Chapter 4: Reprinted by permission, Maria R. Ibanez, Jonathan R. Clark, Robert S. Huckman, Bradley R. Staats, “Discretionary Task Ordering: Queue Management in Radiological Services”, Management Science, Published online in Articles in Advance 30 Aug 2017. Copyright 2017, the Institute for Operations Research and the Management Sciences, 5521 Research Park Drive, Suite 200, Catonsville, Maryland 21228 USA.

Abstract

Behavioral drivers of process deviations and

the effects on productivity and quality:

Evidence from the field

By

Maria R. Ibanez

Doctor of Business Administration

in the subject of

Technology and Operations Management

Harvard Business School

This dissertation provides empirical evidence from high-stakes field settings of how productivity and quality are affected by workers' deviations from prescribed processes. The first essay of the dissertation explores the role of experimentation in field settings to investigate the drivers of performance and how to implement this methodology to answer relevant operational questions rigorously. This dissertation then uses field data from proprietary sources to investigate the behavioral drivers of process variation and their effects on productivity and quality. In particular, the next two essays of this dissertation consider the effects of (1) how workers' decisions *are influenced by* task schedules and (2) how workers' decisions *exert influence on* task schedules.

Many tasks are decisions, which are thus subject to human decision errors. How does scheduling affect how humans—in contrast to machines—perform these tasks? To explore this question, the second essay focuses on one critical task: quality evaluations. The accuracy of quality evaluations is critical to their being a useful input to key managerial decisions, to penalize compliance failures, and to motivate quality improvements. Yet, task-scheduling factors that are related to the workers' work structure (but unrelated to the task itself) could shape workers' predisposition toward the task and subsequent performance. We explore how inspection scheduling can affect inspection quality by influencing bias. Analyzing thousands of food safety inspections, we find that inspection results are affected by when the inspection occurs within an inspector's daily schedule and by inspectors' experience at their prior inspection of a different establishment. For example, the more compliance deterioration found in an inspector's prior inspected establishment, the more violations cited in the inspector's next inspection (of a different establishment). Consistent with negativity bias, this effect is asymmetric, applying when compliance at the inspector's prior establishment deteriorates but not when it improves. Overall, by identifying factors that bias inspections, our work contributes to the literature on monitoring, quality improvement, and scheduling. Our work also suggests a cost-effective lever: exploiting the behavioral effects of the organization of work.

Task scheduling is not always a managerial decision. Those who execute tasks often have discretion over the order in which to perform them. How do these choices affect productivity and quality? The third essay of this dissertation focuses on the drivers and consequences of exercising discretion to "deviate" from a prescribed task sequence. Analyzing 2.4 million decisions, we find that radiologists prioritize similar tasks (grouping tasks into batches) and those tasks they expect to complete faster (shortest expected processing time). Exploiting random assignment of tasks to doctors' queues, instrumental variable estimates reveal that both of these types of deviations erode productivity. Actively grouping similar tasks reduces productivity, in stark contrast to productivity gains from exogenous grouping, indicating deviation costs outweigh benefits from repetition. We also find learning-by-doing in exercising discretion, with doctors deviating more often and more productively over time. Our results highlight the tradeoffs between the time required to exercise discretion and the potential gains from doing so, which has implications for managers deciding task sequence assignments and system design.

Together, these essays generate new scholarly insights regarding the connections between operational factors, decision-making, and performance by analyzing data from high-stakes field settings. In doing so, this research seeks to contribute to theory while also improving management practice.

Table of Contents

List of Tables

List of Figures

Dedicated to my parents.

Acknowledgements

I gratefully acknowledge the support of the members of my dissertation committee. This research would not have been possible without the selflessness of my advisor Professor Ananth Raman, who has been a continuous source of inspiration and has changed my thinking profoundly. Professor Mike Toffel has also been an exceptional mentor, generously helping me through the program. It has been delightful to conduct research together and to learn from him and his dedication. Professor Rob Huckman supported my research ideas and also gave me the opportunity to be a Teaching Fellow and start learning how to teach MBA students, which brought joy to my days at Harvard. Professor Brad Staats and I met during my doctoral program when he was visiting Harvard, and I am very grateful he saw the potential in my research ideas; while we did not get much face time, his support was instrumental.

Other members of the Harvard community, the University of Chicago and Marquette University were also very helpful at many points in time. I was fortunate to have the support from many fellow students at Harvard and outside. I would also like to thank my family and friends, particularly my mother, for her support along the way.

I owe many thanks to the people at my research partner organizations as well as all the radiologists and inspectors across the U.S. who allowed me to shadow them and learn about their work, which gave meaning (and hopefully relevance) to my work. It was my privilege to learn from so many dedicated people.

This thesis has sections that are common to the following papers:

(i) Ibanez, Maria R., and Bradley R. Staats. “Behavioral Empirics and Field Experiments,” Handbook of Behavioral Operations, edited by Karen Donohue, Elena Katok, and Stephen Leider, Wiley (forthcoming).

(ii) Ibanez, Maria R., and Michael W. Toffel. “How Scheduling Can Bias Quality Assessment: Evidence from Food Safety Inspections.” Working Paper.

(iii) Ibanez, Maria R., Jonathan R. Clark, Robert S. Huckman, and Bradley R. Staats. “Discretionary Task Ordering: Queue Management in Radiological Services,” Management Science (forthcoming).

Chapter 1

Introduction: Using data from the field to identify how productivity and quality are affected by workers’ deviations from prescribed processes

With increased access to data and research insights, companies are increasingly interested in making data-driven decisions. As a result, managers seek to apply operations' insights to enhance workers' discretion. This dissertation explores how to use data to identify causal relationships between operational management practices and operational performance, which can in turn inform future managerial policies. In particular, this research investigates how to improve performance by designing data-driven systems that lead individuals to make better decisions. Combining operations management with economic theory and the psychology of decision-making, these studies analyze large-scale field data to identify causal relationships that generate new scholarly insights regarding the connections between operational factors, decision-making, and performance. In doing so, this research seeks to contribute to theory while also improving management practice.

The first essay of the dissertation, "**Field Experiments in Operations Management**" (with B. Staats), explores the role of experimentation in field settings to investigate the drivers of performance and to answer operations management questions more broadly. We build on field experiments conducted in economics and psychology, and we propose how to use these methodologies to answer relevant operational questions rigorously. Reviewing the emerging literature in operations management that involves field experiments, we discuss best practices and opportunities for future investigations.

The rest of this dissertation uses field data from proprietary sources to investigate the behavioral drivers of process variation and the effects on productivity and quality. Focusing on task scheduling, these papers examine the operational implications of workers' decisions regarding the allocation and/or completion of tasks. By considering these implementation issues, this research extends traditional scheduling research that identifies the best schedules for managers to implement. In particular, the next two essays of this dissertation consider how productivity and quality are affected by (1) how workers' decisions *are influenced by* task schedules and (2) how workers' decisions *exert influence on* task schedules.

Many tasks are decisions, subject to human decision errors. How does scheduling affect how humans—in contrast to machines—perform these tasks? To explore this question, we focus on one critical task: quality evaluations. The accuracy of quality evaluations is critical to their being a useful input to key managerial decisions, to penalize compliance failures, and to motivate quality improvements. Yet, task-scheduling factors that are related to employees' work structure (but unrelated to the task itself) could shape workers' predisposition toward the task and subsequent performance. In the second essay, "**How Scheduling Biases Quality Assessments**" (with M. Toffel), we explore how inspection scheduling can affect inspection quality by influencing bias. Analyzing thousands of food safety inspections, we find that inspection results are affected by the timing of the inspection within an inspector's daily schedule and by inspectors' experience at the prior establishment they inspected. For example, the more compliance deterioration an inspector finds at the last establishment he/she inspected, the more violations cited in the inspector's next inspection (of a different establishment). Consistent with negativity bias, this effect is asymmetric, applying when compliance at the inspector's prior establishment deteriorates but not when it improves. Overall, by identifying factors that bias inspections, our work contributes to the literature on monitoring and quality improvement. Our focus on how scheduling affects inspector stringency introduces the operational lens of scheduling to the literature examining inspector bias, which has otherwise largely focused on experience or other sociological and economic factors. Understanding these biases can enable managers and regulators to make better decisions when using inspection report data, to create more reliable information for managers and consumers, and to provide fairer results (and higher motivations for compliance) for inspected establishments. By examining data from actual decisions with important consequences for public health, we contribute to the nascent literature that is exploring high-stakes decision-making in field settings. With managers across many industries seeking to monitor and improve quality, our research suggests a cost-effective lever: exploiting the behavioral effects of the organization of work.

Task scheduling is not always a managerial decision. Those who execute tasks often have discretion over the order in which to perform them. How do these choices affect productivity and quality? In the third essay of this dissertation, "**Discretionary Task Ordering: Queue Management in Radiological Services**" (with J. Clark, R. Huckman, and B. Staats, forthcoming in *Management Science*), we focus on the drivers and consequences of exercising discretion to "deviate" from a prescribed task sequence. Analyzing 2.4 million decisions, we find that radiologists prioritize similar tasks (grouping tasks into batches) and those tasks they expect to complete faster (shortest expected processing time). Exploiting random assignment of tasks to doctors' queues, instrumental variable estimates reveal that both of these types of deviations reduce productivity. Actively grouping similar tasks reduces productivity, in stark contrast to productivity gains from exogenous grouping, indicating deviation costs outweigh benefits from repetition. We also find learning-by-doing in exercising discretion, with doctors deviating more often and more productively over time. Our results highlight the tradeoffs between the time required to exercise discretion and the potential gains from doing so, which has implications for managers deciding task sequence assignments and system design. Methodologically, we present a novel strategy to identify instrumental variables to measure the effects of discretion in queuing settings and show that considering the time required to *exercise* discretion may reverse prescriptions in data analytics, which is illustrated by the fact that the time required to reorder a queue exceeds the beneficial effects of batching in our setting.

Together, the essays in this dissertation provide empirical evidence from high-stakes field settings of how productivity and quality are affected by workers' deviations from prescribed processes. By collaborating closely with the individuals and organizations in the field settings related to the data analyzed, these studies seek to provide relevant scholarly and managerial insights.

Chapter 2

Field Experiments in Operations Management

Abstract

Field experiments are controlled interventions in the real world that enable researchers to measure the effects of a treatment on a randomly assigned subset of subjects. In this paper, we review the advantages and disadvantages of field experiments and provide some practical prescriptions to attain and evaluate a field experiment’s relevance—in other words, the theoretical implications of understanding the effects of the treatment— and rigor, based on many methodological considerations.

2.1. Introduction

Operations management (OM) research is currently seeking to more often incorporate the role of human behaviors in traditional operational problems. The result is an expansion of the field into new empirical approaches that can accommodate new, increasingly relevant research questions. To study real decisions made by workers in their regular tasks, field studies analyze real world data. Nonetheless, using observational data from the field often presents identification challenges, with causal links frequently difficult to establish.[1] Econometric techniques can often, but not always, address these identification issues. In this paper, we propose that the field experiment is a promising tool to study OM research questions.

To more easily claim a causal interpretation of the relationships between variables, lab experiments create a controlled environment where the researcher ("experimenter") changes the variable of interest (X) for a random set of subjects or units (the treatment group) and not others (the control group), which allows the conclusion that whatever change is observed in outcome (Y) was caused by the change in the variable of interest (X). Though this methodology is ideal for addressing some research questions, the artificial nature of the lab results in several drawbacks, including the possibility that the individuals behave differently than in the real world and that the active decision of the subjects to come to the lab to participate in the experiment increases their perception of being observed and attracts particular types of people. This leads to non-representative behaviors or samples. These and other drawbacks raise doubts about external validity.

[1] For example, if examining the effects of X on Y, the research goal is to claim that a given change in X causes some change in Y. Nonetheless, if simply considering the observed changes in X and Y, there may be concerns that the observed change in Y is causing the change in X (reverse causality) or that the observed change in Y is caused by changes in a third variable rather than by changes in X.

Combining the real-world practicality of field studies with the causal interpretation of lab experiments, field experiments have the potential to answer novel research questions in operations management. They are a way for researchers to design the data generation process so that the results have a causal interpretation. In contrast to those of a lab experiment, subjects of field experiments are always familiar with the context studied, like doctors in a medical decision-making study or shop goers in a study on queuing. Thus, field experiments feature the same internal validity benefits as lab experiments but may also lead to greater external validity. Experiments outside the lab can be classified into *natural* or *framed* field experiments, depending on whether they happen without the researcher's intervention or the subjects' awareness (Harrison and List 2004). In this paper, we describe the methodology and provide examples of applications, focusing on *framed* field experiments, because natural field experiments are analogous to non-experimental archival field studies in terms of execution. Though field experiments have many advantages, they also come with their own challenges, which should be addressed and weighed against the benefits. Field experiments can be novel and relevant (when the treatment generates scholarly insights) and rigorous (when design, execution, and analysis are properly carried out), but they can also be neither. In the next sections, we review the advantages of field studies and field experiments. In the final sections, we discuss the methodological aspects of field experiments—ethics, experimental design, and partners—and conclude with a discussion of research takeaways.

2.2. Why Go to the Field

Rigorous empirical research is necessary to apply the scientific method to OM and, for this research to be relevant, it has to have bearing on or a connection with the subject at hand (Van

Mieghem 2012). A direct way to conduct research that is relevant to practitioners is to collaborate with them to conduct research, experimental or not. Such field research includes field case studies, archival field studies and field experiments. Case studies can provide evidence that a phenomenon exists in practice (i.e., an existence proof) or explore 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. Archival field studies and field experiments can use data from the field to test theory-driven hypotheses.

When and how does going to the field enhance research? Compared with research based on analytical models, simulations, or data from the lab, field research (experimental or not) has five main advantages: (1) the opportunity to establish external validity and identify effect sizes; (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 go beyond individual decision-making. Researchers should go to the field when these advantages can make their research rigorous and relevant. In the following subsections, we describe each of these methodological strengths and provide examples of work that has benefited from them.

2.2.1. External Validity and Identification of Effect Sizes

Because field research is conducted "externally" in the real world, it tends to have greater external validity or generalizability than studies conducted in the lab or using simulated data. External validity can be established through empirical studies that analyze data produced within a company by workers conducting their normal activities; consider, for example, Taylor (1911)'s study of workers' actions to find tools and methods that could be used to improve overall productivity.

Moreover, field work is typically required to determine the size of the effects. 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 even larger or relatively small (Cachon 2012). While the lab identifies effects, the field brings the context to estimate effect sizes.

2.2.2. Overcome Observer Bias

Interactions with the field or bringing subjects to a lab have the undesirable consequence of *Hawthorne effects*, whereby the individuals under investigation change their behaviors as a result of the scrutiny and procedures associated with participation in the study, the reminders that they are being observed provided by the experimental treatment itself, and the desire to please the experiment resulting from "experimenter demand effects" (Levitt and List 2011). Hawthorne effects can be eliminated by analyzing data collected before the researcher's involvement (e.g., via observational field studies or natural experiments).

Compared to lab experiments, field studies mitigate observer bias, since data are from subjects, often unaware of their participation in a study, in their natural environments. In addition, even in those cases in which subjects are notified ahead of time, an observer's presence grows less salient over time as participants remain in their familiar environment. When turning to the field to study human behavior, researchers are increasingly able to rely on technology to remove observer bias. For example, Singh, Teng, and Netessine (2017) collaborated with an online taxi booking platform to study the effects of charity-linked and discount-based promotions. They used text messages (SMS) to implement their treatments and then obtained the booking data from the partner organization.

Hawthorne effects can also be addressed through research design (e.g., having a second control group that receives the same attention and monitoring as the treatment group but does not receive the actual treatment). For example, Singh, Teng, and Netessine (2017) include two control groups where individuals do not receive a promotional code for taxi service, where one control group has SMS but no promotional code and the other control group has no SMS at all.

2.2.3. Context

Field research contributes to our understanding of a phenomenon by incorporating context. When our theories are forced to predict what happens in practice, then the inevitable gaps that exist in a theory are exposed. 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. It is this realistic context that enables identifying important moderators and interrelationships with other variables. Furthermore, only the field can bring the full contextual detail, including incentives and high stakes, to understand complex behaviors. For example, field studies focused on retail have been conducted to evaluate how sales are affected by price (Gaur and Fisher 2005) and to examine external audits of on-shelf inventory positions (Chuang, Oliva, and Liu 2016).

2.2.4. Time-based Effects

Laboratory experiments can last several hours or ask the subjects to return in the future, but have the downsides that they cannot observe the subjects between visits and many subjects won't follow up. In contrast, one advantage of field studies is that it is easier to study longer time periods. For example, Singh, Teng, and Netessine (2017) obtained data from their partner company on all taxi bookings made during 2015, which enabled them to study patterns before and after each intervention. Though they find new demand from promotions, they find little evidence of lasting treatment effects after the promotion period, which suggest that these types of promotions may be undesirable for firms. Thus, in this case, the longer time frame reversed the practical implications, providing a more comprehensive picture of the phenomenon studied.

Scholars have benefited from the long time scales that exist in practice, but not the lab, to study habit formation in gym membership (Charness and Gneezy 2009, Milkman, Minson, and Volpp 2014), energy usage (Allcott and Rogers 2014), or process compliance (Staats et al. 2017). 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. With a time scale of years before the effect is observed, this impact would be challenging to replicate within the lab.

2.2.5. Beyond Individual Decision-making

The fifth advantage of field studies is the flexibility in terms of unit of analysis or subjects. In contrast, lab experiments' subjects are individuals or occasionally teams. This move up in levels is important because, while factories, organizations, and countries are made up of individuals, the macro factors that affect them are not always decomposable into individual-level studies. For instance, field studies found that lower proximity (resulting from new airline routes) of manufacturing plants to headquarters increased 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) increased profits by 60% per year for Sri Lankan microenterprises, implying marginal returns above the market interest rates (de Mel, McKenzie, and Woodruff 2008).

Bloom et al. (2013) recruited multi-plant textile (woven cotton fabric) manufacturers 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 the treatment of 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., regular meetings between sales and operations managers), human resources (e.g., performance-based rewards), and sales and orders (e.g., order-wise production planning). The experiment showed that treatment increased productivity by 17%. Reflecting on their findings, the authors argued that the main reason for the lack of implementation of managerial practices in the past was that managers did not believe the practices would be profitable, suggesting that information constraints explain differences in productivity across firms and countries.

2.3. Why Run Field Experiments

As described in the previous section, field studies have many advantages. In this section, we focus on one particular type of field study: field experiments. When and how does conducting an experiment improve field research? When investigating which factors drive the outcomes of interest, empirical researchers rely on causal inference to identify causal relationships between the variables that go beyond mere correlation. Though correlations among variables can be informative at times (usually combined with qualitative analyses), the goal of the majority of empirical research is to identify how a change in X causes a change in Y, holding everything else constant. Simply measuring how Y changes when X changes does not achieve this goal unless it can be argued that the change in Y was indeed caused by the change in X. Accordingly, researchers must understand their data and the data generating process to identify the associated challenges and take the necessary steps to address them. Depending on the characteristics of the data and the goal of the study, different empirical methods should be used, often in combination. Common methods include difference-in-differences (e.g., Levine and Toffel 2010, Gallino and Moreno 2014, Pierce, Snow, and McAfee 2014, Gallino, Moreno, and Stamatopoulos 2016), regression discontinuity (e.g., Lacetera, Pope, and Sydnor 2012), structural estimation (e.g., Musalem et al. 2010), or instrumental variables (e.g., Ibanez et al. 2017). For a review of common empirical methods and causal inference models, see Ho et al. (2017).

These empirical methods are the tools for analyzing data on past events, which enable the researcher to discover the patterns hidden within the data. This data may come from secondary sources (e.g., Cachon and Olivares 2009) or firms' workflow digital records (e.g., Ibanez et al. 2017), traffic counters (e.g., Perdikaki, Kesavan, and Swaminathan 2012), video cameras (e.g., Lu et al. 2013), RFID devices (e.g., Staats et al. 2017), or email records (e.g., Aral, Brynjolfsson, and Van Alstyne 2012). Sometimes, however, the data available may not permit causal identification even through sophisticated econometrics and other data may not exit or be accessible to the researcher. When data to answer a particular research question is not available, the researcher could *generate* the data—through a lab or field experiment.

The main advantage of controlled experiments is that, when properly designed and executed, they generate data that address some of these data analysis challenges. The experimental approach enables the construction of a control group via randomization, which sharpen our measurement of effects. Through the proper experimental design, experimenters are also often able to obtain more controls and to measure intervening steps to decompose causal effects more effectively, resulting in a better understanding of the phenomenon.

Nonetheless, experiments may have the same challenges as other empirical methods and can introduce entirely new difficulties. As a result, the researcher should always be mindful about causal inference methods, even if using experimental data. For example, Caro and Gallien (2010) collaborated with Spanish fast-fashion retailer Zara to evaluate the impact of a proposed inventory management policy. Zara stores display clothing articles for which the store has inventory for key sizes and colors; when this requirement is not met, all units of the article are moved to the backroom, where they cannot be sold. This creates dependencies across articles that should be considered when allocating inventory across stores in the network. The authors ran a pilot with Zara to test a proposed inventory distribution policy based on an optimization model that considers these dependencies across different sizes and colors of each article. They then analyzed the resulting experimental data using a difference-in-differences design. Based on the positive results, Zara decided to expand the policy broadly.

In addition to testing policies, field experiments can generate data that can be used as input in traditional operations research methods. For example, Barnett et al. (2001) studied aviation security, investigating the feasibility of positive passenger bag-match (PPBM) for U.S. domestic flights. Required for international flights, PPBM removes unaccompanied baggage from aircrafts to prevent terrorism. Collaborating with the Federal Aviation Administration, the authors led a two-week test (May 6 through 19, 1997) involving 8,000 flights of 11 airlines with 50 city-pairs, which accounted for 4% of the domestic flights. The trials provided data such as the proportion of passengers that check luggage for a flight but don’t board, the time it takes to pull bags from the plane, and the proportion of bags slated for loading onto a plane that fail bottom-up security screen requirements. The field also captured the richness of the context in several ways; notably, airlines adjusted labor in different degrees, which allowed cost-effectiveness analyses. The authors then used computer simulations to estimate system wide effects, addressing issues such as estimating the impact on delays, which involved estimating the delays not only from removing bags from planes but also from verification tasks related to PPBM, adjusting for hidden delays, and accounting for delay propagation (when a flight delay causes delays on its next departures or by other flights, which in turn, cause other delays). Overall, their estimates suggest that, under usual operating conditions, domestic PPBM would only impose approximate delays of 1 minute per flight and airline costs of 40 cents per passenger enplanement, without restricting the number of flights.

Lab and field should be seen as complements rather than substitutes. Though the lab tends to precede the field, researchers can go back to the lab after field experiments (Harrison and List 2004). For example, Plott and Levine (1978) wanted a flying club to choose a particular aircraft to add to its fleet. When the vote was held, the authors altered the agenda to sway the vote and their aircraft was indeed chosen. The authors later conducted a laboratory experiment to rule out the possibility that the result was 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. They then conducted a lab experiment to confirm that the mechanism was authentic self-expression—being able to act as oneself—. Staats, KC, and Gino (2017) used a similar approach when they analyzed which cardiac stent cardiologists choose to use after a warning announcement about cardiac stent performance from the US Food and Drug Administration (FDA).

2.4. Why Not to Run Field Experiments

Attracted by the method, the opportunity to interact with practitioners, and the ability to identify externally valid, causal links, researchers are often tempted to run experiments without reflecting on why. Field experiments should be driven by theoretical motivation (Card, DellaVigna, and Malmendier 2011, List 2011). Experimental treatments should be grounded in theory—testing or evaluating theory—or aspire to lead to new theory building. Similar to other types of field studies, field experiments won't be relevant if simply reporting observations from the field and failing to provide new theoretical or practical insights.

Although intellectually exciting, field experiments also have disadvantages and thus may not be the best choice methodologically. Field experiments typically involve higher cost in terms of both time and money than other types of research and are likely to run into problems in execution. Moreover, experimental approaches tend to have larger risk of Hawthorne effects and smaller sample sizes. Archival field studies with proper causal inference methods often achieve the goal of identifying the effects of interest without the drawbacks of experiments.

Additionally, the complexity of the field should be avoided when the lab can answer the question (Al-Ubaydli and List 2015). Sometimes a hybrid can provide the right balance. For example, to measure the impact of workaround difficulty on frontline workers' response to operational failures, Tucker (2016) studies nurses, who face many such failures and whose behaviors can have a critical impact on patients health. Running experiments in hospitals, however, would be too risky for patients, and so she conducts laboratory experiments at exhibitor space at national nursing conventions.

2.5. How to Run Field Experiments

In this section, we discuss the creation of data through field experiments. Figure 2.1. provides a checklist for a successful field experiment. When researchers decide to conduct a field experiment after considering all the trade-offs, the next steps involve ethical considerations, experimental design, execution challenges, and field partners.

2.5.1. Ethics and Human Subject Protocol

Because subjects might suffer as a result of their direct or indirect participation in the study, the first step of any field project should be to minimize the potential risks and to get supervision and authorization from the Institutional Review Board (IRB) at their institutions before conducting the study (Levitt and List 2009, List 2011). 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. More particularly, the decision should be fair and based on science, avoiding discrimination. When possible, researchers should obtain informed consent. For example, students explicitly consented to participate in an experiment when signing up for the course in Zhang, Allon, and Van Mieghem (2017). However, in some cases, informing the subject would invalidate the research results; in such cases, a researcher should look for ways to mitigate the risks and receive guidance from the IRB to identify situations in which it could be argued that informed consent is not needed (Levitt and 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 is 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 researchers with validation that their ideas were properly vetted, but also the IRB has resources to help respond to and address any outstanding concerns.

2.5.2. Experimental Design

Once the researchers believe that a field experiment could answer a rigorous and relevant question, their 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, there are four additional practical issues to consider.

2.5.2.1. Treatment. The relevance of the field experiment will depend on the *treatment*, which is the intervention whose effects the researcher wants to evaluate. This intervention will be implemented in a controlled way so that it affects only some individuals or units (those in the so-called treatment or treated group). Broadly, a comparison of the treated group with the control group will reveal the effects of the treatment (average treatment effect, ATE). Projects may involve a single treatment, multiple treatments, and/or multiple degrees of treatment intensity. While the rest of the considerations will ensure rigor, the treatment will determine the scholarly insights and thus the relevance of the experiment. Thus, researchers should choose the treatments carefully.

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 and Gallien 2010, Gallien et al. 2015) and the Cornell bookstore (Lee et al. 2015), pricing with Zara (Caro and Gallien 2012) and Rue La La (Ferreira, Lee, and Simchi-Levi 2015), and scheduling policies with Italian judges (Bray et al. 2016).

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, and Mullainathan 2011). To date, mechanism experiments have been rarely used in OM. 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 experiment. Since the two types of experiments inform each other, it will be desirable to conduct both over time.

Many experiments use an *encouragement design*, whereby subjects receive an encouragement to take the treatment rather than the treatment itself. With an encouragement design, the comparison of subjects receiving the encouragement with those in the control group represents the intent-to-treat (ITT) effect, which may differ from the average treatment effect because of noncompliance. For example, to study how social interaction affected learning outcomes of students in two offerings of an online course in the Coursera platform, Zhang, Allon, and Van Mieghem (2017) encouraged students in the treated group to visit the course discussion board by adding text and four questions related to the discussion board to a survey. In this case, ITT represents the impact of the encouragement on learning outcomes, and there are two non-compliance problems: Encouraged students may not go to the discussion board and those in the control group may go. Exploiting the random encouragement assignment as an instrumental variable (IV), the authors estimate ATE: One additional board visit causally increases the probability that a student finishes the quiz in the subsequent week by 0.5% to 4.3%.

2.5.2.2. Randomization. Researchers must now focus on the goal of any given experiment: to create an appropriate counterfactual. To do so, researchers use *randomization* to assign a condition to what become the treatment group and the *control* group (Harrison and List 2004). A *randomized block design* should be used when a covariate could predict the potential outcome. 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.

In almost any field experiment, a researcher must worry about *contagion effects*, or the sharing of information about the treatment across conditions. Unlike the laboratory where participants are more easily kept separate and supervised, in the field there is substantial risk that participants may talk. 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 or the treatment group disliked the unfairness or felt particular motivated knowing that they received better treatment, resulting in an observed effect that no longer was 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 potential impact. This may mean randomizing across work groups or facilities, for example.

Social experiments may also suffer “randomization bias”: the bias that occurs when randomization itself leads to samples that are not representative of the population (Levitt and List 2009). When individuals or organizations choose to participate in an experiment, the knowledge that randomization would determine whether or not they receive the treatment will influence their decision on whether to participate in the experiment (and join the experiment’s sample).

Finally, researchers should be cautious about implementation issues. For example, a company might simply randomize by 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 most companies assign employee identification numbers in sequential, nonrandom order as individuals join the firm; thus, in such case, the researcher would have a wonderful field experiment on how newer employees who get a treatment compare with older employees who do not. Whenever possible, randomization should be carried out by the researchers rather than by their organizational partner. By taking on this task as a researcher, it prevents finding out about problems later on that could invalidate the prior work.

2.5.2.3. Measures. Before starting a field experiment, researchers should determine the measures to assess dependent variables and controls. 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 causal mechanism. If a gap is identified, 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).

2.5.2.4. Power analyses. 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 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. Speaking from experience, discovering high attrition rates after the fact is painful for the researcher.

2.5.3. Execution Challenges

Researchers (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 the challenges faced by the authors of field experiments-based papers and how they addressed them.

One lesson from challenges faced by these authors is that researchers must be ready to make trade-offs. One type of trade-off results from applicability because research that is relevant for practice often must make sacrifices in theoretical contributions to some extent. 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 still has a positive influence on scholars and practitioners.

Other type of trade-offs relate to experimental design. Seeking to evaluate the performance impact of improving management practices of Indian textile firms, Bloom et al. (2013) decided to provide free consulting on management practices to randomly chosen plants. One challenge was that consulting is expensive. Given their limited funding, the researchers had to make the tradeoff between spending more per plant (enabling them to provide higher quality consulting, encourage participation and retention, and have a higher impact necessary to study large firms) or spending less per plant in more plants (enabling them to have a 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).

A second lesson from experimenting in the field is that the researchers should discuss data availability and collection with the partner organizations to address the problem that some data may not be available in the analysis phase. For example, in Caro and Gallien (2010)'s work with Zara on inventory distribution, 15 articles were initially selected for the inventory distribution test but only 10 of the 15 had data for more than three weeks, forcing the authors to limit the analysis to those 10. Fortunately, those 10 articles remained a representative proportion of "basic" vs. "fashion" items of clothing, yet the already small sample size was reduced. Moreover, the forecasts used during the pilot were not saved and thus could not be used during the data analysis phase.

A third lesson is that experimenters should carefully reflect on whether any other factor could drive the effect that the experiment intends to measure and then take precautionary measures to address those challenges. In the case of Bloom et al. (2013), 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 created 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. Given the size of the experiment, there was a risk of Hawthorne effects, as treated plants had greater interactions with the consultants, but this was unavoidable given the high cost of the intervention.

A fourth lesson is that researchers should micro-manage the execution of the experiment because even small details of the implementation of the treatment can affect results. Consider Singh, Teng, and Netessine (2017)'s assessments of philanthropic campaigns and discount-based promotions sent via text messages in a taxi platform. When analyzing the data from their first experiment, the responses to some of the text messages were not as expected and the authors reasoned that a plausible explanation could be the differences in the actual promotional codes sent for each treatment, which had been chosen by the company's marketing team (page 6). Specifically, the responses may have been lower for the code "4sgd" because four is an unlucky number in Chinese culture, which is predominant in Singapore where the experiments were run, and for the code "Nepal3sgd" because of its higher complexity compared to "2sgd". With this in mind, in the other two experiments, the authors avoided the number four and chose more standardized codes such as "1give" and "1off".

Overall, the experimenters overcame many challenges but were able to answer difficult research questions through the experiments. While there are specific lessons from their experiences, new field experiments will likely face new challenges. The broad lesson is that execution is important. When thinking about 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 set up 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. Common pilots include going to the lab to test questionnaires or conducting a *conference room pilot* where the researchers gather with their sponsors to walk through each step of the proposed field experiment to confirm validity.

