Summer 8-2021

An Examination of Organizational Decision-Making: A Case Approach

Olivia Reinecke

AN EXAMINATION OF ORGANIZATIONAL DECISION-MAKING: A CASE APPROACH

by

Olivia Reinecke, B.S., M.A.

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

COLLEGE OF EDUCATION
LOUISIANA TECH UNIVERSITY

August 2021

LOUISIANA TECH UNIVERSITY

GRADUATE SCHOOL

June 18, 2021
Date of dissertation defense

We hereby recommend that the dissertation prepared by

Olivia Reinecke, B.A., M.A.

entitled **An Examination of Organizational Decision-Making: A Case Approach**

be accepted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Industrial/Organizational Psychology

Tilman Sheets
Supervisor of Dissertation Research

Donna Thomas
Head of Psychology and Behavioral Science

Doctoral Committee Members:
Steven Toaddy
Frank Igou

Approved:
Don Schillinger
Dean of Education

Approved:
Ramu Ramachandran
Dean of the Graduate School

GS Form 13a
(01/20)

ABSTRACT

This applied dissertation offered insights about the scientist-practitioner gap and examined the process of organizational decision making using a real case from a large United States manufacturing company. Historical records of employee work hours were extracted and compiled from several of the company's databases, along with recent plant performance metrics. These archival data were used to evaluate the relationship between the percentage of each plant's population working overtime (i.e., overtime utilization) and each plant’s performance on various manufacturing metrics. This comparison was made using a median split approach, such that the performance of plants with an overtime utilization above the median (i.e., the high-overtime group) was compared to the performance of pants with an overtime utilization falling below the median (i.e., the low-overtime group).

Post-hoc analyses were conducted to determine if a more rigorous methodological approach would have produced different findings. As such, a series of bivariate regressions (with overtime utilization as the predictor variable) was conducted, allowing the overtime variability that exists across plants to be taken into account in order to better understand the relationship between overtime utilization and the outcome metrics. Additionally, when available, financial metrics were leveraged instead of performance metrics so that results could be interpreted through the lens of the company’s bottom line. Ultimately, the more rigorous approach produced richer insights regarding how risky the

decision to enact an overtime-reduction policy might be from a financial perspective and exemplified the value of leveraging confidence intervals to guide organizational decision making.

APPROVAL FOR SCHOLARLY DISSEMINATION

The author grants to the Prescott Memorial Library of Louisiana Tech University the right to reproduce, by appropriate methods, upon request, any or all portions of this Dissertation. It was understood that "proper request" consists of the agreement, on the part of the requesting party, that said reproduction is for his personal use and that subsequent reproduction will not occur without written approval of the author of this Dissertation. Further, any portions of the Dissertation used in books, papers, and other works must be appropriately referenced to this Dissertation.

Finally, the author of this Dissertation reserves the right to publish freely, in the literature, at any time, any or all portions of this Dissertation.

Author______________________________

Date ______________________________

G.S. Form 14
(5/03)

TABLE OF CONTENTS

ABSTRACT iii

APPROVAL FOR SCHOLARLY DISSEMINATION v

LIST OF TABLES viii

LIST OF FIGURES ix

ACKNOWLEDGEMENTS x

CHAPTER 1 INTRODUCTION 1

- The Scientist-Practitioner Gap 2
  - Why the Gap Exists 3
  - Proposals to Bridge the Gap 4
- Conducting Research in Real-World Organizations 7
- Overtime 16
  - Work Hour Trends and Overtime in the United States 16
  - Implications of Long Work Hours 17
  - Motivation to Work Long Hours 28
- Present Study 31

CHAPTER 2 ORIGINAL PROJECT SUMMARY 35

- Participants 35
- Procedure 35
  - Measures 38

Hypotheses 39

Results 41

Business Outcome 42

CHAPTER 3 POST HOC ANALYSIS SUMMARY 44

Procedure 44

Financial Implications 45

Leveraging Confidence Intervals to Examine Level of Risk. 46

Results 48

Assumptions 48

Hypothesis 1 49

Hypothesis 2 52

Hypotheses 3-4 55

Department-Level Analysis 58

Hypotheses 5 60

Plant-Level Analysis 61

Department-Level Analysis 63

CHAPTER 4 DISCUSSION 68

More Data Points Are Better than Fewer Data Points 69

Confidence Intervals Are a Useful Tool for Decision-Making 69

Do Not Forget about Outliers 70

Conclusion 71

REFERENCES 73

LIST OF TABLES

Table 1. *Calculation of Plant X's 60-Hour Overtime Utilization*......................................37

Table 2. *Median Split of Low- and High-Overtime Plants* 37

Table 3. *Performance Across Low- and High-Overtime Groups* 41

Table 4. *Confidence Intervals for Hypothesis 1*..33

Table 5. *Confidence Intervals for Hypothesis 2*..34

Table 6. *Confidence Intervals for Hypotheses 3-4: Plant-Level Analysis* 35

Table 7. *Confidence Intervals for Hypotheses 3-4: Department-Level Analysis*...............38

Table 8. *Confidence Intervals for Hypothesis 5: Plant-Level Analysis* 63

Table 9. *Confidence Intervals for Hypothesis 5: Department-Level Analysis*...................66

Table 10. *Cost Implications for Enactment of Overtime Reduction Policy* 67

LIST OF FIGURES

Figure 1 Normal P-P Plot of Regression Standardized Residual for H1 50

Figure 2 Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H1 50

Figure 3 Normal P-P Plot for H2 53

Figure 4 Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H2. 53

Figure 5 Normal P-P Plot for H3-4 (Plant-Level)......................56

Figure 6 Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H3-4 (Plant-Level) 56

Figure 7 Normal P-P Plot for H3-4 (Department-Level) 59

Figure 8 Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H3-4 (Department-Level)......................59

Figure 9 Normal P-P Plot for H5 (Plant-Level). 61

Figure 10 Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Plant-Level)......................62

Figure 11 Normal P-P Plot for H5 (Department-Level). 64

Figure 12 Original Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Department-Level) 64

Figure 13 Log Transformation Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Department-Level). 65

Figure 14 Square Root Transformation Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Department-Level). 65

ACKNOWLEDGEMENTS

There are so many people who I'd like to thank for making the completion of my doctorate possible. First, I'd like to thank my dissertation committee. Their wisdom, guidance, and (many) edits challenged and transformed my thinking about this project for the better and ultimately enabled me to make what was once only a dream, a reality. Most of all, I want to thank my advisor and committee chair, Dr. Tilman Sheets, for seeing my potential as an eager but nervous college graduate in 2014, taking me under his wing, teaching me to get comfortable with ambiguity, and guiding me through my graduate school journey; and of course for all of the time, energy, and brain power he put into this dissertation. I absolutely could not have done this without him.

I'd also like to thank my family and friends for their continuous support and optimism that at some point, I would eventually, finally, earn my PhD, especially during the times I wasn't quite convinced of that myself. I absolutely would not be Dr. Olivia Reinecke without my incredible parents, who raised me with a strong work ethic and a deep, internal knowing that I could accomplish anything I set my mind to, even if it was way more difficult than I anticipated. I would also be remiss if I didn't thank Samantha and Milo their unwavering support (via puppy snuggles and wet kisses) and barks of encouragement throughout this process. Without them, dissertation writing would have been so much more agonizing.

And finally, to my best friend and husband Seth: thank you for being my number one cheerleader both personally and professionally, for continuously pushing me, inspiring me, and motivating me throughout college, graduate school, and beyond, for calling me out on my procrastination antics and forcing me to write when it was the last thing I wanted to do, and for ultimately molding me into the woman I am today. I love you.

CHAPTER 1

INTRODUCTION

Organizations are messy, complex, and volatile entities, and the imperfect data they produce reflect this reality. Using those data to generate meaningful, relevant insights that are understood *and* implemented by organizational decision-makers is a key component (and challenge) of successfully practicing industrial-organizational (I-O) psychology. This challenge fits squarely within the ever-relevant, decades-old topic of the *scientist-practitioner gap*, a term used to describe the disconnect between the academics who publish research about organizations and the practitioners who operate within them (Aguinis et al., 2017; Dunnette, 1990). As described by Rupp and Beal (2007), the scientist-practitioner model that I-O psychology subscribes to entails a reciprocal relationship, where "practitioners should look to the scientific literature for guidance on setting up effective workplace systems; scientists should take their cues from practitioners in identifying issues relevant to employee well-being and organizational effectiveness" (p. 36). While this concept may seem straightforward at first glance, Rupp and Beal (2007) acknowledge that bringing the scientist-practitioner model to life is easier said than done, and as such, the challenges inherent to the model have been fiercely debated within the I-O community.

This applied dissertation will add to the conversation by taking a reflective approach on a real-world business project that sought to better understand the relationship between working overtime hours and several key metrics at a United States manufacturing company in an effort to provide insights focused on making key policy decisions. In addition to providing insight on how overtime hours might impact key business metrics, this project also offered an opportunity to perhaps better understand the scientist-practitioner gap in a real-world, business problem, as well as an opportunity to determine if a more rigorous, scientific approach would have produced different findings.

The Scientist-Practitioner Gap

The scientist-practitioner gap is nothing new. Consider this quote from Susman and Evered dating back to 1978:

> There is a crisis in the field of organizational science. The principal symptom of this crisis is that as our research methods and techniques have become more sophisticated, they have also become increasingly less useful for solving the practical problems that members of organizations face. (p. 582)

While the principle symptom may have changed over the years and today might be more related to the sophistication of statistical methods and techniques, the struggle to come to terms with scientist-practitioner differences is not new. More recent commentary on the topic (e.g., Bansal et al., 2012; Cascio, 2008; Cascio & Aguinis, 2008; Halfhill & Huff, 2003; Ones et al., 2017; Rotolo et al., 2018; Rynes et al., 2002) suggests the gap is still alive and well. It is also worth noting that the notion of a disconnect between science and practice is not unique to I-O psychology, or even psychology alone for that matter (Belli, 2010; Rotolo et al., 2018). As pointed out by Rynes et al. (2001), a science-practice gap is evident in “nearly all fields in which there are both researchers and practitioners” (p. 340).

Why the Gap Exists

Much has been written about why the science-practice gap exists in organizational science. An early theoretical rationale for the gap was presented by Shrivastava and Mitroff (1984), who argue that researchers and practitioners possess markedly different frames of reference and therefore make different assumptions about what constitutes useful knowledge. From this perspective, well-meaning researchers are doing their best to produce relevant research that practitioners will find valuable and easy to apply, but they see the world differently than their practitioner counterparts and therefore, their best efforts fall short. Johns (1993) approaches the gap from a slightly different perspective, focusing on organizations' tendency to adopt (or not) personnel innovations advocated by I-O psychology. In Johns' view, whether organizations choose to adopt I-O's technically-meritorious practices depends on three things: (1) managers' technical orientation – for example, they consider personnel processes (e.g., selection and assessment) as administrative, rather than technical practices and therefore do not believe technically-sound, psychologically-based methods are necessary; (2) the nature of I-O theory and research, in that it is technically oriented, involves long causal chains, and typically fails to address the larger organizational context; and (3) the external environment, which powerfully shapes how organizations operate, including politics, government regulation, crises, and competing sources of innovative solutions.

More recently, Banks et al. (2016) suggested that the science-practice gap stems from three issues: (1) a transfer problem: scientific knowledge in the format of scholarly articles is not easily available to or understood by practitioners; (2) a collaboration problem: theory and practice are developed in isolation of each other, as research partnerships between academics and practitioners are scarce; (3) a relevance problem: the research topics pursued by academics are not useful for practice. Yet another perspective is offered by Ones et al. (2017), who argue that a number of "troubling trends" in I-O psychology are fueling the science-practice gap and hurting the discipline overall: "an overemphasis on theory; a proliferation of, and fixation on, trivial methodological minutiae; a suppression of exploration and repression of innovation; an unhealthy obsession with publication while ignoring practical issues; a tendency to be distracted by fads; and a growing habit of losing real-world influence to fields" (para. 4).

A final perspective worth discussing is that of Olenick et al. (2018). They built on the typology of organizational psychologists originally proposed by Rupp and Beal (2007; i.e., the pure scientist, the scientist-practitioner, the practitioner-scientist, the pure practitioner) and propose that each of these roles operates in different environments that are impacted by different systemic pressures and disconnected incentive structures — publication standards and tenure systems for pure scientists and scientist-practitioners, and the need to add value to organizations for practitioner-scientists and pure practitioners — regardless of the personal motivations each group possesses.

Proposals to Bridge the Gap

Clearly, the existence of the science-practitioner gap is due to a myriad of complex and interwoven factors; as such, a variety of solutions have been proposed to bridge the gap. An early theory, called the different-frames-of-reference theory, was offered by Shrivastava and Mitroff (1984) as a means to understand and close the science-practice gap. They propose that organizational researchers have much to learn from their practitioner counterparts in terms of how to develop truly useful research. As such, they recommend researchers put less emphasis on perfectly controllable variables and linear causal relationships. Instead, researchers should employ more clinical and qualitative research methods and expand the scope of their research topics to those that make sense to managers and are perceived as actionable. Mohrman et al. (2001) also propose that researchers shift the way they conduct research so that it is more useful to practitioners. Their findings using an extensive qualitative research design suggest that practitioners consider research results most useful when: (a) the research offers opportunities for joint interpretative forums (e.g., practitioners are invited to interact and collaborate with researchers) and (b) the results can be used to inform self-design activities (i.e., the behavioral and cognitive processes used by employees to reorganize and redesign their organizations). Focusing on a different potential cause for the gap, Cascio (2008) and Cascio and Aguinis (2008) argue that without the modification of academic reward systems, the science-practice gap will remain ajar; like others before him (e.g., Mohrman et al., 2001; Münsterberg, 1913; Shrivastava & Mitroff, 1984), Cascio (2008) highlights the need for closer collaboration between researchers and practitioners so that academic research topics are congruent with real-world business problems.

However, as pointed out by Rynes et al. (2002), it is important to recognize that failure to implement scholarly research can occur for a variety of reasons, not necessarily due to the research itself. Their thinking is congruent with Johns (1993), as both articles suggest that organizational factors play a powerful role. Rynes et al. (2002) believe that factors such as organizational politics, risk aversion, lack of bandwidth, and organizational inertia can contribute to the scientist-practitioner gap, rather than the translation/knowing gap proposed by other researchers (e.g., Banks et al., 2016). To test their proposition that lack of implementation results from a lack of *doing* rather than a lack of *knowing*, Rynes et al. (2002) surveyed 5,000 human resource (HR) professionals about the extent to which they agreed with various research findings. Their findings did not support their original proposition, as results showed substantial variance in knowledge of HR best practices across HR professionals employed by different companies. They also acknowledged that their sample may have underrepresented this variance, as they surveyed more senior HR leaders who could possess more knowledge than their less experienced counterparts; additionally, non-respondents may also be likely to possess less HR knowledge. The biggest gaps between research knowledge and real-world practice existed for selection and engagement practices. As a result of their findings, Rynes et al. (2002) recommend that researchers work to publish more of their research content, particularly with regard to best practices, in formats commonly consumed by HR managers (e.g., SHRM, HR Magazine).

Another perspective worth considering is that of Bansal et al. (2012), who make up a balanced mix of researchers and practitioners. Upon reflection of their own experiences collaborating with both sides of the scientist-practitioner continuum, they argue that the relationship between research and practice is too muddied with complexities and paradoxes to easily bridge the two worlds. Instead, they advocate for intermediary organizations such as the Network for Business Sustainability, an organization founded in 2005 to facilitate the transfer of knowledge on business sustainability between researchers and practitioners. The organization has accomplished quite a lot since its inception, including the publication of over 100 research briefs on the topic of business sustainability that practitioners can understand and leverage. This was achieved through a formal process; each year, ten of the most pressing research questions in business sustainability are identified, and then research teams are created to systematically review the research related to these questions, with a focus on incorporating both scientific and practitioner perspectives. The research team then synthesizes and publishes this research into briefs specifically targeted for a practitioner audience.

