what would have happened to the treatment group if it had not received the treatment.

4.2.3. Program or Policy Evaluations

Often researchers are presented with a change and are interested then in examining what happened after the change compared to what might have been expected to happen without the change. Examples could include the deployment of a new human resources policy, the announcement of negative news by a governmental organization (e.g., the F.D.A), or
announcement of a merger. This type of shock can be evaluated in several different ways. One approach, commonly referred to as an event study, measures the effect of a specific event on the value of a firm using data from the stock market (MacKinlay, 1997). The event is given by a shock, which might be endogenous or exogenous. The main advantage of this approach is that it provides an objective, easy-to-obtain measure of performance, as the security prices of the firms affected by the event are expected to reflect the economic value immediately. The impact on performance is given by the abnormal return of the security, computed as the subsequent actual return minus the normal return, where the normal return is the return that was expected without conditioning on the event occurring (MacKinlay, 1997). The procedure also involves defining the event, the event window or time period of interest, the inclusion criteria to select firms into the sample, and the estimation window (time period studied) (see MacKinlay, 1997 for details). This type of methodology is frequently used in Finance and Accounting, and has also been applied to Operations Management (e.g., Girotra, Terwiesch, & Ulrich, 2007; Hendricks & Singhal, 1996, 1997, 2001, 2005), studying questions such as the role of portfolio-level projects interactions on the effect of project failure on the stock market valuation of firms in the pharmaceutical industry (Girotra et al., 2007).

In a series of papers, Hendricks and Singhal applied this methodology to quantify the impact of different operational practices (e.g., supply chain disruptions and Total Quality Management (TQM) programs). With respect to disruptions, they identified supply chain disruptions (glitches that resulted in production or shipment delays) publicly announced during 1989–2000 by searching for key words in full text articles in Wall Street Journal (WSJ) and the Dow Jones News Service (DJNS). Using respective samples of 519 (Hendricks & Singhal, 2003) and 827 (Hendricks & Singhal, 2005) announcements made during this period, they concluded
that the mean abnormal return was -10% over the two-day period that spanned the day of the announcement and the day before the announcement, while the long-term stock price effect corresponded to -40.7% mean abnormal return over the three years from one year before through two years after the disruption announcement date (Hendricks & Singhal, 2005). With respect to TQM programs, the authors found that the winning of a quality award (used as a proxy for effective TQM implementation) tended to improve the market value of firms—with a stock market reaction on the day of the public announcement of about 0.59%-0.67% mean abnormal returns (Hendricks & Singhal, 1996) and positive long-run effects (Hendricks & Singhal, 2001)—, as well as to improve the firm operating income (Hendricks & Singhal, 1997).

4.2.4. Regression discontinuity

When a threshold for a given observable determines assignment to treatment, a regression discontinuity (RD) design can be used to estimate causal effects. This approach consists on focusing on the neighborhood of the threshold (thus, estimating “local” average treatment effects) where treatment is as good as randomized (Lee, 2008). RD was introduced by Thistlethwaite and Campbell (1960), who looked at students who nearly-won a national scholarship competition. These awards were assigned according to a rule based on test scores, which were observed. Students with test scores above a given cutoff won the award and those with scores below such cutoff did not win. Students with scores around this threshold were quite similar in baseline characteristics, and hence differences in future academic outcomes could be reasonably expected to be due to the award treatment.

RD methods are a way to obtain credible causal inferences from non-experimental data as long as individuals are not able to precisely manipulate the variable that determines treatment assignment and baseline characteristics have the same distribution on both sides of the threshold,
which can be tested (Lee & Lemieux, 2010). As a result, RD designs can be used to study a variety of questions. For example, suppose one wants to study the effects of losing a competition on the outcome of future, independent competitions. Simply comparing those who lost with those who won would not work, since those who lost would tend to be worse regardless of the impact of losing the previous competition. However, restricting the sample to those who were borderline, those who lost (but almost won) and those who won (but almost lost) are much more similar, and so one can assume that the differences in subsequent performance are due to the “losing” treatment. That is what Berger and Pope (2011) did. Using data from professional (National Basketball Association, NBA) and collegiate (National Collegiate Athletic Association, NCAA) basketball, they showed that teams behind by a point at halftime were more likely to win the game. These teams scored more points immediately after the break, consistent with an increase in effort. To confirm effort as a mechanism, the authors turned to the lab and showed that feedback of being slightly behind competitors increased effort, especially among individuals with higher levels of self-efficacy, i.e., those who believed they could achieve their goals.