2.5.4. Field Sites and Organizational Partners

Perhaps the biggest key to the success of the execution of any field experiment lies with the organizational partner and 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 very open to finding answers to questions they care about, but start-ups lack the resources to explore on their own. On the other hand, big firms bring the benefits of size and infrastructure—with legal departments and bureaucracy 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 not something a researcher should fear. 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 an 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 the intellectual capital value, as well as potential competitive and 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. In fact, engaging with partners has been one of the most exciting and fulfilling parts of our own research, and we have learned more from working with our collaborators than we would have had we approached the studies as typical lab or data exercises.

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, and 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, researchers need to communicate the need for randomization.

Some companies 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 the researcher can identify companies that are curious about the proposed topic, or who have a real pain point, the researcher can then explain clearly why a given experiment could bring benefits to this partner. From this foundation, a relationship can be built, and an experiment may be able to be implemented.

These 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, as the company continued to grow at an incredibly rapid pace, the executives realized they were not able to continue operating with the same manual systems they had relied on. This timing probably facilitated the successful collaborations that in turn facilitated scalable processes and resulted in 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, to successfully complete the project.

2.6. 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 and List 2004). Given the above review, we highlight five possibilities for future work to consider.

First, we 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. Thus, allocating a portion of your time towards field experiments may have outsized impact on your field of study.

Second, researchers should be aware of the possibility to exploit natural experiments. Often, organizations may make internal changes that can be used as the identification strategy for an empirical analysis. 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. 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 we hope this paper will also help reviewers to hold field studies to appropriate standards.

Fourth, we strongly encourage authors to use field and lab studies together. Not every paper requires both, but when there are weaknesses of note, a lab study may prove to be an excellent complement to a field study. Moreover, 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 suggest that turning to the field may prove valuable in helping the operations management 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. In other words, they studied 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 paper will aid operations management scholars in the use and application of a variety of empirical approaches to conduct field research that advances both scholarship and practice.

2.7. References

Al-Ubaydli O., List J.A. 2015. Do Natural Field Experiments Afford Researchers More or Less Control Than Laboratory Experiments? *American Economic Review* **105**(5) 462-466.

Allcott H., Rogers T. 2014. The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. *American Economic Review* **104**(10) 3003-3037.

Aral S., Brynjolfsson E., Van Alstyne M. 2012. Information, Technology, and Information Worker Productivity. *Information Systems Research* **23**(3-part-2) 849-867.

Bandiera O., Barankay I., Rasul I. 2011. Field Experiments with Firms. *Journal of Economic Perspectives* **25**(3) 63-82.

Barnett A., Shumsky R., Hansen M., Odoni A., Gosling G. 2001. Safe at home? An experiment in domestic airline security. *Operations Research* **49**(2) 181-195.

Bloom N., Eifert B., Mahajan A., McKenzie D., Roberts J. 2013. Does Management Matter? Evidence from India. *The Quarterly Journal of Economics* **128**(1) 1-51.

Bray R.L., Coviello D., Ichino A., Persico N. 2016. Multitasking, Multiarmed Bandits, and the Italian Judiciary. *Manufacturing & Service Operations Management* **18**(4) 545-558.

Cable D.M., Gino F., Staats B.R. 2013. Breaking them in or revealing their best? Reframing socialization around newcomer self expression. *Administrative Science Quarterly* **58**(1) 1-36.

Cachon G. 2012. What is interesting, in operations management? *Manufacturing & Service Operations Management* **14**(2) 166-169.

Cachon G.P., Olivares M. 2009. Drivers of Finished-Goods Inventory in the U.S. Automobile Industry. *Management Science* **56**(1) 202-216.

Card D., DellaVigna S., Malmendier U. 2011. The Role of Theory in Field Experiments. *Journal of Economic Perspectives* **25**(3) 39-62.

Caro F., Gallien J. 2010. Inventory Management of a Fast-Fashion Retail Network. *Operations Research* **58**(2) 257-273.

Caro F., Gallien J. 2012. Clearance Pricing Optimization for a Fast-Fashion Retailer. *Operations Research* **60**(6) 1404-1422.

Charness G., Gneezy U. 2009. Incentives to exercise. *Econometrica* **77**(3) 909-931.

Chuang H.H.-C., Oliva R., Liu S. 2016. On-Shelf Availability, Retail Performance, and External Audits: A Field Experiment. *Production and Operations Management* **25**(5) 935-951.

de Mel S., McKenzie D., Woodruff C. 2008. Returns to Capital in Microenterprises: Evidence from a Field Experiment. *The Quarterly Journal of Economics* **123**(4) 1329-1372.

Ferreira K.J., Lee B.H.A., Simchi-Levi D. 2015. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & Service Operations Management* **18**(1) 69-88.

Gallien J., Mersereau A.J., Garro A., Mora A.D., Vidal M.N. 2015. Initial Shipment Decisions for New Products at Zara. *Operations Research* **63**(2) 269-286.

Gallino S., Moreno A. 2014. Integration of Online and Offline Channels in Retail: The Impact of Sharing Reliable Inventory Availability Information. *Management Science* **60**(6) 1434-1451.

Gallino S., Moreno A., Stamatopoulos I. 2016. Channel Integration, Sales Dispersion, and Inventory Management. *Management Science* **63**(9) 2813-2831.

Gaur V., Fisher M.L. 2005. In-Store Experiments to Determine the Impact of Price on Sales. *Production and Operations Management* **14**(4) 377-387.

Giroud X. 2013. Proximity and Investment: Evidence from Plant-Level Data. *The Quarterly Journal of Economics* **128**(2) 861-915.

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

Hayes R.H. 2002. Wick Skinner: A life sailing against the wind. *Production and Operations Management* **11**(1) 1-8.

Ho T.-H., Lim N., Reza S., Xia X. 2017. OM Forum—Causal Inference Models in Operations Management. *Manufacturing & Service Operations Management* **19**(4) 509-525.

Ibanez M.R., Clark J.R., Huckman R.S., Staats B.R. 2017. Discretionary task ordering: Queue management in radiological services. *Management Science* (forthcoming).

Kuhn T.S. 1962. *The structure of scientific revolutions*. University of Chicago Press, [Chicago].

Lacetera N., Pope D.G., Sydnor J.R. 2012. Heuristic Thinking and Limited Attention in the Car Market. *American Economic Review* **102**(5) 2206-2236.

Lee J., Gaur V., Muthulingam S., Swisher G.F. 2015. Stockout-Based Substitution and Inventory Planning in Textbook Retailing. *Manufacturing & Service Operations Management* **18**(1) 104-121.

Levine D.I., Toffel M.W. 2010. Quality management and job quality: How the ISO 9001 standard for quality management systems affects employees and employers. *Management Science* **56**(6) 978-996.

Levitt S.D., List J.A. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review* **53**(1) 18.

Levitt S.D., List J.A. 2011. Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. *American Economic Journal: Applied Economics* **3**(1) 224-238.

List J.A. 2011. Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off. *Journal of Economic Perspectives* **25**(3) 3-16.

Lu Y., Musalem A., Olivares M., Schilkrut A. 2013. Measuring the Effect of Queues on Customer Purchases. *Management Science* **59**(8) 1743-1763.

Ludwig J., Kling J.R., Mullainathan S. 2011. Mechanism Experiments and Policy Evaluations. *Journal of Economic Perspectives* **25**(3) 17-38.

MacDuffie J.P. 1997. The Road to "Root Cause": Shop-Floor Problem-Solving at Three Auto Assembly Plants. *Management Science* **43**(4) 479-502.

Marx M., Strumsky D., Fleming L. 2009. Mobility, skills, and the Michigan non-compete experiment. *Management Science* **55**(6) 875-889.

Milkman K.L., Minson J.A., Volpp K.G.M. 2014. Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management Science* **60**(2) 17.

Musalem A., Olivares M., Bradlow E.T., Terwiesch C., Corsten D. 2010. Structural Estimation of the Effect of Out-of-Stocks. *Management Science* **56**(7) 1180-1197.

Perdikaki O., Kesavan S., Swaminathan J.M. 2012. Effect of retail store traffic on conversion rate and sales. *Manufacturing & Service Operations Management* **14**(1) 145-162.

Pierce L., Snow D., McAfee A. 2014. Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Sci.*

Plott C.R., Levine M.E. 1978. A Model of Agenda Influence on Committee Decisions. *The American Economic Review* **68**(1) 146-160.

Singh J., Teng N., Netessine S. 2017. Philanthropic Campaigns and Customer Behavior: Field Experiments on an Online Taxi Booking Platform. *Management Science*.

Staats B.R., Dai H., Hofmann D., Milkman K.L. 2017. Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science* **63**(5) 1563-1585.

Staats B.R., KC D.S., Gino F. 2017. Maintaining Beliefs in the Face of Negative News: The Moderating Role of Experience. *Management Science* **64**(2) 804 - 824.

Taylor F.W. 1911. *The Principles of Scientific Management*. Harper & Brothers, New York.

Tucker A.L. 2016. The Impact of Workaround Difficulty on Frontline Employees' Response to Operational Failures: A Laboratory Experiment on Medication Administration. *Management Science* **62**(4) 1124-1144.

Van Mieghem J.A. 2012. OM Forum—Three Rs of Operations Management: Research, Relevance, and Rewards. *Manufacturing & Service Operations Management* **15**(1) 2-5.

Zhang D.J., Allon G., Van Mieghem J.A. 2017. Does Social Interaction Improve Learning Outcomes? Evidence from Field Experiments on Massive Open Online Courses. *Manufacturing & Service Operations Management* **19**(3) 347-367.

Figure 2.1. Checklist for a Successful Field Experiment

- □ Is the field experiment academically relevant?
    - □ Grounded in theory?
    - □ Answering novel research questions?
- □ Is a field experiment the best approach? Don’t run a field experiment if the questions can be better answered via lab experiments or archival studies.
- □ Ethics and Human Subject Protocol: mandatory
- □ Experimental Design
    - □ Treatment: well-defined and connected to the research question
        - □ Ensure the chosen treatment is really the only difference between the treatment and control groups
    - □ Randomization
    - □ Measures: ensure data on all relevant factors is collected
    - □ Power analyses: ensure the sample size is large enough
    - □ Ensure the benefits from the field
        - □ External Validity
        - □ Overcome Observer Bias
        - □ Know the context and use it in the experimental design and analysis
- □ Prevent and Respond to Execution Challenges
    - □ Know the setting well and anticipate problems
    - □ Identify and make trade-offs
- □ Field Sites and Organizational Partners: communication, use your own resources

Chapter 3

How Scheduling Can Bias Quality Assessment:

Evidence from Food Safety Inspections

Abstract

Many production processes are subject to inspection to ensure they meet quality, safety, and environmental standards imposed by companies and regulators. Inspection accuracy is critical to inspections being a useful input to assessing risks, allocating quality improvement resources, and making sourcing decisions. This paper examines how the scheduling of inspections risks introducing bias that erodes inspection quality by altering inspector stringency. In particular, we theorize that inspection results are affected by (a) the inspection outcomes at the inspector's prior inspected establishment and (b) when the inspection occurs within an inspector's daily schedule. Analyzing thousands of food safety inspections of restaurants and other food-handling establishments, we find that inspectors cite more violations after inspecting establishments that exhibited worse compliance or greater deterioration in compliance and that inspectors cite fewer violations in successive inspections throughout their day and when inspections risk prolonging their typical workday. Our estimates suggest that, if the outcome effects were amplified by 100% and the daily schedule effects were fully mitigated (that is, reduced by 100%), the increase in inspectors' detection rates would result in their citing an average of 9.9% more violations. Scaled nationwide, this would yield 19.0 million fewer foodborne illness cases per year, reducing annual foodborne illness costs by $14.2 billion to $30.9 billion. Understanding these biases can help managers develop alternative scheduling regimes that reduce bias in quality assessments in domains such as food safety, process quality, occupational safety, working conditions, and regulatory compliance.

3.1. Introduction

Many companies inspect their own and their suppliers' operations to ensure they are meeting quality, labor, and environmental standards. Various government agencies also inspect for regulatory compliance. The accuracy of inspections is critical to their being a useful input to key managerial decisions, including how to allocate quality improvement resources, which suppliers to source from, and how to penalize noncompliance. Inaccurate assessments can prevent managers, workers, customers, and neighbors from making well-informed decisions based on the risks imposed by an establishment's operations. Moreover, inspections that miss what they could have caught can undermine the inspection regime's ability to deter intentional noncompliance. In this study, we theorize and find evidence of several sources of bias that lead to inaccurate inspections. We also propose solutions—including alternative inspection scheduling regimes—that can improve inspection accuracy without increasing inspection costs.

Several studies have revealed various sources of inspection inaccuracy. Yet little is known about inspector bias. We consider an unexplored type of bias that results from an operational decision: scheduling. Building on work from the behavioral sciences, we hypothesize how the sequence of inspections might affect the number of violations cited. Specifically, inspector stringency on a particular inspection may be influenced by (a) the outcomes of the inspector's prior inspection (prior inspection outcome effects, or, simply, outcome-effects) and (b) its position within the day (daily schedule effects). Throughout this paper, we refer to *an inspector's* preceding inspection as his or her "prior" inspection and *an establishment's* preceding inspection as its "previous" inspection. Figure 3.1 illustrates this distinction.

We study the influence of scheduling on inspection accuracy in the context of local health department food safety inspections of restaurants and other food-handling establishments. While these inspections need to accurately assess compliance in order to protect consumer health, the number of violations cited in these reports is a function of both the facility's actual hygiene and the inspector's stringency in detecting and recording violations. Because citing violations requires supporting documentation, inspector bias takes the form of underreporting the violations that are actually present. Using data on thousands of inspections, we find strong evidence that inspectors' schedules affect the number of violations cited in their reports.

We hypothesize three ways in which an inspector's experience at one inspection affects the number of violations cited at his or her next inspection. First, we hypothesize that an inspector's stringency will be influenced by the number of violations at his or her prior inspection. Those violations will affect the inspector's emotions and perceptions about the general compliance of the community of inspected establishments (via the salience of those recent inspection results), in turn altering his or her expectations and attitudes when inspecting the next establishment. This leads us to predict that having just conducted an inspection that cites more violations will lead the inspector to also cite more violations in the next establishment he or she inspects. As predicted, we find that *each additional violation* cited in the inspector's prior inspection (of a different establishment) increases by 1.5% the number of violations that inspector cites at the next establishment.

Second, we hypothesize that trends matter, too: discovering more compliance deterioration (or less improvement) at one inspection affects inspectors' emotions and perceptions in ways that lead them to cite more violations at the next establishment. Supporting this hypothesis, we find that inspectors cite 1.3% more (fewer) violations after having inspected another establishment whose violation trend worsened (improved) by one standard deviation.

Finally, we hypothesize, based on negativity bias, that this trend effect will be stronger following an inspection that found deterioration than following one that found improvement. Indeed, we find empirical evidence that the trend effect is asymmetric, occurring when compliance at the inspector's prior establishment deteriorates but not when it improves.

We then hypothesize two daily schedule effects. We first theorize that each additional inspection over the course of a day causes fatigue that erodes inspectors' stringency and leads them to cite fewer violations. We find empirical evidence to support this, observing that each subsequent inspection during an inspector's day yields 3.2% fewer citations, an effect that our supplemental analysis demonstrates is not due to inspectors' scheduling presumably cleaner establishments later in their workday. Second, we hypothesize that inspections that risk prolonging an inspector's workday will be conducted less stringently, which will lead to fewer violations being cited. We find empirical support for this, too, in that potentially shift-prolonging inspections yield 5.1% fewer citations.

Overall, our findings reveal that currently unreported violations would be cited if the outcome effects (which increase scrutiny) were triggered more often and the daily schedule effects (which erode scrutiny) were reduced. Our estimates suggest that, if the outcome effects were amplified by 100% and the daily schedule effects were fully mitigated (that is, reduced by 100%), the increase in inspectors' detection rates would result in their citing 9.9% more violations. Scaled nationwide, this would result in 240,999 additional violations being cited annually, which would in turn yield 50,911 fewer foodborne illness related hospitalizations and

19.01 million fewer foodborne illness cases per year,[1] reducing annual foodborne illness costs by $14.20 billion to $30.91 billion.

Our work contributes to both theory and practice. By identifying factors that bias inspections, we contribute to the literature on monitoring and quality improvement (e.g., Gray, Siemsen, and Vasudeva 2015). Our focus on how scheduling affects inspector stringency introduces the operational lens of scheduling to the literature examining inspector bias, which has otherwise largely focused on experience or other sociological and economic factors (e.g., Short, Toffel, and Hugill 2016, Ball, Siemsen, and Shah 2017). Our examination of how operational decisions affect inspector behavior also contributes to the literature on behavioral operations, which emphasizes the importance of human behavior in operations management decisions (Bendoly, Donohue, and Schultz 2006). Our findings show that fatigue can affect performance of primary tasks even during normal shift hours. Moreover, by examining data from actual decisions with important consequences for public health, we contribute to the recent attempts to explore high-stakes decision-making in field settings (e.g., Chen, Moskowitz, and Shue 2016). We also go beyond previous work by not only estimating the magnitude of bias but also estimating their real-world consequences. With managers across many different industries seeking to monitor and improve quality, our research suggests a cost-effective lever: exploiting the behavioral effects of the organization of work.

[1] These estimated reductions in foodborne illness cases and hospitalizations assume establishments would remediate the newly cited violations at the same rate as other cited violations, and that the two sets of violations are equally likely to result in these health outcomes. In extensions, we relax these assumptions to derive alternative estimates.

3.2. Related Literature

Our research builds on two streams of literature: (a) quality management and monitoring and (b) scheduling and task performance.

3.2.1. Quality Management and Monitoring

Decades of scholarship has explored various approaches to ensuring that operations adhere to quality specifications. Prior research has, for example, examined total quality management (e.g., Lapré, Mukherjee, and Van Wassenhove 2000), programs that encourage self-disclosure of process errors and regulatory violations (e.g., Gawande and Bohara 2005, Kim 2015), and electronic monitoring systems (Staats et al. 2017). A primary approach remains physical inspections, such as internal quality control departments assessing manufacturing processes (Shah, Ball, and Netessine 2016), internal auditors assessing inventory records (Kök and Shang 2007), and third-party monitors assessing the conformance of supplier operations to buyers' codes of conduct (e.g., Handley and Gray 2013, Short and Toffel 2016) and to management standards such as ISO 9001 (Corbett 2006, Levine and Toffel 2010, Gray, Anand, and Roth 2015).

An extensive literature has highlighted the role of inspections in fostering organizational learning (Hugill, Short, and Toffel 2016, Mani and Muthulingam 2017), and promoting operational routines and adherence to Good Manufacturing Processes (e.g., Anand, Gray, and Siemsen 2012, Gray, Siemsen, and Vasudeva 2015), occupational health and safety regulations (e.g., Ko, Mendeloff, and Gray 2010, Levine, Toffel, and Johnson 2012), and environmental regulations (for a review, see Shimshack 2014). Prior research has found compliance to be a function of an establishment's inspection history (including how many inspections it had undergone and the time lag between inspections) and inspector characteristics (including their training and experience and their familiarity with a particular establishment) (Ko, Mendeloff, and Gray 2010, Toffel, Short, and Ouellet 2015). In contrast, we examine the extent to which an establishment's inspection report is influenced by the *inspector's* schedule, including (a) the inspector's experience at his or her prior inspection of a different establishment and (b) when during the inspector's day the inspection is conducted.

The usefulness of inspections is contingent on their accuracy. Researchers have long been interested in how to conduct quality control inspections (e.g., Ballou and Pazer 1982), recognizing inspectors' fallibility and variability (Feinstein 1989). The limited number of studies of the heterogeneity across inspectors' propensity to report violations has identified the importance of their tenure, training, gender, and former exposure to the establishment (Macher, Mayo, and Nickerson 2011, Short, Toffel, and Hugill 2016, Ball, Siemsen, and Shah 2017). Inspector accuracy among third-party inspection firms has been shown to be influenced by (a) whether the establishment or its buyer hires the inspection firm and pays for the inspection (Ronen 2010, Duflo et al. 2013, Short and Toffel 2016), (b) the level of competition among inspection firms (Bennett et al. 2013), and (c) whether the inspecting firm has cross-selling opportunities (Koh, Rajgopal, and Srinivasan 2013, Pierce and Toffel 2013). In contrast to these demographic aspects of individual inspectors and structural dimensions of the relationship between the inspection firm and the inspected establishment, we explore a very different potential source of inspection bias: where the inspection falls within an inspector's schedule.

3.2.2. Scheduling and Task Performance

Our study also relates to research that has examined how work schedules affect task performance. This literature has, for example, proposed optimal scheduling of workforces (e.g., Green, Savin, and Savva 2013) and of periodic tasks such as machine inspections (e.g., Lee and

Rosenblatt 1987). Studies of the sequencing of individual workers' tasks have shown that scheduling similar tasks consecutively to increase task repetition can improve performance by reducing delays incurred from switching tasks (e.g., Staats and Gino 2012, Ibanez et al. 2017) and that healthcare workers work more quickly later in a service episode of finite duration (Deo and Jain 2015). We extend this work by focusing on the effects of work schedules on task quality in a setting that purports to provide inspections that are of consistent quality as the basis for a fair and objective monitoring regime.

A few studies have examined the relationship between work schedule and task quality. Dai et al. (2015) finds that healthcare workers become less compliant with handwashing rules over the course of their shift. That study focused on adherence to a secondary task that was largely unobservable to others, where noncompliance was common, and where fatigue might lead workers to shift their attention from this secondary task toward their primary tasks. In contrast, our study focuses on primary tasks, where the outcome of such tasks (violations cited) is explicitly observable to others and where such visibility could deter variation. Moreover, whereas Dai et al. (2015) measured adherence dichotomously, we use a more nuanced scalar measure. Another study examined the decisions of eight judges and found that they were more likely to deny parole as they issued more judgments throughout the course of their day but that taking a break attenuated this bias, suggesting that repeated decisions might have caused mental depletion (Danziger, Levav, and Avnaim-Pesso 2011). Whereas judges became harsher as they made more decisions throughout the day, inspectors might behave differently, given that for an inspector, greater harshness (manifested as stringency) requires more work.

Finally, two studies examined how workers adjust their decisions based on their prior decisions. A study of MBA application assessments found that the higher the cumulative average of the scores an interviewer had given to applicants at a given moment on a given day, the lower he or she scored subsequent applicants that day, suggesting that decision makers adjust their scores to maintain a consistent daily acceptance rate (Simonsohn and Gino 2013). Another study found that judges, loan reviewers, and baseball umpires were more likely to make “accept” decisions immediately after a “reject” decision (and vice versa), a form of decision bias (Chen, Moskowitz, and Shue 2016). Whereas these two studies find that subsequent decisions typically oppose prior ones, inspectors do not have explicit or self-imposed quotas or targets and, as we explain below, their emotions and perceptions may be affected by their prior tasks in ways that encourage subsequent decisions to be similar to prior ones. Additionally, we go beyond what prior work has considered by proposing that the magnitude of the effects from prior task outcomes will be asymmetric and will depend on whether the prior outcome was positive or negative.

3.3. Theory and Hypotheses

Quality assurance audits and inspections have detailed procedures to be followed in pursuit of accuracy. Yet, in practice, behavioral biases may influence an inspector’s stringency. Whereas inspections are typically assumed to yield the same results no matter when they occur on the inspector’s schedule, we hypothesize that inspection results will indeed be influenced by the type of experience inspectors have at their immediately prior inspection—which we refer to as *prior inspection outcome-effects*—and by when an inspection occurs during an inspector’s daily schedule—which we refer to as *daily schedule effects*.

Figure 3.1 depicts the relationships we hypothesize below.

3.3.1. Prior Inspection Outcome-Effects on Quality Assessment

3.3.1.1. Violation level of the inspector's prior inspection. We theorize that inspectors will be influenced by the results of prior inspections. One such outcome-effect is driven by whether the establishment an inspector just visited had many or few violations. There are two reasons why inspecting an establishment with many violations can imbue inspectors with a negative attitude that leads them to inspect more diligently at their next inspection, whereas inspecting a more compliant establishment can lead them to be less stringent in their subsequent inspection. First, an inspector's prior inspection can affect him or her emotionally. When more violations are cited at that prior establishment, its personnel are more likely to be dissatisfied and resentful, which can lead to hostile interactions with inspectors that can erode inspectors' goodwill and thus heighten their stringency during the next inspection. Merely observing their dissatisfaction and resentfulness can similarly affect inspectors via emotional contagion (Barsade 2002). Conversely, inspectors experience at their prior inspections that result in fewer violations is more likely to bolster their goodwill at the next inspection. Second, the experience at the inspector's prior inspection can shape his or her perceptions of the overall behavior of establishments, which can influence his or her stringency at the subsequent inspection. Recently experiencing an event (such as compliance) increases its salience and results in more rapid recall. An inspector may therefore use the results of that inspection to update his or her estimate of typical compliance levels, relying on the availability heuristic (Tversky and Kahneman 1974) and seeking evidence at his or her next inspected establishment that supports these expectations, consistent with confirmation bias (Nickerson 1998). This becomes a self-fulfilling prophecy, where experiencing poor (good) compliance at their prior establishment leads inspectors to reduce (heighten) scrutiny at their next inspection that results in their detecting fewer (more) violations.

Other types of decisions might exhibit the opposite bias whereby successive decisions are negatively autocorrelated, akin to the law of small numbers, the gambler's fallacy, sequential contrast effects, or quotas (Chen, Moskowitz, and Shue 2016). These effects are likely weak in the case of inspections. First, the law of small numbers and the gambler's fallacy, in which the decision-maker underestimates the likelihood of sequential streaks occurring by chance, are less likely to apply in the case of inspectors. Instead, inspectors can be expected to predict a high likelihood that the establishments they sequentially inspect will exhibit similar compliance (that is, sequential streaks) because they share external factors that affect their compliance, including their competition, regulatory knowledge, and requirements about whether they must disclose their inspection results (such as restaurants in Los Angeles, New York, and Boston being required to post restaurant grade cards). Moreover, at least in our setting, inspectors monitor a set of establishments over time and are therefore somewhat responsible for their evolution. As such, inspectors may believe that their own past behavior—including how stringently their past inspections were and how they balanced their dual roles of teaching and enforcement—will influence establishments' compliance trends, which would lead to their establishments exhibiting similar trends. Second, sequential contrast effects, in which the decision-maker's perception of the quality of the current establishment is negatively biased by the quality of the previous one, are ameliorated because inspectors are extensively trained to evaluate quality based on what they observe and thus have well-defined evaluation criteria that reduces the influence of prior inspections as temporary reference points. Moreover, each inspection takes significant time and often involves additional time traveling across inspected entities, so decisions are farther apart than sequential instantaneous decisions that may lead to unconscious contrasts of establishments. Third, quotas for the number of positive or negative decisions (in terms of violations cited or overall assessments of an establishment) would imply that fewer positive decisions could be made after a prior positive decision. Though the immediate prior decision would not directly matter, the cumulative prior decisions could. However, inspectors typically lack quotas or targets.

We therefore hypothesize:

Hypothesis 1: *The more (fewer) violations an inspector cites at one establishment, the more (fewer) violations he or she will cite in the next establishment.*

3.3.1.2. Violation trend of the inspector's prior inspection. An inspector's behavior is shaped not only by the prior establishment's *level* of compliance, but also by its *change* in compliance relative to its previous inspection. This second type of outcome-effect also results from how the prior inspection affects the inspector's emotions and perceptions.

The inspector's emotional response (through emotional contagion and interactions) at his or her prior establishment will depend on the trend there because the expectations of the establishment's personnel will be based on its previous inspection; they will be pleased or displeased according to whether their violation count has decreased or increased. After visiting an establishment with greater improvement, the inspector will exhibit a more positive temperament and will approach his or her next inspection with greater empathy and less stringency.

An inspector's perceptions, too, may be biased by the change in violations at the prior establishment. Many inspectors view inspections as a cooperative endeavor with the regulated entity to help improve business operations and safeguards stakeholders (e.g., May and Wood 2003, Pautz 2009, Pautz 2010). Improved compliance may therefore be attributed to management taking the rules and regulations seriously—that is, cooperating—whereas worsened compliance may be attributed to management ignoring or deliberately flouting the rules—definitely not cooperating. Improved compliance therefore confirms a cooperative relationship, which can lead inspectors to believe that the overall community of inspected establishments is cooperating and thus to be less stringent in the next inspection. Worsened compliance can lead inspectors to believe that the overall community of inspected establishments is not cooperating and thus to be more stringent in the next inspection. We therefore hypothesize:

Hypothesis 2: *The more an establishment's compliance has deteriorated (improved), the more (fewer) violations an inspector will record at the next establishment.*

3.3.1.3. Violation trend at the inspector's prior inspection: Asymmetric effects of deterioration versus improvement. According to the principle of *negativity bias*, negative events are generally more salient and dominant than positive events (Rozin and Royzman 2001). Negative events instigate greater information processing to search for meaning and justification, which in turn strengthens the memory and tends to spur stronger and more enduring effects in many psychological dimensions (Baumeister et al. 2001).

Negativity bias can affect the impact of the prior inspection's violation trend on the inspector's emotions and perceptions. First, negativity bias implies that for the inspected establishment's staff, the negative emotional effect of a drop in compliance may be stronger than the positive emotional effect of an improvement. This would result in stronger conveyance to inspectors of negative emotions associated with a drop in compliance and weaker conveyance of positive emotions associated with an improvement. An inspector will then absorb more negative emotions after the negative finding than positive emotions after the positive finding. Moreover, as argued by Barsade (2002), mood contagion might be more likely for unpleasant emotions because of higher attention and automatic mimicry. These asymmetries in the extent to which declining versus improving conditions affect inspectors' emotions will lead, in turn, to asymmetric effects on the strength of the resulting positive or negative outcome-effects.

Second, the salience of negative outcomes may have a stronger effect on inspectors' perceptions of how all of the establishments they monitor generally think about compliance, which can shape their stringency in a subsequent inspection. This is due to the *status-quo bias*: with the status quo acting as the reference point, negative changes are perceived as larger than positive changes of the same magnitude (Samuelson and Zeckhauser 1988, Kahneman 2003). We therefore hypothesize:

Hypothesis 3: *Observing deteriorated conditions at an establishment will increase the inspector's stringency at the next establishment to a greater extent than observing improved conditions will reduce his or her stringency.*

3.3.2. Daily Schedule Effects on Quality Assessment

3.3.2.1. Inspector fatigue. Inspectors are influenced not only by the results of prior inspections, but also by the sequencing of inspections within the day. Their work typically consists of a sequence of evaluative tasks that include physical tasks (such as manually examining the dimensions of a part or the temperature of a freezer) and mental tasks (such as interviewing an employee or determining whether or not a set of observations is within acceptable standards). As these tasks are executed, physical and mental fatigue will increase (Brachet, David, and Drechsler 2012). Furthermore, experimental evidence indicates that mental fatigue increases physical fatigue (Wright et al. 2007, Marcora, Staiano, and Manning 2009).