While some experts argue progress has been made and the science-practice gap is beginning to shrink (e.g., Aguinis et al., 2017; Latham, 2001), others argue it is continuing to grow (e.g., Hulin, 2001; Ones et al., 2017). This applied dissertation will contribute to the discussion by examining a real-world business problem, which involved many of the organizational factors discussed by Rynes et al. (2002) such as organizational inertia, lack of bandwidth, and organizational politics.

Conducting Research in Real-World Organizations

Boehm (1980) argues that the traditional model of scientific inquiry (i.e., the scientific method), which is based on research practices conducted in the physical sciences, has limited the progress and utility of applied I-O psychology research. In her experience working for The Standard Oil Company (Ohio), "when attempts are made to utilize this model in 'real world' settings (as in-house I/O psychologists or as consultants to organizations), it very rapidly becomes apparent that there is a disparity between the 'state of the science' in I/O psychology and the 'state of the art' regarding application of behavioral science principles to organizational reality" (p. 497). She also notes that the typical reaction of scientists when they encounter the realities of an organizational research environment is to either attempt to modify the environment to better fit the traditional model of scientific inquiry (anyone who has worked inside an organization knows this is near impossible), or to opt out altogether. Neither option bodes well for the progression of scientific research or the betterment of organizations.

As such, Boehm (1980) proposes an alternative model that she considers to be a better fit for the realities of real-world organizational research, a model that embraces the complexity of organizations rather than trying to change or avoid them. Boehm’s (1980) model looks quite different from the traditional scientific method, as it begins with an organizational problem that creates concern or unrest, and then requires an investigation of the organizational context and restraints surrounding the research. It also includes the question, “Are results valuable to the organization?” to determine the utility and overall success of the research (p. 498). If yes, results are reported and then pitched or sold to business leaders, again taking into account organizational context and restraints, and ultimately implemented. Boehm (1980), in describing her organizational research model, points out that though it is “scientifically, distinctly messy in terms of methodology, complexity, statistical analysis, and the conclusions can be drawn… it seems to present a more accurate picture of organizational reality than a model based on *what should be*” (p. 497).

Boehm (1980) notes three key differences between the traditional, *what should be* scientific research model and her proposed, *what is* organizational research model; as compared to the traditional scientific model, the Boehm’s model: (1) contains far more processes/steps; (2) involves more interactions between steps; and (3) originates from a different need – that is, a need to solve a current or anticipated performance problem, rather than a decontextualized “scientific quest to learn what the facts are” (p. 497). All in all, this model grants a great deal of flexibility, allowing one to embrace organizational messiness and complexity rather than trying to modify, refine, or dismiss it.

In addition to proposing a new model for organizational research, Boehm (1980) calls for a change in how organizational research is evaluated by those in the scientific community. If one views the research conducted inside of organizations through the lens of the traditional scientific method, it is indeed methodologically flawed: “real world research is messy – uncontrolled variables abound, predictor and criterion measures interact, alternative hypotheses cannot be ruled out; standard statistical procedures cannot be applied without massive violation of assumptions” (Boehm, 1980, p.499). And yet, these flaws are an accurate reflection of the organization’s environment. As such, Boehm (1980) argues that these “flaws” should be considered necessary and even desirable when conducting research that can be realistically applied to real-world organizations. Further, she laments the sparsity in the published organizational literature describing these methodological difficulties and the reality of real-world research.

Even in 2021, published research on this topic is still scarce. As such, one of the objectives of this applied dissertation is to candidly describe the challenges this project’s research team faced, rather than glossing over them, in addition to reflecting upon what would have been different if a more rigorous methodology had been applied. Boehm (1980) concludes her discussion with a warning:

> Organizations do not exist primarily as research laboratories for behavioral scientists. If this basic fact of life is recognized and a model based on “what is” is accepted as being as equally legitimate as the model of “what should be” there are

> exciting possibilities for the advancement of I/O psychology, both scientifically and empirically. If, however, I/O psychologists continue to adhere solely to a model based on the ideal, in which organizational realities are viewed solely as problems to be circumvented, stagnation of the field is a strong possibility, and widening of the communications gap between academic and organizational I/O psychologists a virtual certainty. (pp. 502-503)

Daft (1983) echoes many of the arguments made by Boehm (1980), calling organizational research a craft rather than a precise science. He explains that, after graduate school, many practitioners recognize that the research techniques they learned are insufficient for conducting research in real-world organizations. Instead, Daft (1983) recommends learning to practice what he calls “research craftsmanship,” which is a significant departure from the “formal, prescriptive approach to research that is frequently taught in graduate school” (p. 540). For example, Daft (1983) recommends that those conducting research in organizations “build in plenty of room for error and surprise” (p. 540); he believes that hard logic and previous evidence should not always be required to justify organizational research hypotheses. In his view, the most informative research to-date has approached an organizational problem with an open-ended question to be answered, rather than with specific hypotheses. He argues against the notion that successful research must include hypotheses justified by previous research and provide predictable results, as this can restrict the discovery of knowledge. Instead, Daft (1983) advocates for asking research questions without having the answer in advance. He boldly states, “If experiments are perfectly designed and the results come out as expected, then they probably are a waste of time” (p. 541).

Daft (1983) also promotes simplicity, arguing that good research does not try to answer all possible questions about a given issue. Instead, “the best research provides an utterly imperfect model of organizational reality” that explains “a tiny piece of organizational reality… [that] can be used to raise new questions for future research” (p. 542). Much like Boehm (1980), Daft (1983) describes the decision processes involved in research craftsmanship as non-linear, random, and messy. Organizational researchers must expect and embrace this.

A more recent discussion on this topic is offered by Swanson and Holton (2005), who in the first chapter of their book *Research in Organizations: Foundations and Methods of Inquiry* discuss the intentionality of the book’s title, highlighting the distinction between conducting research *on* organizations as compared to *in* organizations. Throughout the book, they emphasize that conducting effective research in organizations requires traditional research methods to be adapted and modified to fit organizational realities. In their discussion of the challenges inherent to conducting research in organizations, Swanson and Holton (2005) urge researchers to recognize that organizations are complex, open, and dynamic systems that exist alongside other, external systems and serve as hosts for a number of behavioral phenomena. While these characteristics make organizations fascinating settings for research, they also make them messy and multifaceted, which makes conducting research inside them all the more difficult. As a result, Swanson and Holton (2005) argue that conducting research in organizations requires “harmony or logical trade-offs among the chosen research question, research paradigm, research method, and research context. These considerations are not linear, and tentative decisions in one realm will influence the other three realms" (p. 24).

So, it is clear that conducting research in organizations departs substantially from the traditional research model prescribed as the gold standard for science. Indeed, Stone-Romero (2011) describes randomized experimental design as the ideal research methodology and argues that there is "*no reason* [emphasis added] why such research cannot be conducted using traditional, rigorous research methods" (p. 48). However, she later points out a few potential reasons, acknowledging that conducting research within any organization requires support from senior leadership. She further explains that "it is typically the case that managers are very reluctant to allow experimental research to be conducted in their organizations. In addition, a host of other factors militate against the conduct of randomized experiments in organizations (p. 48).

While the previously discussed sources provide a high-level overview of the challenges inherent to conducting organizational research, it is also helpful to understand the specific limitations organizational researchers encounter in their research. Brutus et al. (2010) examined this systematically, conducting a content analysis of the self-reported limitations reported in 1,903 empirical articles published by the *Journal of Applied Psychology*, *Personnel Psychology*, and the *Academy of Management Journal* between 1995 and 2008. Overall, 75% of the articles they examined mentioned at least one limitation. The most common were: threats to internal validity (41% of articles), due to an inability to infer causality (21%) and inability to measure an important variable (19%); threats to construct validity (38%), caused by less than ideal operationalization of constructs (17%) and common-method variance (16%); and threats to external validity (36%), due to an inability to generalize across people (21%).

Brutus et al. (2010) also examined the relationship between study limitations and methodological choices using phi coefficients. Not surprisingly, their results indicate that research methods more easily employed in organizations – surveys, observation, and the use of archival data – were more likely to suffer from limitations regarding inability to infer causality. Specifically, their results show that studies using survey methodology were positively linked to limitations due to common-method variance and lack of causality; while survey research was negatively related to limitations due to generalizability across situations. Archival data was negatively related to statistical-conclusion validity threats, and observation was linked to low power. Experimental research, a method less commonly leveraged in organizations (Mitchell, 1985; Stone-Romero, 2011) and one that accounted for less than 20% of studies in this sample, was positively linked to limitations concerning generalizability across situations, but negatively linked to lack of causality. Interestingly, their results also show that threats to external validity decreased over time, while threats to statistical validity increased. They saw no changes in the frequency of internal or construct validity limitations. Overall, the results of Brutus et al. (2010) show that a fairly narrow set of (reported) limitations occur in I-O psychology. Of the 36 discrete threats to validity that were studied, only 5 accounted for half of all limitations observed: causality, omission of an important variable, less than ideal operationalization of constructs, common-method variance, and generalizing across people.

These findings are echoed by previous discussions on the validity threats commonly encountered in organizational research. Mitchell (1985), for example, reviewed a sample of 126 empirical articles published between 1979 and 1983 in the *Journal of Applied Psychology,* the *Academy of Management Journal,* and the *Organizational Behavior and Human Performance* journal. She describes her findings as "unsettling" (p. 201), as they showed that most cross-sectional, correlational studies administered a questionnaire to a convenience sample without reporting response rates or exploring the differences between respondents and non-respondents. Further, reliability and validity coefficients were weak, with construct validity rarely checked. While researchers were aware of and sensitive to the threat of common-method variance, cross-validation efforts or the use of holdout samples was virtually nonexistent.

When considering what lead to these unsettling results, Mitchell (1985) suggests the "notion of applied versus basic research" could be part of the problem, as "the field researcher often is more interested in simply showing that some variable (e.g., goal setting) correlates with performance than in demonstrating that the goal setting measure really reflects the goal setting construct… Thus, if it works, no matter what the reason, the result is important from an applied perspective" (p. 202). Clearly, the tension between what is deemed as acceptable in traditional research as compared to that of the applied world is evident here.

Highhouse (2009) offers a different rationale for practitioners' research methods. He believes that experimental research is not conducted within organizations due to external validity concerns and instead argues that "organizational research has relied too heavily on methods characterized by passive observation, likely because there is a widespread belief that experimental research has limited generalizability. However, this is often because researchers (and reviewers or editors) misunderstand the nature of generalizability and what it requires" (p. 554). While this may be true for researchers on the scientist side of the scientist-practitioner continuum, practitioners likely have other, more organizational-specific reasons for avoiding experimental research. Highhouse (2009) does not appear to acknowledge that there are other reasons organizational researchers choose not to conduct experimental research, such as those mentioned by Stone-Romero (2011; e.g., lack of management support).

Grant and Wall (2009) note that true field experiments overcome the external validity concerns associated with laboratory experiments, as they maintain fidelity of the employees and organizational phenomena of interest. And yet, they also recognize that maximizing internal validity in these studies requires random assignment and the deliberate manipulation of some variable, both of which can be difficult to accomplish in the field. Even so, when random assignment and deliberate manipulation exist in organizational research, they have the capacity to alter the nature of the phenomena under study and compromise the authenticity of the independent variable, thereby undermining the study's external validity. Grant and Wall (2009) postulate that two primary obstacles stand in the way of conducting true field experiments in real-world organizations: (1) inability to control random assignment to treatment conditions; and (2) inability to control key variables of interest including tenure, team composition and diversity, and human resource practices. They also note that even in cases where these variables can be controlled by the researcher, "organizations do not hold still, making it difficult for researchers to isolate key causal factors and rule out alternative explanations for changes observed" (p. 654).

The discussion above makes it clear that conducting research inside organizations is challenging and messy, to say the least. When viewed through the lens of the traditional scientific method, which remains the gold standard for many scientists, organizational research is admittedly imperfect. And yet, as others have described and as this paper will showcase, this imperfection is reality. As noted by Boehm (1980), organizations exist independently from behavioral science; they were not created as research laboratories. And yet, we must conduct organizational research anyway, even if it is imperfect. So rather than harping on the methodological flaws and limitations of this research, science-practitioners must recognize and accept that organizations and their data are messy. As a result, organizational research will reflect that messiness. This is precisely what makes the findings of such research important, relevant, and useful for practice.

Overtime

Work Hour Trends and Overtime in the United States

Throughout the early twentieth century, there was much debate on the appropriate workweek length and hourly pay for United States employees. This debate culminated in the passing of the Fair Labor Standards Act of 1938, which mandated a maximum workweek of 40 hours, with any hours worked beyond this maximum requiring time-and-a-half pay (Elder & Miller, 1979). In 2019, the average number of hours worked each week across all United States employees fell well below this maximum, at 34.4 hours (Bureau of Labor Statistics, 2020). However, this number is higher for manufacturing employees, which make up the key population in the present study. The number of hours worked each week by US manufacturing employees has steadily increased over the past decade, with an average of 39.0 hours in 2009 versus an average of 40.6 hours per week in 2019 (Bureau of Labor Statistics, 2020).

Implications of Long Work Hours

Before presenting research findings regarding the impacts of long work hours and overtime, it is important to discuss the requirements of making causal inferences (i.e., the ability to conclude that working a certain number of hours legitimately *causes* a given outcome). According to Cook and Campbell (1979), three conditions must be present to conclude causation: (1) temporal precedence, meaning the cause must precede the effect; (2) covariation of cause and effect, meaning when the cause is present, the effect occurs; and (3) elimination of alternative explanations. These conditions are clearly met via randomized experimental design; however, as previously stated, experimental studies are difficult to conduct in organizations (Grant & Wall, 2009; Stone-Romero, 2011). As such, true experimental research regarding the causal impacts of overtime in organizations is rare.

In the late twentieth century, Spurgeon et al. (1997) reviewed the empirical evidence on working hours and employee safety available at that time. They identified several gaps in the literature, noting that very few health outcomes beyond cardiovascular disease and mental health had been examined and that most research studied employees working more than fifty hours per week. Knowledge of the potential impacts of working fewer than that were largely unknown at that time. Spurgeon et al. (1997) concluded that although sufficient evidence existed to merit some concern about the negative health and safety impacts of long work hours, additional research was needed "to define the level and nature of those risks" (p. 367).

In the same year, Sparks et al. (1997) conducted the first meta-analysis of the research to-date on work hours and employee health. They found a statistically significant correlation of .1302 (considered a small effect size; Cohen, 1992) between employee health (psychological and physiological) and work hours, also noting the need for additional research in this area. A few years later, van der Hulst (2003) conducted a systematic review of 27 empirical studies on the relationship between long work hours and health. Examining both physiological recovery and behavioral lifestyle mechanisms that could explain the hours-health relationship, her review showed support for a modest relationship between long work hours and adverse health outcomes such as cardiovascular disease, type two diabetes, disability-related retirement, as well as subjective reports of poor physical health and fatigue. Her review also provided limited support for two causal pathways that could explain the relationship between long work hours and adverse health: (a) insufficient time for physiological recovery and (b) unhealthy lifestyle behaviors such as smoking, coffee and alcohol consumption, poor diet, and lack of exercise. However, she ultimately deemed her results as inclusive due to a number of potential confounding variables that were not controlled, as well as several studies showing nonsignificant effects.