4.2.5. Structural Estimation

Structural analysis combines the formalized approach of analytical models with the field-based approach of empirical work. A researcher starts by building an analytical model – for example, by treating a customer’s decision to wait for service or abandon a queue as an optimal stopping problem, the researcher can formally model the different elements of the decision (e.g., Aksin et al. 2013). Researchers can then estimate specific model parameters using field data. The structural approach has the advantage of enabling the generalization of results as well as conducting counter-factual analyses. For example, structural models can be used to predict
responses to changes not-yet-observed using data from different changes in the past (Nevo & Whinston, 2010). As a result, structural estimation has answered important operations management questions that could have not been answered with alternative methods (e.g., Akşin, Ata, Emadi, & Su, 2013; Kim et al., 2014; Li, Granados, & Netessine, 2014; Lu, Musalem, Olivares, & Schilkrut, 2013; Olivares, Terwiesch, & Cassorla, 2008; Terwiesch, Savin, & Hann, 2005). Furthermore, structural models can incorporate other techniques such as instrumental variables. For example, Li et al. (2014) study strategic behavior of consumers using data from air travel, and how airlines should incorporate such behavior in their pricing strategies. A challenge is that price may be endogenous, even after controlling for market and date characteristics. To solve this problem, previous prices and demands are used as instrumental variables.

One area in which this methodology has been extensively employed is menu choice. Among these studies, Musalem, Olivares, Bradlow, Terwiesch, and Corsten (2010) present a structural model of demand to estimate the impact of out-of-stocks using product-store-level data over time on sales and partial information about product availability on the shelf. It is hard to know the effects of out-of-stocks, as the costs depend on customer choice. A customer who encounters an out-of-stock can choose to wait until it is back in stock, decide not to buy it (lost sale), or choose to buy an alternative product (substitution). The proposed procedure is tested and found to be accurate through a numerical experiment using simulated data. Using store-day data on sales, availability, prices, and promotions of 24 stock keeping units (SKUs) of shampoo from six Spanish supermarkets during 15 days, the authors applied their methodology to quantify the financial implications of out-of-stocks in terms of substitution and lost sales. In addition, the structural model can be used to estimate the effectiveness of policies aimed to mitigate the effects of stock-outs; for example, the authors illustrated the evaluation of implementing a
temporary price-reduction of other products (reducing lost sales and increasing substitution but at the expense of lower margin for the discounted product and cannibalization of other products), providing a tool to aid store managers’ decisions.

4.3. Field Experiments (creating the data)

The empirical methods reviewed in the prior section are the tools for analyzing data. These tools are valuable to use when researchers are presented with data on the past (whether collected by the researchers themselves, a company, a third party, etc.) and wish to understand the patterns hidden within. However, although archival data analysis is a valuable tool in the researcher’s toolkit, there are many times when data to answer a given question does not exist. When confronted with this problem, what should a researcher do? Find a way to create the data. One possibility is to run a lab experiment. In this section, we discuss the creation of the data through field experiments. Experiments outside the lab can be classified into natural or framed field experiments depending on whether they happen naturally without the researcher intervention and subjects’ awareness (Harrison & List, 2004). In either case, subjects are familiar with the context studied (e.g., they are doctors in medical decision-making studies). Whether a researcher identifies a natural experiment or decides to run a field experiment, there are a number of areas related to experimental design, partners, and ethics that deserve significant attention.

4.3.1. Experimental design

Of particular importance for field experiments is the consideration of the theoretical motivation (Card, DellaVigna, & Malmendier, 2011; List, 2011). Attracted by the method and its ability to identify externally valid, causal links, researchers are often tempted to run experiments

---

3 The data can either literally not be available or alternatively the existing data may not permit causal identification.
without reflecting on why they are doing them. Experimental treatments should be based on theoretical grounds. And before jumping into the complexity of the field, one should ask, what can the field teach you that the lab cannot? (Al-Ubaydli & List, 2015). The opportunity to run a field experiment, although intellectually exciting, may not be the proper choice. For example, if the field experiment fails to provide new theoretical or practical insight, then it is not clear what would be accomplished by running it. In addition, field experiments typically involve higher cost in terms of both time and money than the laboratory and so researchers should make sure that they are subjecting the idea to a sufficiently high bar, prior to moving the project forward.