Over the course of a day, inspectors' physical and mental fatigue will reduce their physical and cognitive effort. This undermines stringency, which requires physical and cognitive efforts such as moving throughout the facility, interviewing personnel, waiting to observe work, executing procedures such as taking measurements, and conducting unpleasant tasks (such as observing storage practices in a walk-in freezer). Once an attribute is observed, inspectors need to recall and interpret the relevant standards to decide whether there is a violation and, if so, to document it. Each step must be executed according to rules that increase the complexity even of tasks that might appear simple to the untrained eye. Moreover, mental effort is required to make decisions against the status quo; as inspectors grow more tired during the day, they may become more willing to accept the status quo (Muraven and Baumeister 2000, Danziger, Levav, and Avnaim-Pesso 2011), which, in the context of inspections, can take the form of passing inspection items. Finally, mental effort is required to withstand the social confrontations that can erupt when a finding of noncompliance is disputed by those working at the establishment, who may genuinely disagree and for whom, in any case, much may be at stake in terms of reputation and sales. Citing violations can also provoke threats of appeals and lawsuits. Anticipating such responses, inspectors who are growing fatigued will exert less effort and seek to avoid confrontation, both of which increase leniency. For all these reasons, we hypothesize:

Hypothesis 4: *Inspectors will cite fewer violations as they complete more inspections throughout the day.*

3.3.2.2. Potential shift prolonging. In many settings, workers have discretion over their pace, which can lead them to prolong tasks to fill the time available (Hasija, Pinker, and Shumsky 2010) and to conduct work more quickly when facing higher workloads (KC and Terwiesch 2012, Berry Jaeker and Tucker 2017). Beyond these workload-related factors, we propose that inspectors will inspect less stringently when they expect to work later than usual (that is, beyond when they typically end work for the day). We hypothesize that inspectors' reluctance to suspend an inspection once underway, which would require them to bear the travel cost again the next day to finish the inspection, combined with a desire to finish at their typical time, will create pressure to speed up and inspect less thoroughly. As workers approach their typical end-of-shift time, accomplishing whatever remaining work cannot be postponed can become increasingly pressing as their perceived opportunity cost of time increases. The desire to speed up in these circumstances can result in the increased reliance on workarounds and cutting corners (Oliva and Sterman 2001) which, in turn, can reduce the quality of the work performed. Because properly conducted inspections require carefully evaluating a series of individual elements to identify whether each is in or out of compliance, omitting or expediting tasks to avoid prolonging the shift will result in a less comprehensive inspection with fewer violations detected and cited. We therefore hypothesize:

Hypothesis 5: *Inspectors will cite fewer violations at inspections when they are at risk of working beyond the typical end of their shift.*

3.4. Empirical Analysis

3.4.1. Empirical Context: Food Safety Inspections

Our hypotheses are ideally tested in an empirical context in which inspectors work individually, which avoids the challenge of discerning individuals' behaviors from those of co-inspectors. Food safety inspections conducted by local health departments fulfill this criterion because environmental health officers are individually responsible for the inspection of restaurants, grocery stores, and other food-handling establishments to protect consumers by monitoring compliance and educating kitchen managers in their assigned geographical area. Moreover, food safety inspections, commonly known as restaurant health inspections despite their broader scope, are designed to minimize foodborne illness; noncompliance can jeopardize consumer health. The quality of these assessments—and their ability to safeguard public health—depends on the accuracy of inspectors.

Foodborne disease in the United States is estimated to cause 48 million illnesses resulting in 128,000 hospitalizations and 3,000 deaths each year, imposing billions of dollars of medical costs and costs associated with reduced productivity and with pain and suffering (Scallan et al. 2011, Scharff 2012, Minor et al. 2015). Violations can affect firms' reputations and revenues and can trigger organizational responses that range from additional training for responsible personnel to legal representation to refute citations.

Several prior studies have examined food safety inspections. For example, Lehman, Kovács, and Carroll (2014) found that consumers are less concerned about food safety at restaurants that they perceive to be more "authentic." Others have investigated the extent to which restaurants improved hygiene practices once they were required to disclose their inspection results to consumers via restaurant grade cards (Jin and Leslie 2003, Simon et al. 2005, Jin and Leslie 2009). More recent studies have found that online customer reviews of restaurants contain text related to hygiene conditions that can predict health inspection results (Kang et al. 2013. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews) and can increase inspector effectiveness if health inspection agencies take them into account when prioritizing establishments for inspection (Glaeser et al. 2016).

Because inspectors need evidence to justify citing violations (and thus can only cite violations if they are truly present), studies of inspection bias (e.g., Bennett et al. 2013, Duflo et al. 2013, Short, Toffel, and Hugill 2016) are based on the assumption that deviations from the true number of violations are only due to underdetection and that bias does not lead inspectors to cite nonexistent violations. This assumption was validated in our interviews with inspectors and underlies our empirical approach. Moreover, because violations are based on regulations that are based on science-based guidance for protecting consumers, each violation item is relevant.

We purchased data from Hazel Analytics, a company that gathers food safety inspections from several local governments across the United States, processes the information to create electronic datasets, and sells these datasets to researchers and to companies—such as restaurant chains—interested in monitoring their licensees. These datasets include information about the inspected establishment (name, identification number, address, city, state, ZIP code), the inspector, the inspection type, the date, the times when the inspection began and ended, the violations recorded, and, where available, the inspector's comments on those violations.

We purchased all of Hazel Analytics' inspection datasets that included inspection start and end times as well as unique identifiers for each inspector, all of which are necessary to observe inspector schedules. This included all food safety inspections conducted in Lake County, Illinois, from September 4, 2013, to October 5, 2015; in Camden County, New Jersey, from September 4, 2012, to September 24, 2015; and in Alaska from December 8, 2007, to October 4, 2015. (These date ranges reflect all inspections from these domains that Hazel Analytics had coded.) Our estimation sample omits (a) inspector-days for which we cannot adequately calculate relevant variables based on what appear to be data entry errors that we were unable to correct (for example, when there was ambiguity about inspection sequence) and (b) inspections that are dropped by our conditional fixed-effects Poisson specification. This results in an estimation sample containing 12,017 inspections of 3,399 establishments conducted by 86 inspectors on 6,880 inspector-days in Camden County, New Jersey (1,402 inspections), Lake County, Illinois (8,962 inspections), and Alaska (1,653 inspections). These sample restrictions do not affect our inferences, as all of our hypothesized results continue to hold when using alternative specifications estimated on all inspections in the raw dataset (results not reported).

Our interviews with managers and inspectors at health inspection departments represented in our dataset indicate that inspectors have limited discretion over scheduling. Each inspector is responsible for inspecting all establishments within his or her assigned geographic territory. Inspectors are rotated to different territories every two or three years. Inspectors are instructed to schedule their inspections by prioritizing establishments based on their due dates, which are computed based on previous inspection dates and the required inspection frequency for each establishment type (which varies based on the riskiness of their operations). To minimize travel time, inspectors are instructed to group inspections with similar due dates by geographic proximity.

Though inspectors also carry out many administrative duties (such as reviewing records, answering emails, and attending department meetings at the office), the bulk of their work is inspections and the associated travel. As they prepare to conduct inspections, inspectors review the establishments' most recent inspections. Traveling between their office and establishments to inspect often accounts for a substantial portion of inspectors' days because of the geographical dispersion in the areas covered by our data. Inspectors are discouraged from working overtime.

When inspectors arrive to an establishment, they ask to speak to the person in charge and encourage this person to accompany them during the inspection. During the inspection, they inspect the establishment (e.g., taking temperatures), observe workers' behaviors (e.g., whether and how they use gloves and wash their hands), and ask many questions to understand the processes (e.g., receiving or the employee health policy). As they walk through the establishment, the inspectors point out the violations they find, explain the public health rationale, and ask the personnel to correct them straightaway when possible. Though any immediately-corrected violations are still marked as violations on the inspection form, this approach ensures that (a) the violations are corrected as soon as possible to improve food safety and (b) the personnel learn how to be compliant. Because of the immediate corrections, the instruction about regulations and how to improve the processes in the future, and the incentive for compliance resulting from effective monitoring and enforcement, whether each violation is cited or not has a real impact on public health. Thus, reducing the underreporting of violations resulting from the effects we identify would improve actual compliance and health outcomes.

3.4.2. Measures

3.4.2.1. Dependent and independent variables. We measure *violations* as the number of violations cited in each inspection, a typical approach used by others (e.g., Helland 1998, Stafford 2003, Langpap and Shimshack 2010, Short, Toffel, and Hugill 2016).

Prior inspected establishment's violations is the number of violations the inspector cited at the establishment inspected prior to the focal inspection, whether minutes or days earlier.

Prior inspected establishment's violation trend is calculated as the percentage change in the number of violations at that establishment between that day's inspection and its previous inspection (we added one to the denominator to avoid dividing by zero).

We create two indicator variables to distinguish whether the inspector's prior establishment had improved, deteriorated, or not substantially changed in its number of violations compared to its previous inspection. We classify an establishment's violation trend as *improved saliently* (or *deteriorated saliently*) if its current inspection yielded at least two fewer (more) violations than its previous inspection. (The intermediate case, in which the number of violations differed by only one or remained constant, is the baseline condition.) We create the dummy variables *prior inspected establishment saliently improved*, coded 1 when the inspector's prior inspected establishment *improved saliently* and 0 otherwise, and *prior inspected establishment saliently deteriorated*, coded 1 when the inspector's prior inspected establishment *deteriorated saliently* and 0 otherwise. An inspection conducted immediately after the inspection of an establishment whose performance change was only one or no violations is considered the baseline condition.

We measure an inspector's schedule-induced fatigue at a given inspection as the *number of prior inspections today*, which is the number of inspections that the inspector had already conducted before the focal inspection on the same day. Thus, this variable is coded 0 for an inspector's first inspection of the day, 1 for the second, and so on. (Our results are robust to measuring schedule-induced fatigue in three alternative ways, as described in the robustness test section.)

To measure whether an inspection might reasonably be anticipated to conclude after the inspector's typical end-of-shift time, we created an indicator variable, *potentially shift-prolonging*, coded 1 when the anticipated end time of an inspection (calculated as the inspection start time plus the duration of that establishment's previous inspection conducted by any inspector) falls after the inspector's running average daily clock-out time based on all of that inspector's preceding days in our sample, and coded 0 otherwise.

3.4.2.2. Control variables. We measure *inspector experience* as the number of inspections the inspector had conducted (at any establishment) since the beginning of our sample period by the time he or she began the focal inspection.

We create an indicator variable, *returning inspector*, coded 1 when the inspector of the focal inspection had inspected the establishment before, and 0 otherwise.

We create two indicator variables to designate the time of day the inspection began: *breakfast period* (midnight to 10:59 am) and *dinner period* (4:00 pm–11:59 pm), with the remaining *lunch period* (11:00 am–3:59 pm) as the (omitted) baseline condition. We also create a series of indicator variables specifying the month and the year of the inspection.

We create a series of indicator variables to control for whether the inspection is the *establishment's* n*th inspection (second through tenth or more)*, each of which indicates whether an inspection is the establishment's first, second, third (and so on) inspection in our sample period.

We create a series of inspection-type dummies to indicate whether the inspection was (a) routine, (b) routine-education, (c) related to permitting, (d) due to a complaint, (e) an illness investigation, or (f) a follow-up. *Routine inspections* are conducted to periodically monitor establishments; *routine-education inspections* are particular cases of routine inspections in which an educational presentation is conducted to train establishment staff. These two types make up 79% of the inspections in our estimation sample. *Permit inspections* are conducted when establishments change ownership or undergo construction, upgrades, or remodeling. *Complaint inspections* are triggered by the local health department receiving a complaint. Because Camden logs complaints dates and the inspectors assigned to investigate complaints, but does not classify particular inspections as triggered by complaints, *complaint risk inspections* refers to all inspections those inspectors conducted the day—and the day after—they were assigned to investigate a complaint. *Illness investigation inspections* are those conducted to investigate a possible foodborne illness (food poisoning). A *follow-up inspection* (or re-inspection) is conducted to verify that violations in a preceding inspection have been corrected and thus is of limited scope. *Other inspections* includes visits to confirm an establishment's deactivation/closure and inspections of mobile establishments, vending machines, and temporary events such as outdoor festivals; this is the omitted category in our empirical specifications.

Table 3.1 reports summary statistics. Additional descriptive statistics are provided in Appendix 3.A in the online supplement.

3.4.3. Empirical Specification

We test our hypotheses by estimating the following model:

$$Y_{ijen} = F(\beta_1\, \rho_{i,j-1} + \beta_2\, \delta_{i,j-1} + \beta_3\, \eta_{ij} + \beta_4\, \lambda_{ij} + \beta_5\, \varphi_j + \beta_6\, \mu_{ijen} + \beta_7\, \tau_{ijen} + \beta_8\, \nu_n + \beta_9\, \gamma_{ijen} + \beta_{10}\, \mathrm{IE}_{ie} + \varepsilon_{ijen}),$$

where Y_{ijen} is the number of *violations* cited in the *n*th inspection of establishment *e* that was conducted by inspector *i* and that was his or her *j*th inspection in our sample. F(·) refers to the Poisson function.

$\rho_{i,j-1}$ is the inspector's *prior inspected establishment's violations*; that is, the number of violations that inspector *i* cited at the immediately preceding inspection of another establishment. $\delta_{i,j-1}$ refers to the *prior inspected establishment's violation trend* or, in some specifications, the two variables that indicate particular ranges of that variable: *prior inspected establishment saliently improved* and *prior inspected establishment saliently deteriorated*. η_{ij} is inspector *i*'s *number of prior inspections today*. λ_{ij} refers to whether the inspection was *potentially shift-prolonging.*

We include φ_j to control for *inspector experience* (Macher, Mayo, and Nickerson 2011, Short, Toffel, and Hugill 2016). We control for *returning inspector* (μ_{ijen}) because inspectors who return to an establishment they had inspected before tend to behave differently than inspectors who are there for the first time (Short, Toffel, and Hugill 2016, Ball, Siemsen, and Shah 2017).

The vector τ_{ijen} includes *breakfast period* and *dinner period* to control for the possibility that an establishment's cleanliness might vary over the course of a day and because prior research indicates that many individual behaviors are affected by time of day (Linder et al. 2014, Dai et al. 2015). τ_{ijen} also includes two sets of fixed effects for the month and for the year of the inspection.

We include a series of fixed effects, ν_n, to control for the *establishment's* n*th inspection (second through tenth or more)* because research has shown that other types of establishment improve compliance over subsequent inspections (Ko, Mendeloff, and Gray 2010, Toffel, Short, and Ouellet 2015).

Because different types of inspection might mechanically result in different numbers of violations (e.g., due to different scopes), the model includes *inspection type* dummies (γ_{ijen}).

Finally, we include fixed effects for every inspector-establishment combination (IE_{ie}). These inspector-establishment dyads control for all time-invariant inspector characteristics (such as gender, formal education, and other factors that might affect their average stringency) and all time-invariant establishment characteristics (such as cuisine type and neighborhood). Thus, our specification identifies changes in the number of violations that a particular inspector cited when inspecting a given establishment on different occasions. Including inspector-establishment fixed effects also avoids concerns that our results are driven by spatial correlation; specifically, the concern that proximate establishments that inspectors tend to visit sequentially might exhibit similar violation counts because they share neighborhood characteristics that might affect the supply of and demand for compliance. Our including fixed effects for inspector-establishment dyads is more conservative than including separate sets of fixed effects for inspectors and for establishments; a robustness test that includes these separate sets of fixed effects yields similar results.

3.4.4. Identification

We took several steps to ensure that our empirical approach tests our hypothesized relationships, controlling for or ruling out alternative plausible explanations. For example, the positive correlation between the number of violations that inspectors cite at a focal establishment and at their prior establishment could result not only from the mechanism represented in H1 but also if inspectors clustered on their schedules the establishments they expected to yield many (or few) violations. Our inspector interviews revealed that they in fact tended to cluster inspections of establishments near each other in order to minimize travel time. While violations might be spatially correlated due to demographic clustering, our inclusion of establishment-inspector-dyad fixed effects controls for such time-invariant establishment characteristics.

We test our hypothesis that inspector fatigue reduces inspector stringency by looking for evidence that fewer violations are cited at inspections conducted later in an inspector's daily sequence. But inspectors citing fewer violations as their schedule proceeds could have two other explanations. First, daily trends in customer visits, staffing levels, and staff cleaning effort could result in establishments exhibiting better hygiene conditions later in the day, when inspections are more likely to be conducting their second and subsequent inspections of their day's schedule. Our inspector interviews indicated that many violations reflect longer-term problems whose propensity does not change throughout the day (e.g., sinks functioning improperly) and that hygiene conditions often get worse (not better) as establishments serve more customers, which would bias against our hypothesized effect. Our specifications nonetheless include fixed effects for time of day to control for potential variation in establishments' cleanliness at different time periods of the day. Second, inspectors might intentionally schedule "dirtier" establishments—those with historically more violations and thereby expected to have more violations—earlier in their daily schedule, leaving "cleaner" establishments for later in their schedule. However, two supplemental analyses yielded no evidence that inspectors constructed their schedules this way. A simple correlation analysis reveals that an establishment's previous inspection violation count is not significantly related to when in an inspector's daily sequence its focal inspection is conducted (Pearson's $\chi^2 = 139$, $p = 1.00$; also see Figure 3.B1). Moreover, Poisson regression results enable us to rule out that inspectors intentionally sequenced, to any meaningful degree, their day's inspections based on establishments' prior violations. Specifically, a Poisson regression that predicts an establishment's place in the inspection sequence based on that establishment's prior violation count and inspector-day fixed effects yielded a tiny positive effect ($\beta = 0.009$, S.E. = 0.003, with standard errors clustered by inspector-day), indicating that more violations in a prior inspection predicts that their subsequent inspection will be scheduled slightly *later* in the inspector's shift, which biases against our hypothesized effect.

Finally, test our hypothesis that an inspection being potentially shift-prolonging reduces inspector stringency by assessing whether shift-prolonging inspections yield fewer violations. However, shift-prolonging inspections might also yield fewer violations if, as an inspector's normal shift end-time approaches, he or she intentionally chooses to inspect establishments anticipated to yield fewer violations in order to minimize how late he or she will need to work, presuming "cleaner" establishments can be inspected more quickly. Two supplemental analyses, however, rule that out. First, establishments that had shift-prolonging inspections averaged 3.1 violations in their prior inspection, significantly *more* than the 2.3 average prior violations among establishments whose inspections were not potentially shift-prolonging (Pearson's $\chi^2 = 243$, $p < 0.01$). Second, a logistic regression indicates that the probability of an establishment's inspection being potentially shift-prolonging slightly increases if its previous inspection yielded more violations. Specifically, regressing a dummy indicating whether an establishment's inspection is *potentially shift-prolonging* on the violation count from its previous inspection and inspector-day fixed effects yield a significant positive coefficient on the violation count ($\beta = 0.104$, S.E. = 0.013, clustered by inspector-day). Both results bias against our hypothesized effect.

3.4.5. Results

3.4.5.1. Model results. We estimate the count model using fixed-effects Poisson regression and report standard errors clustered by establishment (Table 3.2). Poisson panel estimators are consistent even if the data are not Poisson distributed, provided the conditional mean is correctly specified (Azoulay, Graff Zivin, and Wang 2010, Cameron and Trivedi 2010). Because of the weaker distributional assumption of the Poisson panel estimators, they may be more robust than negative binomial regression (Cameron and Trivedi 2010).

Our results are robust to several alternatives: clustering standard errors by inspector, estimating the model with negative binomial regression with conditional fixed effects, and estimating the model using ordinary least squares regression predicting log violations. Multicollinearity is not a serious concern, given that variance inflation factors (VIFs) are less than 1.7 for all hypothesized variables and less than 6.1 for all variables except three of the inspection-type indicators. Because our specifications control for a variety of factors that affect the number of violations cited, we interpret coefficients on the hypothesized variables as evidence of bias, as done in prior studies (e.g., Chen, Moskowitz, and Shue 2016, Short, Toffel, and Hugill 2016). Because deviations from the true number of violations are assumed to result only from underdetection (as described above), we interpret negative coefficients to indicate the extent of underdetection occurring, whereas positive coefficients indicate the extent to which underdetection is avoided. We interpret effect sizes based on incidence rate ratios (IRRs).

We test Hypotheses 1, 2, 4, and 5 using Model 1. We begin by interpreting the coefficients on our control variables. The estimated coefficient on *inspector experience* is positive and statistically significant, suggesting that, all else constant, the number of violations cited per inspection increases as the inspector conducts inspections over time, albeit by a small amount on an inspection-by-inspection level. The negative and statistically significant coefficient on *returning inspector* (β = -0.116, p < 0.01) indicates that inspectors who return to an establishment cite 11% fewer violations than inspectors who had not inspected that establishment before, which is also consistent with prior studies. Considering time-of-day effects, we note that, on average, inspections conducted earlier in the day cite 6% more violations than inspections conducted during the lunch period, whereas inspections conducted during the dinner and lunch periods cite statistically indistinguishable numbers of violations. The estimated coefficients on the *establishment's* n*th inspection* (not reported) indicate that fewer violations were cited at successive inspections of a given establishment, a result consistent with prior research on other types of inspection. For example, the estimated coefficient on the dummy variable denoting an establishment's third inspection (β = -0.209, p < 0.01) indicates that those inspections cite 19% fewer violations on average than its initial inspection.

To explore the influence of the outcome at the inspector's prior inspected establishment, we first consider the number of violations cited in that inspection. The coefficient on *prior inspected establishment's violations* is positive and statistically significant (β = 0.015, p < 0.01), which supports H1. Each additional citation at the establishment inspected immediately before the focal inspection increases the number of violations cited in the focal inspection by 1.51%. The statistically significant positive coefficient on *prior inspected establishment's violation trend* (β = 0.013, p < 0.05) supports H2. A one-standard-deviation increase in this trend increases the number of citations in the focal inspection by 1.31%. Note that this is in addition to the effect of the number of violations (H1).

To test H3, Model 2 replaces *prior inspected establishment's violation trend* with the indicator variables *prior inspected establishment saliently improved* and *prior inspected* *establishment saliently deteriorated*. The baseline condition occurs when the prior inspected establishment exhibited no more than one violation more or less than it did in its previous inspection. Compared to this baseline condition, we find that inspectors cite more violations in inspections conducted after their prior inspected establishment exhibits salient deterioration ($\beta = 0.075$, $p < 0.01$). The IRR indicates that, on average, an inspector who has just inspected an establishment with salient deterioration will report 8% more violations in the focal inspection. However, we find no evidence that observing salient improvement in the prior inspected establishment has any effect on the number of violations cited in the focal inspection. A Wald test indicates that these effects significantly differ (Wald $\chi^2 = 4.21$, $p < 0.05$), which supports H3: the spillover effect on the focal inspection of having observed salient deterioration in the prior inspected establishment is statistically significantly stronger than the spillover effect of having observed salient improvement.

Model 1 also supports both of our hypothesized daily schedule effects. The negative, statistically significant coefficient on *number of prior inspections today* ($\beta = -0.032$, $p < 0.01$) indicates that each subsequent inspection during the inspector's workday cites 3.15% fewer violations, which supports H4. Applying this 3.15% effect to the 2.42 average violations per inspection yields an average marginal effect of 0.08 fewer violations being cited per inspection for each subsequent inspection conducted throughout the day. This amounts to 80 violations not being cited for every 1,000 "second inspections of the day," 160 violations not being cited for every 1,000 "third inspections of the day," and so on.

The negative statistically significant coefficient on *potentially shift-prolonging* ($\beta = -0.052$, $p < 0.05$) indicates that inspections that risked extending an inspector's workday result in 5.07% fewer citations, as predicted by H5. Applying this 5.07% to the 2.42 average violations yields an average marginal effect of 0.14 fewer violations being cited in each potentially shift-prolonging inspection. This amounts to 140 violations not being cited for every 1,000 potentially shift-prolonging inspections, a substantial number given 26% of inspections in our sample are *potentially shift-prolonging*.

3.4.5.2. Results interpretation. Our main results indicate that inspectors, despite their effort and training, are vulnerable to decision biases that lead them to underreport violations in predictable ways. These biases represent monitoring failures. If inspectors' detection rates were improved so they cited the violations that are currently unreported due to the scheduling biases we identify, those additional citations would lead establishments to improve their food safety practices. Thus, these behavioral effects have real implications because they affect citation rates of actual violations. These additional citations can improve compliance in two ways. They can improve compliance immediately because those violations that can be rectified instantly are. These additional citations can also improve compliance in the future because they motivate establishments to improve processes that not only prevent the cited violations from reoccurring but also and more broadly motivates compliance effort to prevent other violations, which is the deterrent intent of monitoring and enforcement. Thus, citations prompt behavioral responses that improve compliance, which in turn prevent foodborne health incidents.

Prior research that reveal decision biases tend to focus only on quantifying the magnitudes of such bias. Improving the accuracy of inspectors' citations of violations is in itself a very important outcome, one that organizations and governments care deeply about. We go beyond that typical approach to contributing to the literature by not only quantifying new sources of decision biases but also by estimating their real-world consequences. Our efforts to translate our primary findings (how scheduling affects the citations of violations) into its broader societal impacts (health consequences) would be equivalent to, for example, Chen, Moskowitz, and Shue (2016) not only quantifying a source of bias among umpires in calling balls and strikes (which that paper does), but also estimating how a team's win/loss record would be affected if the baseball umpires were to call pitches more accurately without the identified bias (which that paper does not consider). Similarly, it would be akin to that paper not only revealing an important source of asylum judges' decision bias, but also estimating the social injustice created by those erroneous decisions.

Thus, to better understand the potential benefits that would arise by addressing these biases, we develop nationwide estimates of how many fewer violations would be underreported—and the consequent healthcare outcomes and costs that would be avoided—if inspection managers implemented measures such as better awareness, new training, and different scheduling regimes that would somewhat attenuate these biases. We consider the impact of interventions that would exploit outcome-effects and ameliorate daily schedule effects that would lead inspectors to cite violations that currently go underreported. We estimate the effects of such interventions on the average inspection based on our sample, scale up the results to estimate how many currently undetected violations would be cited nationwide, and then estimate how many fewer foodborne illness cases and hospitalizations would result and by how much that would reduce associated healthcare costs. Our methodology and results (including assumptions and caveats) are described in Appendix 3.D, but we briefly report some key results here.

In the ideal scenario, the outcome effects (which increase scrutiny) would be fully triggered all the time and the daily schedule effects (which erode scrutiny) would be entirely eliminated. In practice, different interventions would have different degrees of effectiveness. Figures 3.2a-c illustrate a range of scenarios. If the drivers of outcome and daily schedule effects were respectively amplified and reduced by 100%, inspectors would cite 9.9% more violations, yielding 240,999 additional violations being cited annually nationwide, which would result in 50,911 fewer foodborne illness related hospitalizations and 19.01 million fewer foodborne illness cases, and would reduce foodborne illness costs by $14.20 billion to $30.91 billion. A 50% scenario, which amplifies the outcome effects by 50% and mitigates the daily schedule effects by 50%, would generate half of these gains, resulting in 115,571 additional violations being cited, 24,415 fewer hospitalizations, 9.1 million fewer foodborne illness cases, and savings of $6.81 billion to $14.83 billion in foodborne illness costs. Even a very conservative scenario, which diminishes the daily schedule effects by 10% and triggers the outcome effects by 10%, elicits substantial benefits. It would yield 22,376 additional violations cited annually nationwide, 4,727 fewer foodborne illness related hospitalizations and 1.77 million fewer foodborne illness cases, and would reduce foodborne illness costs by $1.32 billion to $2.87 billion.

3.4.6. Robustness Tests

We conduct several analyses to confirm the robustness of our findings. Our primary results are based on a conservative approach that includes establishment-inspector–dyad fixed effects. We find similar results whether we instead include establishment fixed effects or separate sets of fixed effects for inspectors and for establishments (estimating the latter with Poisson regression led to convergence problems that led us to instead use OLS regression to predict *log (violations+1)*). Also, to assess whether unusually busy days, which might lead inspectors to become especially fatigued, might be driving our schedule-induced fatigue (H4) results, we reestimated our models on the subsample of inspector-days with no more than six inspections (the 99th percentile). Our hypothesized results are robust to these subsample tests.

Our results regarding the effects of schedule-induced fatigue (H4) hold even when we measure this construct using any of the following three alternative approaches rather than the number of prior inspections on the day of the focal inspection. In our first alternative, we calculate the *actual cumulative minutes* inspectors spent onsite in their prior inspections that day to better account for the fact that some inspections take longer than others and that longer (and not just more numerous) inspections are likely to cause more fatigue. Our second alternative approach accommodates the potential concern that fatigue increased the duration of prior inspections. Here, we calculate the *anticipated cumulative minutes* inspectors would expect to have spent onsite in their prior inspections that day, computed as the average of the durations of those establishments' previous two inspections (or their single previous inspection if only one is available). In our third alternative approach, we compute the *predicted cumulative minutes* inspectors would spend onsite in their prior inspections that day, using the predicted durations derived from an ordinary least squares regression model, with a log-transformed outcome variable and including the covariates from the corresponding main specification.

Our results are also robust to including, as additional controls in our primary models, indicator variables denoting the day of the week the inspection occurred. Our results are mostly robust to substituting our three time-of-day periods (*breakfast period*, *lunch period*, and *dinner period*) with indicator variables for each hour of the day at which the inspection occurred. Only the *potentially shift-prolonging* coefficient is no longer statistically significant, likely due to the higher multicollinearity introduced by this approach.

Finally, our results are robust to controlling for *weekly workload* or *monthly workload,* measured as the number of inspections the inspector conducted the week or the month of the focal inspection. As an aside, the estimated coefficient on workload is not significant, suggesting that despite the extensive prevalence of workload effects in other settings, inspectors in our sample are resilient to them; that is, we find no evidence that workload pressures affect inspector accuracy. This shows that inspection outcomes are difficult to influence and makes our identified effects even more impressive (Prentice and Miller 1992).

3.4.7. Extensions

We conduct additional analysis to examine the persistence of some of our outcome-effects. To explore whether these outcome-effects persist beyond the next inspection, we added two additional variables to our models: the *penultimate inspected establishment's violations* (that is, two establishments ago) and then also the *antepenultimate inspected establishment's violations* (that is, three establishments ago). The significant positive coefficients on both of these variables indicates that the number of violations cited in an inspection is significantly affected not only by the violations at the inspector's immediately preceding inspection, but also by each of the two inspections before that; the declining magnitudes of these coefficients indicates that the effect successively dissipates (Table 3.3, columns 1-2).

We also explore whether the presence or magnitude of outcome-effects relied on inspectors conducting inspections in rapid succession. First, we assessed whether the outcome effects were affected by how much time had lapsed since the inspector's prior inspection by adding to our primary model *tens of hours since the inspector's prior inspection* (top coded at its 99th percentile to avoid outliers influencing results) and its interaction with the two outcome-effects variables, *prior inspected establishment's violations* and *prior inspected establishment's violation trend*. Neither interaction coefficient is statistically significant, which yields no evidence that time between inspections attenuates outcome-effects (Table 3.3, column 3). (As an aside, the non-significant coefficient on *tens of hours since the inspector's prior inspection* provides no evidence that more time between inspections increases the inspector's violation detection rate, in contrast with prior research that longer breaks "recharge" decision makers in other settings (Danziger, Levav, and Avnaim-Pesso 2011)). Second, we assessed whether the outcome effects attenuated if an inspector's successive inspections occur across different days, as opposed to on the same day. To test this, we replaced *prior inspected establishment's violations* with two variables: *prior inspected establishment's violations for the first inspection of the day* (coded as *prior inspected establishment's violations* for the first inspection of the day, and 0 otherwise) and *prior inspected establishment's violations for the second+ inspection of the day* (coded 0 for the first inspection of the day, and as *prior inspected establishment's violations* otherwise). Finding nearly identical significant negative coefficients on both variables that are statistically indistinguishable (Wald $\chi^2 = 0.03$, $p = 0.86$) indicates that an overnight break did not attenuate the outcome-effects (column 4). Together, these results rule out the inspector's mood or other temporary factors as driving the outcome-effects.

We also investigate the extent to which our hypothesized effects influence the citing of two types of violation: (a) *critical violations*, which are related to food preparation practices and employee behaviors that more directly contribute to foodborne illness or injury, and (b) *noncritical violations*, which are overall sanitation and preventative measures to protect foods, such as proper use of gloves, that are less risky but also important for public health. We find that our daily schedule effects only influence citing noncritical violations, whereas our outcome-effects influence citing both types; results are reported in Appendix 3.B in the online supplement.

We also examine whether our hypothesized effects influence other aspects of inspections that might be linked to scrutiny. We find that inspectors conduct inspections more quickly as they progress through their shifts: inspection duration (the number of minutes between its start time and end time) decreases by 3.5% for each subsequent inspection of the day; results are reported in Appendix 3.C in the online supplement. Moreover, the inspector's citation pace—*violation citations per hour*, a measure of productivity in this setting, representing the net of the effects on violations and inspection duration—decreases by 1.3% for each subsequent inspection of the day. *Potentially shift-prolonging* inspections are conducted 3.6% more quickly but citation pace remains largely unaffected; thus, our main finding that *potentially shift-prolonging* inspections result in fewer violations is likely due to inspectors' desire to avoid working late, rather than to fatigue eroding their citation pace. We find no evidence of outcome-effects on inspection duration and conclude that our main outcome-effect findings—that more violations and worsening trends at an inspector's prior establishment increase the inspector's citations at his or her next inspection—result from inspectors increasing their citation pace rather than from spending more time onsite.

Finally, we examine whether our hypothesized effects are associated with documentation effort. We find no evidence that average violation comment length (in characters or words) is influenced by *number of prior inspections today*, *potentially shift-prolonging*, or *prior inspected establishment's violations* (results not reported). Inspectors document the focal inspection with shorter comments when the prior establishment exhibited worsening violation trends. Thus, a potential mechanism by which such trends might increase citation pace (that is, improve inspectors' productivity in citing violations) is by shifting some effort from documentation to detection. Because each violation citation references the regulatory code infringed and only on some occasions does customization of violation comments provide additional value, we interpret these results as inspectors redirecting their attention to important matters.