As noted by Allen et al. (2007), many work hour studies largely ignore the individual characteristics (e.g., health status, job type, compensation type) of the employees studied. The practical utility of work hour research is further limited because studies often include a select sample of employees within a specific industry, focus on a narrow subset of risk factors, and/or define *long hours* inconsistently (Caruso, et al., 2006). As a result, organizations can find it difficult to leverage this research to determine sensible, data-driven limits on the number of hours employees should be allowed to work or even the particular circumstances that pose excessive risk to employees. Some organizations have taken matters into their own hands, conducting internal research to investigate the relationship between hours worked and key metrics (e.g., injuries, worker productivity). For example, Shepard and Clifton (2000) discuss a study conducted by the Kellogg Company in the 1930s. The company reduced work hours to a maximum of six per day and saw an increase in productivity, as well as improvements in employee morale and decreases in accidents and insurance costs.

In a more recent example, the International Truck and Engine Corporation (ITEC), a heavy machinery manufacturer, partnered with health and safety experts from the United Auto Workers and the National Institute for Occupational Safety and Health to better understand the relationship between long work hours and various health and safety outcomes in their workforce. While designing this study, Allen et al. (2007) lament the lack of methodological rigor of previous overtime research and set out to provide a more substantial basis for making causal claims about the true impact of overtime. One way they do this is by controlling for employees' prior health status, including medical conditions, health risks, and prior work injuries. Additionally, their analyses explored not only the direct effects of work hours on various outcome measures, but also the potential for work hours to mediate the relationship between various antecedent characteristics and outcomes. They then compared the strength of these direct and indirect effects to the direct effect between the antecedent characteristics and outcomes, bypassing work hours altogether.

Their results showed that the likelihood of an employee experiencing an adverse event (e.g., work-related injury) after working long hours varied substantially based on employee characteristics such as prior work injuries and illnesses, health risks such as smoking and being overweight, type of compensation/work (hourly vs. salary), and age. For example, working long hours (in this case, more than 48 per week) was related to a higher likelihood of injury for hourly males with a history of disease and short-term disability episodes. Similarly, working long hours (more than 60 hours per week) was linked to greater odds of new disease onset for older employees with a history of at least one medical condition. They note, "all of the direct and indirect effects involving workhours [sic] considered together were overwhelmed by the combined contribution of the direct effects associated with the prior (non-work related) health, job and demographic characteristics alone" (p. 165).

In another manufacturing-specific study, Shepard and Clifton (2000) examined the relationship between overtime hours and employee productivity using aggregate panel data across 18 United States manufacturing industries from 1956 through 1991. Using a factor-augmented production function model, the researchers showed that overall, overtime was linked to lower productivity, with the model indicating that a 10% increase in overtime hours was related to 2-4% lower productivity, on average. However, their results varied substantially by industry, even in cases where different industries exhibited similar trends in average weekly work hours and overtime usage. For example, companies in the paper products and transportation equipment industries did not exhibit any overtime-productivity relationship, while companies in the fabricated metals industry (which exhibited similar overtime and work hour trends compared to paper products and transportation companies) showed a significant negative relationship. Shepard and Clifton (2000) note that technology differences as well as the type of work being done across these industries have the potential to differentially impact employee fatigue, on-the-job stress levels, and the amount of unproductive or idle time, all of which could explain the inconsistent relationship between overtime hours and productivity.

Vegso et al. (2007) looked at another potential outcome of working overtime in a manufacturing setting, acute occupational injuries. They leveraged a case-crossover design, which involves matching cases to themselves at an earlier point in time, to examine the number of hours employees worked prior to a shift where an injury occurred as compared to the number of hours worked prior to a non-injury shift. Additionally, they utilized paired *t*-tests and conditional logistic regression to evaluate the significance of the difference between the two instances. Their results showed that there was small, but highly significant differences between hours worked directly preceding the injury and the control, which was each employee's work hours four weeks prior to the injury. Employees who worked more than 64 hours in the week prior to the shift in which an injury occurred had an 88% excess risk when compared to those who worked 40 or fewer hours during the week prior. As a result, Vegso et al. (2007) concluded that controlling overtime in manufacturing settings may reduce employees' risk of injury on the job.

Seeking to study overtime outcomes beyond a single industry or occupation, Dembe et al. (2005) examined responses from 10,793 Americans who participated in the

National Longitudinal Survey of Youth (NLSY) from 1987 to 2000. The NLSY was sponsored by the US Bureau of Labor Statistics and targeted a representative sample of young people living in the United States. The survey collected information on sociodemographic characteristics, as well as job and employer characteristics including work hours and incidents of work-related injuries and illnesses. Their study was based on a theoretical model proposed by Schuster and Rhodes (1985), which posits that overtime and long work hours impact employees' risk of workplace accidents by precipitating the intermediary conditions of drowsiness, fatigue, and stress. The pathway that links a demanding work schedule with these intermediary conditions and ultimately to an accident can be mediated by various personal (e.g., age, gender, health status), job (e.g., hazards, work intensity), and organizational (e.g., overtime policy, amount of supervision) factors.

Dembe et al.'s (2005) results showed that jobs with an overtime schedule were linked to a 61% higher injury hazard rate as compared to jobs with no overtime and that working at least 60 hours per week was associated with a 23% higher hazard rate. Even more, a clear trend was evident, whereby the injury rate increased as the number of hours per day (beyond 8) and week (beyond 40) increased. Nonetheless, this study, as well as those previously discussed, cannot be used to infer that a causal relationship exists between overtime and negative outcomes.

However, one experimental study was conducted by Fang et al. (2015) that does offer support for a causal relationship between fatigue, (a possible outcome of long work hours; the relationship between work hours and fatigue is discussed below) and safety performance in construction workers. In this study, ten participants were required to complete the task of transporting heavy materials across a 10-meter distance, which included a *risky* area intended to mimic the risks construction workers must navigate during a standard workday. The risky area was designed to measure two types of errors: (1) perception, which required participants to visually discriminate between two light signals indicating either a safe or hazardous step, and (2) motor control, which required participants to avoid areas signaled as hazardous.

The researchers designed the task to test if participants could identify and correctly respond to potential hazards while performing a physically demanding activity, and each participant was required to complete the task 200 times. Participants' fatigue level was measured via self-report using the Fatigue Assessment Scale for Construction Workers (FASCW), which includes three dimensions (physical inactiveness, mental fatigue, and discomfort). After every 50 trials, participants completed the FASCW and their error rate was calculated. This enabled four different pairs of fatigue-error data to be collected for each participant. To assess performance under a range of different fatigue levels, participants were required to complete the experiment task on two separate days — once after a half day of normal work, and once after a full day of normal work. Afterwards, the researchers interviewed each participant.

Results indicated that throughout the experiment, participants' fatigue levels ranged from 12 to 38, while error rates ranged from 4% to 38%. To determine if average error rates significantly differed under varying levels of fatigue, *t*-tests were calculated. When participants' fatigue levels were below 20, no significant differences appeared; however, when fatigue levels exceeded 20, there were significant differences between error rates. As such, Fang et al. (2015) deemed a fatigue level of 20 as the critical point at which the impact of fatigue began to emerge. Next, they conducted a linear regression using all sample data with fatigue levels greater than 20, which produced an R^2 value of .82 and a slope of 1.44%. So, on average, each unit increase in fatigue (beyond 20) was associated with a 1.44% increase in error rate. Examining each participants' individual data, all but one worker's error rates increased significantly once their fatigue exceeded 20, while changes in error rates were not significant when fatigue was below 20. Fang et al. (2015) did not discuss how or if participants' baseline fatigue varied after a half day compared to a full day of work, nor did they explicitly mention the use of counterbalancing in their study design. Thus, their study provides limited support for the causal connection between fatigue and safety in the construction industry.

Regarding the relationship between overtime work and fatigue, previous studies offer mixed findings (r ranging from .00 to 0.12), showing that working long hours can sometimes be related to fatigue. Beckers (2008) suggests these findings could be due to range restriction in both overtime and fatigue levels, as well as failure to account for the quality and type of overtime work. She also suggests that in cases of *high-quality overtime*, working additional hours can be fun and therefore have positive outcomes (e.g., increased motivation and job satisfaction).

As previously discussed, researchers such as Harma (2003) and van der Hulst (2003) argue that overtime can create an imbalance between the effort employees expend at work and the amount of time they have to recover from work, as overtime requires not only additional effort investment at work, but also results in less recovery time (Meijman & Mulder, 1998). This theoretical model is supported by the research of Dahlgren et al. (2006), who conducted a within-person experimental field study with sixteen white-collar workers. Participants were exposed to two, counterbalanced conditions: a standard workweek (i.e., 8 hours per day, 40 hours total) or an overtime workweek (i.e., 12 hours per day, 60 hours total). During the extra four hours per day in the overtime condition, participants were instructed to perform their normal tasks, so that workload and stress levels remained consistent between the two conditions. They were also asked to maintain the same sleep schedules throughout both conditions.

Throughout the study, participants underwent extensive surveying and testing. Every morning, participants completed a sleep diary and answered a variety of questions about their sleep quality the previous night. At six points throughout each day of the experiment, participants made self-ratings on their current feelings of stress and sleepiness. At the end of each workday, participants answered additional questions regarding their feelings of tension, irritation, exhaustion, sense of time pressure, ease of thinking, and ability to recuperate throughout the day. They were also asked to rate their workload, pace of work, and satisfaction with their work performance, as well as indicate the activities they did after work (e.g., caring for children or relatives, housework, leisure activities, watching television, rest or sleep).

Results indicated that during the overtime condition, participants reported greater feelings of exhaustion, irritation, and difficulty thinking, and they experienced less sufficient recuperation. Additionally, participants reported less sleep, more fatigue symptoms, and increased sleepiness at the end of the week in the overtime condition as compared to the regular workweek condition. While Dahlgren et al. (2006) could not determine if participants' increased fatigue and sleepiness resulted from overtime hours alone or decreased sleep, previous research does support an association between long work hours and sleep deprivation (e.g., Kageyama et al., 2001; Liu & Tanaka, 2002).

Åkerstedt et al. (2002) also studied the relationship between overtime and fatigue. They examined a Swedish sample of 58,000 working individuals and found that fatigue was significantly predicted by both overtime work (measured as more than 50 hours per week) and physically strenuous work. In another field survey study, Park et al. (2001) examined fatigue complaints and work hours across a sample of 238 male engineers working in South Korea. Park et al. (2001) used an ANCOVA to compare three different groups of participants: those working fewer than 60 hours per week, those working between 60 and 70 hours weekly, and those working more than 70 hours per week, while controlling for age. Their results showed that drowsiness and dullness differed significantly across all three groups, with participants who worked more than 70 hours per week reporting the highest levels of drowsiness and dullness. It should be noted, however, that the additional dimensions of fatigue they examined did not differ significantly across the three work hour groups.

In another field study, Proctor et al. (1996) examined the relationship between overtime, cognitive performance (via a series of neuropsychological tests), and fatigue in 248 hourly employees at a fuel-injector manufacturing plant. Of this sample, 66% of participants worked some overtime during the week prior to testing; 28.6% worked more than 8 hours of overtime, and 9.3% worked more than 20 hours of overtime (i.e., more than 60 hours that week). First, Proctor et al. (1996) compared cognitive performance across employees who worked overtime (i.e., any hours over 40) in the week before testing versus those who did not work any overtime in the prior week. The overtime group scored significantly lower than the no-overtime group on 63% (15 of 24) of the cognitive tasks. Additionally, they leveraged a series of multiple regressions to examine the relationship between overtime and a variety of outcome variables. After controlling for demographic characteristics (including gender, age, education, repeated grade in school) and work characteristics (including job type, shift worked, number of consecutive days worked prior to test day, number of hours worked on the test day prior to testing, acute petroleum naphtha exposure), results showed that overtime was negatively associated with scores on tests of attention and executive function, as well as greater ratings of fatigue.

Wong et al. (2019) conducted the most recent meta-analysis on the topic of long work hours and occupational health, which included 46 studies spanning 814,084 employees across 13 countries. Their analysis produced an overall odds ratio of 1.245 (95% CI: 1.195-1.298) between long working hours and occupational health, with their category of *related health* producing the strongest odds ratio (1.465, 95% CI: 1.332-1.611); this category included sleep disturbance, short sleep duration, sleep problems, exhaustion, and injuries. While these results do suggest that a negative relationship exists between long work hours and employee health, this is a relatively small effect, as an odds ratio of 1.68 is equivalent to a Cohen's *d* of 0.2 (Chen et al., 2010). This meta-analysis also identified four moderators: health measure (longer hours associated with greater risk of cardiovascular heart diseases, metabolic syndrome, fatigue, injury, poor sleep quality, short sleep duration, sleep disturbance); cut-point for long hours (significantly higher health risks when working more than 50 hours per week or 10 hours per day); study method (stronger negative relationship for case-control studies versus cross-sectional studies); and country of origin (strongest negative relationship in China). Wong et al.'s (2019) meta-analysis did not consider employees' prior medical history, preexisting medical conditions, or any other individual employee characteristics beyond gender. As noted by Allen et al. (2007), this is a critical component of understanding the true relationship between working hours and employee outcomes. This may also explain why Wong et al. (2019) found a relatively small odds ratio.

In summary, research to-date suggests that the relationship between working long hours and employee outcomes is more complex than it appears. The majority of research findings on the topic do not employ rigorous enough methodology to infer causality, though there is theoretical support for causal pathways through which long hours could negatively impact employees (Harma, 2003; van der Hulst, 2003) as well as empirical support that fatigue can indeed cause more frequent mistakes on the job (Fang et al., 2015) and that fatigue can be a product of overtime work (Dahlgren et al., 2006). Correlational research shows that long work hours, especially beyond 50 hours per week, is often related to negative health and safety outcomes for some employees, but individual demographics such as age and prior health status can moderate this relationship (Allen et al., 2007), as can characteristics such as job type, industry, and even workplace technology (Shepard & Clifton, 2000).

Motivation to Work Long Hours

Another important topic to address in the study of overtime is why employees make the decision to work long hours in the first place. In the mid-twentieth century, economists noticed a somewhat abrupt halt in the steady decline of hours worked by the average American employee; around 1970, the number of working hours began to rise

(Schor, 1991). While some researchers suggest this increase in hours occurred as a result of employers' power to require additional work from their employees (e.g., Schor, 1991), George (1997) argues that modern marketing and advertising tactics could be to blame, as Americans need additional income to support their desire to purchase more and more products. Others, however, have noted that the notion of American consumerism and/or materialism causing longer work hours is not supported by the available data. Feldman (2002), for example, notes that many of the managers who are working high amounts of overtime can already afford the luxuries they want. Furthermore, if working long hours was motivated exclusively by economic need, we would see employees with more children working longer hours than those without, yet the relationship between hours worked and family size is not strong (Filer et al., 1996; Schor, 1991).

To explore why some managers work long hours, Feldman (2002) constructed a theoretical framework consisting of individual-level, job-level, organizational-level, and economic-level factors that explain when and why managers choose to work long hours. At the individual level, Feldman suggests that employees' tendency to work long hours varies across demographic and personality factors (i.e., males, those who are single, those without children, those in the breadwinner role, those with high self-monitoring tendencies, those without hobbies/leisure activities, and those with higher levels of conscientiousness and achievement motivation are likely to work longer hours).