Once a researcher believes that a field experiment could answer a rigorous and relevant question, one’s attention turns to the design of the experiment. While the principles of experimental design of field experiments are generally similar to those of lab experiments (see Chapters 1 and 2), there are four major, additional, practical issues to consider.

First, one needs to identify the treatment. This is the intervention the researcher cares about, to be executed, and that will affect only some individuals or units (the treated group), so that the researcher can study how the treated group differs from the control group, as in a standard lab experiment. One possibility is to implement a particular policy (and carry out a policy evaluation). Most experiments in OM fall under this category; consider, for example, the studies testing inventory policies with Zara (Caro & Gallien, 2010; Gallien, Mersereau, Garro, Mora, & Vidal, 2015) or pricing with Zara (Caro & Gallien, 2012) and Rue La La (Ferreira, Lee, & Simchi-Levi, 2015). An alternative is to directly assign as treatment what is believed to be the mediator of the effect of the policy on the outcome (and carry out a mechanism experiment) to identify the causal mechanism through which a policy affects the outcome (Ludwig, Kling, & Mullainathan, 2011). Mechanism experiments have been rarely used in OM, as of the writing of
this; instead, OM research has focused on testing policies strongly founded on prior knowledge in order to discover factors from the field that were not thought to play a role or even known to exist by researchers, or to compare relative effects of multiple mechanisms believed to be associated with a policy (rather than testing the existence of a particular mechanism). We expect mechanism experiments to become more common over time, as the scope of OM evolves and field experiments grow more common. Overall, the appropriateness of policy versus mechanism experiments depends on the research question, the related body of knowledge, and the costs of conducting either type of experiments, and since they inform each other, it will be desirable to conduct both types over time.

When thinking about the experimental design, a researcher should recognize that it is rare to get the opportunity to repeat a field experiment. Typically, organizations are not willing to rerun an experiment when the researcher finds a design flaw. To make matters worse, it may be difficult to impossible to pilot a field experiment, at least in the field at the partner organization. Therefore, researchers should recognize that the up-front time investment required to setup a field experiment is likely significantly more than in the lab. Researchers who cannot pilot in the field should creatively consider how to get similar feedback, for example, by going to their own lab to test questionnaires or by asking for a conference room pilot where the researchers gather with their sponsors to walk through each step and confirm validity.

Second, the goal of the experiment is to create an appropriate counterfactual, and to do so, the researcher uses randomization to assign a condition to what becomes the treatment group and the control group (Harrison & List, 2004). There are several important differences that need consideration when conducting a field experiment, however. In almost any field experiment, a researcher must worry about contagion effects - the sharing of information about the treatment
across conditions. Unlike the laboratory where participants are more easily kept separated, in the field there is substantial risk that participants may talk. For example, imagine a researcher ran an experiment inside an organization where the treatment group received a lump sum cash payment and the control group did not. If individuals in the treatment and control groups knew each other, then this information would likely be shared. This could create serious organizational difficulties for one’s research sponsor. In addition, it could bias the results if, for example, the control group was demotivated by not receiving a bonus and thus the observed effect was no longer solely a story of providing incentives to one group.

To address contagion, researchers must randomize at a level to ideally prevent, or at least minimize, its impact. This may mean randomizing across work groups or facilities, for example. More generally, when a covariate could predict the potential outcome, a randomized block design should be used. In such case, subjects are divided into “blocks”, and treatment is randomly assigned within each block. For example, if researchers studying attrition at a firm knew that men and women departed at different rates and that there were few men in the firm, then a block design would address this concern.

Finally, researchers should be cautious about “randomization bias”: the bias that occurs when randomization itself leads to samples that are not representative of the population (Levitt & List, 2009). For example, a company might simply randomize by alphabetical letter or employee identification number (e.g., employees above 10,000 get the treatment while those below 10,000 do not). This can result in a non-random sample since, in the example of employee identification number, most companies do not assign such numbers randomly, but rather assign them in sequential order as individuals join the firm. Thus, in the example above a researcher would have a wonderful field experiment on how newer employees who get a treatment compare to older
employees who do not. As a general rule, researchers should ask if they can conduct randomization, rather than allowing their organizational partner to conduct it. By taking on this task as a researcher, it prevents finding out about problems later on that could invalidate the prior work.