3.5. Discussion

We find strong evidence that inspectors' evaluations are affected by their daily schedules and by their experience at the prior establishment they inspected. As inspectors conduct inspections throughout the day, their scrutiny is eroded by increasing fatigue and by the perceived time pressure to complete their inspections before the typical end of their shift. We also find strong evidence that inspectors' scrutiny is influenced by their experience at their prior inspected establishment. The effect magnitudes that we identify, ranging from 1.3% to 7.8% individually and 9.9% overall, are large compared to decision bias among professionals in other field settings—such as a 0.5% effect size regarding decision bias exhibited by judges, 0.9% by baseball umpires, and 2.1% to 6.9% by social auditors—and are similar in magnitude to experimental results yielding biases of 0 to 8 percentage points by loan review officers (Chen, Moskowitz, and Shue 2016, Short, Toffel, and Hugill 2016).

3.5.1. Contributions

Our work contributes to three literature streams. First, this study is among the first to bring an operational lens to the literature on monitoring and assessment of standards adherence. In particular, we identify important scheduling effects on the scrutiny and thus the accuracy of those who monitor establishments' adherence to standards. We contribute to this literature's focus on improving monitoring schemes' effectiveness by analyzing how inspection outcomes are affected by outcomes of prior inspections at other establishments and by inspectors' daily schedules.

Second, by identifying spillover effects between inspections, our findings contribute to a related literature on the spillover effects of regulatory sanctions (e.g., Cohen 2000, Shimshack and Ward 2005). While that literature focuses on how an inspection agency's monitoring efforts and enforcement actions affect its reputation for stringency, which has a spillover influence on other establishments' compliance, our study focuses on how inspectors' experiences at one establishment have spillover effects on their scrutiny at others. Ours is thus the first study of which we are aware that identifies spillover effects on inspector stringency associated not only with the outcomes of the immediately preceding inspection, but also with how many prior inspections an inspector had already conducted that day and with the inspector's apparent desire to avoid working late. Moreover, our work contributes to the nascent literature on the accuracy of inspections—specifically, of regulatory regimes and third-party monitoring of labor conditions in supply chains—that has largely focused on inspector bias due to economic conflicts of interest, team composition, and site-specific experience (e.g., Duflo et al. 2013, Short and Toffel 2016, Short, Toffel, and Hugill 2016, Ball, Siemsen, and Shah 2017). To our knowledge, our study is the first to bring the operational lens of scheduling to this literature by showing how work schedules can drive inaccuracies.

Third, we contribute to the literature on the performance implications of scheduling and task sequencing. By examining actual decisions with important consequences for consumers, we contribute to the recent attempts to explore high-stakes decision making in field settings (e.g., Chen, Moskowitz, and Shue 2016). The idiosyncrasies of quality-evaluation decisions result in biases that are different from those for other types of decision. In contrast to a prior study that finds that judges, loan reviewers, and baseball umpires are more likely to make an "accept" decision following a "reject" decision (and vice versa) (Chen, Moskowitz, and Shue 2016), we find the opposite relationship among inspectors' decisions over time, suggesting that their prior tasks affect their emotions and perceptions. This disparity could be due to inspectors monitoring rather than making predictions, so that their beliefs about the underlying probabilities are likely different from those in the settings that Chen, Moskowitz, and Shue (2016) examined. Moreover, inspectors can affect future compliance through their interactions during inspections (e.g., by offering ideas on how to remedy violations). Further research is needed to identify circumstances under which decisions are similar or opposite to prior decisions. Further, we find that this effect was asymmetric: prior negative outcomes are much more influential than prior positive ones.

In contrast to prior research that finds judges becoming more stringent as they make more decisions over the course of a day (Danziger, Levav, and Avnaim-Pesso 2011), we find that inspectors become less stringent. Perhaps one way to resolve this apparent contradiction is to note that in both studies, decision-makers appear to increasingly exhibit status-quo bias as they make decisions over a day. Though judges can reject without justification, inspectors must find proof of violations, which requires physical and mental effort to engage in social interactions with establishment staff (and the resulting impact on emotions and perceptions). Fatigue associated with additional inspections can impede violation detection and thus inspection quality.

Our daily schedule effects findings also complement the literature that has found that increased worker fatigue after long hours has led to accidents among nuclear and industrial plant operators, airline pilots, truck drivers, and hospital workers (Dinges 1995, Landrigan et al. 2004). In response to such findings, industry standards and regulations have capped the number of consecutive work hours in some of these professions; our results indicate that such policies might also improve inspection accuracy. We contribute to this debate by providing evidence of the negative effects of fatigue on work quality during normal shifts (rather than the very long work periods others have examined) in a different setting (health inspections), focusing on primary tasks (rather than secondary ones). Moreover, we investigate a different performance dimension (accuracy of quality assessments) and identify potential remedies. To the best of our knowledge, we are the first to provide evidence of the negative impact on quality assessment of within-day fatigue and potentially-shift-prolonging tasks. The results of our extension analysis suggest that inspectors themselves might attempt to ameliorate these effects by focusing on critical violations at the expense of detecting fewer noncritical violations and producing less documentation.

In addition, our finding that inspectors inspect less stringently as they approach the time they typically end their workday contributes to a broader understanding of how workers alter their procedures as they approach their end-of-shift (e.g., Chan's (2017) finding that hospital physicians concluding their shift accept fewer patients and make different decisions about patient care). More broadly, our work shows that even in a setting in which workers lack formal shifts and have some flexibility to schedule their own hours, workers behave differently as they approach the time they usually end their workday, suggesting that work is done differently toward the end of a workday in more settings than previously conceived. Our work also responds to the call for behavioral research in the operations management field (Bendoly, Donohue, and Schultz 2006) by identifying ways in which task sequencing affects worker behavior. Our finding that inspectors' experiences at prior inspections bias their subsequent inspections shows that the outcome of tasks can affect how humans—unlike machines—perform their next task.

3.5.2. Managerial Implications

Extrapolating our study's results to the approximately one million food-handling establishments monitored annually across the United States suggests that hundreds of thousands of violations of food safety regulations are likely being systematically overlooked each year. The effect is far larger if one considers food safety inspections conducted around the world, as well as inspections worldwide in other domains such as environmental, quality, and financial management. Moreover, overlooking even a single food safety violation matters beyond the correction of that violation because citing that violation could have increased the salience of food safety to the establishment’s personnel and thus spurred improvements. Personnel can get very upset if even one violation is cited (Rossen 2017). In our observations of food safety inspections, we witnessed several occasions in which the personnel of the inspected establishment were frustrated when an inspection yielded even one violation. Compared to the average of 2.42 violations cited per inspection, citing one fewer violation constitutes a 41% decrease, a large change that creates unfairness across facilities and impedes accurate decisions being taken in response to inspection reports. Thus, more accurate inspections that result in fewer violations being overlooked could prompt more effort to fully comply with food safety standards. For example, franchisors could be better equipped to interpret inspection reports so as to know which franchisees require more (or less) oversight.

Moreover, regulators and private-sector inspectors across industries can also take steps to mitigate these biases to create more accurate inspection reports, which would yield fairer and more comparable results across inspected establishments, generate more reliable information for consumers, and better motivate compliance. For example, one way to reduce the extent to which these biases erode inspection accuracy is to impose a cap on the number of inspections conducted by a given inspector each day in order to limit fatigue effects, although this risks reducing inspection capacity. Another approach, which can be used at the same time, is to minimize the number of shift-prolonging inspections by reallocating an inspector’s weekly schedule to reduce variation in the predicted completion time of their final inspection each day or by shifting administrative tasks (such as office meetings) from the beginning to the end of the day. Reorganization of inspectors’ schedules could eliminate these negative outcomes, which —

according to our interviews with health inspectors in the areas covered in our data as well as in other areas across the United States— might be possible without adding cost.

Our identified outcome-effects imply that increasing the salience of noncompliance and thus the need to enforce regulation could increase the number of violations detected. This suggests that reminders or other ways to increase such salience could be a lever for inspection managers to increase the stringency of inspectors, even if the information is already available to them and despite their innate desire to protect consumers.

Managers can also use our findings to develop policies to reduce the *consequences* of inspector biases eroding inspection accuracy. For example, understanding that scrutiny typically declines as inspectors (a) conduct successive inspections during the day and (b) conduct inspections that risk prolonging their shift, the inspectors themselves could be required to schedule establishments that pose greater risks earlier in the shift. By improving inspector effectiveness in the case of food safety, these changes could reduce risk to public health.

3.5.3. Limitations and Future Research

Our study has several limitations that could be explored in future research. Though our data contain details of inspections and citations, we do not observe inspectors' beliefs or their interactions with the establishment personnel. We find that inspectors cite fewer violations after inspecting establishments that had fewer violations. Perhaps they make less effort to find hidden violations and are more willing to take a coaching approach—emphasizing education over enforcement and training operators to operate with better hygiene for borderline violations—rather than writing citations. Possible extensions of our study could use observations of these actions to quantify how they are affected by scheduling. In addition, although our research context—food safety inspections—is common worldwide, it is just one of many types of inspection conducted by companies and governments. Future research should examine whether the relationships we identified hold in other contexts.

3.6. References

Anand G., Gray J., Siemsen E. 2012. Decay, shock, and renewal: Operational routines and process entropy in the pharmaceutical industry. *Organization Science* **23**(6) 1700-1716.

Azoulay P., Graff Zivin J.S., Wang J. 2010. Superstar extinction. *Quarterly Journal of Economics* **125**(2) 549-589.

Ball G., Siemsen E., Shah R. 2017. Do plant inspections predict future quality? The role of investigator experience. *Manufacturing & Service Operations Management* **19**(4) 534-550.

Ballou D.P., Pazer H.L. 1982. The impact of inspector fallibility on the inspection policy in serial production systems. *Management Science* **28**(4) 387-399.

Barsade S.G. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly* **47**(4) 644-675.

Baumeister R.F., Bratslavsky E., Finkenauer C., Vohs K.D. 2001. Bad is stronger than good. *Review of General Psychology* **5**(4) 323-370.

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

Bennett V.M., Pierce L., Snyder J.A., Toffel M.W. 2013. Customer-driven misconduct: How competition corrupts business practices. *Management Science* **59**(8) 1725-1742.

Berry Jaeker J.A., Tucker A.L. 2017. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* **63**(4) 1042-1062.

Brachet T., David G., Drechsler A.M. 2012. The effect of shift structure on performance. *American Economic Journal: Applied Economics* **4**(2) 219-246.

Cameron A.C., Trivedi P.K. 2010. *Microeconometrics using Stata, revised edition*. Stata Press, College Station, TX.

Chan D.C., Jr. 2017. The efficiency of slacking off: Evidence from the emergency department. *Econometrica* (forthcoming).

Chen D.L., Moskowitz T.J., Shue K. 2016. Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Quarterly Journal of Economics* **131**(3) 1181-1242.

Cohen M.A. 2000. Empirical research on the deterrent effect of environmental monitoring and enforcement. *Environmental Law Reporter News and Analysis* **30**(4) 10245-10252.

Corbett C.J. 2006. Global diffusion of ISO 9000 certification through supply chains. *Manufacturing & Service Operations Management* **8**(4) 330-350.

Dai H., Milkman K.L., Hofmann D.A., Staats B.R. 2015. The impact of time at work and time off from work on rule compliance: The case of hand hygiene in health care. *Journal of Applied Psychology* **100**(3) 846-862.

Danziger S., Levav J., Avnaim-Pesso L. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* **108**(17) 6889-6892.

Deo S., Jain A. 2015. *Slow first, fast later: Empirical evidence of speed-up in service episodes of finite duration*. SSRN Working Paper.

Dinges D.F. 1995. An overview of sleepiness and accidents. *Journal of Sleep Research* **4(S2)** 4-14.

Duflo E., Greenstone M., Pande R., Ryan N. 2013. Truth-telling by third-party auditors and the response of polluting firms: Experimental evidence from India. *Quarterly Journal of Economics* **128**(4) 1499-1545.

Feinstein J.S. 1989. The safety regulation of U.S. nuclear power plants: Violations, inspections, and abnormal occurrences. *Journal of Political Economy* **97**(1) 115-154.

Gawande K., Bohara A.K. 2005. Agency problems in law enforcement: Theory and application to the U.S. Coast Guard. *Management Science* **51**(11) 1593-1609.

Glaeser E.L., Hillis A., Kominers S.D., Luca M. 2016. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review* **106**(5) 114-118.

Gray J.V., Anand G., Roth A.V. 2015. The influence of ISO 9000 certification on process compliance. *Production and Operations Management* **24**(3) 369-382.

Gray J.V., Siemsen E., Vasudeva G. 2015. Colocation still matters: Conformance quality and the interdependence of R&D and manufacturing in the pharmaceutical industry. *Management Science* **61**(11) 2760-2781.

Green L.V., Savin S., Savva N. 2013. "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10) 2237-2256.

Handley S.M., Gray J.V. 2013. Inter-organizational quality management: The use of contractual incentives and monitoring mechanisms with outsourced manufacturing. *Production and Operations Management* **22**(6) 1540-1556.

Hasija S., Pinker E., Shumsky R.A. 2010. Work expands to fill the time available: Capacity estimation and staffing under Parkinson's Law. *Manufacturing & Service Operations Management* **12**(1) 1-18.

Helland E. 1998. The enforcement of pollution control laws: Inspections, violations, and self-reporting. *Review of Economics and Statistics* **80**(1) 141–153.

Hugill A.R., Short J.L., Toffel M.W. 2016. *Beyond symbolic responses to private politics: Examining labor standards improvement in global supply chains*.

Ibanez M.R., Clark J.R., Huckman R.S., Staats B.R. 2017. Discretionary task ordering: Queue management in radiological services. *Management Science* (forthcoming).

Jin G.Z., Leslie P. 2003. The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Quarterly Journal of Economics* **118**(2) 409-451.

Jin G.Z., Leslie P. 2009. Reputational incentives for restaurant hygiene. *American Economic Journal: Microeconomics* **1**(1) 237-267.

Kahneman D. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review* **93**(5) 1449--1475.

Kang J.S., Kuznetsova P., Choi Y., Luca M. 2013. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, WA. 1443-1448.

KC D.S., Terwiesch C. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50-65.

Kim S.-H. 2015. Time to come clean? Disclosure and inspection policies for green production. *Operations Research* **63**(1) 1-20.

Ko K., Mendeloff J., Gray W. 2010. The role of inspection sequence in compliance with the US Occupational Safety and Health Administration's (OSHA) standards: Interpretations and implications. *Regulation & Governance* **4**(1) 48-70.

Koh K., Rajgopal S., Srinivasan S. 2013. Non-audit services and financial reporting quality: Evidence from 1978 to 1980. *Review of Accounting Studies* **18**(1) 1-33.

Kök A.G., Shang K.H. 2007. Inspection and replenishment policies for systems with inventory record inaccuracy. *Manufacturing & Service Operations Management* **9**(2) 185-205.

Landrigan C.P., Rothschild J.M., Cronin J.W., Kaushal R., Burdick E., Katz J.T., Lilly C.M., Stone P.H., Lockley S.W., Bates D.W., Czeisler C.A. 2004. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine* **351**(18) 1838-1848.

Langpap C., Shimshack J.P. 2010. Private citizen suits and public enforcement: substitutes or complements? *Journal of Environmental Economics and Management* **59**(3) 235–249.

Lapré M.A., Mukherjee A.S., Van Wassenhove L.N. 2000. Behind the learning curve: Linking learning activities to waste reduction. *Management Science* **46**(5) 597-611.

Lee H.L., Rosenblatt M.J. 1987. Simultaneous determination of production cycle and inspection schedules in a production systems. *Management Science* **33**(9) 1125-1136.

Lehman D.W., Kovács B., Carroll G.R. 2014. Conflicting social codes and organizations: Hygiene and authenticity in consumer evaluations of restaurants. *Management Science* **60**(10) 2602-2617.

Levine D.I., Toffel M.W. 2010. Quality management and job quality: How the ISO 9001 standard for quality management systems affects employees and employers. *Management Science* **56**(6) 978-996.

Levine D.I., Toffel M.W., Johnson M.S. 2012. Randomized government safety inspections reduce worker injuries with no detectable job loss. *Science* **336**(6083) 907-911.

Linder J.A., Doctor J.N., Friedberg M.W., et al. 2014. Time of day and the decision to prescribe antibiotics. *JAMA Internal Medicine* **174**(12) 2029-2031.

Macher J.T., Mayo J.W., Nickerson J.A. 2011. Regulator heterogeneity and endogenous efforts to close the information asymmetry gap. *Journal of Law & Economics* **54**(1) 25-54.

Mani V., Muthulingam S. 2017. Does Learning from Inspections Affect Environmental Performance? – Evidence from Unconventional Well Development in Pennsylvania. *Manufacturing & Service Operations Management* forthcoming.

Marcora S.M., Staiano W., Manning V. 2009. Mental fatigue impairs physical performance in humans. *Journal of Applied Physiology* **106**(3) 857-864.

May P.J., Wood R.S. 2003. At the regulatory front lines: Inspectors' enforcement styles and regulatory compliance. *Journal of Public Administration Research and Theory* **13**(2) 117-139.

Minor T., Lasher A., Klontz K., Brown B., Nardinelli C., Zorn D. 2015. The per case and total annual costs of foodborne illness in the United States. *Risk Analysis* **35**(6) 1125-1139.

Muraven M., Baumeister R.F. 2000. Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin* **126**(2) 247-259.

Nickerson R.S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* **2**(2) 175-220.

Oliva R., Sterman J.D. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* **47**(7) 894-914.

Pautz M.C. 2009. Trust between regulators and the regulated: A case study of environmental inspectors and facility personnel in Virginia. *Politics & Policy* **37**(5) 1047-1072.

Pautz M.C. 2010. Front-line regulators and their approach to environmental regulation in Southwest Ohio. *Review of Policy Research* **27**(6) 761-780.

Pierce L., Toffel M.W. 2013. The role of organizational scope and governance in strengthening private monitoring. *Organization Science* **24**(5) 1558-1584.

Prentice D.A., Miller D.T. 1992. When small effects are impressive. *Psychological Bulletin* **112**(1) 160-164.

Ronen J. 2010. Corporate audits and how to fix them. *Journal of Economic Perspectives* **24**(2) 189-210.

Rossen J. 2017. *12 Secrets of restaurant health inspectors*. Retrieved 03-27-2018, http://mentalfloss.com/article/500853/12-secrets-restaurant-health-inspectors.

Rozin P., Royzman E.B. 2001. Negativity bias, negativity dominance, and contagion. *Personality & Social Psychology Review* **5**(4) 296-320.

Samuelson W., Zeckhauser R. 1988. Status quo bias in decision making. *Journal of Risk and Uncertainty* **1**(1) 7-59.

Scallan E., Griffin P.M., Angulo F.J., Tauxe R.V., Hoekstra R.M. 2011. Foodborne illness acquired in the United States—unspecified agents. *Emerging Infectious Disease journal* **17**(1) 16-22.

Scharff R.L. 2012. Economic burden from health losses due to foodborne illness in the United States. *Journal of Food Protection* **75**(1) 123-131.

Shah R., Ball G.P., Netessine S. 2016. Plant operations and product recalls in the automotive industry: An empirical investigation. *Management Science* **63**(8) 2439-2459.

Shimshack J.P. 2014. The economics of environmental monitoring and enforcement. *Annual Review of Resource Economics* **6**(1) 339–360.

Shimshack J.P., Ward M.B. 2005. Regulator reputation, enforcement, and environmental compliance. *Journal of Environmental Economics and Management* **50**(3) 519-540.

Short J.L., Toffel M.W. 2016. The integrity of private third-party compliance monitoring. *Administrative & Regulatory Law News* **42**(1) 22-25.

Short J.L., Toffel M.W., Hugill A.R. 2016. Monitoring global supply chains. *Strategic Management Journal* **37**(9) 1878–1897.

Simon P.A., Leslie P., Run G., Jin G.Z., Reporter R., Aguirre A., Fielding J.E. 2005. Impact of restaurant hygiene grade cards on foodborne-disease hospitalizations in Los Angeles County. *Journal of environmental health* **67**(7) 32-36.

Simonsohn U., Gino F. 2013. Daily horizons: Evidence of narrow bracketing in judgment from 10 years of M.B.A. admissions interviews. *Psychological Science* **24**(2) 219-224.

Staats B.R., Dai H., Hofmann D., Milkman K.L. 2017. Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science* **63**(5) 1563-1585.

Staats B.R., Gino F. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science* **58**(6) 1141-1159.

Stafford S.L. 2003. Assessing the effectiveness of state regulation and enforcement of hazardous waste. *Journal of Regulatory Economics* **23**(1) 27–41.

Toffel M.W., Short J.L., Ouellet M. 2015. Codes in context: How states, markets, and civil society shape adherence to global labor standards. *Regulation & Governance* **9**(3) 205-223.

Tversky A., Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* **185**(4157) 1124-1131.

Wright R.A., Junious T.R., Neal C., Avello A., Graham C., Herrmann L., Junious S., Walton N. 2007. Mental fatigue influence on effort-related cardiovascular response: Difficulty effects and extension across cognitive performance domains. *Motivation and Emotion* **31**(3) 219-231.

Table 3.1. Summary Statistics

Variable	Description	Mean	SD	Min	Max
Violations	Number of violations cited in the inspection	2.42	2.73	0	25
Prior inspected establishment's violations	Number of violations cited at the establishment inspected by the inspector immediately prior to the focal inspection	2.11	2.62	0	25
Prior inspected establishment's violation trend	Percentage change in the number of violations at that establishment between that day's inspection and its previous inspection (adding one to the denominator to avoid dividing by zero)	0.42	1.58	-0.95	23
Prior inspected establishment saliently improved	Indicates if the inspector's prior inspected establishment *improved saliently* (i.e., its current inspection yielded at least two *fewer* violations than its previous inspection)	0.24	0.43	0	1
Prior inspected establishment saliently deteriorated	Indicates if the inspector's prior inspected establishment *deteriorated saliently* (i.e., its current inspection yielded at least two *more* violations than its previous inspection)	0.21	0.41	0	1
Number of prior inspections today	Number of inspections that the inspector had already conducted before the focal inspection on the same day	0.94	1.10	0	9
Potentially shift-prolonging	Indicates if the anticipated end time of an inspection (calculated as the inspection start time plus the duration of that establishment's previous inspection conducted by any inspector) falls after the inspector's running average daily clock-out time based on all of that inspector's preceding days in our sample	0.26	0.44	0	1
Inspector experience	Number of inspections the inspector had conducted (at any establishment) since the beginning of our sample period by the time he or she began the focal inspection	520.09	303.30	1	1429
Returning inspector	Indicates if the inspector of the focal inspection had inspected the establishment beforehand	0.84	0.37	0	1
Establishment's *n*th inspection (second through tenth or more)	Indicators that indicate whether an inspection is the establishment's first, second, third (and so on) inspection in our sample period	4.04	2.15	1	20
Breakfast period (midnight to 10:59 am)	Indicates if the inspection began midnight to 10:59 am	0.32	0.47	0	1
Lunch period (11:00 am–3:59 pm)	Indicates if the inspection began 11:00 am–3:59 pm (omitted category)	0.66	0.47	0	1
Dinner period (4:00 pm–11:59 pm)	Indicates if the inspection began 4:00 pm–11:59 pm	0.02	0.15	0	1

N = 12,017 inspections

Table 3.2. How Inspectors' Schedules Influence Inspection Outcomes

	Dependent variable:	*violations*	
		(1)	(2)
H1	Prior inspected establishment's violations	0.015*** (0.004)	0.014*** (0.004)
H2	Prior inspected establishment's violation trend	0.013** (0.006)	
H3	Prior inspected establishment saliently improved		0.012 (0.023)
H3	Prior inspected establishment saliently deteriorated		0.075*** (0.027)
H4	Number of prior inspections today	-0.032*** (0.011)	-0.032*** (0.011)
H5	Potentially shift-prolonging	-0.052** (0.025)	-0.051** (0.025)
	Inspector experience	0.001*** (0.000)	0.001*** (0.000)
	Returning inspector	-0.116*** (0.035)	-0.118*** (0.035)
	Breakfast period (midnight to 10:59 am)	0.056** (0.025)	0.056** (0.026)
	Dinner period (4:00 pm–11:59 pm)	0.000 (0.078)	-0.002 (0.078)
	Month fixed effects	Included	Included
	Year fixed effects	Included	Included
	Establishment's *n*th inspection (second through tenth or more) fixed effects	Included	Included
	Inspection-type fixed effects	Included	Included
	Establishment x Inspector fixed effects	Included	Included
	Number of observations (inspections)	12,017	12,017

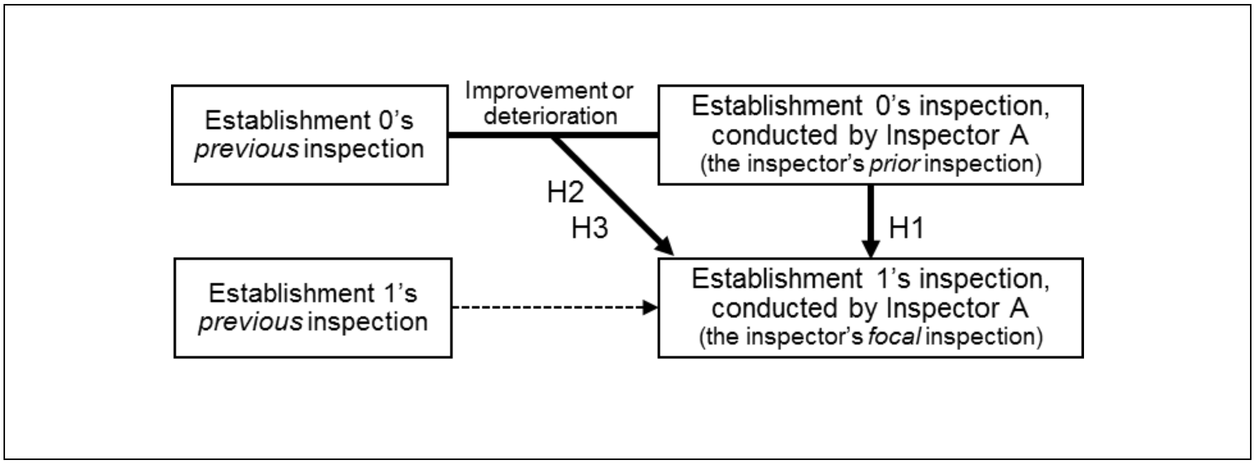
Notes: Poisson regression coefficients with robust standard errors clustered by establishment.
*** $p < 0.01$, ** $p < 0.05$

Table 3.3. Persistence of Outcome-effects

Dependent variable:	*Violations*			
	(1)	(2)	(3)	(4)
Prior inspected establishment’s violations	0.015***	0.015***	0.014***	
	(0.004)	(0.004)	(0.004)	
Penultimate inspected establishment’s violations	0.010***	0.009***		
	(0.003)	(0.003)		
Antepenultimate inspected establishment’s violations		0.006*		
		(0.004)		
Prior inspected establishment’s violation trend	0.013**	0.013**	0.017***	0.013**
	(0.006)	(0.006)	(0.006)	(0.006)
Number of prior inspections today	-0.030***	-0.029***	-0.026**	-0.031***
	(0.011)	(0.011)	(0.011)	(0.012)
Potentially shift-prolonging	-0.057**	-0.058**	-0.052**	-0.052**
	(0.025)	(0.025)	(0.025)	(0.025)
Tens of hours since the inspector’s prior inspection			0.003	
			(0.003)	
Tens of hours since the inspector’s prior inspection			0.001	
* Prior inspected establishment’s violations			(0.001)	
Tens of hours since the inspector’s prior inspection			-0.002	
* Prior inspected establishment’s violation trend			(0.001)	
Prior inspected establishment’s violations				0.016***
for the first inspection of the day				(0.006)
Prior inspected establishment’s violations				0.015***
for the second+ inspection of the day				(0.005)
Inspector experience	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Returning inspector	-0.119***	-0.121***	-0.118***	-0.116***
	(0.034)	(0.034)	(0.035)	(0.035)
Breakfast period (midnight to 10:59 am)	0.055**	0.056**	0.054**	0.055**
	(0.026)	(0.026)	(0.026)	(0.026)
Dinner period (4:00 pm–11:59 pm)	-0.004	-0.008	-0.004	-0.000
	(0.077)	(0.077)	(0.078)	(0.078)
Month fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included
Establishment's nth inspection (second through tenth or more) fixed effects	Included	Included	Included	Included
Inspection-type fixed effects	Included	Included	Included	Included
Establishment x inspector fixed effects	Included	Included	Included	Included
Number of observations (inspections)	12,011	12,000	12,017	12,017

Notes: Poisson regression coefficients with robust standard errors clustered by establishment. Models 1 and 2 have fewer observations than Models 3 and 4 because *Penultimate inspected establishment’s violations* is missing for establishments’ first inspections and *Antepenultimate inspected establishment’s violations* is missing for establishments’ first and second inspections.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Figure 3.1. Prior Inspection Outcome-Effects


Establishment 0's *previous* inspection
Improvement or deterioration
Establishment 0's inspection, conducted by Inspector A (the inspector's *prior* inspection)
H2
H3
H1
Establishment 1's *previous* inspection
Establishment 1's inspection, conducted by Inspector A (the inspector's *focal* inspection)

Notes. This diagram represents the inspector's history (downward arrow) and the establishment's history (left to right). We refer to *an inspector's* preceding inspection as his or her "prior" inspection and *an establishment's* preceding inspection as its "previous" inspection. In this diagram, Inspector A inspected Establishment 0 immediately before inspecting Establishment 1. Hypothesis 1 refers to the relationship between those two inspections. Hypotheses 2 and 3 refer to how Inspector A's inspection of Establishment 1 is associated with the change in compliance between Establishment 0's focal and previous inspections. Prior research has focused, in contrast, on the relationship between an establishment's prior and focal inspections (depicted by the dashed arrow), such as an establishment's improvements as it undergoes successive inspections, the lag between inspections, inspectors' familiarity with an establishments from having inspected it before, and other factors related to an establishment's inspection history (e.g., Ko, Mendeloff, and Gray 2010, Macher, Mayo, and Nickerson 2011, Toffel, Short, and Ouellet 2015, Ball, Siemsen, and Shah 2017).