With respect to the job itself, Feldman (2002) proposes that managers whose work outputs are less visible/tangible are likely to work more hours, as are those in jobs with less specific appraisal criteria and those who are evaluated on contextual rather than task-specific performance. Feldman also suggests that managers whose work is more intrinsically rewarding work more hours. Schedule, working conditions, and telecommuting opportunities also play a role. Finally, the organizational factors that can result in managers working longer hours include leadership support of long hours/overtime, strong organizational norms and expectations of working long hours, a cultural perception of time as monochronic (i.e., viewing time as a fixed, measurable commodity), and even economic pressures such as competition with other organizations, declining profits, and the threat of layoffs.

Like Feldman (2002), Lambooij et al. (2007) argue that the relationship between employee and employer is characterized by more than finances and economic principles. While time and money may be the most obvious components of the relationship, Lambooij et al. (2007) argue that a social exchange relationship also exists and that this relationship is governed by the exchange of social goods such as cooperation, time, effort, career prospects, and social capital. For example, in exchange for an employee's time, in the form of working long hours, the employer may provide the employee with better career prospects, in the form of a special assignment or promotion (Lambooij et al., 2007).

The concept of a social exchange relationship, as well as the monetary benefits discussed by George (1997), are both evident in the research of Goldenhar et al. (2003), who conducted focus groups with construction workers in the United States. Throughout the discussion of why these employees choose to work overtime hours, three primary themes emerged: expectations of management, long-term career prospects, and money. Management expectations and career potential point to the existence of a social exchange relationship (Lambooij et al., 2007), while monetary gains point to employees' desires to fund new cars, vacations, even gifts for their children – all of which point to modern consumerism and the success of marketing and advertising practices noted by George (1997). In summary, available research suggests that employees are motivated to work long work hours and/or overtime hours for a variety of reasons, and the choice to work more is not always strictly monetary in nature.

Present Study

The data for this applied dissertation are from a large, US-based manufacturing company. The company's Supply Chain leadership team (which oversees manufacturing) was considering a company-wide reduction in the number of voluntary overtime hours that manufacturing employees were permitted to work in a given week. At the time of this project, all hourly employees could work unlimited overtime hours. Some members of the Supply Chain leadership team wanted to better understand the relationship between overtime and various metrics including absenteeism, downtime, quality holds (i.e., product deficiencies), workplace accidents, and dollar amount of workers' compensation claims. Specifically, they wanted to understand if locations with employees working a lot of overtime performed worse on these metrics, which would support the implementation of a policy limiting the total number of hours employees could work each week.

The Vice President of Human Resources (HR) for the Supply Chain function approached the company's internal talent analytics team – composed of one I-O Psychology PhD practitioner, one I-O Psychology MA, ABD practitioner, and one Master of Human Resources practitioner – to help with this work. A project team was formed, consisting of the talent analytics team, several subject matter experts (specifically, engineers who worked at various manufacturing facilities across the country and were very familiar with day-to-day operations and the nature and nuances of plant data/metrics), and HR leaders in the Supply Chain function.

As discussed by Rynes et al. (2002), a number of contextual factors can impact how or if research is conducted in organizations. In this case, as the team started working on the project, several challenges became apparent: data quality and availability, the company's lack of ability/willingness to conduct research, and significant time constraints. The team recognized that they could provide some insight about overtime, but the ability to meticulously research the true causal impact of overtime would require experimental methodology, substantial data extraction and cleaning, advanced modeling techniques, and the most precious and limited resource of all, time. Time limitations were arguably the most influential constraint, as budgetary meetings were set to occur within the next month, and Supply Chain leadership needed to know whether to set aside funds to hire additional hourly headcount, which would be necessary if an overtime-reduction policy was enacted.

After a series of conversations with the project team, it became apparent that the HR Vice President who originally commissioned this work was the stakeholder most willing to slow down and leverage data-driven insights before making the decision to enact an overtime policy; the majority of stakeholders (i.e., business leaders) did not believe it was necessary to examine the available data before making a decision and were very much comfortable relying on anecdotal evidence alone. Their lack of interest in examining the available data prior to implementing the policy was likely due to three primary factors: (a) the company's overarching culture, which values quick, agile decision-making and action over slow, careful deliberation and inaction; (b) anecdotal evidence, gut feelings, and/or company narratives suggesting that due to the demanding nature of working in a manufacturing plant, allowing employees to work more than 60 hours per week could be dangerous; and (c) the messiness of (and therefore lack of trust in) the Supply Chain function's data, especially at a national level. With this context in mind, the project team decided on an analysis plan that required less data extraction and cleaning, as well as significantly less time and methodological rigor: using a median split to assign plants to high- and low-overtime groups and then comparing the two groups' mean performance on several important business metrics.

This applied dissertation describes that methodology including how the data were aggregated, analyzed, and examined. It is worth noting that the original analyses did not involve any statistical tests; instead, mean values were compared based on practical significance. This approach was chosen because it made the results simple to explain to stakeholders, who were able to evaluate mean differences based on practical importance and financial implications and were less concerned with evaluating the statistical significance of these differences. Many researchers have discussed the shortcomings of an overreliance on statistical significance (e.g., Amrhein et al., 2019; Hurlbert et al., 2019) and instead advocate for "a more holistic view of the evidence that includes the consideration of … related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of findings, and other factors that vary by research domain" (McShane et al., 2019, p. 238); however, these researchers do not recommend entirely abandoning the use of statistics as was done in this project. As such, after briefly summarizing the original work done by the project team, the same data will be statistically examined and interpreted through a scientific lens. These results will be compared with the original team's findings, and the project work will be evaluated and discussed in the broader context of the science-practice gap, including how and why internal I-O practitioners often make different methodological decisions than their research-scientist counterparts.

CHAPTER 2

ORIGINAL PROJECT SUMMARY

Participants

The study's sample consisted of twenty-nine manufacturing plants based in the United States that employ approximately 16,700 hourly employees.

Procedure

Because two of the five outcome metrics, downtime and destroyed quality holds, cannot be broken out by department, all analyses were conducted at the plant level. Data were extracted by a subject matter expert (SME), an engineer highly familiar with Supply Chain data and the various systems and software used to record and store important plant metrics. Before examining any of the plant metrics, the company's 29 plant locations were classified into one of two groups: high or low overtime (OT). This classification method was selected primarily for its simplicity – the Supply Chain leadership team already recognized that plant locations differed on the amount of overtime worked, so breaking them out into two groups offered a simple method of comparison. Furthermore, this type of analysis was quick and easy to explain to others, both of which were important to the project's stakeholders.

To determine whether each plant would be assigned to the high- or low-overtime group, the team used company logs of the number of hours each employee worked per week in 2018. These logs along with total headcount at each plant were used to assign each plant an overtime percentage, referred to as its *overtime (OT) utilization*. This OT utilization percentage indicates the average proportion of a plant's total headcount that is spending 50% or more of the year working more than 60 hours per week.

To calculate a plant's overtime utilization, the number of employees at each plant who worked more than 60 hours per week for at least 6 weeks in each quarter (50% of the time) in 2018 were counted. The decision to use a threshold of 60 hours per week (rather than 40, which would be considered standard overtime) was guided by the company's proposed policy, which would limit overtime to no more than 20 hours in a week (i.e., a maximum of 60 hours worked each week). Next, the data across the four quarters were used to calculate the average number of employees that worked 60 or more hours per week, for at least six weeks per quarter, in 2018. Finally, this number was divided by the total headcount at that plant.

It is important to note that the team felt that this methodology would allow each plant's population size, which varied substantially across the 29 plant locations, to be taken into account; instead of just looking at total hours worked at each plant, this analysis identified the percentage of total headcount that was working overtime, as well *spread* of overtime (i.e., whether a plant has *many* people working over 60 hours consistently or just a few people working very high hours and driving plant total hours up). This allowed the team to identify plants that had a high percentage of their plant population working overtime hours consistently – these are the high-overtime plants. See Table 1 for an example of how each plant's overtime utilization was derived.

Table 1

Calculation of Plant X's 60-Hour Overtime Utilization

# Employees per Quarter Working >60 Hours/Week 6 or More Weeks				2018 Avg.	Total Headcount	60-Hr OT Utilization
(Q1) 22	(Q2) 15	(Q3) 11	(Q4) 11	15	300	5%

Next, the 29 plants were divided into two categories based on their OT utilization. Plants whose OT utilization fell below the median were assigned to the low-overtime group, while those whose OT utilization fell above the median were assigned to the high-overtime group (see Table 2). A median-split approach, common in applied research (Altman & Royston, 2006), was selected for its simplicity, as it enabled quick and easy comparison across two groups.

Table 2

Median Split of Low- and High-Overtime Plants

Low OT Plants		High OT Plants	
Plant	OT Utilization	Plant	OT Utilization
ST	0.34%	CU	1.88%
CAN	0.76%	KIR	1.92%
WO	0.77%	RO	2.16%
WV	0.82%	JO	2.47%
BE	0.85%	PE	2.61%
WI	0.88%	SA	2.66%
YO	0.88%	AB	2.97%
MO	1.03%	KE	2.98%
KIL	1.11%	FA	3.41%
FRC	1.13%	TO	3.49%
PU	1.38%	DE	3.72%
FRE	1.41%	CHA	3.87%
CAS	1.58%	LY	4.36%
OR	1.85%	VA	6.31%
		IR	6.94%

Finally, the two groups were compared across several important metrics: absenteeism, downtime, quality holds (i.e., product deficiencies), workplace accidents, and dollar amount of workers' compensation claims.

Measures

Absenteeism. Absenteeism data indicated the total number of hours employees were absent from work in 2018, by plant. To control for the population size of each plant, the total number of absent hours was divided by each plant's total headcount. This number was then divided by eight (an estimate of the average number of hours in a workday), to represent each plant's average number of absent days per employee.

Downtime. Downtime is the percentage of time, based on total operating hours, that a plant's manufacturing processes are not fully operating. Downtime can be planned (i.e., scheduled) in the case of shift changes or routine sanitation, or it can be unplanned (i.e., not scheduled) in the case of machinery malfunction or employee errors. In this study, downtime was measured as each location's total downtime percentage in 2018. In addition to raw downtime percentage, a financial estimate of the cost of downtime was also utilized. This metric will be explained in greater detail when the results of the downtime hypotheses are presented.

Destroyed Quality Holds. Quality holds occur when a product is flagged as not meeting the company's quality standards. These products are pulled off the line and manually examined in order to determine if they can still go to market (i.e., there is a minor flaw) or if they must be discarded (i.e., there is a severe flaw that deems the product unsellable). The company labels these severe flaws or defects as *destroyed quality holds*. In this study, destroyed quality holds were measured as each location's total number of destroyed quality holds in 2018, divided by total headcount to control for each plant's population size.

Workplace Accidents. Employees are required to report all workplace accidents, regardless of severity. Details of each accident are stored in a company-wide database and routinely monitored by the company's Health and Safety team. In this study, workplace accidents were measured as each location's total number of workplace accidents recorded in 2018, divided by the total headcount to control for each plant's population size.

Workers' Compensation Claims. Workers' compensation claims are paid out to employees who have been injured on the job to cover the medical costs associated with their injury. In this study, workers' compensation claims were measured as each location's total dollar amount of claims paid out to injured employees in 2018, divided by the total headcount to control for each plant's population size.

Hypotheses

The project team hypothesized that locations with higher overtime would have a greater number of fatigued employees due to the higher number of hours they were working. Fatigued employees could be more likely to make errors while working on the production line, which could result in increased downtime and increased quality holds at high-overtime locations. While previous research does suggest that working long hours is related to greater fatigue (e.g., Åkerstedt et al., 2002; Dahlgren et al., 2006; Park et al., 2001; Rosa, 1995; van der Hulst, 2003), as well as lower productivity (e.g., Shepard & Clifton, 2000) and cognitive performance (Proctor et al., 1996), there is a lack of published research that explicitly discusses relationships with manufacturing downtime and quality holds, perhaps due to their proprietary nature and/or businesses' unwillingness to publish research on them. Therefore, the first two hypotheses were based primarily on the project team's own observations and business knowledge rather than published empirical research. It should be noted that this is somewhat commonplace in applied research, as hypotheses are often formed inductively based on organizational knowledge (e.g., Aguinis, 1993; Daft, 1983).

Hypothesis 1. High-overtime plants will experience more downtime than low-overtime plants.

Hypothesis 2. High-overtime plants will have a higher number of destroyed quality holds than low-overtime plants.

Fatigued employees are more likely to be injured on the job (Dembe et al., 2005; Fang et al., 2015; Folkard & Tucker, 2003; Schuster & Rhodes, 1985; Swaen et al., 2003; Vegso et al., 2007), potentially leading to over 50% of employees filing workers' compensation claims to cover costs associated with their accident (Shannon & Lowe, 2002).

Hypothesis 3. High-overtime plants will have a higher number of workplace accidents than low-overtime plants.

Hypothesis 4. High-overtime plants will have higher dollar amounts of workers' compensation claims than low-overtime plants.

Finally, the project team hypothesized that high-overtime plants would have more absenteeism as compared to low-overtime plants. This trend is supported by Ose (2005), who found that overtime has a positive relationship with the frequency of short-term absences.

Hypothesis 5. High-overtime plants will exhibit higher employee absenteeism than low-overtime plants.

Results

As shown in Table 3, downtime was lower for plants in the low-overtime group ($\overline{x}$ = 6.32%) as compared to the high-overtime group ($\overline{x}$ = 7.00%). While this difference of less than 1% between the low- versus high-overtime groups was not statistically examined and may seem small, the financial impact provides a starker contrast. The project team was given access to financial estimates (i.e., a cost-of-downtime metric) by Supply Chain SMEs within the company to monetize this difference, suggesting that the < 1% difference across the two groups costs over $3,000,000. This cost implication occurs because when a product line is down (i.e., not operating), the business is unable to produce and subsequently sell and thus profit from new product.

Table 3

Performance Across Low- and High-Overtime Groups

	Low OT Plants	High OT Plants
Avg. OT Utilization	1.06%	3.45%
Avg. Downtime	6.32%	7.00%
Est. Cost of Downtime	$28,831,462.27	$31,950,721.26
Avg. Destroyed Quality Holds	36,479.26	53,596.90
Avg. Number of Accidents	9.29	16.08
Avg. Workers' Compensation	$103,585.17	$309,632.25
Avg. Number of Absences	29.17 days	30.36 days

Concerning the Hypothesis 2, the average number of destroyed quality holds was lower for plants in the low-overtime group ($\overline{x}$ = 36,479.26) as compared to plants in the high-overtime group ($\overline{x}$ = 53,596.90). This suggests that product deficiencies are more frequent and more severe in plants with a higher proportion of employees working overtime.

Regarding Hypotheses 3 and 4, low-overtime plants exhibited fewer accidents ($\bar{x}$ = 9.29) as compared to plants in the high-overtime group ($\bar{x}$ = 16.08). Workers' compensation costs follow the same trend; plants in the low-overtime groups paid out fewer dollars ($\bar{x}$ = \$103,585.17) as compared to plants in the high-overtime group ($\bar{x}$ = \$309,632.25).

Finally, the results concerning Hypothesis 5 showed that high-overtime plants averaged about one additional day of absenteeism as compared to low-overtime plants (low-overtime $\bar{x}$ = 29.17 days; high-overtime $\bar{x}$ = 30.36 days). While this comparison was not examined statistically, the project team expected to see a larger difference in the average number of absent days between the two overtime groups and therefore considered Hypothesis 5 as only partially supported.