Third, one should establish the measures to be used, identifying clear outcomes for dependent variables, as well as measures for controls prior to beginning the field experiment. Clearly articulating measures up front addresses two purposes. It first makes sure that a researcher will be able to study the phenomenon of interest. In working with field partners on various projects we have more than once been told that only if we had asked for something sooner the company could have provided it (e.g., by saving data from its computer logs, rather than purging the data, as was their practice). Thus, if the right data is not available, then perhaps the experiment will not be launched, or the company may be able to provide more data than if they were not asked. Relatedly, by specifying the necessary data up front it is possible to identify gaps that new data collection could address. For example, archival data sources often provide rich information on operational performance. They typically provide less on actual behavior, which may explain a mechanism. Thus, if a gap is identified then the researcher may realize that a survey over time could provide meaningful value in a study. Alternatively, a company may be willing to add questions to their own internal data collection efforts (e.g., annual reviews or employee satisfaction surveys).

Fourth, while it might be possible to repeat an experiment to increase the sample size in the lab, this is typically not the case in the field. It is typically impractical for an organizational sponsor to go and collect more data after the initial intervention has been run. As a result, before running any experiment, one must carefully assess the minimum sample size required through
power calculations, including such considerations as clustering (List, 2011). An additional consideration in sample size in the field is that researchers will need to estimate the rate at which participants may drop from the sample. For example, if a field experiment takes place over many months, then employees may leave. If it takes place over a smaller time frame, then absenteeism or business travel could affect the sample size. In addition, if surveys are collected, then the lack of 100% response should be anticipated and addressed in initial sizing. Participant attrition is often not a concern in a laboratory environment, or in archival data analysis, and, speaking from experience, discovering it after the fact, is a painful experience for a researcher.

Authors are not the only parties that must think carefully about experimental design for field experiments. Readers (and reviewers) should understand the challenges that even the most successful field experiments faced. Researchers running experiments must be prepared to jump over obstacles and respond to complications. Some challenges result in design trade-offs to be made (often before executing the experiment), while others are natural complications from “doing business” in the field. To illustrate, let’s consider two well-known experiments. Caro and Gallien (2010)’s work with Zara on inventory distribution highlights some of the trade-offs of field experiments. First, while 15 articles were initially selected for the test, only 10 of the 15 had data for more than three weeks, and the authors were forced to limit the analysis to those 10. Fortunately, these 10 articles remained a representative proportion of basic vs. fashion, yet the already small sample size was reduced. Second, the forecasts used during the pilot were not saved, and thus could not be used during the data analysis phase. Finally, research that is relevant for practice often must make sacrifices in theoretical contributions; for example, Caro and Gallien (2010) write, “The forecasting model considered takes as input from store managers their shipment requests, which is the very input they provide in the legacy process. This approach was
believed to constitute the easiest implementation path, because it does not require any changes in the communication infrastructure with the stores or the store managers’ incentives” (page 258). While the resulting model “sacrifices analytical tractability for realism,” their research has a positive influence on scholars and practitioners.

Similarly, Bloom et al. (2013)’s experiment to evaluate the impact of management practices on Indian textile firms’ performance also illustrates the challenges of experimenting in the field. First, consulting is expensive. Given their limited funding, the researchers had to trade-off spending more per plant (higher quality consulting, to encourage participation and retention, and have a higher impact necessary to study large firms) versus spending less per plant in more plants (higher sample size). Because they wanted to study large firms, they ended up with a small sample size of only 14 treatment plants (in 11 treatment firms) and 6 control plants (in 6 control firms), for a total of 20 plants (in 17 firms). Second, having consultants involved in the treatment delivery and the data collection in an experiment to measure the value of their work (i.e., management consulting through management practices) naturally creates a conflict of interest. To ameliorate this risk, the authors had graduate students overseeing data collection and created opportunities for the plant directors to see the data received so that they could raise questions if they found it inaccurate. Third, given the size of the experiment, there was a risk of Hawthorne effects, as treated plants had greater interactions with the consultants. Overall, the experimenters overcame many challenges, but again were able to answer a difficult question through an admirable experiment.

The takeaway for readers, and reviewers, in particular, is not to immediately reject a field experiment paper for its flaws, nor to accept one uncritically. Rather, as with any research methodology it is important to evaluate the challenges and understand whether the paper makes a
contribution in its chosen field. In our experience in the role of author, reviewer, or editor, as well as in conversations with others, problems arise when individuals uncritically move the standards of one research approach to another. Rejecting all lab papers because they do not occur in the field is foolish in the same way that expecting identical control of the environment in a field experiment would be. Similar analogies can be drawn across analytical and archival, empirical work. We hope that this chapter aids authors in the proper conduct of field experiments, but also that it offers value for editorial teams in evaluating such papers responsibly.