Figure 3.2a. Estimated nationwide increase in food safety violations being cited as biases are attenuated

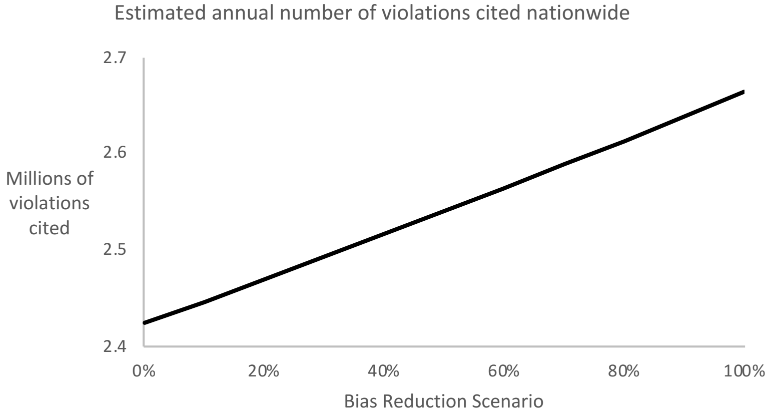


Figure 3.2b. Estimated reduction in healthcare cases associated with more food safety violations being cited as biases are attenuated

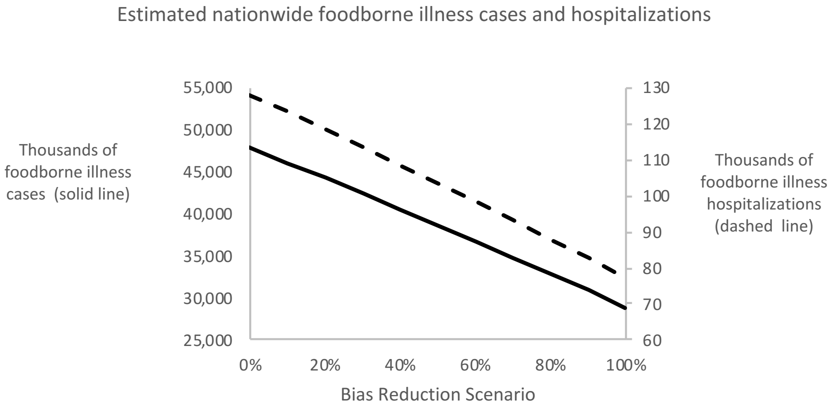
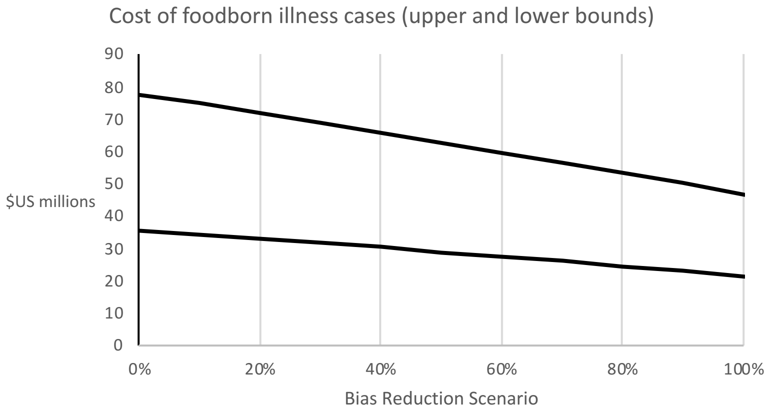


Figure 3.2c. Estimated cost reductions associated with improved health impacts resulting from more food safety violations being cited as biases are attenuated



Notes. These figures graph data from Table 3.D1, which is based on the methodology described in Appendix 3.D. The horizontal axes represent different bias reduction scenarios. For example, the 20% scenario illustrates the results of reducing bias by amplifying by 20% the outcome effects (which increase scrutiny) and mitigating by 20% the daily schedule effects (which erode scrutiny).

Appendix 3.A. Supplemental Descriptive Statistics and Correlations

Table 3.A1. Inspection sequence within day

1st inspection of the day	5,328
2nd inspection of the day	3,618
3rd inspection of the day	1,971
4th inspection of the day	763
5th inspection of the day	248
6th inspection of the day	61
7th+ inspection of the day	28
Total number of inspections	12,017

Table 3.A2. Number of inspector-days

1 inspection days	1,790
2 inspection days	2,226
3 inspection days	1,637
4 inspection days	801
5 inspection days	295
6 inspection days	83
7+ inspection days	48
Total number of inspector-days	6,880

An *inspector-day* refers to a particular day during which an inspector conducts at least one inspection.

Table 3.A3. Inspections by hour begun and corresponding meal period

7 am or earlier	39	3,856 during breakfast period
8 am	222	
9 am	972	
10 am	2,623	
11 am	1,986	7,888 during lunch period
12 pm	1,331	
1 pm	2,331	
2 pm	1,653	
3 pm	587	
4 pm	171	273 during dinner period
5 pm	59	
6 pm or later	43	
Total number of inspections:	12,017	

Table 3.A4. Correlations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) Violations	1.00										
(2) Prior inspected establishment's violations	0.18	1.00									
(3) Prior inspected establishment's violation trend	0.11	0.56	1.00								
(4) Prior inspected establishment saliently improved	0.01	-0.17	-0.37	1.00							
(5) Prior inspected establishment saliently deteriorated	0.12	0.60	0.69	-0.29	1.00						
(6) Number of prior inspections today	-0.02	-0.03	-0.01	0.02	0.00	1.00					
(7) Potentially shift-prolonging	-0.01	0.03	0.01	0.00	0.01	0.36	1.00				
(8) Inspector experience	-0.05	0.05	-0.02	-0.01	0.01	-0.03	0.01	1.00			
(9) Returning inspector	-0.14	-0.01	0.10	-0.03	-0.01	-0.01	-0.01	0.33	1.00		
(10) Establishment's nth inspection (second through tenth or more)	0.05	0.00	0.06	0.01	0.02	-0.01	0.02	0.47	0.34	1.00	
(11) Breakfast period (midnight to 10:59 am)	0.01	-0.47	-0.41	-0.03	-0.02	0.01	-0.02	0.04	0.02	0.03	1.00
(12) Dinner period (4:00 pm–11:59 pm)	0.01	0.15	0.21	0.01	-0.01	0.01	0.00	-0.10	-0.07	-0.05	-0.10

N = 12,017 inspections

Appendix 3.B. Supplemental Analysis: Critical versus Noncritical Violations

To assess whether our hypothesized relationships differentially influence inspectors' behavior across different types of violation, we estimated our models on two subsets of violations. First, we predict the number of *critical violations*, which are related to food preparation practices and employee behaviors that more directly contribute to foodborne illness or injury. These factors are prioritized in Alaska and in Camden County by being displayed on the first page of the inspection report and in Lake County by being tagged in the reports. Second, we estimated our models on the number of *noncritical violations* (that is, violations of procedures often referred to as "good retail practices"). While less risky than the other type, these are also important for public health and include overall sanitation and preventative measures to protect foods, such as proper use of gloves. Inspections averaged 0.93 *critical violations* and 1.49 *noncritical violations*.

More noncritical violations are cited in inspections conducted during the breakfast period than in other periods, but the results yield no evidence that time of day affects critical violations (see Table 3.B1). The latter is consistent with critical violations being related to longer-term establishment practices that are insensitive to the number of customers being served or the staff's ability to respond to the inspector's presence. These results also indicate that the daily schedule effects identified in our primary results are driven by noncritical violations rather than critical ones. In particular, we find no evidence that citations of critical violations are affected by daily schedule effects: the coefficients on *number of prior inspections today* and *potentially shift-prolonging* are not statistically significant when predicting critical violations (Columns 1 and 2). This suggests that fatigue does not affect inspectors' ability to discover and report critical violations. In contrast, an inspector's daily schedule has large statistically significant effects on noncritical violations (Columns 3 and 4). Each subsequent inspection during the day results, on average, in 4.02% fewer noncritical violations cited and *potentially shift-prolonging* inspections result in 5.82% fewer citations.

Outcome-effects are more ubiquitous, affecting critical and noncritical violations alike. Each additional violation cited at the inspector's prior inspected establishment is associated with 1.82% more critical violations (Column 1: $\beta = 0.018$, $p < 0.01$) and 1.41% more noncritical violations (Column 3: $\beta = 0.014$, $p < 0.01$) cited in the focal inspection.

As with total violations, there is no evidence of critical and noncritical violations being affected when the *prior inspected establishment saliently improved*. When the *prior inspected establishment saliently deteriorated*, inspections yield, on average, 7.36% more critical violations (Column 2: $\beta = 0.071$, $p < 0.10$) and 7.79% more noncritical violations (Column 4: $\beta = 0.075$, $p < 0.05$).

Overall, these results indicate that inspectors' schedules have somewhat different effects on citing critical versus noncritical violations. Citing noncritical violations appears to be influenced by both daily schedule effects and outcome-effects, while citing critical violations appears to be influenced only by outcome-effects.

Table 3.B1. Critical and Noncritical Violations

	Dependent variable:	*critical violations*		*noncritical violations*	
		(1)	(2)	(3)	(4)
H1	Prior inspected establishment's violations	0.018***	0.015***	0.014***	0.013***
		(0.006)	(0.006)	(0.005)	(0.005)
H2	Prior inspected establishment's violation trend	0.013*		0.012*	
		(0.008)		(0.007)	
H3	After salient improvement		-0.016		0.029
			(0.031)		(0.028)
H3	After salient deterioration		0.071*		0.074**
			(0.039)		(0.031)
H4	Number of prior inspections today	-0.014	-0.014	-0.042***	-0.041***
		(0.015)	(0.015)	(0.013)	(0.013)
H5	Potentially shift-prolonging	-0.037	-0.036	-0.060**	-0.059**
		(0.036)	(0.036)	(0.030)	(0.029)
	Breakfast period	0.046	0.046	0.063**	0.063**
	(midnight to 10:59 am)	(0.035)	(0.035)	(0.030)	(0.030)
	Dinner period	0.041	0.040	-0.039	-0.041
	(4:00 pm–11:59 pm)	(0.097)	(0.097)	(0.100)	(0.100)
	Inspector experience	0.001***	0.001***	0.001***	0.001***
		(0.000)	(0.000)	(0.000)	(0.000)
	Previous inspector	-0.113**	-0.114**	-0.105**	-0.107***
		(0.049)	(0.049)	(0.041)	(0.041)
	Month fixed effects	Included	Included	Included	Included
	Year fixed effects	Included	Included	Included	Included
	Establishment's *n*th inspection (second through tenth or more) fixed effects	Included	Included	Included	Included
	Inspection-type fixed effects	Included	Included	Included	Included
	Establishment x Inspector fixed effects	Included	Included	Included	Included
	Number of observations (inspections)	10,298	10,298	10,624	10,624

Notes: Poisson regression coefficients with robust standard errors clustered by establishment.
*** p < 0.01, ** p < 0.05, * p < 0.10.

Appendix 3.C. Supplemental Analysis: Inspector Speed and Citation Pace

Our primary results show how daily schedules and inspections of prior establishments are associated with the number of violations cited. To assess whether such results might be driven by inspectors spending less time and exhibiting less scrutiny in the subsequent (focal) inspection, we estimate our primary models on the log of *inspection duration*, the number of minutes between an inspection's start time and end time. Moreover, to assess the net of these two effects, we explore the inspector's citation pace—a measure of productivity in this setting—and estimate our primary models on the log (after adding 1) of *violation citations per hour*. The results are reported in Table 3.C4.

We find that inspectors conduct inspections more quickly as they progress through their shift: *inspection duration* decreases by 3.5% for each inspection of the day (Column 1: *number of prior inspections today* β = -0.035, p < 0.01). For context, recall that our primary results indicate that each subsequent inspection during the day cites an average of 3.15% fewer violations. The model reported in Column 3 of Table 3.C4 indicates that the net effect is that inspector citation pace decreases by 1.3% for each subsequent inspection of the day (*number of prior inspections today* β = - 0.013, p < 0.10).

Turning to *potentially shift-prolonging* inspections, recall that our primary results indicated that these had 5.07% fewer citations. Column 1 of Table 3.C4 reveals that inspectors conduct such inspections 3.6% more quickly (*potentially shift-prolonging* β = -0.036, p < 0.01). Column 3 reveals that the effect of *potentially shift-prolonging* on citation pace is not statistically significant. These results suggest that the diminishments in citations result from shorter inspection durations rather than slower inspector speed, with inspectors' citation pace remaining largely unaffected by the risk of working late. This suggests that our earlier finding that *potentially shift-prolonging* inspections result in fewer violations is likely due to inspectors' desire to avoid working late, rather than to fatigue eroding their citation pace.

Turning to potential outcome-effects, we find no evidence that the outcome of the inspector's prior inspection affects inspection duration, as the coefficients on the variables related to the prior inspected establishment's violations and violation trend are not statistically significant (Columns 1 and 2 of Table 3.C4). Recall that our primary results found that more violations and worsening trends at an inspector's prior establishment predicted more violations cited at the focal inspection. Results reported in Column 3 of Table 3.C4 indicate that inspectors' citation pace increases by 1.0% for each additional violation at the prior establishment ($\beta = 0.010$, $p < 0.01$) and by 1.9% for each one-standard-deviation increase in the *prior inspected establishment's violation trend* ($\beta = 0.012$, $p < 0.05$). Column 4 indicates that, as was the case with the number of violations, the latter effect is asymmetric and driven by negative trends: whereas we find no change in citation pace after inspecting an establishment with salient improvement, it does increase by 3.9% after inspecting an establishment with salient deterioration. This indicates that our earlier outcome-effect findings—that more violations and worsening trends at an inspector's prior establishment increase the inspector's citations at his or her next inspection—result from inspectors increasing their citation pace rather than spending more time onsite.

Table 3.C1. Effects of Inspectors’ Schedules on Speed and Citation Pace

	Inspector speed		Inspector citation pace	
Dependent variable:	log *inspection duration*		log (*violation citations per hour* + 1)	
	(1)	(2)	(3)	(4)
Number of prior inspections today	-0.035***	-0.035***	-0.013*	-0.014*
	(0.005)	(0.005)	(0.008)	(0.008)
Potentially shift-prolonging	-0.036***	-0.036***	-0.025	-0.024
	(0.011)	(0.011)	(0.018)	(0.018)
Prior inspected establishment’s violations	0.002	0.002	0.010***	0.009***
	(0.002)	(0.002)	(0.003)	(0.003)
Prior inspected establishment’s violation trend	-0.001		0.012**	
	(0.003)		(0.005)	
Prior inspected establishment saliently improved		0.010		-0.017
		(0.009)		(0.016)
Prior inspected establishment saliently deteriorated		0.004		0.038*
		(0.012)		(0.020)
Inspector experience	0.000	0.000	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Returning inspector	0.079***	0.079***	-0.100***	-0.100***
	(0.016)	(0.016)	(0.024)	(0.024)
Breakfast period (midnight to 10:59 am)	0.036***	0.036***	-0.013	-0.013
	(0.011)	(0.011)	(0.018)	(0.018)
Dinner period (4:00 pm–11:59 pm)	-0.036	-0.036	0.053	0.051
	(0.032)	(0.032)	(0.053)	(0.053)
Month fixed effects	Included	Included	Included	Included
Year fixed effects	Included	Included	Included	Included
Establishment's nth inspection (second through tenth or more) fixed effects	Included	Included	Included	Included
Inspection-type fixed effects	Included	Included	Included	Included
Establishment x Inspector fixed effects	Included	Included	Included	Included
Number of observations (inspections)	12,017	12,017	12,017	12,017
R-squared	0.45	0.45	0.20	0.20

Notes:Ordinary least squares coefficients with robust standard errors clustered by establishment.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Appendix 3.D. Interpretation of Results

To illustrate the magnitude of the estimated effects, we consider interventions that ameliorate daily schedule effects and exploit outcome-effects that would lead inspectors to cite violations that currently go underreported. In particular, we consider various scenarios that *mitigate the daily schedule effects* in order to attenuate the reduced scrutiny that accompany successive inspections and potentially shift-prolonging inspections, while also *amplifying the outcome effects* in order to more routinely trigger the heightened inspector scrutiny that ensues after inspections reveal many violations and worsening compliance trends. We estimate the effects of such interventions on the average inspection based on our sample, scale up the results to estimate the impact across the entire United States, and translate how such an increase in violations being citing would translate to fewer foodborne illness cases and associated healthcare costs.

In the best-case scenario, outcome effects (which increase scrutiny) would be fully triggered all the time and daily schedule effects (which erode scrutiny) would be entirely eliminated. The full consequence of these biases is reflected by the difference in inspection outcomes between this best-case scenario and the status quo, which quantifies the number of unreported violations and excess illnesses and costs that could be avoided if steps were taken to address these biases. Our discussions with inspectors suggest that some interventions are feasible, such as limiting the number of inspections each inspector conducts, often without imposing any additional costs. We estimate a range of scenarios that consider the impacts associated with the daily schedule effects being attenuated by, and the outcome effects being actuated by, varying amounts.

We first consider the average impact on violations cited per inspection. Specifically, we compare the status quo (that is, the current practice with its associated scheduling effects) with alternative scenarios that consider various percentage changes (10% to 100% in 10% increments) of the effects we identified that would increase inspectors' detection rate (i.e., decrease by 10% the daily schedule effects and increase by 10% the outcome effects). We make all these comparisons based on Model 1 in Table 3.2. Specifically, we calculate average predicted values under each scenario based on the model's estimates after recoding the estimated coefficients on *number of prior inspections today*, *potentially shift-prolonging*, *prior inspected establishment's violations*, and *prior inspected establishment's violation trend* by the percentage specified and report results in Column 1 of Table 3.D1. This is equivalent to preserving the estimated coefficients and instead recoding the values of the variables by that same percentage; thus, the results can be interpreted as altering the per-unit bias represented by the estimated coefficients (e.g., raising inspectors detection rates to the heightened levels associated with the identified outcome effects) or as altering the factors that generate the bias (e.g., reducing the number of prior inspections conducted by the inspector per day). For the status quo, we use the model's estimates to calculate the average predicted number of violations per inspection, based on actual values of all variables, to be 2.42365. Column 2 reports the percent change (from the status quo) in the average predicted violations for each of these scenarios. This shows the average percent change in violations cited per inspection compared to the status quo.

For example, consider the very conservative scenario depicted in the second row of Table 3.D1, in which we estimate the effects of amplifying the outcome effects by increasing by 10% the actual values of *prior inspected establishment's violations* and *prior inspected establishment's violation trend* while also mitigating the daily schedule effects by decreasing by 10% the actual values of *number of prior inspections today* and *potentially shift-prolonging*. Applying these recoded values to the coefficient estimates from Model 1 of Table 3.2, we calculate the average predicted number of violations to be 2.446 (Column 1). This indicates that this "10% scenario" would result in 0.92% more violations being cited per inspection than the status quo of 2.42365 (Column 2).[2]

We then estimate the potential nationwide implications of our calculations, based on the assumptions that the estimated one million food establishments that are monitored by state, local, and tribal agencies in the United States (US Food and Drug Administration 2016) are each inspected annually and that our sample of inspections is representative of those conducted across the country. To calculate a nationwide figure, we take the difference between the average predicted values from each scenario and the status quo (that is, the Column 1 figure minus 2.42365) and multiply that by the one million inspections conducted annually across the country, and report results in Column 3. Column 3 figures can be interpreted in the context of an estimated 2.4 million violations cited in the status quo scenario.[3] Continuing the example of the 10% scenario, we scale the difference in average predicted violations per inspection that arise in this scenario compared to the status quo (2.44603 - 2.42365) by the one million inspections conducted annually nationwide, to estimate that this scenario would yield 22,376 additional violations (currently undetected) being cited nationwide per year (Column 3).

Citing violations leads establishments to improve their food safety practices, which in turn mitigates foodborne illness cases and associated hospitalizations. To calculate the health impacts that would result from these formerly undetected violations now being cited, we translate the estimated nationwide changes in violation counts into health outcomes and their

[2] The 0.92% figure is calculated as (2.44603 - 2.42365) / 2.42365.

[3] The 2.4 million figure is calculated by multiplying 2.424 violations cited per inspection in the status quo scenario by the one million inspections conducted annually nationwide.

associated costs. We attempt to be as conservative as possible but acknowledge that there are uncertainties associated with these conversions.[4]

First, we consider how the increased violations cited per inspection beyond the status quo translates into fewer foodborne illness hospitalizations (Columns 4 and 5). We do so by multiplying the percent change in the average predicted number of violations between the scenario and the status quo (Column 2) by the ratio of 20% decrease in foodborne illness hospitalizations per 5% improvement in restaurant compliance scores based on prior research on Los Angeles restaurants (Jin and Leslie 2003).[5] To calculate the impact of the 10% scenario, we multiply the 0.92% increase in the number of violations cited per inspection (Column 2) by the ratio of 20% decrease in hospitalizations per 5% improvement in restaurant compliance scores, which indicates that hospitalizations would decrease by 3.69% (Column 4). This would correspond to 4,727 fewer foodborne illness hospitalizations occurring each year across the United States (Column 5), based on applying the 3.69% decline to the estimated 128,000 annual

[4] The estimates we construct should be considered as an illustration of the possible implications of the biases. We acknowledge the possibility that our estimates might *overestimate* the effects if the conversion factors we use overestimate the benefits of citing a particular violation, and that they might *underestimate* the effects because we do not incorporate spillovers and system-wide benefits of citing a particular violation, as each citation may encourage establishments to improve health practices more broadly and thus, failing to cite one violation not only carries the health risks associated with that violation but may also feed noncompliance—an effect similar to the broken window phenomenon. That said, while developing a more comprehensive methodology to estimate the health impacts of citing more food safety violations is a necessary and worthy endeavor, it is beyond the scope of this paper.

[5] We are aware of little research that has estimated the effect of each food safety violation on health outcomes and we rely on Jin and Leslie (2003), which we believe presents the best estimate. Jin and Leslie (2003) shows that introducing restaurant grade cards—signs posted outside restaurants that report the establishment's letter grade based on its most recent food safety inspection results—affects food safety inspection violation scores and health outcomes, so restaurant grade cards can be viewed as an instrument that reveals the relationship between violations cited and health outcomes. Because violations are supposed to be corrected when cited, we assume that the new citations resulting from reducing the bias translate into fewer actual violations. (To be conservative, we are not accounting for how citations motivate compliance more broadly.) The relationship Jin and Leslie (2003) identified between compliance and health outcomes applies to our setting because it is based on a similar type of inspection and a compliance measure based on total violations, which implicitly controls for the heterogeneous effects of different types of violation on health.

foodborne illness hospitalizations that occur nationwide under the status quo (Scallan et al. 2011).

We also estimate the impact of citing more violations on annual nationwide foodborne illness cases (Column 6). Based on the ratio of Scallan et al.'s (2011) two nationwide annual estimates of 47.8 million foodborne illness cases and 128,000 annual foodborne illness hospitalizations, there are 373.4 foodborne illness cases per foodborne illness hospitalization. Therefore, the 10% scenario, estimated earlier to reduce foodborne illness hospitalizations by 4,727, would also reduce foodborne illness cases by 1.77 million cases (calculated as 4,727 * 373.4).

Finally, we estimate the impact of citing more violations on the costs associated with foodborne illness cases based on two alternative estimates of the average cost per foodborne illness case of $747 (Minor et al. 2015) and $1,626 (Scharff 2012), which we use to construct lower and upper bounds of our cost estimates (Columns 7 and 8). In the 10% scenario, applying these figures to the estimated 1.77 million fewer foodborne illness cases compared to the status quo, yields a range of $1,319 million to $2,870 million of reduced nationwide annual costs associated with foodborne illness cases.

As noted, there are many assumptions and caveats associated with these analyses, and one can consider alternative scenarios. Our estimations above assume that mitigating bias would yield citations of violations that are as correlated with foodborne incidents as the violations currently cited. But what if newly cited violations are less "important," meaning they impose less health risk? For example, suppose that remediating a newly cited violation would prevent half as many foodborne incidents as remediating a currently cited violation. Estimating the health impacts would then require adjusting Jin and Leslie's (2003) finding that a 5% improvement in restaurant compliance yields a 20% decline in foodborne illness hospitalizations to a 10% decline. In that scenario, if the drivers of outcome and daily schedule effects were respectively amplified and reduced by 100%, the 9.9% increase in citations (last row, Column 2) would translate into a 19.89% decline in hospitalizations [= 9.94*(-20/5)] (compared to our original estimate of a 39.77% decline, calculated as 9.94*(-20/5) and reported in the last row of Column 2]), which nationwide would result in 25,456 fewer foodborne illness related hospitalizations and 9.51 million fewer foodborne illness cases, saving $7.10 billion to $15.46 billion in foodborne illness costs.

Table 3.D1. Estimates of Nationwide Effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Impact on citations of violations			Impact on health and associated costs				
Bias Reduction Scenario	Average predicted number of violations cited per inspection	Percent change in average predicted number of violations cited per inspection compared to the status quo	Change in nationwide annual number of violations cited compared to the status quo	Percent change in foodborne illness hospitalizations compared to the status quo	Change in nationwide annual number of foodborne illness hospitalizations compared to the status quo	Change in nationwide annual number of foodborne illness cases compared to the status quo, **in millions**	Change in nationwide annual costs of foodborne illness cases compared to the status quo, **in millions**	
							Lower estimate	Upper estimate
0%	2.424	0.00%	0	0.00%	0	0.00	$0	$0
10%	2.446	0.92%	22,376	-3.69%	-4,727	-1.77	-$1,319	-$2,870
20%	2.469	1.86%	45,113	-7.45%	-9,530	-3.56	-$2,659	-$5,787
30%	2.492	2.81%	68,219	-11.26%	-14,411	-5.38	-$4,020	-$8,751
40%	2.515	3.78%	91,703	-15.13%	-19,372	-7.23	-$5,404	-$11,763
50%	2.539	4.77%	115,571	-19.07%	-24,415	-9.12	-$6,811	-$14,825
60%	2.563	5.77%	139,834	-23.08%	-29,540	-11.03	-$8,240	-$17,937
70%	2.588	6.79%	164,498	-27.15%	-34,750	-12.98	-$9,694	-$21,101
80%	2.613	7.82%	189,575	-31.29%	-40,048	-14.96	-$11,172	-$24,318
90%	2.639	8.87%	215,072	-35.50%	-45,434	-16.97	-$12,674	-$27,588
100%	2.665	9.94%	240,999	-39.77%	-50,911	-19.01	-$14,202	-$30,914

Each row represents a bias reduction scenario (0% scenario is the status quo). For example, the 10% scenario (row 2) illustrates the results of reducing bias if the outcome effects (which increase scrutiny) were amplified by 10% and the daily schedule effects (which erode scrutiny) were mitigated by 10%.

Column 1 is the average predicted number of violations per inspection, based on Model 1 of Table 3.3, under each scenario.

Column 2 is calculated as the percent change in the average predicted number of violations per inspection, comparing each scenario (Column 1) to the status quo value of 2.42365.

Column 3 is calculated as the difference in the average predicted number of violations per inspection, comparing each scenario (Column 1) to the status quo value of 2.42365 and multiplying this by the one million food safety inspections that are conducted nationwide each year.

Column 4 is calculated by multiplying the percent change in average predicted number of violations compared to the status quo (Column 2) by the ratio of the change in hospitalizations to the change in compliance (derived from the 20% hospitalizations decline per 5% improvement in restaurant compliance relationship reported by Jin and Leslie (2003); that is, -20%/5% = -4).

Column 5 is the difference in hospitalizations between (a) the estimated number that would have occurred under each scenario and (b) the 128,000 that actually occurred (Scallan et al. 2011). Specifically, we multiply the percent change in hospitalizations (Column 4) by the 128,000 nationwide annual hospitalizations.

Column 6 is calculated by multiplying the change in nationwide annual number of foodborne illness hospitalizations compared to the status quo (Column 5) by 373.4, the number of illness cases per hospitalization (calculated as the ratio between Scallan et al. (2011)'s two estimates of the 47.8 million annual foodborne illnesses and the resulting 128,000 hospitalizations).

Columns 7 and 8 are calculated by multiplying the estimated change in illness cases (Column 6) by $747 (the weighted average from Minor et al. (2015)) and $1,626 (the enhanced model estimate from Scharff (2012)) in estimated costs per illness case, respectively.

Chapter 4

Discretionary Task Ordering:

Queue Management in Radiological Services

Abstract

Work scheduling research typically prescribes task sequences implemented by managers. Yet employees often have discretion to deviate from their prescribed sequence. Using data from 2.4 million radiological diagnoses, we find that doctors prioritize similar tasks (batching) and those tasks they expect to complete faster (shortest expected processing time). Moreover, they exercise more discretion as they accumulate experience. Exploiting random assignment of tasks to doctors' queues, instrumental variable models reveal that these deviations erode productivity. This productivity decline lessens as doctors learn from experience. Prioritizing the shortest tasks is particularly detrimental to productivity. Actively grouping similar tasks also reduces productivity, in stark contrast to productivity gains from exogenous grouping, indicating deviation costs outweigh benefits from repetition. By analyzing task completion times, our work highlights the tradeoffs between the time required to exercise discretion and the potential gains from doing so, which has implications for how discretion over scheduling should be delegated.

4.1. Introduction

The scheduling of work is a key driver of operational performance in many settings, including factories (Berman, Larson and Pinker 1997), trucking (Roberti, Bartolini and Mingozzi 2015), healthcare (KC and Terwiesch 2009), and financial services (Staats and Gino 2012). Accordingly, a rich line of research investigates task-scheduling policies identifying optimal schedules that managers can then implement (Pinedo and Yen 1997; Pinedo 2012). In many settings, however, those who execute the tasks often have discretion over the order in which to perform their assigned duties. Yet little is known about the drivers of workers' decisions to exercise such discretion and how scheduling should be managed when discretion exists. In this paper, we consider the operational drivers and implications of "discretion over task ordering," defined as an individual's ability to select which task to complete next from a work queue.

Worker discretion can improve system performance (van Donselaar et al. 2010; Campbell and Frei 2011; Kim et al. 2015; Phillips, Şimşek and Ryzin 2015) but can sometimes enable workers to "choose the 'wrong' task (operationally)" (Boudreau et al. 2003, p. 186). We consider a worker's decision about which task to execute next among a queue of pending, independent tasks; although the assigned order would suggest choosing the first task in the queue, the worker may choose to exercise discretion by selecting a task from the rest of the queue. Hereafter, "deviation" denotes the exercise of discretion over task ordering by selecting a task that is not the next one in the queue. As technological advances are facilitating the delegation and monitoring of decisions made by front-line workers in manufacturing, services, and knowledge work (Pierce, Snow and McAfee 2014), understanding the operational implications of discretion over task sequence is increasingly important.

We address two research questions. First, what are the drivers of deviations? Second, what are their performance implications? To identify the drivers of deviations, we consider the circumstances under which workers are more likely to exert discretion over the order in which they execute tasks. We posit that the ability of workers to identify an alternative task sequence that they perceive as superior to the assigned sequence will depend on the characteristics of the individual as well as the characteristics of the individual's queue of pending tasks. With respect to the individual, we examine the role of worker experience. As for queue characteristics, we examine whether an individual has an opportunity to deviate to pursue a shortest expected processing time (SEPT) policy (i.e., select the task in the queue that is expected to be completed most quickly) or a batching policy (i.e., repeat the prior case type) by deviating.

We investigate these questions using data on doctors reading radiological images of different types (e.g., chest X-rays, head CT scans) at a company where images are randomly assigned to the individual queues of qualified doctors. These radiologists deviate from the assigned first-in-first-out scheduling policy 42% of the time. Our findings show that doctors deviate more often when they are more experienced, when there is an opportunity to follow a SEPT policy by deviating, and when there is an opportunity to batch by deviating. When doctors deviate, however, their average read time tends to increase by about 13%. Other performance dimensions, including quality, are mostly unaffected. Overall, our calculations suggest that forgoing deviations would have led to faster reading times that could have saved 2,494 hours per year, which would have increased annual profits by 3%.

We also find that different types of deviations have varied effects on performance. First, the deviation penalty is lower when the worker is more experienced. Second, although SEPT may create the illusion of working faster precisely because it selects shorter cases, it tends to impair speed and is a particularly detrimental type of deviation. Third, consistent with theory, prior empirical work, and the beliefs of the radiologists we interviewed, batching is associated with superior performance when it occurs naturally, yet this is not the case when batching results from a deviation because of the search costs and other time costs associated with actively choosing and selecting the case from the queue. Individuals may seek to group their tasks to achieve the benefits of batching, but this may not be worth it if they need to do so themselves; this provides evidence of the potential harmful performance effects of exercising discretion when an individual may underestimate the costs of deviating in relationship to the potential gain.

Our paper makes several contributions to both theory and practice. Our work is among the first to focus attention on the role of discretion over task sequence in queue management, recognizing that workers may choose their own approach to sequencing or prioritizing work. Whether in call centers, software companies, or doctors' offices, technology increasingly allows managers to choose how much discretion to grant employees with respect to the order in which they complete tasks. This element of system design merits greater theoretical and empirical attention to understand its performance implications, and we provide important evidence related to this goal. Second, we identify conditions under which individuals are more likely to deviate from the assigned queue. Examining the role of experience and the contents of the queue in this decision provides insight into the design of work systems. Third, we make important methodological contributions by identifying a novel approach to discover valid instrumental variables by exploiting exogenous queue contents to evaluate discretion in queuing settings. Finally, we evaluate the performance implications of these choices. Though attention has been given to the performance effects of discretion, the efficiency of discretion—which incorporates the time invested to *exercise* it—has been overlooked. Our analysis suggests that there is a cost of exercising discretion, which managers should take into account when evaluating the effects of delegation. This time cost of reorganizing the queue may make queue improvements inefficient, underscoring the value of having a centralized individual perform queue management, rather than dividing it across workers. Deviations, even those that lead to batching and would thus be recommended *a priori*, may have a higher execution cost than the resulting benefit. Though deviation is unlikely to be detrimental to performance in all situations, our findings illustrate that it can be and that managers must carefully evaluate the full operational implications of allowing discretion.