Business Outcome

These results were compiled, visualized in a PowerPoint deck, and presented to the project stakeholders. The project team emphasized the non-causal nature of the results, which prompted discussion about other potential drivers of plant performance including managerial quality, plant culture, and even the age and functioning of various pieces of machinery in a given plant. Although the project team was unable to conclude that overtime was the cause of differing levels of performance across high- versus low-overtime groups, stakeholders were impressed by the overarching trend showing that plants with a higher proportion of employees working overtime had poorer performance (i.e., higher downtime, more destroyed quality holds, more accidents, greater workers' compensation claim amounts, and more absenteeism) across the business metrics examined. As such, the leadership team decided to invest in additional headcount and enact a policy that permitted hourly manufacturing employees to work a maximum of 60 work per week.

CHAPTER 3

POST HOC ANALYSIS SUMMARY

Procedure

The team's decision to use a median split to transform continuous data into categorical/dichotomous data (high- versus low-overtime groups), while simple and commonly done in applied research (see Altman & Royston, 2006), has clear drawbacks. The richness of the data was lost, and the variability within the high and low groups was discarded (Austin & Brunner, 2004; Cohen, 1983). Furthermore, as Hunter and Schmidt (1990) explain, correlations between an artificially dichotomized independent variable and any dependent variables of interest will systematically understate the true correlation by 20%, assuming the underlying relationship is linear. Kelley and Preacher (2012) go as far as to say this practice is "almost always inappropriate to do" (p. 146). So, although a *t*-test would most closely replicate the original analysis conducted by the project team, a series of bivariate regressions (where the predictor variable is overtime utilization) allows us to utilize the overtime variability that exists across plants in order to better understand the relationship between overtime utilization and the outcome metrics.

Additionally, when available, financial metrics were leveraged instead of performance metrics so that results could be interpreted through the lens of the company's bottom line. This methodology offers greater insight than a *t*-test, as it enables stakeholders to understand the level of financial risk associated with an overtime reduction policy, rather than comparing performance across two, artificially defined high- and low-overtime groups. A third methodological improvement for this analysis was the addition of department-level metrics, where available. The original analyses were conducted only at the plant level to maintain consistency across outcome metrics, as two of the five metrics, downtime and destroyed quality holds, cannot be broken out by department. Each plant contains two primary departments: manufacturing and warehouse. As such, department-level metrics provide a larger sample size ($n = 58$) and subsequently more power. Note that when department-level breakdowns are available, plant-level regressions will still be conducted and interpreted so that a viable comparison with the original analysis is possible.

Financial Implications

If the company limits the number of hours employees can work each week and wants to maintain the same operational schedule and production as was achieved prior to an overtime-reduction policy being in place, it will need to invest in additional headcount (i.e., hire new employees). While it is certainly possible to avoid hiring additional headcount by requiring current employees who work less than 60 hours per week to work more hours, Supply Chain leadership did not pursue this option based on HR and Legal advisement. A review of the company's 2018 payroll records indicated that roughly 232,000 hours beyond 60 per week were worked across all hourly manufacturing employees that year. As such, in the presence of a maximum overtime policy limiting hours to no more than 60 per week, additional work hours (and employees to work them) would be required.

To hire new employees, the company incurs a number of costs including recruiting, screening, training, and insurance/benefits. Using an upper-end estimate of 232,000 hours to calculate these costs, it is projected that the company would need to hire an additional 108 employees, which costs roughly $7.7 million based on available estimates of recruiting, screening, hourly pay, insurance/benefits, and training. However, enacting an overtime-reduction policy would also result in cost savings, because the company would no longer have to pay any employee for working above 60 hours per week. Based on the amount of overtime worked in 2018, this could result in about $3 million in savings. So, when considering the implications of enacting an overtime-reduction policy, $4.7 million (a $7.7 million investment minus potential savings of $3 millions) represents a rough estimate of the break-even point. In other words, if the company can recoup that cost by reducing downtime, destroyed quality holds (i.e., product deficiencies that are so severe that the product is discarded and never brought to market), absenteeism, accidents, and workers' compensation claims, then the cost incurred by reducing overtime would pay for itself.

It should be noted that this monetary break-even point, while a helpful tool to evaluate the results of these ad-hoc regression analyses, was not discussed by the original project team because the primary stakeholders (i.e., Supply Chain leadership) were well aware that their proposed overtime-reduction policy would require some investment. However, the ability to provide stakeholders with a sense of how financially risky the policy implementation clearly would have been a value-add to the original project. Furthermore, evidence suggesting the company would actually profit from the policy would serve as additional support for its implementation.

Leveraging Confidence Intervals to Examine Level of Risk

By looking at the range of estimates across three different confidence intervals (95%, 85%, 75%), we can better understand how much financial risk the company could incur by implementing an overtime-reduction policy. While these three intervals are arbitrary, they enable a quick comparison across different ranges of the certainty of financial risk. In this way, they will not serve as dichotomous indicators of statistical significance, but rather reflect confidence intervals as a fluid risk indicator of the potential error in the point estimates.

While the vast majority of social science research (including applied research) uses a threshold of $\alpha = .05$, thus reflecting a 95% confidence interval (Miller & Ulrich, 2019), this threshold makes Type I Errors less likely and Type II Errors more likely (Rice & Trafimow, 2010; Wuensch, 1994). As noted by McShane et al. (2019), the tradeoff between Type I and Type II Errors “should depend on the costs, benefits, and probabilities of all outcomes which in turn depend on the problem at hand and which vary tremendously across studies and domains” (p. 238). When trying to develop and test a new scientific theory, setting α at 0.05 or even 0.01 may be justified, as this is a case where a Type I Error is more harmful than a Type II Error. However, in other contexts such as airbag testing or life-saving drug development, committing a Type II Error can be much more serious and more dangerous than committing a Type I Error. In an article titled *How many people have to die over a type II error?* Rice and Trafimow (2010) discuss precisely this dilemma, illustrating how often Type II Errors can occur and the impact this can have on applied research.

Clearly the present study presents less risk than airbag testing or life-saving drug development, but it is likewise an instance in which committing a Type II Error (i.e., incorrectly concluding there is no relationship between overtime and work accidents) is more harmful than a Type I error, as, previous experimental research has demonstrated that overtime can result in fatigue (Dahlgren et al., 2006), which may make accidents more likely (Fang et al., 2015). Making a Type I Error in this situation (i.e., concluding there *is* a relationship between working overtime and accidents even though no relationship actually exists) may be far less harmful than making a Type II Error (i.e., concluding there *is not* a relationship between working overtime and accidents when there actually is a relationship). While there is potential that tenured employees will be frustrated by a newly imposed overtime policy (because it will limit the amount of overtime pay they can earn) and therefore choose to leave the company, previous research does not suggest there are negative health consequences of working fewer than 60 hours per week (i.e., still permitting 20 hours of overtime); however, some research does suggest that working fewer than 40 hours per week is negatively related to employee mental health (Allen et al., 2007). Given the context presented above, it is reasonable to examine confidence intervals as a fluid indicator of risk.

Results

Assumptions

Prior to conducting any analyses, all assumptions required for bivariate linear regression were evaluated: linearity, homoscedasticity, independence of errors, residual normality, and absence of outliers. Outliers were identified as any data point with a standardized residual greater than ± 3 standard deviations. Within this project's dataset, the likelihood of an outlier being a legitimate, accurate value or a data-entry error depends on the source from which the data originated. For example, the company's mechanisms for recording employee absenteeism, accidents, and workers' compensation claims are dependent on employee record keeping, vary significantly by location, and are not validated at a national level. As such, absenteeism, accident, and workers' compensation outliers are likely to be erroneous data points and therefore should be removed from the analysis (Iglewicz & Hoaglin, 1993; Osborne & Overbay, 2004). Downtime, on the other hand, is a meticulously tracked, electronically-measured metric with national processes in place to ensure accuracy and consistency in reporting across plants. As a result, downtime outliers are very likely accurate, legitimate values that should not be removed. Decisions regarding outliers will be presented in more detail when the results of each hypothesis are discussed.

Residual normality was assessed by inspecting a normal probability plot (i.e., a Normal P-P Plot) of the standardized residuals. Some plots exhibited minor departures from the line of best fit, which will be presented and discussed with the results of each hypothesis. The homoscedasticity assumption was met for all variables, assessed via inspection of a scatter plot of the standardized residual values against the standardized predicted values. Finally, the independence of errors assumption was tested by examining the Durbin-Watson value; all values were near 2, suggesting this assumption was met.

Hypothesis 1

The first hypothesis posited that downtime would be significantly higher in the high-overtime group compared to the low-overtime group. A bivariate linear regression was conducted to test this hypothesis, with overtime utilization as the predictor variable and each plant's downtime cost as the outcome variable. Supply Chain SMEs provided the team with an estimate that one hour of downtime was equivalent to $10,000 in lost product, which enabled a conversion of downtime percent to the cost incurred as a result of that downtime. Downtime costs the company money because when a line is down (i.e., not operational), product cannot be produced and therefore cannot be sold to customers. As such, the company loses sales as a result of downtime. By using this cost-of-downtime estimate instead of raw downtime, the regression coefficient and confidence intervals can serve as a means of examining the company's risk of financial loss based on the strength of the relationship between overtime utilization and downtime. No outliers were detected for this hypothesis, and all other assumptions were met (see Figures 1-2).

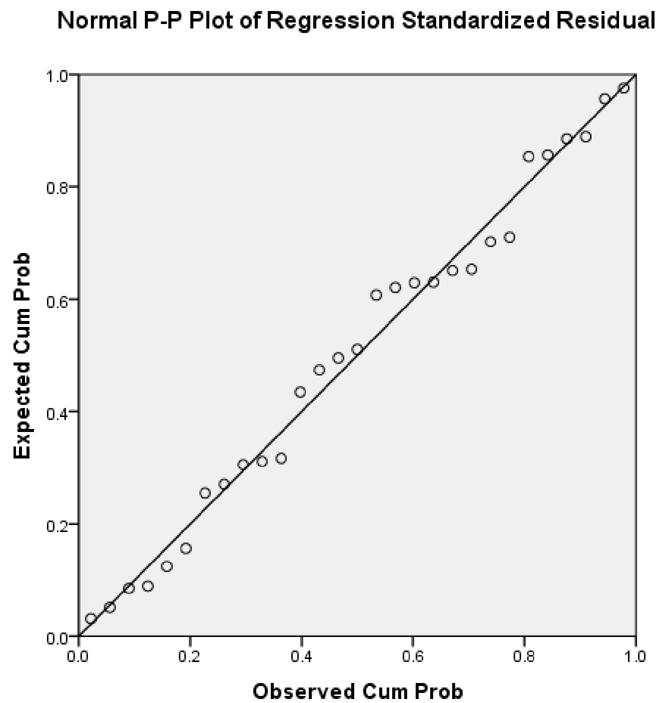

Figure 1: Normal P-P Plot of Regression Standardized Residual for H1

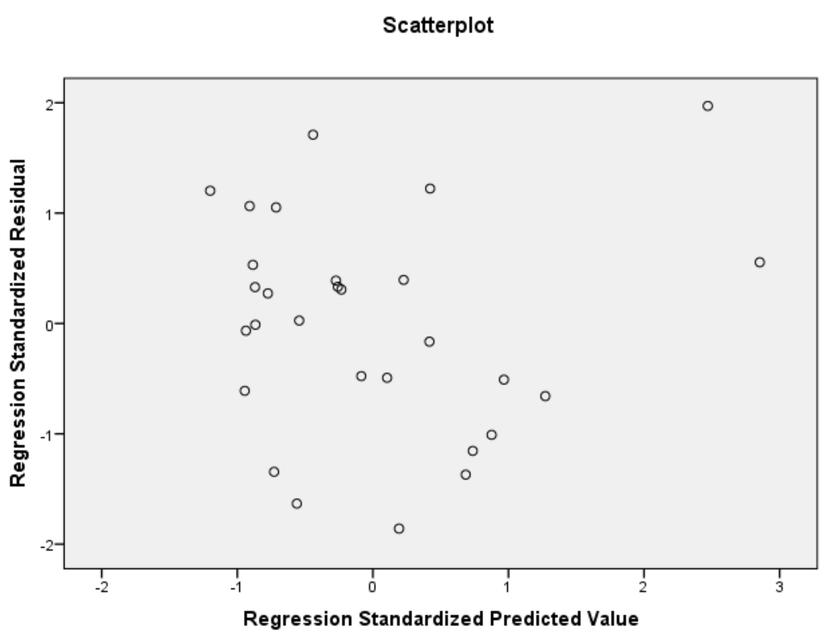

Figure 2: Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H1

Results showed $R^2 = 0.09$, *Adjusted* $R^2 = 0.05$, and $B = \$4,841.03$, $\beta = 0.30$, which suggests overtime utilization accounts for roughly 9% of the variance in downtime costs. Additionally, by decreasing overtime utilization by 1%, the company may save $4,841 (via decreased downtime, assuming that the causal connection hypothesized is indeed the source of the downtime). In order to better understand the relationship between overtime and downtime costs, 2018 headcount and hours data were used to estimate the average number of employees at each plant who would need to reduce their hours to fewer than 60 per week in order to observe a 1% decrease in overtime utilization. The result was about five employees, and so the beta value can be interpreted as follows: for every five employees who reduce their hours to fewer than 60 per week, the company may save $4,841. Company data indicates that 1,285 employees worked over 60 hours per week at least 50% of the time in 2018. If these 1,285 employees could work only 60 hours per week, as would be the case if an overtime reduction policy was enacted, this regression model suggests that the potential cost savings of the subsequent reduced downtime could total $1,244,145.

However, this dollar amount is based on a single sample with only 29 data points that represent the average cost downtime across an entire year. By examining the confidence interval of *B*, we can better understand the range of potential monetary outcomes that might occur if we were to repeat the study with a very large sample. Further, by examining three different levels of confidence (i.e., 95%, 85%, 75%), we can examine varying levels of financial risk. As shown in Table 4, the risk of losing money as a result of enacting a maximum-overtime policy is likely minimal; even in the most conservative 95% CI, the upper range is far larger than the lower range, suggesting there are more favorable odds that the company will save versus lose money. The less conservative 85% and 75% CIs indicate virtually no risk of losing money.

Table 4

Confidence Intervals for Hypothesis 1

CI for B	CI Lower Bound	CI Upper Bound	Policy Impact Lower Bound	Policy Impact Upper Bound
95%	-$1,337	$11,019	-$343,709	$2,831,999
85%	$379	$9,303	$97,503	$2,390,784
75%	$1,301	$8,381	$334,347	$2,153,943

As compared to the original project findings, which leveraged the same cost-of-downtime estimates and estimated potential savings of about $3 million (based on reducing the average downtime of all plants in the high-overtime group to the same amount of downtime reported by plants in the low-overtime group), this ad-hoc analysis suggest savings may be lower, at only $1.2 million.

Hypothesis 2

Next, the team hypothesized that high-overtime plants would have a higher number of destroyed quality holds (i.e., product deficiencies so severe that the product cannot be brought to market and instead must be thrown away/destroyed) than low-overtime plants. Using each plant’s estimated cost of production, a financial metric representing the total dollar amount lost as a result of destroyed quality holds was calculated. This metric was divided by each plant’s total headcount to control for the size of the plant. A bivariate linear regression was conducted to test this hypothesis, with each plant’s 60-hour overtime utilization as the predictor variable and each plant’s cost of destroyed quality holds per person as the outcome variable. No outliers were detected for this hypothesis, and one case was removed due to missing data (this location uses different software to track quality holds, so the national database did not include their counts). While the assumptions of linearity, independence of errors, and homoscedasticity were met, the Normal P-P plot shows some deviation from the line of best fit (see Figures 3-4). Note that if residual normality is the only assumption violated in linear regression, the resulting coefficients will remain consistent and unbiased (Ernst & Albers, 2017; Williams et al., 2013).