4.3.2. Field Sites and Organizational Partners

Perhaps the biggest key to the success of the execution of any field experiment lies with the organizational partner—the field site where the experiment takes place. Like any collaboration, the potential of the partnership depends on the match between the researcher’s goals and those of the partner organizations, however there are several practical issues to take into account. In terms of the types of organizations that make good partners, start-ups tend to be more flexible, faster, and welcoming to finding answers to questions they might care about but lack resources to explore themselves, while, on the other side, big firms bring the benefits of size and infrastructure, however often also have legal departments to match.

Unlike the lab, organizational partners will often ask for non-disclosure agreements (NDAs) to protect their own interests. These agreements are not uncommon, and therefore are not something a researcher must be afraid of. However, one should seek proper advice. First, understand your university’s policy on NDAs. Can a researcher sign an NDA on her own or must it be done through the school’s legal department? If it can be done on your own then make sure to still seek advice from a knowledgeable individual (an experienced colleague or possibly legal
counsel). Make sure that the NDA gives you the right to publish the results of the study, with a proper review period to protect the partner and make sure that no confidential information is disclosed. An often asked question is whether to identify the company in the research. Different researchers seem to have different preferences on this front. Our standard tact has been to always start with anonymity, as is often required by a NDA, but point out to the company that we will give them an opportunity to self-reveal at the end of the project, if they wish. In almost all cases, the company has chosen to reveal its own name as, once an experiment has been successfully run, then the company recognizes that there are intellectual capital and perhaps even competitive or internal benefits to disclosing their name.

A successful field experiment partnership typically requires both senior and front-line engagement. The senior engagement is important because the experiment will require organizational approval—typically the domain of top management. However, front-line engagement is necessary for execution purposes, since these are the individuals that will help roll out the intervention and provide the data. Each level of engagement typically requires time to cultivate trust and a common understanding. Researchers who are used to handing materials off to a research assistant or a lab manager to just get the study done will need to take a different tact. This likely involves in-person visits and time spent relationship building. For example, even if senior management says to move forward with an idea, a front-line employee can effectively kill the project by only meeting the letter of a request. We have experience with one partner who shared that this was exactly what he did when an executive in another area forced him to take part in a project he wasn’t interested in and the researcher involved simply emailed instructions as if dealing with a research assistant. As a result, we believe that field experiments are particularly appropriate for management researchers who enjoy interacting with practitioners—
engaging with partners has been one of the most exciting and fulfilling parts of our research, and we have learned more from working with our collaborators than we otherwise would have.

Having honest and clear communication with the organization will ensure no (or at least fewer) surprises. Researchers should explain what they can and cannot do, and seek to address the partners’ concerns. For instance, a frequent concern is the notion of fairness when treatment directly benefits or hurts subjects; in such case, one could offer to switch all subjects between control and treatment groups at exogenously predetermined times (Bandiera, Barankay, & Rasul, 2011). Alternatively, if a treatment is seen as beneficial to everyone (e.g., providing a report on an individual’s strengths), then the researcher could commit to providing reports to the control group as well – after the study is completed. On the other side, one needs to communicate the need for randomization, and “make clear that you do not have all the answers” (List, 2011).

When approaching partners, some are convinced that they have all of the answers. These types of companies rarely make good partners, since they do not see the need to work together. However, if one can identify companies that are curious on the proposed topic, or who have a real pain point, and then the researcher can explain clearly why your particular experiment could bring benefits to this partner, then a relationship can be built and an experiment may be able to be implemented. Benefits can vary tremendously, from increasing profit to improving employee satisfaction to improving reputation to applying cutting-edge practices and satisfying their own curiosity. For example, Caro and Gallien (2010) proposed a project to Zara that was beneficial to this company because, growing at an incredible rapid pace, the executives realized they were not able to continue operating with the same manual systems they were using. This timing probably facilitated the successful collaborations that facilitated scalable processes, and resulted in their cutting-edge research.
Finally, a practical tip to keep in mind is to use your own resources rather than those from the company; this is important both to preserve your objectivity as a researcher as well as to keep control over the experiment. In particular, we believe that it is helpful to have your own implementation team, if possible, rather than relying on company employees, who are busy with their jobs and likely do not know about how to run experiments without contaminating results. Overall, everything you can do to reduce the burden on your partner can help not only the partner, but also the researcher in successfully completing the project.