4.2. Related Literature

A long line of research on scheduling investigates the optimal allocation of scarce resources (e.g., a machine or a worker) to tasks over time (Pinedo 2012). Problems considered include project scheduling (e.g., Goh and Hall 2013), transportation scheduling (e.g., Zhu, Crainic and Gendreau 2014), appointment scheduling (e.g., Bassamboo and Randhawa 2015; Truong 2015), and workforce scheduling (e.g., Berman et al. 1997). An influential area of research since the 1950s, the optimization problem can have multiple objectives and typically assumes a central planner. Empirical research has studied the effects of task sequence on performance. Among this work, Schultz et al. (2003) provide experimental evidence of the negative effects of work interruptions, showing that changing machines leads to a performance penalty beyond just the time cost of moving locations. Examining data entry clerks, Staats and Gino (2012) find that repeating the same task is associated with superior shift performance, suggesting that managers should provide variety across days or weeks but minimize task switches within a day.

In this work, the often-implicit assumption is that scheduling is a managerial decision and that workers will execute the schedule chosen by the central planner. In many settings, however, this is not the case; front-line workers have autonomy regarding which task to complete next and, therefore, can deviate from the assigned task schedule. We consider the role of worker discretion with respect to task sequencing. If the exercise of discretion by workers yields better performance, managers should encourage such behavior. At the same time, if there are costs of exercising discretion, managers should look for ways to lessen these negative outcomes. It is thus important for managers to understand how workers behave when given the freedom to deviate from an assigned task order, to know whether workers deviate frequently and in predictable ways, and, if so, whether the choices add value. Our paper addresses these questions.

Though little is known about discretion over task sequence, research has examined discretion with respect to other work dimensions, including capacity allocation (Kim et al. 2015), routing a task to a specialist (Shumsky and Pinker 2003; Saghafian et al. 2014; Freeman, Savva and Scholtes 2016), processing time (Schultz et al. 1998; Schultz, Juran and Boudreau 1999), and balancing the speed-quality tradeoff when quality increases with the duration of the interaction (Hopp, Iravani and Yuen 2007; Anand, Paç and Veeraraghavan 2011; Powell, Savin and Savva 2012). These studies show that worker discretion has important operational implications (Lu, Hechling and Olivares 2014; Tan and Netessine 2014; Berry Jaeker and Tucker 2015) and can help improve system performance (Kim et al. 2015). Research has also examined when individuals make decisions different from those that analytical models assume or recommend. Sometimes these deviations are suboptimal and indicate bias (for reviews see Bendoly, Donohue and Schultz 2006; Gino and Pisano 2008), such as those identified in inventory management (Schweitzer and Cachon 2000), forecasting (Kremer, Moritz and Siemsen

2011), or contract structure (Davis, Katok and Santamaría 2014). Despite the potential for bias, the management coefficients theory (Bowman 1963) postulates that managers should be allowed to modify decision rules periodically because managers possess valuable information regarding the current environment. For example, van Donselaar et al. (2010) find that store managers at a supermarket chain deviate from the automated inventory order recommendations because of system inadequacy and misaligned incentives, and these deviations add value by diminishing the costs of managing workload and stock-outs. We contribute to and extend this line of work by studying discretion over a different operational variable (scheduling), by introducing task and individual dimensions that may lead to deviation, and by incorporating the time cost of making decisions. By evaluating the efficiency, not just the effectiveness, of discretion, our work highlights the tradeoffs between the additional time required to exercise discretion and the potential gains from doing so and enables us to understand better the role of discretion in worker productivity.

4.3. Discretionary Task Ordering

4.3.1 Drivers of Deviations from the Assigned Task Order

Many jobs consist of executing a series of sequential, independent, and previously ordered tasks. Examples include doctors seeing patients, mechanics fixing cars, or back-office processors completing claims. In such settings, workers often have visibility into the queue and the ability to deviate from its assigned order, resulting in discretionary task ordering. Although the assigned sequence would suggest choosing the next task in the queue, they may choose a task from the rest of the queue. We refer to the selection of a task other than the next as a “deviation.” Workers may choose to exercise discretion over task ordering when they believe that the assigned order is not optimal for performance.[1] We posit that the tendency to deviate will depend on attributes of the worker and the queue.

With respect to the attributes of the worker, we focus on an individual's work experience. First, with experience may come the ability to identify queue inefficiencies and opportunities to improve upon the assigned order. Significant attention has been given to the relationship between experience and process improvement, finding that additional experience typically leads to learning (Lapré and Nembhard 2010; Argote and Miron-Spektor 2011). One reason why experienced individuals show improvement may be that they recognize more opportunities to change their work by altering the order in which they complete tasks. Accordingly, as long as the assigned task sequence is not optimal, individuals should deviate more often as they gain experience. Second, in addition to identifying more improvement opportunities, individuals may be more likely to act upon such opportunities as they acquire more experience. Notably, workers gain confidence through experience (Bandura 1977), and this confidence could encourage them to take action. These arguments lead to our first hypothesis.

Hypothesis 1: *The probability that an individual deviates from the next task in the queue increases with that individual's level of experience.*

We next turn to the attributes of the queue. One task sequencing strategy that individuals may pursue is a shortest expected processing time (SEPT) policy, in which the task that is expected to take the least time to perform is completed next. There are operational and behavioral reasons to follow this policy. Operationally, this scheduling discipline minimizes the

[1] Workers may also choose to reorder tasks based on personal incentives. In this paper, we focus on operational drivers, and present an empirical setting without personal incentives in conflict with performance.

average number of jobs in the system and the average job wait time. Because of its operational benefits, workers might opt to improve on these dimensions, even if these metrics are not used to evaluate performance. Behaviorally, individuals may exhibit a preference for completing easier tasks first, even in settings where their self-interest would be better served by completing tasks in a different order. For example, Amar et al. (2011) find that individuals choose to pay back smaller debts with lower interest rates, to accomplish completion, instead of paying back the same amount of a larger debt with a higher interest rate (and thus saving money). In terms of scheduling policies, that means that individuals might first take on what are expected to be the shortest tasks. If workers reorganize the queue according to SEPT, they will be more likely to deviate when the remaining queue (i.e., the queue excluding the first item) contains the task type in the queue with the shortest expected processing time.

Hypothesis 2: *The probability that an individual deviates from the next task in the queue is higher when that task is not of the shortest type in the queue.*

A second category of task sequencing strategies involves batching—grouping tasks by their types to increase the repetition of similar activities. After completing a task, an individual could recognize that a task further in the queue is very similar to the just-completed task and so choose to complete it next. This batching could bring benefits in terms of decreased setup time, even if the required setup is just cognitive (Staats and Gino 2012). Further, it could also provide processing time benefits, as the relevant knowledge is still in the individual's working memory, allowing her to complete the work quickly and avoid interruptions (Bendoly, Swink and Simpson 2014; Froehle and White 2014; Wang et al. 2015). Thus, one specific reason to deviate may be the batching of tasks. Batching, however, is not always possible; the opportunity to batch depends on the availability of at least one task of the same type as the predecessor. When the first task in the queue is of the same type as the predecessor, respecting the assigned order would automatically bring the benefits of batching. When the first task in the queue is *not* of the same type as the predecessor, one could deviate towards batching if the rest of the queue contains a repetition of the previous task type. If individuals have an intention to batch, then they will be more likely to deviate when they would not batch by following the first task in the queue but can batch by deviating.

Hypothesis 3: *The probability that an individual deviates from the next task in a queue is higher when that task is different from its predecessor, and the remainder of the queue offers an opportunity to repeat the predecessor task type (i.e., to batch).*

4.3.2 Performance Implications

At least since the origin of the scientific management movement (Taylor 1911), the field of operations has sought to understand the drivers of performance. Over time, improvements have been identified in many areas, from task scheduling (Pinedo and Yen 1997) to product variety (Fisher and Ittner 1999) to queuing system design (Gans, Koole and Mandelbaum 2003). Research has increasingly considered human aspects of operations management problems (Boudreau et al. 2003), incorporating such behavioral factors as workload (KC and Terwiesch 2009; Powell et al. 2012; Tan and Netessine 2014; Berry Jaeker and Tucker 2015) and team composition (Huckman, Staats and Upton 2009; Schultz, Schoenherr and Nembhard 2010).

Recent work shows that, under certain conditions, expert discretion over operational variables can improve decisions (Campbell and Frei 2011; Kim et al. 2015; Phillips et al. 2015). In this paper, we study how discretion over task sequence affects task completion time. The completion time for a person carrying out a task is given by the setup time plus the processing time. Under discretionary task ordering, where individuals select which task to execute next, setup time includes both the search time investigating the queue and choosing a task (task selection time) and the time preparing to execute the new task (standard setup or changeover). Processing time represents the "run time" required to complete the task itself.

We begin by exploring how the exercise of discretion might improve completion time. To the extent that individuals deviate to enhance task sequence, we would expect the exercise of discretion to benefit productivity. Front-line personnel often have information about improvement opportunities that is not available to a central planner (MacDuffie 1997; Tucker 2007; Staats, Brunner and Upton 2011). For example, delivery drivers may adjust their daily route after observing a road under construction during the prior day. Thus, even if we assume that the queue has been optimally organized by the central planner with the knowledge that she has, the worker may recognize opportunities to improve upon that plan. Moreover, many queues are not optimally organized to begin with, creating more improvement chances. Hence, workers may deviate to apply generally accepted best practices for task scheduling. For example, workers can avoid the cost of switching (e.g., either mental or physical setup costs) by selecting a task that repeats the predecessor's task type (i.e., batching). These improved sequences may thus result in superior speed.

Hypothesis 4A: *Task deviation leads, on average, to faster completion time.*

Although exercising discretion could be beneficial, it might instead prove distracting. First, by searching through the queue to choose the next task to complete, a worker is adding a search cost (task selection time) to the setup time. Second, switching back and forth from searching through the queue to executing tasks could generate cognitive distractions (KC and Staats 2012; Staats and Gino 2012; Froehle and White 2014) and slow the worker more than the gain from the deviation. In addition, the improvement of task sequence may be suboptimal if workers do not look for optimality but rather satisfice—selecting an option meeting their minimum requirements (Simon 1978). Given the potential for conflicting performance effects, we offer the following competing hypothesis:

Hypothesis 4B: *Task deviation leads, on average, to slower completion time.*

Regardless of the average net effect of deviations on speed, different types of deviations may have varied effects. We first consider the heterogeneous effect of deviations across levels of worker experience. An extensive literature in learning-by-doing shows that individuals' performance improves with experience (Huckman and Pisano 2006; Narayanan, Balasubramanian and Swaminathan 2009). One activity that individuals may learn through experience is how to exercise discretion over task sequence. That is, workers may learn about how to deviate more effectively and efficiently as they gain experience, which, in turn, could lead to better deviations in terms of speed performance. For example, they may develop better intuition about which task to work on next or learn how to search through the queue faster to execute a preferred strategy. Hence, for positive net effects of deviations on speed, we would expect the performance benefit to grow with experience, and for negative net effects of deviations on speed, we would expect the performance penalty to be smaller with experience. We thus hypothesize:

Hypothesis 5A: *Task deviation leads, on average, to faster completion time when an individual has greater experience.*

Next, we investigate the performance implications of deviations according to the task selected. We categorize task-type deviations based on two dimensions that are consistent with SEPT and batching. Theory does not generate a clear prediction for the direct effect of a SEPT policy on completion time. On one side, research in psychology suggests that completing tasks motivates (Gal and McShane 2012), and hence SEPT could be associated with faster speed. On the other side, individuals might have an expectation regarding the appropriate time to be working on any given task regardless of its actual complexity. If this expectation regarding how long a task should take is affected by the other tasks in their queues, then when selecting the shortest task among a given queue, workers might allow their processing time to expand beyond or relatively to the expected processing time for this particular task type (Hasija, Pinker and Shumsky 2010). Hence, when selecting the shortest task (following SEPT), they might under-adjust their expectation regarding the reasonable time to be working on such task, leading to a slower speed after controlling for the actual complexity of the task. Therefore, following a SEPT policy may affect efficiency either positively or negatively. Disentangling these effects is ultimately an empirical question.

What would be the performance consequences of deviations towards SEPT? Although the direct effect of SEPT may be theoretically unclear, deviations towards SEPT are likely performance-decreasing. In terms of types of deviations, there are reasons to believe that deviations towards SEPT may be particularly time-consuming, as the search cost of exploring the queue not only includes looking at the mix of tasks but also mental estimation of the reading time of each task type and comparison of those expected times across all tasks. For example, while deviations toward batching only require an individual to search through the queue until a particular task type is identified, SEPT involves going through the *entire* queue, determining the expected processing time of each task type and contrasting it to the shortest one identified to that point. In addition, precisely because deviating towards SEPT requires this consideration of expected processing times for all tasks in the queue, it may magnify the salience of those other (longer) tasks in the queue as reference points (Hossain and List 2012), leading individuals to increase their expectation for how long a task should take and subsequently expand the actual processing time to fill this time (Hasija et al. 2010). Thus, we expect deviations to be worse for performance when they are towards SEPT than when they are not.

Hypothesis 5B: *Task deviation leads, on average, to slower completion time when the selected task is of the task type with the shortest expected processing time in the queue.*

The final dimension for categorizing deviations is whether they are toward task-type repetition (i.e., batching) or not. To the extent that repetition of task type is associated with superior performance, one would expect deviations to be more effective when they result in batching than otherwise. Moreover, controlling for the direct effect of repetition on speed, deviations toward repetition are expected to be relatively more efficient. Compared to other sequencing strategies that require evaluating the whole queue, contrasting different options, or computing certain metrics (e.g., expected processing time), the strategy based on task-type repetition only requires searching the queue to find a specific type, and this search stops as soon as the first task meeting this requirement is found. Thus, we expect deviations to be more beneficial when they are toward task-type repetition (i.e., when the selected task is of the same type as the predecessor) than when they are not.

Hypothesis 5C: *Task deviation leads, on average, to faster completion time when there is a task-type repetition.*

4.4. Setting, Data and Models

4.4.1 Empirical Setting – Outsourced Teleradiology Services

We test our hypotheses using transaction-level data from one of the largest outsourced radiological services (teleradiology) firms in the United States. In this setting, radiologists seated at computer workstations—at home or at a reading center—sequentially interpret "cases", each of which corresponds to a set of digital images of a particular technology and anatomical area for a patient. Technologies used in our setting include X-rays (electromagnetic waves, 3.68% of our final sample), computed tomography (CT, 84.26%), nuclear medicine (1.01%), magnetic resonance imaging (MRI, 0.85%), and ultrasound (10.20%). The anatomical areas include abdomen (5.58%), body (combination of areas, 36.31%), brain (33.23%), breast (0.01%), cardio (0.2%), chest (12.78%), gastrointestinal/genitourinary (1.24%), head and neck (2.44%), musculoskeletal (1.31%), obstetrics (2.6%), pelvis (0.53%), spine (3.73%), and other (0.04%). The company receives cases from clients—typically hospitals or physician group practices—and assigns the reading of them to individual radiologists on a round-robin basis following a computer-based algorithm. To be eligible to receive a case, a radiologist must be trained in the technology and anatomical area, licensed by the state and credentialed by the hospital where the radiological image was created. Given these requirements, most radiologists in the company can interpret the majority of cases, are licensed in over 35 states, and are credentialed at one-third of the client hospitals. Finally, an eligible radiologist must be on duty and not too backed up when the study arrives. Conditional on a radiologist meeting the availability and eligibility requirements, case assignment to radiologists is random.

At any point in time, the radiologists see their own queue of pending cases. We observe—and control for—the factors the radiologist observes when deciding which case to select next. In particular, for each case in the queue, the radiologist sees the time the case was assigned to her queue, the technology employed to create the images, the anatomical location of the study, and the number of images. Once a case is assigned to a particular radiologist, it is not reallocated to another radiologist. Radiologists work independently without supervision, and only see cases assigned to them. Because new cases are continually added to this dynamic queue while radiologists are on duty, radiologists make a decision, explicitly or implicitly, regarding which case to read next every time they start a case rather than reordering all cases at the beginning of a shift, as could happen in settings where all tasks to be completed are known initially.

Management expected radiologists to follow a first-in-first-out (FIFO) policy but did not enforce it, thereby leaving the radiologists free to deviate. Therefore, in this setting, individuals seeking to improve performance are given a task schedule based on the random arrival times of cases, are supposed to follow this order, but are allowed to adjust it. Each case represents a well-defined task, and radiologists have the freedom to decide which case to work on next. Because the number of cases in a radiologist's queue is 5.6 on average, they can reasonably inspect the queue to evaluate alternative sequencing strategies. Though the company did not provide access to the radiologists whose work is captured in our data, we interviewed several radiologists at different institutions with similar processes to understand how they approached task deviation. We found that dedicated, individual queues and freedom to alter task sequence were common. These radiologists also indicated that they often chose to deviate from their assigned task ordering. They provided different explanations, including a desire to read faster cases first and an intention to repeat the same technology-anatomy combination (batching), in line with Hypotheses 2 and 3. In particular, the radiologists indicated that they thought that batching was helpful in the interpretation of images and highlighted the importance of focusing their attention, if possible, on a specific anatomical area at a given point in time. In our final sample described below, half of the deviations are consistent with these two reasons to deviate. First, 48% of deviations are toward a case of the shortest type available, consistent with a SEPT scheduling policy. Second, 15% of deviations are toward a case-type repetition. Deviating toward batching is not always possible, as it is conditional on the case types available in the remainder of the queue. When batching is possible, the percentage of deviations consistent with batching is 46%. Of the cases interpreted, 46% correspond to SEPT and 13% are case-type repetitions.

The radiologists in our setting aim to maximize their overall speed subject to delivering the correct clinical interpretation. They seek to maximize speed for both business and clinical reasons. On the business side, teleradiology companies compete for business on the promise of fast service. On the clinical side, unlike some service settings, completion speed is a major determinant of quality, as timely access to the reading report is often critical for the patient's referring doctor to deliver proper treatment. This positive relationship between speed and quality in healthcare has been noted in prior work (Pisano, Bohmer and Edmondson 2001). While radiologists seek to maximize speed, they do so subject to the constraint of maintaining acceptable quality. The teleradiology company tracks reading discrepancies, whereby a customer receiving a radiological report may raise an objection, including minor comments. Such discrepancies are rare, affecting only 0.3% of the images in our final sample, so it is reasonable to assume that the clinical quality of the reads is deemed acceptable.

Numerous features of this research site make it an ideal setting to explore our questions and mitigate concerns about gaming behavior or other reasons unrelated to performance that might cause radiologists to prioritize certain cases. First, in this teleradiology company, there is no preemption. Both the response and reading times are quick, so radiologists do not interrupt the reading of a case once it is in progress. Second, all cases are deemed urgent, so there is no prioritization based on medical emergency. Third, given the time sensitivity of the service, cases are not left in a doctor's queue before any break longer than thirty minutes or by the end of their shift; therefore, postponing a job does not affect the case-mix or the workload, and there is no need to prioritize shorter cases due to time constraints. Fourth, the type (technology-anatomy) corresponding to arriving cases is independent of doctors' speed and the types of cases they have previously received or completed, addressing any remaining concern about prioritizing cases by type to affect case mix. Fifth, there are no financial incentives to prioritize cases. Radiologists are compensated based on hours worked, and must complete each and every case in their queue within the shift.[2] Sixth, there are no mandatory order restrictions (such as required predecessor tasks) or external factors (such as collaboration with other individuals, Halsted and Froehle 2008; Wang et al. 2015) that could limit the doctor's discretion. Finally, there is no need to reorder cases to put the images for a single patient together, as those images are grouped into a single case prior to entering a radiologist's queue.

4.4.2 Data

Our data covers all 2,766,209 cases processed by the teleradiology company between July 2005 and December 2007. We observe both the order in which the jobs are assigned to radiologists and the order in which they are completed, with differences in the two being due to a radiologist's exercise of discretion. For each case, we observe its characteristics (e.g., technology, anatomy, and number of images), the radiologist who interpreted it, the time it was assigned to the radiologist, and the time it was completed. We then reconstruct the set of cases in a radiologist's queue at any point in time.

We impose three restrictions on the initial sample. First, because we seek to study decisions made by radiologists about whether to deviate from the queue, we limit the sample to those cases that were selected from a queue of at least two cases. This restriction eliminates cases that were the only ones in the queue when the radiologist started them, as there would be no potential for a radiologist to deviate in such instances. Second, we drop cases for which the time elapsed since the last case exceeds thirty minutes (the 99.5th percentile), as we assume that it represents a new shift or break. We do not have records of the exact breaks taken by the radiologists, who lack fixed schedules and rules for breaks. Through this restriction, we define a

[2] In this setting, there is no incentive to prioritize cases based on clients, as the client base is extensive and the company does not systematically offer preferential treatment. Consistent with this, and alleviating concerns regarding prioritization of cases based on preferential treatment of certain hospitals, the results are robust to the inclusion of hospital fixed effects.

shift break as a period longer than thirty minutes without completing a case, and drop the first observation for each shift because the estimated reading time would include initial set-ups. No case is left in the queue at the end of these shifts, which is consistent with the company practice of zero backlogs between shifts to ensure quality of care. Our results are robust to alternative cutoffs for shift breaks (20, 40, 60, 90, 480 minutes). Third, we drop the observations corresponding to four radiologists, each of whom had less than 100 cases in the remaining sample. This facilitates convergence of the model estimation and ensures that the fixed effects do not introduce bias into unconditional probit estimates, as we describe in the Econometric Models section. The final sample for our regression analysis includes 2,408,218 cases of 53 unique case types interpreted by 91 radiologists. Table 4.1 displays summary statistics and correlations.

4.4.2.1. Dependent Variables

Deviation from the assigned queue order. DEVIATION is a binary variable indicating whether the individual deviated from the assigned order when selecting the current case. This variable takes the value of one if the job selected was not the first one in the queue and zero otherwise. In our sample, doctors deviate from the order in which the cases were assigned to them 42% of the time. The radiologists who deviate the most and the least deviate 59% and 24% of the time, respectively.

Completion (Read) Time. We capture performance using the amount of time the radiologist spends reading a case (total time to select and interpret a case). We estimate this reading time (READTIME) for a case as the difference between the time when the current case and the prior case of the radiologist are completed. When a case is not available in the queue when the radiologist submits the previous case, we use the time difference between when the case becomes available (i.e., the case is assigned to the radiologist) and when it is completed (i.e., the radiologist submits the report). This calculation could overestimate reading time for the first case in a shift; because of this, we ignore the first case of each shift for each radiologist. If cases were left in the queue between shifts, our method for calculating reading time could overestimate reading times; however, in our setting, no case is available in the queue when the last case of a shift is completed.

The time stamp is at the minute level; that is, for each case, we know the minute in which it is assigned to the radiologist and the minute in which the radiologist completes the reading. When a case exits the system within the same minute it enters or the prior case is submitted, the time difference has a zero value. Because our empirical models use the natural log of read time, and the logarithm of zero is not defined, we add one to all values (Allcott and Sweeney 2016), which is equivalent to rounding up the estimated reading time. The average estimated reading time per case is 3.75 minutes.

4.4.3 Identification Strategy

Our field data allows us to establish external validity, identify effect sizes, overcome observer bias, and study the phenomenon in a rich context over a relatively long time period. One challenge with these data, however, is that the decision to deviate from the assigned order is possibly endogenous. Specifically, there may be unobserved factors that affect both the decision to deviate and performance in terms of the reading time for a case, as illustrated in Figure 4.1. Disregarding this endogeneity could lead to bias. We address this challenge by exploiting a set of instrumental variables and estimating an endogenous treatment-regression model (Heckman 1978; Maddala 1983) composed of an equation for the treatment (i.e., the decision to deviate from the first case in the queue) and an equation for performance. The performance equation is identified if and only if at least one exogenous regressor excluded from this equation has a nonzero coefficient in the other (i.e., deviation) equation. This is known as the rank condition. This simultaneous equation system can be interpreted as using instrumental variables for the endogenous regressor (DEVIATION). A valid instrument must have two characteristics; first, it must be exogenous, that is, contemporaneously uncorrelated with the error, influencing the outcome (i.e., performance) only through the deviation decision; and second, it must be correlated with the endogenous variable, DEVIATION.

We present a novel approach to identify valid instruments to study discretion in queuing settings. Specifically, we propose that the choice to exercise discretion is affected by the composition of the tasks within the queue, and, when exogenous, such composition can generate valid instruments. It is precisely these queue contents that allow the use of discretion, determining the options available to the decision maker. For example, queues with only one item do not allow for discretion regarding which item to select. Queues with only one *type* of item do not allow discretion over which task type to work on. Alternatively, queues that offer a choice of different opportunities (in terms of types of items, sequencing strategies, or other dimensions) provide the opportunity to employ discretion. When queue contents are exogenous, certain queue characteristics may be valid instruments. This approach can be applied to different types of decisions involving queues. For example, the exogenous arrival of a new task in a queue during the processing of a given task could be used as an instrument to study task interruptions and preemption.

We use multiple queue characteristics to instrument for deviation from the next case in the queue. The first instrumental variable is the opportunity to deviate to follow a shortest-expected-processing-time policy (SEPT_OPPTY) due to the availability of a shorter case in the remaining queue. To the extent that a radiologist might pursue a SEPT policy, the opportunity to do so by deviating will affect the decision to deviate. SEPT_OPPTY is positively correlated with DEVIATION (Table 4.1). On average, radiologists deviate 48% of the time when there is an opportunity to follow a SEPT policy by deviating, while they only deviate 31% of the time when such opportunity does not exist (Table 4.2). Thus, there is a 17-percentage-point difference in the deviation rate between the decisions made when the first case is the shortest in the queue versus when it is not. At the same time, SEPT_OPPTY does not directly affect performance.

The second instrumental variable is the opportunity to deviate to repeat the task type (REPEAT_OPPTY). REPEAT_OPPTY is positively correlated with DEVIATION (Table 4.1). On average, when there is an opportunity to repeat case type only by deviating, radiologists deviate 52% of the time, while they only deviate 39% of the time when such opportunity does not exist (Table 4.2). There is thus a 13-percentage-point increase in deviation associated with the presence of an opportunity to batch by deviating. At the same time, REPEAT_OPPTY does not directly affect performance.

Our final set of instrumental variables capture the case type of the first case in the queue. Doctors may be more likely to choose (by deviating or not) certain task types. The more attractive the type of the first case in the queue, the more likely the radiologist would be to select this case next and thus the less likely the radiologist would be to deviate. Therefore, the case type of the first case in the queue is expected to affect the decision to deviate from the queue.[3] At the same time, after controlling for the case type of the selected case, the case type of the first case in the queue does not directly affect performance. This corresponds to the case that was supposed to be read next (according to the assigned order, had the radiologist not deviated), while the speed

[3] The basic assumption behind this set of instruments is that they lead to different propensities to deviate. To confirm this, we regress DEVIATION against first case type fixed effects and a set of controls (Column 4 in Table 4.3). We strongly reject the hypothesis that the fixed effects for the different types of the first case are the same ($p < 0.0001$).

of the read will only depend on the type of case selected and actually read—a factor for which we control. In conclusion, there is substantial support for the validity of these variables as instruments.

Using this instrumental variables approach, the estimates represent the average performance effect of deviation for the subgroup of task selections affected by these instruments. The identification is driven by a comparison of differences in the reading times among cases with similar observable characteristics (e.g., same type, interpreted by the same radiologist) but for which the doctor deviated or not to select them because of the different queue characteristics related to (1) opportunities to follow a SEPT policy by deviating, (2) opportunities to repeat case type by deviating, and (3) the default choices as represented by the type of the first case in the queue. These instruments provide variation that is likely to capture many deviations. First, the large number of instruments employed has the advantage of increasing the subset of tasks considered in estimating the effect, which thus represents a larger portion of the overall population of tasks. Second, and more importantly, the type of the first case of the queue affects a broad set of decisions to deviate, because any such decision will compare the default option (given by the first task in the queue) to the remaining options (given by the rest of the tasks in the queue). Accordingly, choosing a different task from the first one (i.e., deviating) is equivalent to rejecting the first task. Though these instruments might not apply to all deviation decisions, they have the potential to affect the large majority of decisions.

4.4.4 Econometric Models

As discussed in the prior section, we examine workers' decisions to deviate from the assigned queue order and the subsequent performance implications by estimating an endogenous treatment-regression model (Heckman 1978; Maddala 1983) composed of two equations—a probit model for the treatment (i.e., the decision to deviate) and a log-linear regression model for speed performance:

$$DEVIATION_{ij} = \begin{cases} 1 \; if \; \boldsymbol{X_{ij}\beta} + \varepsilon_{ij} > 0 \\ 0 \; otherwise \end{cases} \tag{1}$$

$$\ln\left(READTIME_{ij}\right) = \delta DEVIATION_{ij} + \boldsymbol{X_{ij}^{\prime}} \boldsymbol{\beta}^{\prime} \; + u_{ij}, \tag{2}$$

Here *i* and *j* denote radiologists and cases, respectively; and the random disturbance terms ε_{ijt} and u_{ij} are bivariate normal with mean zero and covariance matrix $\begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}$.

The vectors of covariates $\boldsymbol{X_{ij}}$ and $\boldsymbol{X_{ij}^{\prime}}$ include fixed effects for radiologist (to control for time-invariant radiologist characteristics), day of week, calendar year, and case type (defined by the unique combinations of technology and anatomy of the case, to control for heterogeneity in attractiveness and average reading time across case types); the number of years the radiologist has been working at the company (EXPERIENCE); and controls for (a) the number of cases in the radiologist's queue when the current case is selected (QSIZE), which captures both a radiologist's range of options as well as her workload and is, therefore, expected to affect both the decision to deviate and performance, (b) the case-type variety in the queue (QVARIETY), representing alternative options for the individual to choose among, (c) the number of images in the current case (NUM_IMAGES), as a larger number of images involves additional reading time and could, therefore, affect the likelihood of deviation, (d) the number of cases read by the radiologist since the beginning of the current shift (ORDER_IN_SHIFT), as each additional case contributes to both warm-up and fatigue over the course of the shift, and (e) an indicator for whether the queue was empty when the previous case was finished (RESTART), accounting for the warm-up effects after being idle for a short period within a shift. This restarting is infrequent in our sample, occurring in only 1% of cases, and is not included in the deviation model because it perfectly predicts the outcome.

In addition to these common covariates, $\boldsymbol{X_{ij}}$ in the deviation model includes our instrumental variables: an indicator for whether the first case in the queue does not have the shortest expected processing time within the queue (SEPT_OPPTY), an indicator for whether there is an opportunity to repeat case type but only by selecting a case from the queue other than the first one (REPEAT_OPPTY), and indicators for the type of the first case in the queue. Finally, in the extended model to test Hypotheses 5A-C, $\boldsymbol{X_{ij}^{\prime}}$ in the performance model also includes (a) the indicator variable SEPT, which equals one if the case read corresponds to the shortest case type in the queue and zero otherwise, (b) the indicator variable REPEAT, which equals one if the prior case was of the same type as the current one and zero otherwise, and (c) interaction terms of DEVIATION with EXPERIENCE, SEPT, and REPEAT.

We note four important points regarding the empirical specification. First, the simultaneous-equation system takes into account the fact that the deviation decision is determined within the model rather than predetermined. Given that DEVIATION is potentially endogenous, ordinary least squares should not be applied to estimate the performance equation because the estimators would potentially not only be biased but also inconsistent. The maximum likelihood estimator of the simultaneous-equation system presented is consistent (Heckman 1978; Maddala 1983). Second, because we do not have a measure of radiologists' experience prior to joining the company, Equation (2) uses the exponential learning curve model, which prevents bias from our lack of information about prior experience (Lapré and Tsikriktsis 2006). Third, although the maximum likelihood estimator in the presence of fixed effects in discrete choice models shows a finite sample bias when the number (T) of observations per individual is very small (i.e., the incidental parameters problem), this bias declines rapidly as T increases beyond three and is negligible for large T (Greene 2004). Given that we have a deep panel, with an average of 26,464 cases per radiologist and at least 239 cases per radiologist, the only problem estimating unconditional fixed effects is computational. Finally, though the dependent variable completion time would suggest that a survival model could be used (Lu et al. 2014), we do not have a censoring or truncation problem. Hence, a log-linear model is appropriate.