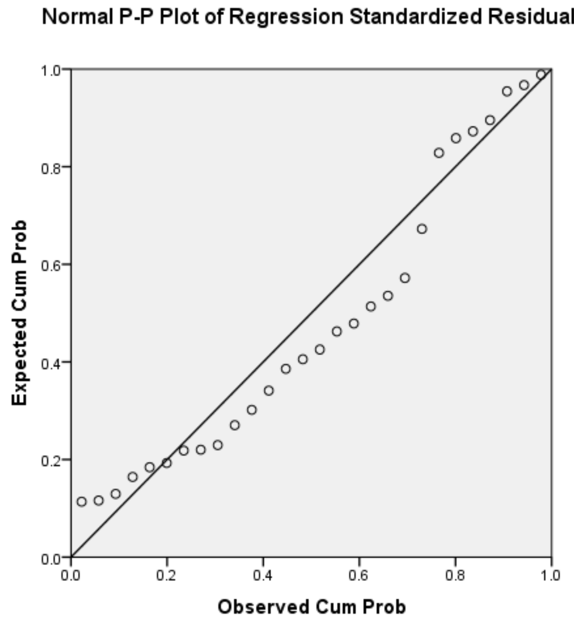

Figure 3: Normal P-P Plot for H2

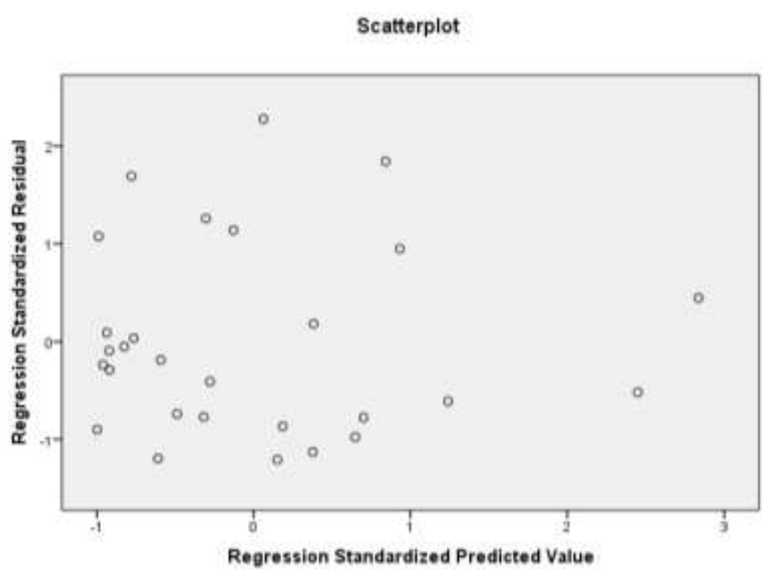

Figure 4: Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H2

The analysis resulted in $R^2 = .13$, *Adjusted* $R^2 = .09$, and $B = \$7.03$, $\beta = .35$, which suggests overtime utilization accounts for roughly 13% of the variance in the cost of destroyed quality holds. Because this regression analysis used cost of destroyed quality holds *per person* to control for the different population sizes across the plants, the *B* value of $7.03 represents only a per-person estimate. To create a plant-level estimate and extrapolate the overall policy impact, this *B* value must be multiplied by the number of employees at each plant regularly working more than 60 hours per week – this number was calculated by multiplying each plant's overtime utilization by its total headcount. This calculation provides a per-plant estimate of the potential cost savings that might result from an overtime-reduction policy because, in theory, each plant would have zero employees working over 60 hours per week once the policy is enacted. Summing this number across all plants, which provides us with an estimate of the policy impact across all plants, suggests a potential cost savings of $225,839. Again, three different confidence intervals were examined (see Table 5), which suggest the company's risk of losing money as a result of decreasing overtime utilization is acceptable.

Table 5

Confidence Intervals for Hypothesis 2

CI for B	CI Lower Bound Per Person	CI Upper Bound Per Person	Policy Impact Lower Bound	Policy Impact Upper Bound
95%	-$.47	$15	-$15,099	$466,455
85%	$2	$12	$52,043	$399,635
75%	$3	$11	$88,023	$363,655

The amount of projected savings provides a very conservative estimate, because it represents *only* the cost of discarded raw materials and does not take into account the loss of profit that occurs because these destroyed products are never brought to market, nor does it account for the employee wages paid by the company to produce product that is ultimately discarded. Unfortunately, the project team did not have access to this type of information. These regression results are consistent with the original project's estimate of $226,000 in cost savings due to decreased destroyed quality holds.

Hypotheses 3-4

Next, the team hypothesized that high-overtime plants would have more accidents and higher workers' compensation claim amounts (i.e., dollars paid out to employees who have been injured on the job to cover the medical costs associated with their injury) compared to low-overtime plants. Due to the low frequency of accidents, only the relationship between overtime and workers' compensation claims, which are the direct result of workplace accidents, will be examined. This metric was divided by each plant's total headcount to control for the size of the plant. Because workers' compensation can be broken down to the department level ($n = 58$), two regressions were used to examine the relationship between overtime and workers' compensation: (1) each plant's overtime utilization predicting each plant's workers' compensation claim amount; and (2) each department's overtime utilization predicting each department's workers' compensation claim amount.

Plant-level analysis. No outliers were detected for the plant-level analysis. The assumptions of linearity, independence of errors, and homoscedasticity were met, but the Normal P-P Plot shows some questionable characteristics; however, it does generally follow the line of best fit (see Figures 5-6). Note that if residual normality is the only assumption violated in linear regression, the resulting coefficients will remain consistent and unbiased (Ernst & Albers, 2017; Williams et al., 2013).

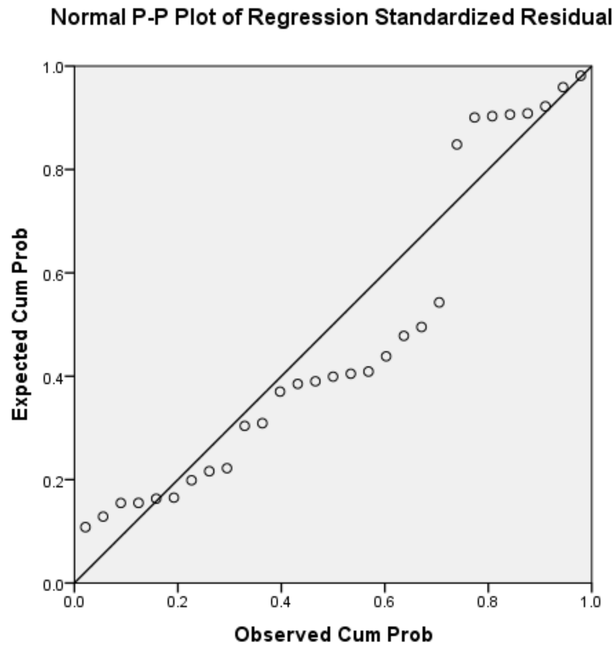

Figure 5: Normal P-P Plot for H3-4 (Plant-Level)

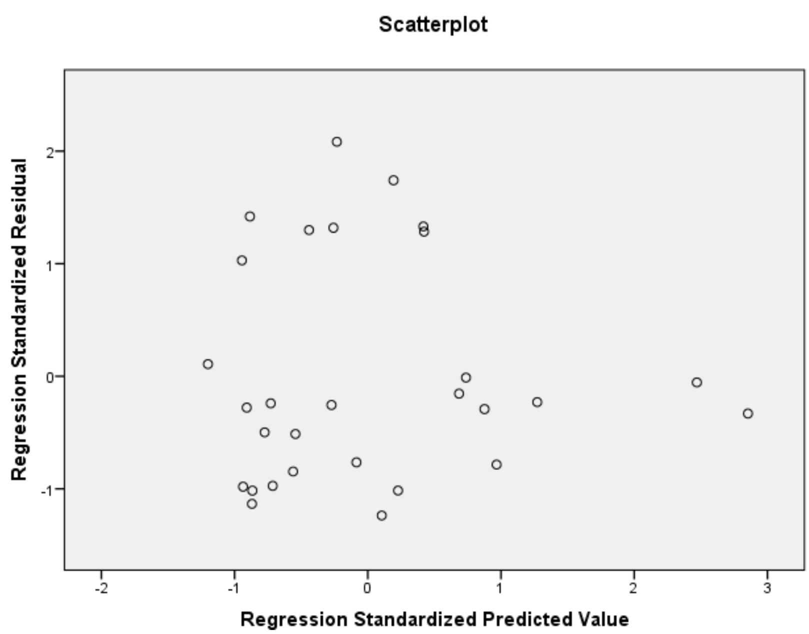

Figure 6: Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H3-4 (Plant-Level)

The regression results were $R^2 = 0.08$, *Adjusted* $R^2 = 0.04$, and $B = \$52.00$, $\beta = 0.27$. This suggests that overtime utilization accounts for roughly 8% of the variance in workers' compensation claims and that for every 1% decrease in overtime utilization, the company may save $52 in paid-out workers' compensation claims. Using the same policy-impact calculation that was applied for Hypothesis 2, a per-plant estimate of projected savings that could occur as a result of reducing overtime utilization to zero was calculated. This calculation suggests that across all plants, a total of $1,831,125 in paid-out workers' compensation claims could be saved, with each plant potentially saving an average of $63,142 per year. These results are consistent with previous literature showing that fatigued employees are more likely to be injured on the job (Folkard & Tucker, 2003; Swaen et al., 2003; Vegso et al., 2007) and subsequently more likely to seek worker's compensation from the company. Again, three confidence intervals were examined (see Table 6).

Table 6

Confidence Intervals for Hypotheses 3-4: Plant-Level Analysis

CI for B	CI Lower Bound	CI Upper Bound	Policy Impact Lower Bound	Policy Impact Upper Bound
95%	-$20	$124	-$653,423	$3,994,101
85%	-$.24	$104	-$7,710	$3,348,389
75%	$11	$93	$338,919	$3,001,760

These confidence intervals suggest greater odds of the company earning money versus losing it, though unlike the CIs of the prior two hypotheses, both the 95% and 85% suggest that the company could lose money. However, the upper ranges remain far larger than the lower ranges. As such, the overall financial risk appears to be relatively low. As compared to the original project findings, which estimated that about $2.9 million could be saved in workers' compensation costs if accidents were reduced, this post-hoc analysis suggests lower cost savings of only $1.8 million.

Department-Level Analysis

Each plant has two primary departments, manufacturing and warehouse, and workers' compensation records indicate the department employees were working in when the accident occurred. These records were used to create a department-level dataset, which indicated the total dollar amount of workers' compensation claims within each plant's two primary departments (n = 58). Before running the regression, each department's overtime utilization, which is based on the number of hours worked by each employee within a department, was converted into a *z*-score to check for outliers. While raw hours records are very likely accurate (due to the Finance team ensuring workers are accurately paid based on the number of hours they worked), there are known data quality errors in the company's records of which department employees work in. Therefore, overtime utilization outliers at the department level are likely erroneous values that should be removed from the analysis (Iglewicz & Hoaglin, 1993; Osborne & Overbay, 2004). One department with very high overtime was identified (overtime utilization = 13.72%, $z = 4.63$) and removed from all department-level analyses (i.e., workers' compensation and absenteeism).

For the department-level workers' compensation regression, all assumptions were met (see Figures 7-8). Four worker's compensation outliers were identified with standardized residuals of 5.50, 3.70, 3.58, and 3.30. As previously discussed, workers' compensation outliers are very likely erroneous values that should be removed from the analysis (Iglewicz & Hoaglin, 1993; Osborne & Overbay, 2004), because the company's mechanisms for recording workers' compensation claims are dependent on employee record keeping, vary significantly by location, and are not validated at a national level. Removal of these outliers produced a larger R^2 value (0.01 to 0.09).

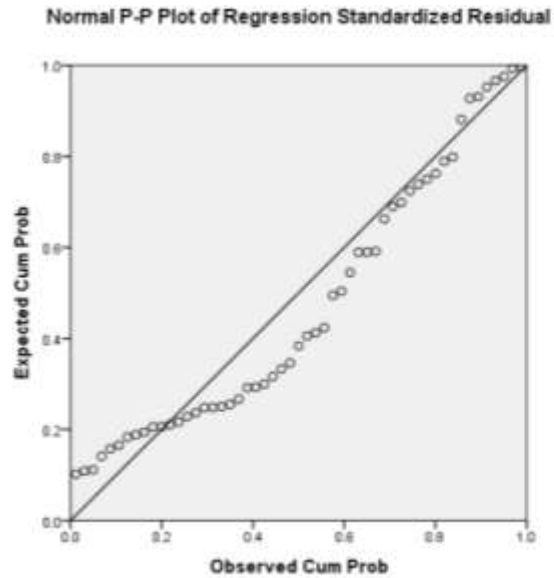

Figure 7: Normal P-P Plot for H3-4 (Department-Level)

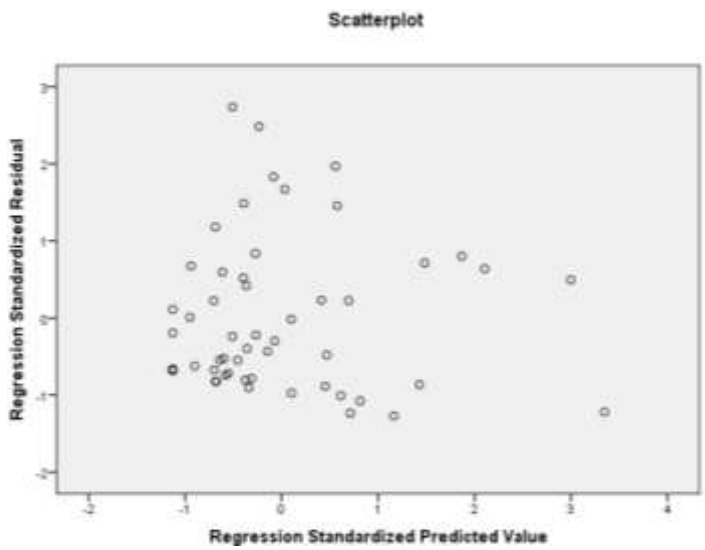

Figure 8: Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H3-4 (Department-Level)

Results were $R^2 = 0.09$, *Adjusted* $R^2 = 0.08$, and $B = \$46.24$, $\beta = 0.31$. This suggests that, at the department level, overtime utilization accounts for roughly 9% of the variance in workers' compensation claims and that a 1% decrease in a department's overtime utilization may result in that department saving \$46.24 in paid out workers' compensation claims per year. Using a per-department calculation to extrapolate the policy impact across the entire company, it was estimated that this decision could result in a total of $1,268,487 in savings. The confidence intervals displayed in Table 7 suggest very little risk of the company losing money as a result of decreasing overtime, even at the most conservative 95% CI.

Table 7

Confidence Intervals for Hypotheses 3-4: Department-Level Analysis

CI for B	CI Lower Bound	CI Upper Bound	Policy Impact Lower Bound	Policy Impact Upper Bound
95%	$6	$87	$153,349	$2,383,626
85%	$17	$76	$456,480	$2,080,220
75%	$23	$70	$621,899	$1,914,801

In summary, plant- and department-level results for workers' compensation were largely consistent, with the department-level model explaining similar estimates of variance (R^2 = 0.09) compared to the plant-level model (R^2 = 0.08).

Hypothesis 5

Finally, the project team hypothesized that high-overtime plants would have more absenteeism as compared to low-overtime plants. Absenteeism data are available at the department level, so two bivariate linear regressions were run to test this Hypothesis: (1) each plant's overtime utilization predicting each plant's number of absences; and (2) each department's overtime utilization predicting each department's number of absences. The team did not have access to a financial metric for absenteeism, so only raw absences were examined. Each plant's total number of absent days was divided by total headcount, as a means to control for the size of the plant.