4.3.3. Ethics and Human Subject Protocol

Lastly, we would like to discuss the first step of any field project. One needs to remember that subjects are people who might suffer during or after the study as a result of their direct or indirect participation. The potential risk depends on the specific study, but a researcher should always get supervision and approval from the Institutional Review Board (IRB) at her organization (Levitt & List, 2009; List, 2011). Similarly to lab experiments (see Chapter 1), field experiments should obtain official authorization from the IRB before being conducted. Outside the United States, the researcher should search for the appropriate guidelines. One factor to consider is the ethics of the inclusion (exclusion) criteria, that is, the rules used to decide which subjects to include in (exclude from) the study; the decision should be fair and based on science, avoiding discrimination. When possible, one should obtain informed consent. However, in some cases, informing the subject would invalidate the research results; in such cases, a researcher should look for ways to lessen the risks and receive guidance from the IRB to identify situations in which it could be argued that informed consent is not needed (Levitt & List, 2009). Following good IRB practices is not only important for the protection of participants, but also for the protection of the researcher. Given that outside parties are involved, sometimes without their
consent, our experience tells us that the risk that concerns get raised are higher in a field experiment than in the lab. As such, the IRB should be seen as a partner. If any concerns are raised, then not only does the IRB provide the researcher with validation that her ideas were properly vetted, but also the IRB has resources to help respond to and address any outstanding concerns.

4.4. Conclusion - the way forward

Field data can bring richness to any study seeking to explain real world phenomena. While field and lab experiments each have their own advantages and disadvantages, together they can be complements (Harrison & List, 2004). Given the above review, we would highlight five possibilities for future work to consider.

First, we would encourage scholars to seek out opportunities to conduct field experiments. As outlined above, field experiments take a significant amount of work. In particular, getting a company to agree to participate is often the hardest part. At the same time, the learning and the impact on practice are both substantial, if one is able to successfully run a field experiment. Thus, allocating a portion of one’s time towards field experiments may have substantial impact.

Second, researchers should be aware of the possibility to exploit natural experiments. Often times organizations may make internal changes that can be used as the identification strategy for an empirical analysis. For example, Marx, Strumsky, and Fleming (2009) exploited an unrelated change in the state of Michigan’s employee non-compete provisions in order to examine how loosening non-compete enforcement affected innovative outcomes. As mentioned above, Staats, KC & Gino (2016) similarly, used an announcement of negative news by the FDA
to examine how individuals update their beliefs. Researchers should continuously scan for such changes either within firms or institutions that govern firms.

Third, reviewers should recognize that field experiments should be analyzed in different ways than lab experiments. As discussed throughout this chapter, different methods have different strengths. A reviewer would not evaluate a qualitative paper in the same way that she reviews an event study or a survey-based study. The same is true for lab and field studies. Flaws should be identified and addressed, but sacrifices in internal validity are often necessary for external validity. It is incumbent upon the author to identify strengths and weaknesses, but hopefully this chapter will also help reviewers to hold field studies to appropriate standards.

Fourth, we would strongly encourage authors to utilize both field and lab studies together. Not every paper requires both, but when there are weaknesses of note (as discussed in the last paragraph) then a lab study may prove to be an excellent complement to a field study. The same is true in the opposite way. Providing a field study to show an existence proof for interesting work from the lab is an impactful way to conduct work that is both rigorous and relevant.

Fifth and finally, we would suggest that turning to the field may prove valuable in helping the behavioral operations field tackle questions beyond decision making. In particular, seminal works in the field by such individuals as Frederick Taylor or Wickham Skinner (Hayes, 2002) were, at their core, studies that involved questions of both operations and human resource management or what eventually became the field of organizational behavior. The behavioral effects on operational performance are as important as they have ever been for understanding business outcomes and yet this intersection remains relatively understudied. By turning to the
field, it is possible to expand the scope of behavioral operations management such that it continues to build theory that is both rigorous and relevant.

Altogether we hope this chapter will aid behavioral operations management scholars in utilizing a variety of empirical approaches to conduct research that utilizes field data to advance scholarship and practice.
References