4.5. Results and Discussion

4.5.1. The Determinants of Deviations

To investigate the drivers of deviations from the queue, probit maximum likelihood estimates of Equation (1) are shown in Table 4.3. Because our dependent variable is binary, we use probit regression but a linear probability model yields similar inferences. Average marginal effects (AME) are provided next to the corresponding coefficients. Standard errors are clustered at the radiologist level. In the baseline model (Column 1), we only include the controls. The estimated coefficient on the length of the queue (QSIZE) is positive and significant; it may be that larger queues offer more opportunities to deviate or that they create workload pressures that lead a radiologist to choose to deviate. It is possible that the use of discretion could be part of what leads to the eventual "speed up" effect that prior literature has observed with larger queues (KC and Terwiesch 2009; Staats and Gino 2012; Delasay et al. 2015). The magnitude of the average marginal effect implies that one more case pending in the queue increases the probability of deviating by 1.84 percentage points (a 4.39% increase when compared to the sample average of 41.9%). Keeping the size of the queue fixed, individuals adhere to the queue order less frequently as the variety of different case types available within the radiologist's queue

(QVARIETY) goes up; this result is consistent with greater options creating more opportunities for radiologists to exercise their discretion over task scheduling. Specifically, all else constant (including queue length), a one-unit increase in the variety of the queue increases the probability of deviating from the queue by about 15.50 percentage points (a 36.99% increase compared to the sample average of 41.9%). In addition, the coefficient estimate for the number of images (NUM_IMAGES) is not statistically significant, while the coefficient estimate for the position within the radiologist's shift (ORDER_IN_SHIFT) is negative and statistically significant.

To test the hypothesis that worker characteristics affect adherence to the assigned work order, radiologist EXPERIENCE is included in Column 2. The results show that longer tenure at the company is associated with a higher likelihood of deviation, which supports Hypothesis 1. An additional year of experience increases the probability of deviating by 7.71 percentage points (a 18.40% increase when compared to the sample average of 41.9%). We measure experience by the number of years that the radiologist has worked at the company, as it allows us to capture experience prior to our sample period and it is a common measure in the literature (Tucker, Nembhard and Edmondson 2007). Using a radiologist's case volume as an alternative measure of experience yields similar results.

We next explore the impact of queue characteristics on deviation, looking at two particular deviation strategies—deviation toward the shortest cases and deviation toward batching of case types (Column 3). Including an indicator for whether the first case in the queue is inconsistent with a SEPT policy, thereby creating an opportunity to follow SEPT by deviating (i.e., SEPT_OPPTY=1), we find that the predicted probability of deviating from the assigned order is higher when the first case in the queue is not the shortest case in the queue, supporting Hypothesis 2. The average marginal effect suggests that having an opportunity to follow a SEPT policy by deviating boosts the probability of deviation by 2.41 percentage points, increasing the average predicted probability of deviation from 40.87% to 43.28%. The results also indicate that individuals are more likely to deviate when the first case in the queue is not of the same type as the predecessor but the remainder of the queue offers an opportunity to repeat (i.e., REPEAT_OPPTY=1), as predicted by Hypothesis 3. The average marginal effect indicates that having an opportunity to batch by deviating from the queue increases the probability of deviation by 1.54 percentage points, increasing the average predicted probability from 42.10% to 43.64%.

4.5.2. The Impact of Deviations on Performance

To study the impact of deviations on performance, we estimate equations (1) and (2) jointly via maximum likelihood (Column 1 of Table 4.4). Using a control-function estimator provides equivalent results. Standard errors are clustered by radiologist. The Wald test of independent equations indicates that we can reject the null hypothesis of exogeneity of deviation ($p < 0.0001$). The estimated correlation between the treatment-assignment errors and the outcome errors, ρ, is negative, indicating that unobservables that increase reading time tend to occur with unobservables that reduce deviation occurrence (negative bias). Accounting for the endogeneity of the deviation decision is important in obtaining consistent estimates of the deviation effect on reading times. For each additional case in the queue (QSIZE), the average reading time decreases, all else equal, by about 2.9% on average. For a one-unit increase in QVARIETY, there is a 20% decrease in completion time. Reading time increases for each additional image (NUM_IMAGES) included in the case by about 18.7% and decreases over the course of a given shift (ORDER_IN_SHIFT), though by a small amount on a case-by-case level. We find evidence of learning-by-doing; on average, an additional year of EXPERIENCE decreases reading time per case by 5.4%, holding all else constant. On average, reading time per case more than doubles after a temporary period of zero workload due to restarting effects (RESTART). Turning to our main independent variable, we find that, on average, DEVIATION is associated with slower reading times. Compared to cases that are first in queue, cases that are deviations take 13.3% longer on average. This supports Hypothesis 4B rather than Hypothesis 4A in our setting. These results provide evidence of the cost of exercising discretion and call for managerial and academic attention.

Though they tend to worsen performance, deviations from the queue are frequent. Why do individuals take actions that ultimately work against their own interests? To understand this question (and test Hypotheses 5A, 5B, and 5C), we estimate a model that distinguishes the effects of different types of deviations. Column 2 of Table 4.4 shows the results from maximum likelihood estimation of the simultaneous equation system. The predicted mean reading times per case (in minutes) derived from this model are shown in Table 4.5.

Worker Experience. We first discuss the heterogeneous effects of deviations by worker experience (Table 4.5.A). We find that tenure ameliorates the negative effect of deviation on performance, consistent with Hypothesis 5A and suggesting that radiologists may be more efficient at deviating or choose better types of deviations as they become more experienced. At each integer level of years of experience, deviations have a higher predicted mean reading time than cases where the radiologist follows the assigned queue order, suggesting that learning about how to deviate does not overcome the net cost of deviating in our sample. The predicted mean reading time for deviations after three years at the company is equivalent to the prediction for adherence to the assigned order by newcomers in their first year ($\chi^2(1)=0.60$, $p = 0.4405$), indicating that the deviation penalty is large enough to suppress the learning from two years of experience.

The overall effect of experience on the performance impact of deviations depends on how experience affects both the individual impact of each deviation and the frequency with which they occur. To evaluate this, we consider how productivity changes for a radiologist over the course of a year. We combine the results from the deviation and the performance models. A one-year increase in experience from the average (two years) increases the frequency of deviation by 7.71 percentage points, from 41.9% to 49.6%. At the same time, it reduces the penalty associated with each deviation, from a 9% increase in reading time (compared to cases that are first in queue) to an 8% increase. Combined, these results suggest that a one-year increase of experience from the average results in a 5% decrease in reading times, on average. Given that, on average, a radiologist interprets approximately 11,000 cases per year, the time saved during a year by an additional year of experience corresponds to 33 hours. If the deviation tendency had not increased, the one additional year of experience would have delivered a productivity boost of 37 hours (i.e., four extra hours). Thus, experience still leads to better performance, but the improvement is lower than what would have happened if the deviation tendency had not increased.

To understand how the contents of the queue affect the performance impact of exercising discretion, we next consider two deviation strategies that depend on queue contents: (1) choosing the shortest case in the queue and (2) repeating case type.

Shortest Expected Processing Time (SEPT). According to Table 4.5.B, following SEPT increases the predicted mean reading time by 2% and 5% for non-deviations and deviations, respectively ((3.83-3.77)/3.77=0.02; (4.26-4.04)/4.04=0.05). The deviation penalty corresponds to 11% and 7% of the reading time when following SEPT and otherwise, respectively, based on the second and first rows of the table ((4.26-3.83)/3.83=0.11; (4.04-3.77)/3.77=0.07). Therefore, our results suggest that following a SEPT policy hurts performance, in general, and increases the performance penalty of deviations, supporting Hypothesis 5B.

Repetition of Case Type. A common reason cited by radiologists for deviation is the desire to batch cases of the same type. As discussed above, research documents that task repetition is associated with improved performance; this is the argument behind why an individual would choose to deviate towards batching (Hypothesis 3). To analyze the impact of batching on performance without accounting for the deviation choice (and hence, without a deviation model), we run an ordinary least squares model of completion time. We find that, all else constant, average reading times per case tend to be 1.7% lower for those cases that are repetitions of the prior case type than for those cases that are not repetitions (Columns 3 and 4 of Table 4.4). This confirms that the general finding of batching being associated with improved performance holds in this setting. Given these results, as well as the aforementioned received wisdom of operations management, knowledgeable individuals might conclude that they should use their discretion to increase the number of repetitions—and reduce the number of case-type "set ups"—by reordering tasks. The question thus becomes whether these deviations towards batching are, in fact, beneficial for performance.

To answer this question, we return to our simultaneous-equations system and compare the predicted mean reading times depending on whether the case is a repeat of the prior case type and whether it represents a deviation from the assigned queue order (Table 4.5.C). When adhering to the assigned sequence, the predicted mean reading time per case is 1% lower when there is a natural (i.e., without deviation) repetition of case type than when there is neither repetition nor deviation (left column, (3.77-3.8)/3.8=-0.01, $\chi^2(1) = 4.49$, $p = 0.0340$). Conditional on deviating, the predicted mean read time is 3% lower when there is a repetition than otherwise (right column, (4.04-4.15)/4.15=-0.03, $\chi^2(1) = 19.00$, $p < 0.0001$). Thus, the results corroborate that batching is generally associated with superior performance. Batching tends to hurt performance, however, when it is the result of queue reordering; the predicted mean reading time is 6% higher when the radiologist deviates from the queue to take a case that creates a repetition (bottom right cell) compared to cases in which a radiologist neither deviates nor batches (top left cell, (4.04-3.8)/3.8=0.06, $\chi^2(1) = 18.42$, $p < 0.0001$). Though detrimental, this deviation penalty is smaller than the 9% increase in predicted mean reading time when deviating from the queue by taking a case that is *not* a repetition (top row, (4.15-3.8)/3.8=0.09, $\chi^2(1) = 43.06$, $p < 0.0001$). Overall, these results support Hypothesis 5C and suggest that deviations towards batching are less detrimental than other deviations but still have a negative net effect on performance compared to adhering to the original order. Hence, contrary to expectations that commonly overlook the costs of exercising discretion, with respect to completion time, we find that radiologists should forgo deviations that are aimed at taking advantage of batching. Radiologists would complete their reads faster, on average, if they did not deviate to take advantage of batching, as the benefit of repetition does not compensate for the cost associated with reordering the queue in this setting.

4.5.3. Evaluating Alternative Policies

We evaluate the overall impact of deviations on productivity using the predicted reading times from Table 4.5.D to compare the status quo (current deviation policy) with three benchmarks involving no deviations. Each benchmark considers a centralized ordering policy with a different assigned task sequence.

Original sequence. We estimate the reading time for each case based on the values that REPEAT and SEPT would have had if the radiologists would have followed the assigned queue order, where SEPT was computed over the average queue (five cases), which includes the current case and the next four cases. The resulting average reading time is 3.79 percent below that corresponding to the status quo. The improvement in speed would have saved 2,494 hours per year for the company. These time savings translate into 39,434 cases per year of additional reading, or an estimated $451,385 salary savings per year for the company.[4] Translated to the bottom line, these savings would have increased annual profits by 3.1%.

REPEAT and SEPT sequence. The second benchmark is a policy that batches tasks of the same type together and sequences the resulting batches within the shift in order of increasing expected processing time. Although SEPT is associated with lower performance in our setting, radiologists show a tendency towards this policy. This policy might gain doctors' acceptance, as it uses their revealed preferences, hence reducing their desire to deviate from the queue. Compared to the status quo, this policy implies a 2.97% decrease in reading times. Over the course of a year, this represents 1,957 hours saved, 30,691 additional cases read, $354,278 of labor cost savings, and a 2.4% increase in annual profits.

REPEAT without SEPT sequence. Though a policy that sequences similar tasks together and longer tasks first could be harder to implement, as radiologists exhibit a tendency towards prioritizing those tasks they expect to complete faster, SEPT is generally associated with lower performance in our setting. Hence, on the basis of our empirical results, this policy is expected to be the best. Compared to the status quo, this policy implies a 4.27% decrease in reading times. Over the course of a year, this accounts for 2,809 hours saved, 44,640 additional cases, and an

[4] We use the estimated median hourly wage for a Radiologist of $181. Source: Physician - Radiology Hourly Wages, http://www1.salary.com/radiologist-hourly-wages.html, accessed on April 27, 2016. The annual profit comparison uses confidential, company data.

estimated $508,426 of labor cost savings. Translated to the bottom line, these savings would have increased annual profits by 3.5%.

4.5.4. Other Dimensions of Performance

Our analysis focuses on short-term speed performance, the primary concern of managers in this company. An important question is how deviations affect other dimensions of performance, specifically, longer-term speed, employee turnover, and quality. We find that other performance dimensions are mostly unaffected.

First, we examine the effects of deviations on longer-term speed performance. Based on extensions of our full simultaneous-equations model, we find that past deviations (measured as either deviations as a proportion of total cases prior to the current case or deviations as a proportion of total cases prior to the current shift) do not affect current speed. In addition, despite the detrimental instantaneous effect of SEPT on the current reading, deviating towards SEPT could have an effect on subsequent tasks by "alleviating" the work of the radiologist. To explore such a delayed effect of SEPT on subsequent tasks, we look at the effect of past SEPT—measured as lagged SEPT (i.e., an indicator for whether the prior case interpreted by this radiologist corresponded to SEPT); two lags; three lags; or the proportion of cases that were consistent with SEPT since the beginning of the shift—on current reading time and find that it is associated with slower speed. Thus, SEPT (and hence deviating towards SEPT) does not help future speed.

Second, we examine employee turnover. Departure is infrequent in our sample, with only 6% of radiologists not interpreting cases by the end of the sample. Based on logistic and survival analysis, the radiologists' deviations do not predict whether they depart. Hence, it is reasonable to conclude that turnover remains largely unaffected by task sequencing choice.

Third, we look at the effect of deviations on quality, measured by whether there was a discrepancy found for the case. We find that quality is not affected by deviations. Furthermore, there is no evidence of the two specific task sequence strategies—SEPT and REPEAT—affecting quality.

4.5.5. Why Deviate When It Hurts

Given these findings, why would individuals deviate from their assigned queue order? In the case of following a SEPT policy, there are likely two main reasons. One is that individuals might pursue an alternative performance metric, perhaps seeking to reduce client waiting time rather than reading time. In our setting, this was not a strategy the company expected (nor wanted) the radiologists to follow, as the waiting time could be kept under control by adjusting the pool of radiologists on duty. An alternative explanation is that individuals misperceived the implications of following a SEPT approach. We note the negative correlation between DEVIATION and LnREADTIME; it is not until we control for case-type fixed effects that the relationship becomes positive. Hence, precisely because doctors are switching toward shorter cases when they choose the shortest cases within the queue, a SEPT policy may create the mental illusion of working faster. Conditional on a particular case type being selected, the reading time tends to be higher when it has the shortest expected processing time within the queue, but this outcome may be difficult for the individual to anticipate or observe.

In the case of batching, our analysis illustrates a phenomenon that could disentangle the paradox surrounding the detrimental exercise of discretion for other types of deviations as well. We argue that one plausible explanation is that individuals have the illusion of improving performance by exercising discretion because they underestimate, or fail to consider entirely, the time required to do so (e.g., the time required to look through the queue to determine the exact case to complete next and the related cognitive distraction). This task selection time is an opportunity cost. As such, one might expect these costs to be overlooked, as evaluating opportunity costs requires decisionmakers to account for unrealized, implicit options (Frederick et al. 2009). Seeking to increase their speed, individuals may exercise discretion in a manner that they believe will improve their workflow but that ends up reducing their speed. These results suggest a possible behavioral challenge, as individuals actively pursue strategies that would in fact have been beneficial for performance were it not for the unrecognized costs of pursuing those strategies (e.g., task selection costs). Our findings provide evidence of the vulnerabilities of self-management and the potential value of using centralized management in settings where the costs of exercising discretion are particularly high relative to the benefits.

The proposition that individuals may underestimate the costs of exercising discretion—and hence believe that deviations help their productivity even in situations when they are detrimental—may explain the fact that deviations rise with experience. As workers deviate more over time, they might erroneously attribute their performance improvements from learning-by-doing to their exercised discretion. Because they only observe their performance improving over time, they may not realize that their deviation behavior is actually hurting them. That is, learning-by-doing may mask any deteriorating effect of deviations, so individuals may increasingly rely on their discretion over time, even in situations when exercising that discretion undermines overall improvement.

4.5.6. Managerial Implications

Our paper contributes not only to the theory of operations management but also to its practice. First, we show that discretion has costs that need to be balanced against its potential benefits. Our finding that deviations are, on average, related to worse performance serves as a warning to managers and workers concerning the costs of exercising certain types of discretion. As noted by a radiologist we interviewed, “I always thought that by reading the easiest cases on the queue first, one might end up reading faster, but it seems that the opposite is true. To lend support to your findings, I have noticed that I read more cases on the days I don't go through my list and choose the order in which to read cases. I do think that incorporating information on how individuals order their task sequences could help speed up the process of reading studies and avoid duplicating time spent on organizing workflow.”

Managers should pay attention to the effects of deviations on productivity in their settings. Although an initial task sequence assignment might not be optimal, allowing front-line workers to take an active role in scheduling might not be advisable in settings where the time required to exercise discretion exceeds the benefits of doing so. In any setting, reorganizing the queue takes time, so managers should look for ways to reduce this time while maintaining the benefits from better ordering of queues. Productive nudges from managers could include recommendations on the task to complete next. Reducing workers’ desire or need for task reordering, through education about the costs of deviation, centralized queue management or more-responsive software for task ordering, can be a way for managers to improve productivity. In our setting, adherence to the assigned sequence of tasks could result in a meaningful increase in firm profits of roughly 3%.

Our findings regarding the ability of experience and certain types of deviations to offset—though only partially—the detrimental average effect of deviations on performance suggest that organizations need to take these variables into account in structuring work. In settings, such as ours, where deviations tend to reduce performance, the benefits from experience of senior employees are reduced by their tendency to deviate from the queue more often. In such contexts, managers have the opportunity to improve performance by creating awareness, finding ways to persuade experienced employees to adhere to the queue sequence or be even more thoughtful about when they deviate. More generally, in any setting, managers can collaborate with workers to improve scheduling strategies. For example, workers can identify when they deviate from the assigned order and why they believe it improves performance. With this knowledge, an individual or organization could test these ideas. Identifying productive deviations could help the individual's productivity and perhaps result in beneficial changes to the organization's recommended task schedule. Finally, managers should consider the time required to exercise discretion when using analytics to inform workplace practices, as illustrated by the fact that the time required to reorder a queue offsets the beneficial effects of batching in our setting.

4.6. Conclusion

Though prior literature studies task ordering from the perspective of a central scheduler, those who execute the tasks often have discretion over the actual order in which they are performed. In practice, either by constraint or by choice, the delegation of task-scheduling decisions is common and results in individuals self-scheduling their work. Due to limited research on discretionary task ordering, little is known about how managers should manage scheduling when such discretion exists. Understanding when and how individuals exercise this discretion informs decisions about system design, whether (or to what extent) to grant discretion, how to nudge behavior toward particular uses of discretion, and how to adjust policies to incorporate responses expected from front-line employees.

We consider this underexplored territory by analyzing the drivers and consequences of exercising discretion over task sequence in a setting where deviations from the queue are observable. Examining a proprietary dataset from a teleradiology company in which radiologists are assigned a queue of cases to interpret but are not restricted to follow that order, we find that individuals are more likely to exercise discretion via deviations from the queue when their experience is greater. This finding is consistent with the view that experience leads to superior ability to identify high-leverage opportunities for deviation and/or higher self-confidence to deviate. Our results highlight both the potential power and the limitations of learning-by-doing. On one hand, an individual can learn over time how to exercise discretion over task sequence more effectively. On the other, this learning may not overcome the costs of exercising discretion and experienced individuals may also fail to assess those costs appropriately. Thus, deviations may remain unnoticeably detrimental, even for experienced individuals. We also show that individuals in our setting have a higher probability of using discretion when doing so creates an opportunity to follow either of two particular strategies—SEPT or batching.

The exercise of discretion via deviations from the queue has a net negative effect in our empirical setting, at least in the short to intermediate term. Doctors often choose the "wrong" tasks (e.g., SEPT, found to be detrimental, despite conflicting theoretical predictions), and even when they choose certain "right" tasks (e.g., repetitions), the resulting benefits are smaller than the time-cost of deviating. Deviations when the individual is more experienced or to repeat task type are less detrimental but are still related to worse performance compared to the case of no deviation. That is, deviations harm performance, even when they are pursued to take advantage of scheduling strategies that are assumed to be—and in the case of batching, actually are—

beneficial to performance. In such cases, the benefits of discretion via deviation may not compensate for the costs of exercising it.

4.6.1. Contributions

Our study offers five main contributions. First, we analyze the implications of discretionary queue management. Scheduling tasks is a critical determinant of employee and organizational productivity (Pinedo 2012). Most studies on task scheduling examine contexts where scheduling is solely a managerial decision. This is often not the case in practice, however, as front-line workers frequently have discretion in scheduling. Future analytical work should incorporate the endogeneity of task sequences. Accounting for deviations may more accurately reflect many situations and may lead to unexpected recommendations on how to structure work when workers may choose not to implement prescribed schedules.

Second, we provide evidence of costs of *exercising* discretion that may, in settings such as ours, outweigh the associated benefits. Individuals consistently deviate from the order of tasks within their queues despite the fact that this reordering results in cases taking 13% longer, on average. Our investigation of batching suggests an important reason why this behavior occurs. When cases are naturally batched in the queue, completion times are faster. When individuals deviate in a manner that is consistent with batching, however, completion times are slower than in situations where they do not deviate. Thus, it is possible that individuals believe that deviations will improve productivity even though they do not. This creates an opportunity to examine ways to encourage individuals to assess the costs of discretion, find ways to reduce such costs, and deviate less when the costs outweigh the benefits. This may be by shifting to centralized ordering of tasks, avoiding duplication of decisions, or encouraging fewer deviations through nudges and information.

Third, we identify conditions under which an individual is more likely to deviate – namely, when that individual has greater experience and when the queue offers shorter cases, more batching opportunities, more tasks, or a greater variety of tasks. Each of these results offers important theoretical grounding for system design. Future research should explore alternative reasons to deviate based on principles of operations management and identify additional strategies – advantageous or deleterious – that individuals follow in task selection.

Fourth, we provide evidence of the difficulties of achieving optimal behavior in practice. Analytical models have considered workers using their discretion to alter the number of tasks performed and their processing speeds, assuming they do so to maximize a given utility function. Our findings, however, call into question whether workers are able to make these calculations and thereby contribute to recent empirical work that illustrates suboptimal behavior.[5]

Finally, from a methodological perspective, we present a novel strategy to identify instrumental variables to measure the effects of discretion in queuing settings. We propose that the choice of exercising discretion is affected by the composition of the tasks within the queue, and, as long as it can be considered exogenous, such composition can provide valid instruments. Exploiting random assignment of tasks to individual queues, together with variation in queue characteristics, we construct instrumental variables based on the expected duration of the pending tasks, the similarity of tasks with the one just finished, and the type of the task in the first position within the queue. More broadly, our approach can be applied to identify different instruments to estimate the impact of operational decisions related to queues.

[5] We thank an anonymous reviewer for this suggestion.

4.6.2. Conclusion

Despite the prevalence of discretion in practice, little is known about how workers use discretion over task sequence and how to manage scheduling when discretion exists. Seeking to fill this gap, we consider the implications of discretionary queue management. We find that reordering queues is common in our setting but tends to have negative implications for performance. We identify conditions that encourage—and document the performance effects of—deviations. Overall, our results highlight the need for both managers and academics to pay careful attention to the use of worker discretion in queue management.

4.7. References

Allcott, H. and R. L. Sweeney (2016). "The role of sales agents in information disclosure: Evidence from a field experiment." *Management Sci.*

Amar, M., D. Ariely, S. Ayal, C. E. Cryder and S. I. Rick (2011). "Winning the battle but losing the war: The psychology of debt management." *Journal of Marketing Research* **48**(SPL): S38-S50.

Anand, K. S., M. F. Paç and S. Veeraraghavan (2011). "Quality–Speed Conundrum: Trade-offs in Customer-Intensive Services." *Management Sci.* **57**(1): 40-56.

Argote, L. and E. Miron-Spektor (2011). "Organizational learning: From experience to knowledge." *Organ. Sci.* **22**(5): 1123-1137.

Bandura, A. (1977). "Self-efficacy: Toward a unifying theory of behavioral change." *Psych. Rev.* **84**(2): 191-215.

Bassamboo, A. and R. S. Randhawa (2015). "Scheduling homogeneous impatient customers." *Management Science*.

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

Bendoly, E., M. Swink and W. P. Simpson (2014). "Prioritizing and monitoring concurrent project work: Effects on switching behavior." *Production and Operations Management* **23**: 847–860.

Berman, O., R. C. Larson and E. Pinker (1997). "Scheduling workforce and workflow in a high volume factory." *Management Sci.* **43**(2): 158-172.

Berry Jaeker, J. and A. L. Tucker (2015). "Past the point of speeding up: The negative effects of workload saturation on efficiency and quality." *Management Sci.*

Boudreau, J., W. Hopp, J. O. McClain and L. J. Thomas (2003). "On the interface between operations and human resources management." *Manufacturing Service Oper. Management* **5**(3): 179-202.

Bowman, E. H. (1963). "Consistency and optimality in managerial decision making." *Management Sci.* **9**(2): 310-321.

Campbell, D. and F. X. Frei (2011). "Market heterogeneity and local capacity decisions in services." *Manufacturing Service Oper. Management* **13**(1): 2-19.

Davis, A. M., E. Katok and N. Santamaría (2014). "Push, pull, or both? A behavioral study of how the allocation of inventory risk affects channel efficiency." *Management Sci.* **60**(11): 2666-2683.

Delasay, M., A. Ingolfsson, B. Kolfal and K. Schultz (2015). "Load effect on service times." *Working Paper.*

Fisher, M. L. and C. D. Ittner (1999). "The impact of product variety on automobile assembly operations: Empirical evidence and simulation analysis." *Management Sci.* **45**(6): 771-786.

Frederick, S., N. Novemsky, J. Wang, R. Dhar and S. Nowlis (2009). "Opportunity cost neglect." *Journal of Consumer Research* **36**: 553-561.

Freeman, M., N. Savva and S. Scholtes (2016). "Gatekeepers at work: an empirical analysis of a maternity unit " *Management Sci.*

Froehle, C. M. and D. L. White (2014). "Interruption and forgetting in knowledge-intensive service environments." *Production and Operations Management* **23**(4): 704-722.

Gal, D. and B. B. McShane (2012). "Can small victories help win the war? Evidence from consumer debt management." *Journal of Marketing Research* **49**(4): 487-501.

Gans, N., G. Koole and A. Mandelbaum (2003). "Telephone call centers: Tutorial, review, and research prospects." *Manufacturing Service Oper. Management* **5**(2): 79-141.

Gino, F. and G. P. Pisano (2008). "Toward a theory of behavioral operations." *Manufacturing Service Oper. Management* **10**(4): 676-691.

Goh, J. and N. G. Hall (2013). "Total cost control in project management via satisficing." *Management Sci.* **59**(6): 1354-1372.

Greene, W. (2004). "The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects." *The Econometrics Journal* **7**: 98–119.

Halsted, M. J. and C. M. Froehle (2008). "Design, implementation, and assessment of a radiology workflow management system." *American Journal of Roentgenology* **191**(2): 321-327.

Hasija, S., E. Pinker and R. A. Shumsky (2010). "OM Practice—Work expands to fill the time available: Capacity estimation and staffing under Parkinson's Law." *Manufacturing Service Oper. Management* **12**(1): 1-18.

Heckman, J. J. (1978). "Dummy endogenous variables in a simultaneous equation system." *Econometrica* **46**: 931–959.

Hopp, W. J., S. M. R. Iravani and G. Y. Yuen (2007). "Operations systems with discretionary task completion." *Management Sci.* **53**(1): 61-77.

Hossain, T. and J. A. List (2012). "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Sci.* **58**(12): 2151-2167.

Huckman, R. S. and G. P. Pisano (2006). "The firm specificity of individual performance: Evidence from cardiac surgery." *Management Sci.* **52**(4): 473-488.

Huckman, R. S., B. R. Staats and D. M. Upton (2009). "Team familiarity, role experience, and performance: Evidence from Indian software services." *Management Sci.* **55**(1): 85-100.

KC, D. and B. R. Staats (2012). "Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance." *Manufacturing Service Oper. Management* **14**(4): 618-633.

KC, D. and C. Terwiesch (2009). "Impact of workload on service time and patient safety: An econometric analysis of hospital operations." *Management Sci.* **55**(9): 1486-1498.

Kim, S.-H., C. W. Chan, M. Olivares and G. Escobar (2015). "ICU admission control: An empirical study of capacity allocation and Its implication for patient outcomes." *Management Sci.* **61**(1): 19-38.

Kremer, M., B. Moritz and E. Siemsen (2011). "Demand forecasting behavior: System neglect and change detection." *Management Sci.* **57**(10): 1827-1843.

Lapré, M. A. and I. M. Nembhard (2010). "Inside the organizational learning curve." *Foundations and Trends in Technology, Information and Operations Management* **4**(1): 1-103.

Lu, Y., A. Hechling and M. Olivares (2014). "Productivity analysis in services using timing studies."

MacDuffie, J. P. (1997). "The road to "Root Cause": Shop-floor problem-solving at three auto assembly plants." *Management Sci.* **43**(4): 479-502.

Maddala, G. S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge, Cambridge University Press.

Narayanan, S., S. Balasubramanian and J. M. Swaminathan (2009). "A matter of balance: Specialization, task variety, and individual learning in a software maintenance environment." *Management Sci.* **55**(11): 1861-1876.

Phillips, R., A. S. Şimşek and G. v. Ryzin (2015). "The effectiveness of field price discretion: Empirical evidence from auto lending." *Management Sci.*

Pierce, L., D. Snow and A. McAfee (2014). "Cleaning house: The impact of information technology monitoring on employee theft and productivity." *Management Sci.*

Pinedo, M. L. (2012). Scheduling: Theory, Algorithms, and Systems. New York, Springer.

Pinedo, M. L. and B. P.-C. Yen (1997). "On the design and development of object-oriented scheduling systems." *Annals of Operations Research* **70**(1): 359–378.

Pisano, G. P., R. M. J. Bohmer and A. C. Edmondson (2001). "Organizational differences in rates of learning: Evidence from the adoption of minimally invasive cardiac surgery." *Management Sci.* **47**(6): 752-768.

Powell, A., S. Savin and N. Savva (2012). "Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient." *Manufacturing Service Oper. Management*.

Roberti, R., E. Bartolini and A. Mingozzi (2015). "The fixed charge transportation problem: An exact algorithm based on a new integer programming formulation." *Management Sci.*

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond and S. L. Kronick (2014). "Complexity-augmented triage: A tool for improving patient safety and operational efficiency." *Manufacturing & Service Operations Management* **16**(3): 329-345.

Schultz, K. L., D. C. Juran and J. W. Boudreau (1999). "The effects of low inventory on the development of productivity norms." *Management Sci.* **45**(12): 1664-1678.

Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain and L. J. Thomas (1998). "Modeling and worker motivation in JIT production systems." *Management Sci.* **44**(12): 1595-1607.

Schultz, K. L., J. O. McClain and L. J. Thomas (2003). "Overcoming the dark side of worker flexibility." *J. of Operations Management* **21**(1): 81-92.

Schultz, K. L., T. Schoenherr and D. Nembhard (2010). "An example and a proposal concerning the correlation of worker processing times in parallel tasks." *Management Sci.* **56**(1): 176-191.

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

Shumsky, R. A. and E. J. Pinker (2003). "Gatekeepers and referrals in services." *Management Sci.* **49**(7): 839-856.

Simon, H. A. (1978). "Rationality as process and as product of thought." *American Economic Review* **68**(2): 1-16.

Staats, B. R., D. J. Brunner and D. M. Upton (2011). "Lean principles, learning, and knowledge work: Evidence from a software services provider." *J. of Operations Management* **29**(5): 376-390.

Staats, B. R. and F. Gino (2012). "Specialization and variety in repetitive tasks: Evidence from a Japanese bank." *Management Sci.* **58**(6): 1141-1159.

Tan, T. F. and S. Netessine (2014). "When does the devil make work? An empirical study of the impact of workload on server's performance." *Management Sci.* **60**(6): 1574-1593.

Taylor, F. W. (1911). The Principles of Scientific Management. New York, Harper & Brothers.

Truong, V. A. (2015). "Optimal advance scheduling." *Management Science* **61**(7): 1584-1597.