Plant-Level Analysis

One outlier was detected for this analysis, with a standardized residual of 4.05; As previously explained, absenteeism outliers are very likely erroneous values that should be removed from the analysis (Iglewicz & Hoaglin, 1993; Osborne & Overbay, 2004), because the company's mechanisms for tracking employee absenteeism are dependent on employee record keeping, vary significantly by location, and are not validated at a national level. Removal of this outlier did not substantially alter the R^2 value (0.02 to 0.01). All assumptions were met (see Figures 9-10).

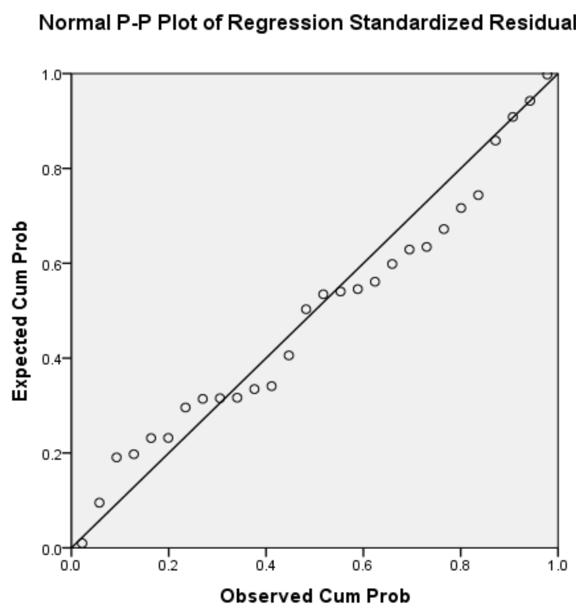

Figure 9: Normal P-P Plot for H5 (Plant-Level)

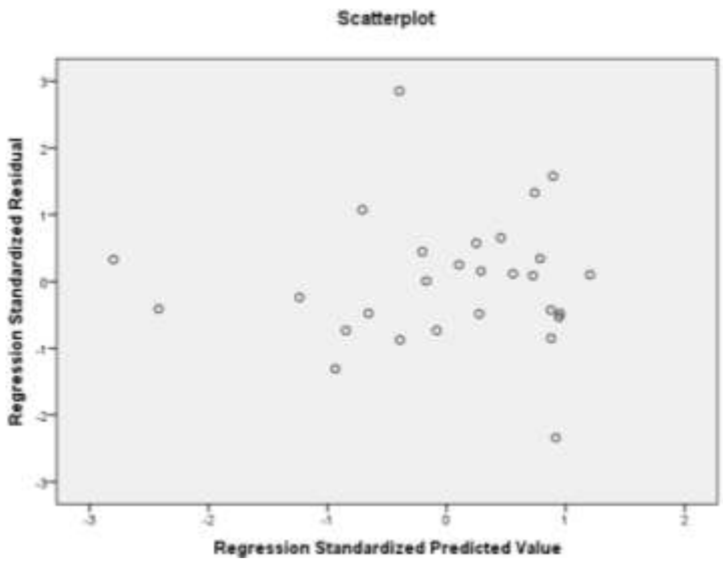

Figure 10: Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Plant-Level)

Results were R^2 = 0.01, *Adjusted* R^2 = -0.03, *B* = -.42, β = -0.11. This suggests that, at the plant level, overtime utilization explains less than 1% of the variance in absenteeism. The negative *B* value, though small, suggests that absenteeism may increase as overtime utilization decreases. Because this regression analysis used absent days *per person* to control for the different population sizes across the plants, a policy impact calculation is necessary to extrapolate the potential company-wide impact. Again, three confidence intervals were examined (see Table 8).

Table 8

Confidence Intervals for Hypothesis 5: Plant-Level Analysis

CI for B	CI Lower Bound	CI Upper Bound	Policy Impact Lower Bound	Policy Impact Upper Bound
95%	-2.01	1.16	-$63,365.25	$36,569.00
85%	-1.57	0.72	-$49,494.25	$22,698.00
75%	-1.33	0.49	-$41,928.25	$15,741.25

These confidence intervals are much wider than the CIs produced by the prior analyses. They also show the greater odds that absenteeism will increase versus decrease if an overtime-reduction policy were enacted. However, previous research and theory (e.g., Ose, 2005) do not support a negative relationship between overtime and employee absenteeism. Moreover, there are known data-quality issues with how absenteeism is tracked across plants in this organization. With all of this in mind, it is reasonable to conclude that there is not a meaningful relationship between overtime utilization and absenteeism and that these results reflect messy data rather than a true negative relationship between overtime and absenteeism.

Department-Level Analysis

One department did not have absenteeism data, so it was removed from the analysis. Additionally, outlier analyses identified one outlier with a standardized residual of -3.08. Its removal did not substantially alter the R^2 value (.01 to .00). The assumptions of linearity, residual normality, and independence of errors were met (see Figure 11), however the scatterplot of standardized residuals versus standardized predicted values exhibits some heteroscedasticity (see Figure 12). Both logarithmic and square root transformations were applied, but the resulting scatterplots did not show marked improvement (see Figures 13-14). Ernst and Albers (2017) note that when heteroscedasticity is the only violation present, the regression coefficients will remain consistent and unbiased; nonetheless, the resulting confidence intervals should be interpreted with caution.

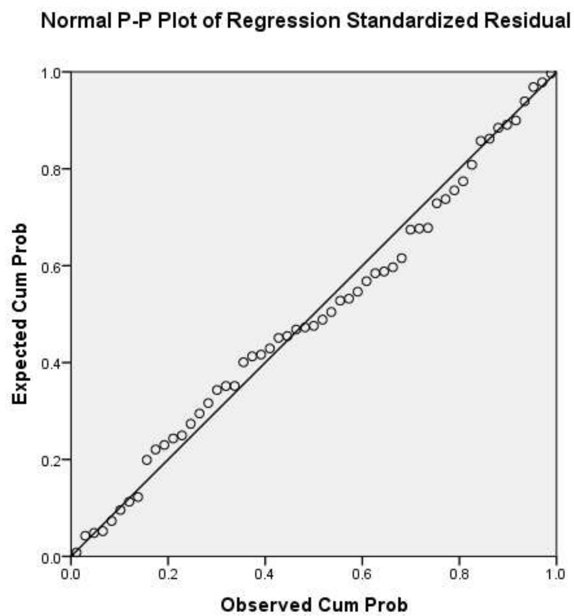

Figure 11: Normal P-P Plot for H5 (Department-Level)

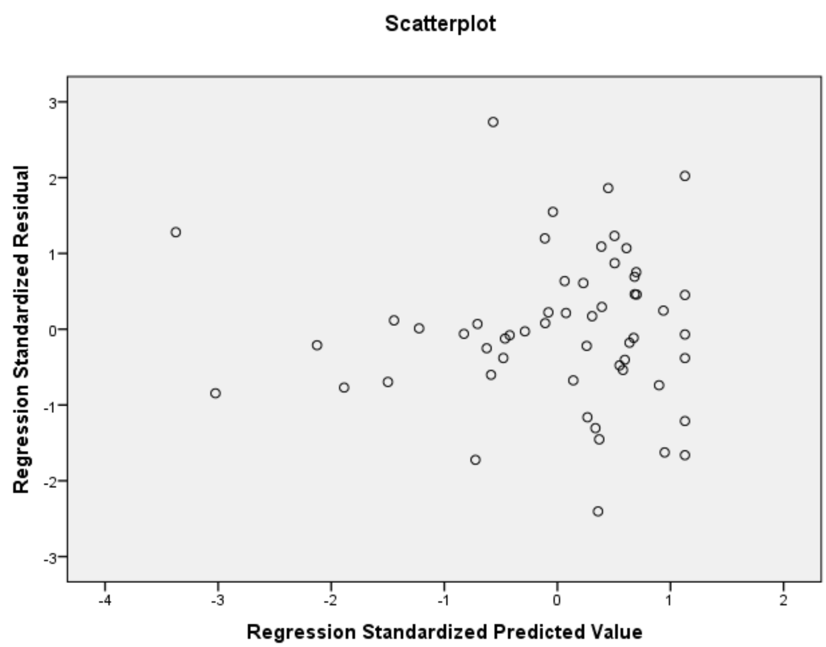

Figure 12: Original Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Department-Level)

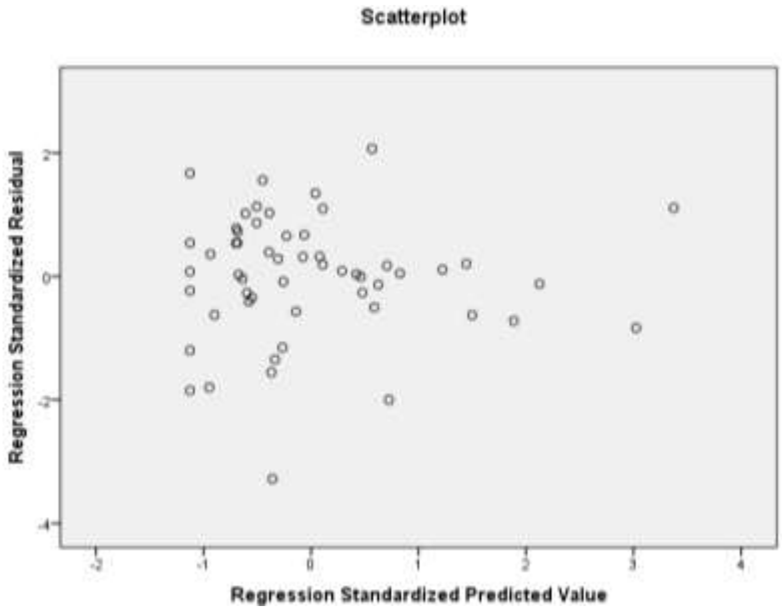

Figure 13: Log Transformation Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Department-Level)

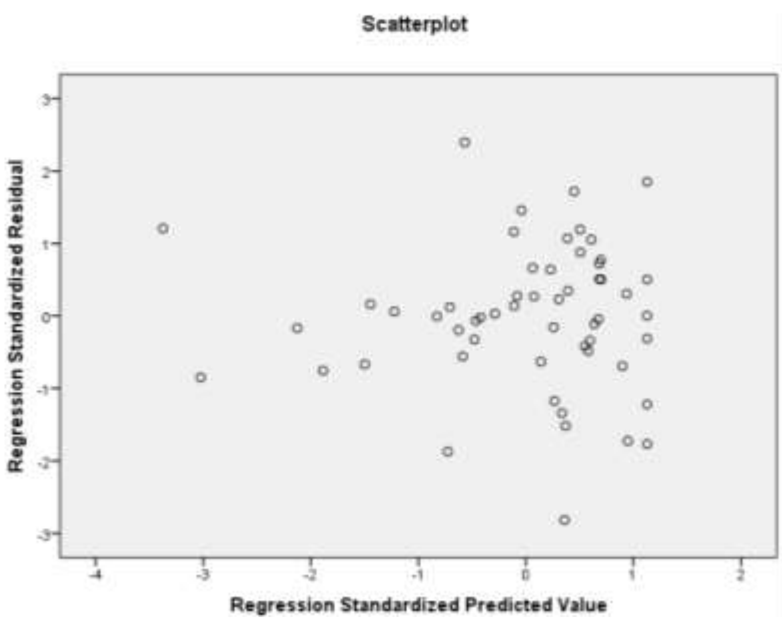

Figure 14: Square Root Transformation Scatterplot of Standardized Residuals vs. Standardized Predicted Values for H5 (Department-Level)

The results were $R^2 < 0.01$, *Adjusted* $R^2 = -0.02$, $B = -.06$, $\beta = -0.02$, which suggests overtime utilization does not explain any of the variance in employee absenteeism. In other words, there is not a meaningful relationship between the two variables. Again, policy impact was calculated, and three confidence intervals were examined (see Table 9).

Table 9

Confidence Intervals for Hypothesis 5: Department-Level Analysis

CI for B	CI Lower Bound	CI Upper Bound	Policy Impact Lower Bound	Policy Impact Upper Bound
95%	-1.13	1.01	-$32,976.43	$29,474.51
85%	-0.84	0.72	-$24,513.45	$21,011.53
75%	-0.68	0.56	-$19,844.22	$16,342.30

Despite the additional power offered by the larger sample size of this analysis as compared to the plant-level analysis, the regression output and CIs are similar. This offers additional support for a lack of meaningful relationship between overtime utilization and absenteeism in this sample.

As shown in Table 10, the results across all hypotheses can be combined to estimate the company's odds of breaking even on the investment needed to enact an overtime reduction policy allowing employees to work a maximum of 60 hours per week. With a break-even point of $4.7 million, point-estimate results indicate the company may only recoup about $3.3 million of the 4.7 million in costs associated with the policy.

Table 10

Cost Implications for Enactment of Overtime Reduction Policy

	B Estimate	95% CI Lower Bound	95% CI Upper Bound	85% CI Lower Bound	85% CI Upper Bound	75% CI Lower Bound	75% CI Upper Bound
H1: Downtime	$1,244,145	-$343,709	$2,831,999	$97,503	$2,390,784	$334,347	$2,153,943
H2: Quality Holds	$225,839	-$15,099	$466,455	$52,043	$399,635	$88,023	$363,655
H4: Workers' Comp.	$1,831,125	-$653,423	$3,994,101	-$7,710	$3,348,389	$338,919	$3,001,760
Total	$3,301,109	-$1,012,230	$7,292,555	$141,836	$6,138,807	$761,288	$5,519,358

CHAPTER 4

DISCUSSION

This study examined two very different approaches to understanding the relationship between overtime and a variety of outcome variables. One approach was a *back-of-the-napkin* analysis, methodologically imperfect and fairly basic (i.e., using a median split and visually comparing means), which some may associate with a *practitioner* problem-solving model; the other was more thoughtful, more methodologically sound, and leveraged a more complex statistical analysis (i.e., bivariate regression). These approaches not only produced different results (i.e., the more rigorous post-hoc analysis projected lower cost savings), but they also provided very different levels of information and insight. Both sets of analyses were underpowered, but the post-hoc analysis allowed us to understand the level of error introduced by the lack of power, while the original analysis completely failed to address it.

Furthermore, the transformation of confidence intervals into dollar amounts offered meaningful insights about the potential financial outcomes of an overtime-reduction policy, which would have allowed stakeholders to make a more informed decision about an overtime reduction policy. This practice allows practitioners to translate statistical analyses into meaningful financial metrics, which stakeholders can clearly understand and articulate. Ultimately, although it can be tempting to dismiss the methodologically rigorous scientific approach when working in an applied setting, this comparison exercise revealed a number of lessons that internal I-O practitioners can benefit from when tasked with answering business questions.

More Data Points Are Better than Fewer Data Points

Though this must be balanced with time constraints and related costs, I-O practitioners should seek to acquire as many meaningful data points as possible when beginning a project. This will allow them to derive more meaningful statistical insights about relationships between variables of interest, which they can share with stakeholders to aid them in effective decision making.

Leveraging the right organizational data is key to accurately answering virtually any business question that arises. In this example, the project team received annual averages of each plant's performance across five key metrics. Person-level data across time would have been the preferred choice and would have allowed a better understanding of the distributions within each plant/department, as well as the seasonality of the business. In addition, monthly or even quarterly averages may have provided deeper insights regarding variation across time without requiring significant additional time or expense related to the analysis. At the very least, the project team should have requested standard deviations along with each plant's mean performance.