Tucker, A. L. (2007). "An empirical study of system improvement by frontline employees in hospital units." *Manufacturing Service Oper. Management* **9**(4): 492–505.

Tucker, A. L., I. M. Nembhard and A. C. Edmondson (2007). "Implementing new practices: An empirical study of organizational learning in hospital intensive care units." *Management Sci.* **53**(6): 894-907.

van Donselaar, K., V. Gaur, T. van Woensel, R. Broekmeulen and J. Fransoo (2010). "Ordering behavior in retail stores and implications for automated replenishment." *Management Sci.* **56**(5): 766-784.

Wang, L., I. Gurvich, J. A. Van Mieghem and K. J. O'Leary (2015). "Collaboration and professional labor productivity: An empirical study of physician workflows in a hospital."

Zhu, E., T. G. Crainic and M. Gendreau (2014). "Scheduled service network design for freight rail transportation." *Operations Research* **62**(2): 383-400.

Table 4.1. Descriptive Statistics

	Variable	Definition	Mean	Std. Dev.	Min	Max
(1)	READTIME (Minutes)	Amount of time (in minutes) spent working on the current case. This is the completion time.	3.755	3.812	0	30
(2)	DEVIATION (Indicator)	Whether the worker deviates from the assigned order and selects a case other than the first one in the queue.	0.419	0.493	0	1
(3)	QSIZE (Count)	Length of radiologist's queue (i.e., count of pending jobs) when selecting the current case to be read next.	5.553	3.791	2	58
(4)	QVARIETY (Index 0-1)	Variety of case types in the queue, measured as one minus the Herfindahl index of different case types in the queue.	0.556	0.198	0	0.918
(5)	NUM_IMAGES (Count)	Number of images in the current case.	1.414	0.598	1	17
(6)	ORDER_IN_SHIFT (Count)	Case number in the shift— the number of cases read by the radiologist since the shift start, including the current case.	57.653	51.612	2	459
(7)	RESTART (Indicator)	Whether the queue was empty when the previous case was finished.	0.013	0.114	0	1
(8)	EXPERIENCE (Years)	Employee job tenure in years (with decimals), measured from the number of days that the radiologist has been working at the firm when interpreting the current case.	1.906	1.234	0	5.501
(9)	SEPT_OPPTY (Indicator)	Whether the first case in the queue is *not* of the Shortest Expected Processing Time (SEPT) type within the queue.	0.635	0.481	0	1
(10)	SEPT (Indicator)	Whether the current case is of the Shortest Expected Processing Time (SEPT) within the queue.	0.455	0.498	0	1
(11)	REPEAT_OPPTY (Indicator)	Whether the first case in the queue is different from the case just finished by this radiologist but it is possible to repeat prior type by choosing another case in the queue.	0.206	0.404	0	1
(12)	REPEAT (Indicator)	Whether the current case is of the same type as the case just finished by this radiologist (batching).	0.125	0.331	0	1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(2)	-0.046										
(3)	-0.242	0.200									
(4)	-0.186	0.157	0.430								
(5)	0.182	-0.088	-0.031	-0.239							
(6)	-0.177	0.042	0.159	0.046	0.043						
(7)	0.159	0.136	-0.084	-0.056	-0.032	-0.019					
(8)	-0.054	0.068	0.109	0.071	-0.007	-0.053	-0.013				
(9)	-0.025	0.169	0.213	0.398	0.136	0.048	0.001	0.043			
(10)	-0.050	0.048	-0.189	-0.335	-0.392	-0.056	0.069	-0.014	-0.545		
(11)	-0.065	0.104	0.193	0.226	-0.002	0.007	-0.016	0.039	0.308	-0.155	
(12)	-0.092	0.070	0.024	0.031	-0.257	-0.040	0.011	0.013	-0.116	0.307	0.203

Notes. The unit of analysis is an individual case. N=2,408,218. **Expected processing time (EPT)** is calculated as the average reading time for the given case type (technology and anatomy) for the focal radiologist. A case is a **repetition** if it is of the same type as the previous case read by the radiologist. We do not consider a repetition of the anatomy categories "body" or "other" as a repetition, since each case in these categories is unique. We compute whether there is a repetition using the full sample, before imposing the restrictions described in the Data section.

Table 4.2. Observed Deviation Rate for Cases Selected among Different Queues

Queue Characteristic (IV)	Percentage of Deviations if No (IV = 0)	Percentage of Deviations if Yes (IV = 1)
Opportunity to follow SEPT (SEPT_OPPTY)	30.9%	48.3%
Opportunity to repeat case-type (REPEAT_OPPTY)	39.3%	52.1%
First case type indicators	41.0%-44.8%	33.0%-78.3%

Note. This table describes the percentage of decisions (n=2,408,218) regarding which case to work on next among queues of different characteristics in which the radiologist deviated from the next case in the queue (DEVIATION=1).

Table 4.3. Drivers of Deviations

Dependent Variable: DEVIATION	(1a) Coefficients	(1b) AME	(2a) Coefficients	(2b) AME	(3a) Coefficients	(3b) AME
QSIZE	0.0524*** (0.0039)	0.0184	0.0521*** (0.0039)	0.0183	0.0512*** (0.0038)	0.0180
QVARIETY	0.4408*** (0.0220)	0.1550	0.4307*** (0.0211)	0.1512	0.3447*** (0.0195)	0.1210
NUM_IMAGES	-0.0103 (0.0100)	-0.0036	-0.0084 (0.0101)	-0.0030	-0.0079 (0.0101)	-0.0028
ORDER_IN_SHIFT	-0.0004*** (0.0001)	-0.0001	-0.0004*** (0.0001)	-0.0001	-0.0004*** (0.0001)	-0.0001
EXPERIENCE			0.2195*** (0.0158)	0.0771	0.2201*** (0.0158)	0.0773
SEPT_OPPTY					0.0685*** (0.0068)	0.0241
REPEAT_OPPTY					0.0436*** (0.0046)	0.0154

Notes. This table reports the results from maximum-likelihood probit estimation. The unit of observation is a radiologist's case. The number of observations is 2,408,218. The dependent variable is whether the case was a deviation (DEVIATION). Robust standard errors (in parentheses) are clustered by radiologist. All specifications include fixed effects for case type (technology and anatomy) of the first case in the queue, case type of the focal case (i.e., the case selected and read), radiologist, day of week, and year fixed effects. AME = average marginal effect. *** p < 0.01, ** p < 0.05, * p < 0.10.

Table 4.4. Performance Implications of Deviations and Task Repetition

	(1)	(2)	(3)	(4)
Dependent Variable:	LnREADTIME	LnREADTIME	LnREADTIME	LnREADTIME
	ML	ML	OLS	OLS
QSIZE	-0.0290***	-0.0280***	-0.0267***	-0.0265***
	(0.0023)	(0.0022)	(0.0020)	(0.0020)
QVARIETY	-0.2230***	-0.1894***	-0.1997***	-0.1643***
	(0.0146)	(0.0141)	(0.0139)	(0.0137)
NUM_IMAGES	0.1714***	0.1711***	0.1711***	0.1710***
	(0.0061)	(0.0062)	(0.0063)	(0.0063)
ORDER_IN_SHIFT	-0.0003***	-0.0004***	-0.0004***	-0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
EXPERIENCE	-0.0560***	-0.0487***	-0.0459***	-0.0461***
	(0.0128)	(0.0124)	(0.0125)	(0.0125)
RESTART	0.8412***	0.8336***	0.8688***	0.8682***
	(0.0195)	(0.0195)	(0.0179)	(0.0179)
DEVIATION	0.1250***	0.0875***		
	(0.0158)	(0.0146)		
EXPERIENCE * DEVIATION		-0.0090***		
		(0.0035)		
SEPT		0.0161***		0.0344***
		(0.0051)		(0.0049)
SEPT * DEVIATION		0.0373***		
		(0.0047)		
REPEAT		-0.0099**	-0.0170***	-0.0171***
		(0.0047)	(0.0048)	(0.0048)
REPEAT * DEVIATION		-0.0170***		
		(0.0048)		
ρ	-0.0850***	-0.0441***		
ρ Standard Error	(0.0129)	(0.0105)		
Test $\rho = 0$ (p-value)	0.0000	0.0000		

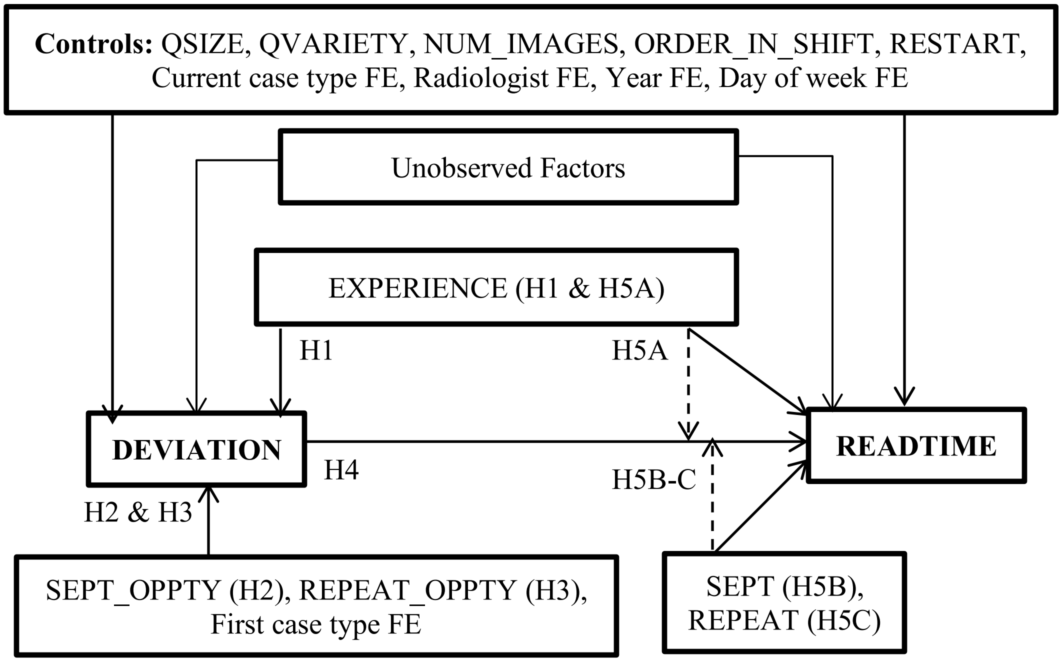
Notes. Columns 1 and 2 report the results from joint maximum-likelihood estimation of the performance model (probit deviation model with full specification, as in Column 3 of Table 4.3, not shown). Columns 3 and 4 report the results from ordinary least squares estimation of the performance model. The unit of observation is a radiologist's case. The number of observations is 2,408,218. Robust standard errors (in parentheses) are clustered by radiologist. All models include fixed effects for case type (technology and anatomy) of the focal case (i.e., the case selected and read), radiologist, day of week, and year fixed effects. The probit models also include fixed effects for the case type of the first case in the queue. *** p < 0.01, ** p < 0.05, * p < 0.10.

Table 4.5. Predicted Mean Reading Time (in Minutes) per Case

	No Deviation (DEVIATION=0)	Deviation (DEVIATION=1)	Difference
A. By Number of Years of Experience			
1 year of EXPERIENCE	3.97	4.35	0.39 (10%)
2 years of EXPERIENCE	3.78	4.11	0.33 (9%)
3 years of EXPERIENCE	3.60	3.88	0.28 (8%)
4 years of EXPERIENCE	3.43	3.66	0.23 (7%)
5 years of EXPERIENCE	3.27	3.46	0.19 (6%)
B. By Whether Following a Shortest Expected Processing Time (SEPT) Policy			
Not the Shortest Case (SEPT=0)	3.77	4.04	0.27 (7%)
Shortest Case (SEPT=1)	3.83	4.26	0.43 (11%)
C. By Whether Repeating the Case Type of the Predecessor			
No Repetition (REPEAT=0)	3.80	4.15	0.35 (9%)
Repetition (REPEAT=1)	3.77	4.04	0.27 (7%)
D. By Whether Following a SEPT Policy and/or Repeating Case Type			
Not the Shortest Case, No Repetition	3.78	4.05	0.28 (7%)
Not the Shortest Case, Repetition	3.74	3.95	0.21 (6%)
Shortest Case, No Repetition	3.84	4.28	0.44 (11%)
Shortest Case, Repetition	3.80	4.16	0.36 (10%)

Notes. These tables report the predicted mean reading time (in minutes) per case based on the performance model in Column 2 of Table 4.4. The difference within row represents deviation versus non-deviation. In 5.A., the difference within column shows learning. In 5.B., the difference within column shows the effect of SEPT. In 5.C., the difference within column shows the effect of task-type repetition; the first column compares repetitions versus non-repetitions when there is no deviation, and the second column compares repetition versus non-repetition when there is deviation. In 5.D., the first column compares task sequences (in terms of SEPT and REPEAT) when there is no deviation, and the second column compares task sequences (in terms of SEPT and REPEAT) when there is deviation.

Figure 4.1. Causal Diagram


Controls: QSIZE, QVARIETY, NUM_IMAGES, ORDER_IN_SHIFT, RESTART,
Current case type FE, Radiologist FE, Year FE, Day of week FE
Unobserved Factors
EXPERIENCE (H1 & H5A)
H1
H5A
DEVIATION
H4
H5B-C
READTIME
H2 & H3
SEPT_OPPTY (H2), REPEAT_OPPTY (H3),
First case type FE
SEPT (H5B),
REPEAT (H5C)

***Note*.** The instrumental variables (IVs) used to account for the endogeneity of the decision to deviate (DEVIATION) on performance (READTIME) are SEPT_OPPTY, REPEAT_OPPTY and First case type fixed eff

Appendix 4. Supplemental Analysis

To confirm the robustness of our findings, we conduct further analyses. We find that the results are robust to the use of a Linear Probability Model (LPM) rather than the probit regression used in our main models (Table 4.A1, Column 1), the use of a radiologist's case volume as an alternative measure of experience (Table 4.A1, Column 2), the inclusion of hospital fixed effects (Table 4.A1, Column 3), and to alternative cutoffs for shift breaks (Table 4.A2). We explore the longer-term speed performance effects of deviations and find that past deviations (measured as either proportions of deviations prior to the current case, not including the focal case, or proportions of deviations prior to the current shift) do not affect current speed (Table 4.A3). We also explore the effect of past SEPT on speed-performance and find that alternative measures of past SEPT are associated with slower reading times (Table 4.A4).

Table 4.A1. Robustness Checks: Linear Probability Model, Experience Measure, and Hospital Fixed Effects

	(1)	(2a)	(2b)	(3a)	(3b)
Dependent Variable:	DEVIATION	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME
		EXPERIENCE:			
Model:	LPM	Volume of Cases		Hospital Fixed Effects	
QSIZE	0.0183***	0.0506***	-0.0276***	0.0515***	-0.0273***
	(0.0013)	(0.0039)	(0.0023)	(0.0038)	(0.0022)
QVARIETY	0.1113***	0.3536***	-0.1905***	0.3460***	-0.1830***
	(0.0074)	(0.0206)	(0.0143)	(0.0193)	(0.0137)
NUM_IMAGES	-0.0030	-0.0088	0.1711***	-0.0067	0.1710***
	(0.0036)	(0.0100)	(0.0062)	(0.0099)	(0.0063)
ORDER_IN_SHIFT	-0.0001***	-0.0004***	-0.0003***	-0.0005***	-0.0004***
	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
EXPERIENCE	0.0779***	0.0046***	-0.0017***	0.2138***	-0.0459***
	(0.0057)	(0.0005)	(0.0005)	(0.0156)	(0.0123)
SEPT_OPPTY	0.0281***	0.0650***		0.0682***	
	(0.0028)	(0.0068)		(0.0068)	
REPEAT_OPPTY	0.0145***	0.0417***		0.0431***	
	(0.0014)	(0.0049)		(0.0048)	
RESTART			0.8329***		0.8330***
			(0.0196)		(0.0196)
DEVIATION			0.0796***		0.0487***
			(0.0143)		(0.0163)
DEVIATION			-0.0003**		-0.0088***
*EXPERIENCE			(0.0001)		(0.0034)
SEPT			0.0163***		0.0143***
			(0.0051)		(0.0051)
SEPT*DEVIATION			0.0358***		0.0443***
			(0.0047)		(0.0049)
REPEAT			-0.0098**		-0.0112**
			(0.0047)		(0.0046)
REPEAT			-0.0175***		-0.0140***
*DEVIATION			(0.0049)		(0.0048)
Rho			-0.0435***		-0.0085
Rho Standard Error			(0.0108)		(0.0116)

Notes. This table reports the results from the linear probability model (LPM) of the deviation equation (Column 1); and maximum-likelihood estimation (Columns 2 and 3) of the probit deviation model (a) and the performance model (b). The unit of observation is a radiologist's case. The number of observations is 2,408,218. R-squared is 0.1336 for the results in Column 1. Robust standard errors (in parentheses) are clustered by radiologist. In Column 2, EXPERIENCE is the number of cases interpreted by this radiologist up to the current case, in thousands. All specifications include case type (modality and anatomy), radiologist, day of week, and year fixed effects. The specifications in Columns 1a, 2a, and 3a also include fixed effects for the case type of the first case in the queue. The specification in Column 3a also includes hospital fixed effects. *10% statistical significance; **5% statistical significance; ***1% statistical significance.

Table 4.A2. Robustness Checks: Alternative Shift-Break Cutoffs

	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
Dependent Variable:	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME
Shift-Break Cutoffs:	20-minutes	20-minutes	40-minutes	40-minutes	60-minutes	60-minutes	90-minutes	90-minutes	480-minutes	480-minutes
QSIZE	0.0519***	-0.0273***	0.0508***	-0.0280***	0.0505***	-0.0280***	0.0504***	-0.0280***	0.0502***	-0.0280***
	(0.0039)	(0.0021)	(0.0038)	(0.0022)	(0.0037)	(0.0022)	(0.0037)	(0.0022)	(0.0037)	(0.0023)
QVARIETY	0.3608***	-0.1820***	0.3363***	-0.1895***	0.3294***	-0.1900***	0.3276***	-0.1900***	0.3260***	-0.1899***
	(0.0193)	(0.0140)	(0.0196)	(0.0141)	(0.0199)	(0.0139)	(0.0199)	(0.0139)	(0.0199)	(0.0139)
NUM_IMAGES	-0.0119	0.1617***	-0.0060	0.1713***	-0.0053	0.1714***	-0.0053	0.1714***	-0.0052	0.1713***
	(0.0100)	(0.0058)	(0.0103)	(0.0062)	(0.0103)	(0.0062)	(0.0103)	(0.0062)	(0.0103)	(0.0062)
ORDER_IN_SHIFT	-0.0004***	-0.0003***	-0.0004***	-0.0004***	-0.0003***	-0.0004***	-0.0003***	-0.0004***	-0.0002**	-0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
EXPERIENCE	0.2208***	-0.0457***	0.2194***	-0.0484***	0.2194***	-0.0480***	0.2195***	-0.0478***	0.2194***	-0.0479***
	(0.0158)	(0.0124)	(0.0158)	(0.0124)	(0.0158)	(0.0124)	(0.0157)	(0.0124)	(0.0157)	(0.0124)
SEPT_OPPTY	0.0707***		0.0634***		0.0618***		0.0614***		0.0615***	
	(0.0066)		(0.0068)		(0.0069)		(0.0069)		(0.0069)	
REPEAT_OPPTY	0.0437***		0.0502***		0.0559***		0.0575***		0.0594***	
	(0.0049)		(0.0053)		(0.0058)		(0.0060)		(0.0062)	
RESTART		0.7994***		0.8683***		0.8808***		0.8880***		0.8975***
		(0.0193)		(0.0200)		(0.0223)		(0.0229)		(0.0237)
DEVIATION		0.0797***		0.0882***		0.0899***		0.0902***		0.0905***
		(0.0140)		(0.0145)		(0.0142)		(0.0142)		(0.0141)
DEVIATION*		-0.0091***		-0.0091***		-0.0094***		-0.0093***		-0.0091***
EXPERIENCE		(0.0035)		(0.0034)		(0.0033)		(0.0033)		(0.0032)
SEPT		0.0142***		0.0165***		0.0167***		0.0168***		0.0167***
		(0.0051)		(0.0051)		(0.0051)		(0.0051)		(0.0051)
SEPT*		0.0380***		0.0369***		0.0365***		0.0363***		0.0360***
DEVIATION		(0.0048)		(0.0047)		(0.0047)		(0.0047)		(0.0047)
REPEAT		-0.0081*		-0.0101**		-0.0103**		-0.0104**		-0.0103**
		(0.0046)		(0.0047)		(0.0047)		(0.0047)		(0.0047)
REPEAT*		-0.0175***		-0.0173***		-0.0173***		-0.0173***		-0.0172***
DEVIATION		(0.0048)		(0.0049)		(0.0049)		(0.0049)		(0.0049)

Notes. This table reports the results from maximum-likelihood estimation of the probit deviation model (a) and the performance model (b). The unit of observation is a radiologist's case. The number of observations is 2,387,032, 2,412,544, 2,415,623, 2,416,535 and 2,417,453 in Columns 1, 2, 3, 4, 5, respectively. Rho (Standard Error) is -0.0384*** (0.0101), -0.0445*** (0.0105), -0.0455*** (0.0105), -0.0459*** (0.0105), -0.0464*** (0.0104) in Columns 1, 2, 3, 4, 5, respectively. Robust standard errors (in parentheses) are clustered by radiologist. All specifications include case type (technology and anatomy), radiologist, day of week, and year fixed effects. The probit specifications also include fixed effects for the case type of the first case in the queue. *** p<0.01, ** p<0.05, * p<0.1

Table 4.A3. Longer-Term Speed Performance Effects of Deviations

	(1a)	(1b)	(2a)	(2b)
Dependent Variable:	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME
QSIZE	0.0511***	-0.0280***	0.0512***	-0.0348***
	(0.0038)	(0.0022)	(0.0038)	(0.0026)
NUM_IMAGES	-0.0079	0.1711***	-0.0080	0.1938***
	(0.0101)	(0.0062)	(0.0101)	(0.0067)
ORDER_IN_SHIFT	-0.0002*	-0.0004***	-0.0004***	-0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
QVARIETY	0.3488***	-0.1894***	0.3472***	-0.2231***
	(0.0195)	(0.0141)	(0.0195)	(0.0167)
SEPT_OPPTY	0.0660***		0.0660***	
	(0.0067)		(0.0067)	
REPEAT_OPPTY	0.0414***		0.0421***	
	(0.0048)		(0.0048)	
EXPERIENCE	0.2090***	-0.0486***	0.2195***	-0.0570***
	(0.0159)	(0.0125)	(0.0160)	(0.0152)
RESTART		0.8336***		0.9725***
		(0.0196)		(0.0247)
DEVIATION		0.0872***		0.1047***
		(0.0146)		(0.0175)
DEVIATION*EXPERIENCE		-0.0090***		-0.0119***
		(0.0034)		(0.0044)
SEPT		0.0161***		0.0156***
		(0.0051)		(0.0059)
DEVIATION*SEPT		0.0374***		0.0422***
		(0.0047)		(0.0057)
REPEAT		-0.0099**		-0.0045
		(0.0047)		(0.0061)
DEVIATION*REPEAT		-0.0169***		-0.0261***
		(0.0048)		(0.0059)
DEV_UP_TO_NOW	0.0000***	-0.0000		
	(0.0000)	(0.0000)		
DEV_PRIOR_SHIFTS			0.0000	0.0000***
			(0.0000)	(0.0000)
Observations	2,408,218	2,408,218	2,406,214	2,406,214
Rho	-0.0438***		-0.0468***	
Rho Standard Error	(0.0106)		(0.0098)	

Notes. This table reports the results from maximum-likelihood estimation of the probit deviation model (a) and the performance model (b). The unit of observation is a radiologist's case. **DEV_UP_TO_NOW** is the proportion of prior cases (not including the focal case) interpreted by this radiologist that were deviations from the first case in the queue. **DEV_PRIOR_SHIFTS** is the proportion of cases interpreted by this radiologist before the start of the current shift that were deviations from the first case in the queue. Robust standard errors (in parentheses) are clustered by radiologist. All specifications include case type (technology and anatomy), radiologist, day of week, and year fixed effects. The probit specifications also include fixed effects for the case type of the first case in the queue. *** p<0.01, ** p<0.05, * p<0.1

Table 4.A4. Panel A. The Effect of Past SEPT on Speed-Performance

	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Dependent Variable:	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME	DEVIATION	LnREADTIME
Model:	One lag of SEPT		Two lags of SEPT		Three lags of SEPT	
QSIZE	0.0512***	-0.0273***	0.0512***	-0.0271***	0.0513***	-0.0270***
	(0.0038)	(0.0022)	(0.0038)	(0.0022)	(0.0038)	(0.0022)
NUM_IMAGES	-0.0080	0.1712***	-0.0080	0.1712***	-0.0080	0.1713***
	(0.0101)	(0.0063)	(0.0101)	(0.0063)	(0.0101)	(0.0063)
ORDER_IN_SHIFT	-0.0004***	-0.0003***	-0.0004***	-0.0003***	-0.0004***	-0.0003***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
EXPERIENCE	0.2200***	-0.0470***	0.2200***	-0.0465***	0.2200***	-0.0463***
	(0.0158)	(0.0125)	(0.0158)	(0.0125)	(0.0158)	(0.0125)
QVARIETY	0.3466***	-0.1752***	0.3465***	-0.1732***	0.3465***	-0.1728***
	(0.0194)	(0.0137)	(0.0194)	(0.0135)	(0.0194)	(0.0134)
SEPT_OPPTY	0.0668***		0.0668***		0.0669***	
	(0.0067)		(0.0067)		(0.0067)	
REPEAT_OPPTY	0.0417***		0.0419***		0.0419***	
	(0.0049)		(0.0049)		(0.0049)	
RESTART		0.8233***		0.8221***		0.8219***
		(0.0192)		(0.0191)		(0.0191)
DEVIATION		0.0764***		0.0753***		0.0750***
		(0.0150)		(0.0150)		(0.0150)
DEVIATION *		-0.0092***		-0.0092***		-0.0093***
EXPERIENCE		(0.0034)		(0.0034)		(0.0034)
SEPT		0.0114**		0.0111**		0.0110**
		(0.0050)		(0.0049)		(0.0049)
DEVIATION *		0.0376***		0.0373***		0.0372***
SEPT		(0.0047)		(0.0046)		(0.0046)
REPEAT		-0.0325***		-0.0316***		-0.0315***
		(0.0048)		(0.0047)		(0.0047)
DEVIATION *		-0.0172***		-0.0172***		-0.0172***
REPEAT		(0.0048)		(0.0048)		(0.0048)
SEPT_L1		0.0337***		0.0329***		0.0328***
		(0.0040)		(0.0039)		(0.0039)
SEPT_L2				0.0108***		0.0104***
				(0.0022)		(0.0021)
SEPT_L3						0.0041**
						(0.0017)
Observations	2,408,218	2,408,218	2,408,176	2,408,176	2,408,130	2,408,130
Rho	-0.0323***		-0.0308***		-0.0304***	
Rho Standard Error	(0.0112)		(0.0112)		(0.0112)	

Notes. This table reports the results from maximum-likelihood estimation of the probit deviation model (a) and the performance model (b). The unit of observation is a radiologist's case. **SEPT_L1**, **SEPT_L2** and **SEPT_L3** are the one-period, two-period, and three-period lagged values of SEPT, respectively. **SEPT_IN_SHIFT** is the proportion of prior cases so far in the shift that were consistent with a SEPT policy. Robust standard errors (in parentheses) are clustered by radiologist. All specifications include case type (technology and anatomy), radiologist, day of week, and year fixed effects. The probit specifications also include fixed effects for the case type of the first case in the queue. *10% statistical significance; **5% statistical significance; ***1% statistical significance.

Table 4.A4. Panel B. The Effect of Past SEPT on Speed-Performance

	(4a)	(4b)
Dependent Variable:	DEVIATION	LnREADTIME
Model:	Proportion of Past SEPT	
QSIZE	0.0512***	-0.0282***
	(0.0038)	(0.0022)
NUM_IMAGES	-0.0080	0.1710***
	(0.0101)	(0.0062)
ORDER_IN_SHIFT	-0.0004***	-0.0004***
	(0.0001)	(0.0001)
EXPERIENCE	0.2200***	-0.0495***
	(0.0158)	(0.0126)
QVARIETY	0.3473***	-0.1905***
	(0.0195)	(0.0141)
SEPT_OPPTY	0.0659***	
	(0.0067)	
REPEAT_OPPTY	0.0419***	
	(0.0048)	
RESTART		0.8344***
		(0.0194)
DEVIATION		0.0882***
		(0.0147)
DEVIATION * EXPERIENCE		-0.0090***
		(0.0035)
SEPT		0.0173***
		(0.0048)
DEVIATION * SEPT		0.0373***
		(0.0047)
REPEAT		-0.0092**
		(0.0046)
DEVIATION * REPEAT		-0.0170***
		(0.0049)
SEPT_IN_SHIFT		-0.0177
		(0.0145)
Observations	2,408,218	2,408,218
Rho	-0.0450***	
Rho Standard Error	(0.0106)	

Notes. This table reports the results from maximum-likelihood estimation of the probit deviation model (a) and the performance model (b). The unit of observation is a radiologist's case. **SEPT_L1**, **SEPT_L2** and **SEPT_L3** are the one-period, two-period, and three-period lagged values of SEPT, respectively. **SEPT_IN_SHIFT** is the proportion of prior cases so far in the shift that were consistent with a SEPT policy. Robust standard errors (in parentheses) are clustered by radiologist. All specifications include case type (technology and anatomy), radiologist, day of week, and year fixed effects. The probit specifications also include fixed effects for the case type of the first case in the queue. *10% statistical significance; **5% statistical significance; ***1% statistical significance.

Chapter 5

Conclusion

With companies increasingly looking to make data-driven decisions, this dissertation illustrate how to use data to gain insights that can be used to enhance workers' discretion. Essays in this dissertation explore the role of experimentation in field settings to answer operations management questions and use causal inference methods to estimate behavioral drivers of process variation. Focusing on task scheduling, these papers examine the operational implications of workers' decisions regarding the allocation and/or completion of tasks.

Analyzing thousands of food safety inspections, "**How Scheduling Biases Quality Assessments**" (Chapter 3) investigates how inspection scheduling can affect inspection quality by influencing bias. By identifying factors that bias inspections, the chapter's findings can enable managers and regulators to make better decisions when using inspection report data, help create more reliable information for managers and consumers, and provide fairer results (and higher motivations for compliance) for inspected establishments. Our results suggest several interventions that could exploit and ameliorate the biases we identify. Future work could consider additional sources of inspection bias and alternative ways to improve monitoring effectiveness.

One way for managers to reduce biases resulting from task scheduling is to centralize and optimize the decision to ameliorate the potential biases. However, task scheduling is not always under the control of managers because, intentionally or not, those who execute tasks often have discretion over the order in which they perform them. Analyzing 2.4 million decisions made by diagnostic radiologists, "**Discretionary Task Ordering: Queue Management in Radiological Services**" (Chapter 4) investigates the drivers and consequences of exercising discretion to "deviate" from a prescribed task sequence. This paper finds that radiologists prioritize similar tasks (grouping tasks into batches) and those tasks they expect to complete faster (shortest expected processing time). Exploiting random assignment of tasks to doctors' queues, instrumental variable estimates reveal that both of these types of deviations decrease productivity. Actively grouping similar tasks reduces productivity, in stark contrast to productivity gains from exogenous grouping, indicating deviation costs outweigh benefits from repetition. These results highlight the tradeoffs between the time required to exercise discretion and the potential gains from doing so, which has implications for managers deciding task sequence assignments and system design. This investigation suggests many opportunities for future research. While we focus on discretion over which task to work on next and whether the worker selects or postpones the next item in a given queue, workers in other field settings may have different ways to exercise discretion. For example, future work could consider how workers exercise discretion over which pool of work to select the next task to work on or how to incentivize workers to look for ways to reduce the costs of deviations and to exercise discretion more efficiently.

Together, the essays in this dissertation provide empirical evidence from high-stakes field settings of how productivity and quality are affected by workers' deviations from prescribed processes. By collaborating closely with the individuals and organizations in the field settings related to the data analyzed, these studies seek to provide relevant scholarly and managerial insights as well as to motivate future work on workers' discretion.