Confidence Intervals Are a Useful Tool for Decision-Making

When making data-driven business decisions, confidence intervals can be particularly informative. They enable us to understand how much uncertainty is present in our point estimates, which is incredibly important when dealing with messy organizational data, so that we can gauge how close they are to the true population parameter. When transformed into financial metrics, as was done in this project, confidence intervals can offer valuable clarity about the level of uncertainty of economic risk. Furthermore, the contextual inspection of confidence intervals can be particularly useful when taking into consideration the circumstantial elements related to the decision at hand (e.g., costs and benefits of the decision, magnitude of the decision's impact). However, as noted by McShane et al. (2019), it's important to also take into account things like "related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain" (p. 235) when interpreting statistical results.

Do Not Forget About Outliers

Outliers have the potential to powerfully shape the outcomes of any analysis; failing to determine if outliers are present can compromise the results derived from those data (García-Pérez, 2012). As such, if outliers are detected, the next step is to determine their legitimacy. Removing outliers that are erroneous (e.g., data entry errors, machinery malfunctions) makes sense, but choosing to retain or remove legitimate outliers requires thoughtful consideration (Iglewicz & Hoaglin, 1993; Osborne & Overbay, 2004). Practitioners must be informed about the sources of their data and whether those sources are reliable and valid. If they are not, data quality will likely suffer.

Outlier identification is relatively simple to do, plus it is easy to explain to stakeholders who do not have a background in statistics. Even if practitioners decide to forego testing other assumptions, which might be common (Hoekstra et al., 2012) but is strongly discouraged (e.g., Choi, 2005; Vardeman & Morris, 2003), testing for outliers should not be dismissed, even in a fast-paced, applied setting.

Conclusion

This applied dissertation presented an opportunity to examine the role of the scientist-practitioner gap in organizational decision making, while also evaluating how results differed between a *back-of -the-napkin* analysis versus a more scientific, methodologically rigorous approach. The more rigorous approach produced richer insights regarding how risky the decision to enact an overtime-reduction policy might be from a financial perspective and exemplified the value of leveraging confidence intervals to guide organizational decision making.

While many I-Os, especially those who work as internal practitioners, may eschew the rigors of a more scientific approach, this dissertation demonstrates the value of combining scientific rigor with the contextual knowledge of an applied approach to organizational decision-making. Bridging the scientist-practitioner gap may require that both parties understand their inherent limitations – scientists performing studies in applied settings must better understand the importance of contextual information related to their results, as the results provided by organizational data must be interpreted through a contextual lens that is developed from business acumen achieved through experience in the organization under study. Practitioners, on the other hand, need to adopt methods that lead to better methodology and cleaner data when possible. Though many organizations are developing metrics and systems that result in more accurate data, both anecdotal and empirical evidence indicate that poor data quality plagues many organizations (Tee et al., 2007). This can make insightful, high-quality organizational research that much more challenging to accomplish, as is evidenced by this project. The present study shines a light on how organizations might benefit from the investment in better data and better analytics. In today's world where there is a growing reliance on information technologies, I-O practitioners have the opportunity to bring additional insights to bear on important organizational decisions.

REFERENCES

Aguinis, H. (1993). Action research and scientific method: Presumed discrepancies and actual similarities. *The Journal of Applied Behavioral Science, 29*(4), 416-431.

Aguinis, H., Ramani, R. S., Campbell, P. K., Bernal-Turnes, P., Drewry, J. M., & Edgerton, B. T. (2017). Most frequently cited sources, articles, and authors in Industrial-Organizational Psychology textbooks: Implications for the science-practice divide, scholarly impact, and the future of the field. *Industrial and Organizational Psychology, 10*(4), 507-557.

Åkerstedt, T., Fredlund, P., Gillberg, M., & Jansson, B. (2002). Work load and work hours in relation to disturbed sleep and fatigue in a large representative sample. *Journal of Psychosomatic Research, 53*, 585-588.

Allen, H., Slavin, T., & Bunn, W. (2007). Do long workhours impact health, safety, and producitivty at a heavy manufacturer? *Journal of Occupational and Environmental Medicine, 49*(2), 148-171.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ, 332*(7549), 1080.

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature, 567*, 305-307.

Austin, P., & Brunner, L. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine, 23*(7), 1159-1178.

Banks, G. C., Pollack, J. M., Bochantin, J. E., Kirkman, B. L., Whelpley, C. E., & O’Boyle, E. H. (2016). Management’s science–practice gap: A grand challenge for all stakeholders. *Academy of Management Journal, 59*(6), 2205-2231.

Bansal, P., Bertels, S., Ewart, T., MacConnachie, P., & O'Brien, J. (2012). Briding the research-practice gap. *The Academy of Management Perspectives, 26*(1), 73-92.

Beckers, J. G. (2008). *Overtime work and well-being: Opening up the black box.* Nijmegen: Radboud University.

Belli, G. (2010). Bridging the researcher-practitioner gap: Views from different fields. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society* (pp. 11-16). Proceedings of the Eight International Conference on Teaching Statistics.

Boehm, V. R. (1980). Research in the "real world": A conceptual model. *Personnel Psychology, 33*, 495-503.

Brutus, S., Gill, H., & Duniewicz, K. (2010). State of science in industrial and organizational psychology: A review of self-reported limitations. *Personnel Psychology, 63*(4), 907-936.

Bureau of Labor Statistics. (2020, January 2). *Data Retrieval: Employment, Hours, and Earnings (CES)PRINT:PrintCES CES Program Links*. Retrieved from U.S. Bureau of Labor Statistics: https://www.bls.gov/webapps/legacy/cesbtab2.htm

Caruso, C., Bushnell, T., Eggerth, D., Heitmann, A., Kojola, B., Newman, K., Rosa, R., Sauter, S., Vila, B. (2006). Long working hours, safety, and health: Toward a national research agenda. *American Journal of Industrial Medicine, 49*, 930-942.

Cascio, W. F. (2008). To prosper, organizational psychology should... bridge application and scholarship. *Journal of Organizational Behavior, 29*, 455-468.

Cascio, W. F., & Aguinis, H. (2008). Research in Industrial and Organizational Psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology*, *93*(5), 1062-1081.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation*, *39*(4), 860-864.

Choi, P. (2005). Statistics for the reader: What to ask before believing the results. *Canadian Journal of Anesthesia*, *52*(6), R1-R5.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *1*(112), 155-159.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analytics issues for field settings.* Chicago: Rand-McNally.

Daft, R. L. (1983). Learning the craft of organizational research. *Academy of Management Review*, *8*(4), 539-546.

Dahlgren, A., Åkerstedt, T., & Kecklund, G. (2006). Overtime work and its effects on sleep, sleepiness, cortisol and blood pressure in an experimental field study. *Scandinavian Journal of Work, Environment & Health*, *32*(4), 318-327.

Dembe, A. E., Erickson, J. B., Delbos, R. G., & Banks, S. M. (2005). The impact of overtime and long work hours on occupational injuries and illnesses: New evidence from the United States. *Occupational and Environmental Medicine, 62*, 588–597.

Dunnette, M. D. (1990). Blending the science and practice of industrial and organizational psychology: Where are we and where are we going? In M. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, pp. 1-27). Palo Alto, CA: Consulting Psychologists Press.

Elder, P. K., & Miller, H. D. (1979). The Fair Labor Standards Act: Changes of four decades. *Monthly Labor Review*, 10-16.

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice - A systematic review of common misconceptions. *PeerJ*, *5*(e3323), 1-16.

Fang, D., Jiang, Z., Zhang, M., & Wang, H. (2015). An experimental method to study the effect of fatigue on construction workers' safety performance. *Safety Science, 73*, 80-91.

Feldman, D. C. (2002). Managers' propensity to work longer hours: A multilevel analysis. *Human Resource Management Review*, *12*, 339-357.

Filer, R. K., Hammermesh, D. S., & Rees, A. E. (1996). *The economics of work and pay*. New York: Harper Vollins.

Folkard, S., & Tucker, P. (2003). Shift work, safety, and productivity. *Occupational Medicine, 53*, 95-101.

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology, 3*(325), 1-11.

George, D. (1997). Working longer hours: Pressure from the boss or pressure from the marketers? *Review of Social Economy,* 33-65.

Goldenhar, L. M., Hecker, S., Moir, S., & Rosecrance, J. (2003). The ‘‘Goldilocks model’’ of overtime in construction: Not too much, not too little, but just right. *Journal of Safety Research, 34,* 215-226.

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

Halfhill, T., & Huff, J. (2003). The scientist-practitioner gap: Present status and potential utility of alternative article formats. *The Industrial-Organizational Psychologist, 40*(4), 25-37.

Harma, M. (2003). Are long workhours a health risk? *Scandinavian Journal of Work Environment Health, 29*(3), 167-169.

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

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*(137), 1-9.

Hulin, C. L. (2001). Applied psychology and science: Differences between research and practice. *Applied Psychology: An International Review, 50,* 225–234.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*(3), 334-339.

Hurlbert, S. H., Levine, R. A., & Utts, J. (2019). Coup de grâce for a tough old bull: “Statistically Significant” expires. *The American Statistician, 73*(1), 352-357.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers* (Vol. 16). American Society for Quality Control Statistics Division.

Johns, G. (1993). Constraints on the adoption of psychology-based personnel practices: Lessons from organizational innovation. *Personnel Psychology, 46*(3), 569-592.

Kageyama, T., Nishikido, N., Kobayashi, T., & Kawagoe, H. (2001). Estimated sleep debt and work stress in Japanese white-collar workers. *Psychiatry and Clinical Neurosciences*, *55*(3), 217-219.

Kelley, K., & Preacher, K. J. (2012). On effect size. *American Psychological Association, 17*(2), 137-152.

Lambooij, M., Flache, A., Sanders, K., & Siegers, J. (2007). Encouraging employees to co-operate: The effects of sponsored training and promotion practices on employees' willingness to work overtime. *The International Journal of Human Resource Management, 18*(10), 1748-1767.

Latham, G. P. (2001). The reciprocal transfer of learning from journals to practice. *Applied Psychology: An International Review, 50*, 201-211.

Liu, Y., & Tanaka, H. (2002). Overtime work, insufficient sleep, and risk of non-fatal acute myocardial infarction in Japanese men. *Occupational and Environmental Medicine*, *59*, 447-451.

McShane, B. B., Gal, D., Gelman, A., & Robert, C. T. (2019). Abandon statistical significance. *The American Statistician, 73*(S1), 235-245.

Meijman, T. F., & Mulder, G. (1998). Psychological aspects of workload. In C. deWolff, P. Drenth, & H. Thierry (Ed.), *Handbook of Work and Organizational Psychology, Vol. 2: Work Psychology* (pp. 5-33). Hove, UK: Psychology Press.

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, *14*(1), 1-13.

Mitchell, T. R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review, 10*(2), 192-205.

Mohrman, S. A., Gibson, C. B., & Mohrman, A. M. (2001). Doing research that is useful to practice a model and empirical exploration. *Academy of Management Journal, 44*(2), 357-375.

Münsterberg, H. (1913). *Psychology and industrial efficiency.* Boston: Houghton Mifflin.

Olenick, J., Walker, R., Bradburn, J., & DeShon, R. P. (2018). A systems view of the scientist–practitioner gap. *Industrial and Organizational Psychology, 11*(2), 220-227.

Ones, D. S., Kaiser, R. B., Chamorro-Premuzic, T., & Svensson, C. (2017). Has Industrial-Organizational Psychology lost its way? *The Industrial-Organizational Psychologist*, *54*(4), 67-74.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation, 9*(1), 1-8.

Ose, S. (2005). Working conditions, compensation and absenteeism. *Journal of Health Economics, 24,* 161-188.

Park, J., Kim, Y., Chung, H. K., & Gisanaga, N. (2001). Long working hours and subjective fatigue symptoms. *Industrial Health, 39*, 250-254.

Proctor, S. P., White, R. F., Robins, T. G., Echeverria, D., & Rocskay, A. Z. (1996). Effect of overtime work on cognitive function in automotive workers. *Scandinavian Journal of Work, Environment and Health, 22*(2), 124-132.

Rice, S., & Trafimow, D. (2010). How many people have to die over a type II error? *Theoretical Issues in Ergonomics Science, 11*(5), 387-401.

Rosa, R. R. (1995). Extended workshifts and excessive fatigue. *Journal of Sleep Research, 4*(2), 51-56.

Rotolo, C. T., Church, A. H., Seymour, A., Smither, J. W., Colquitt, A. L., Shull, A. C., & Foster, G. (2018). Putting an end to bad talent management: A call to action for the field of Industrial and Organizational Psychology. *Industrial and Organizational Psychology, 11*(2), 176-219. doi:10.1017/iop.2018.6

Rupp, D. E., & Beal, D. (2007). Checking in with the scientist-practitioner model: How are we doing? *The Industrial-Organizational Psychologist, 45*(1), 35-40.

Rynes, S. L., Bartunek, J. M., & Daft, R. L. (2001). Across the great divide: Knowledge creation and transfer between practitioners and academics. *Academy of Management, 44*(2), 340-355.

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management, 41*(2), 149–174.

Schor, J. B. (1991). *The overworked American: The unexpected decline of leisure*. Basic Books.

Schuster, M., & Rhodes, S. (1985). The impact of overtime work on industrial accident rates. *Industrial Relations: A Journal of Economy and Society, 24*(2), 234-246.

Shannon, H. S., & Lowe, G. S. (2002). How many injured workers do not file claims for workers' compensation benefits? *American Journal of Industrial Medicine, 42*(6), 467-473.

Shepard, E., & Clifton, T. (2000). Are longer hours reducing productivity in manufacturing? *International Journal of Manpower*, *21*(7), 540-552.

Shrivastava, P., & Mitroff, I. I. (1984). Enhancing organizational research utilization: The role of decision makers' assumptions. *Academy of Management Review*, *9*(1), 18-26.

Sparks, K., Cooper, C., Fried, Y., & Shirom, A. (1997). The effects of hours of work on health: a meta-analytic review. *Journal of Occupational Organizational Psychology*, *21*(3-4), 323-331.

Spurgeon, A., Harrington, J., & Cooper, C. (1997). Health and safety problems associated with long working hours: A review of the current position. *Occupational Environmental Medicine*, *54*(6), 367-375.

Stone-Romero, E. (2011). Research strategies in industrial and organizational psychology: Nonexperimental, quasi-experimental, and randomized experimental research in special purpose and non-special purpose settings. In S. Zedeck (Ed.), *APA Handbook of Industrial and Organizational Psychology* (Vol. 1, pp. 37-72). Washington, DC: American Psychological Association.

Susman, G. I., & Evered, R. D. (1978). An assessment of the Scientific Merits of Action Research. *Administrative Science Quarterly*, *23*(4), 582-603.

Swaen, G., van Amelsvoort, L., Bultmann, U., & Kant, I. (2003). Fatigue as a risk factor for being injured in an occupational accident: results from the Maastricht Cohort Study. *Occupational and Environmental Medicine, 60*(Suppl 1), i88-i92.

Swanson, R. A., & Holton, E. F. (2005). *Research in organizations: Foundations and methods in inquiry*. Berrett-Koehler.

Tee, S. W., Bowen, P. L., Doyle, P., & Rohde, F. (2007). Factors influencing organizations to improve data quality in their information systems. *Accounting and Finance, 47*, 335-355.

van der Hulst, M. (2003). Long workhours and health. *Scandinavian Journal of Work, Environment & Health, 29*(3), 171-188.

Vardeman, S., & Morris, M. (2003). Statistics and ethics: Some advice for young statisticians. *American Statistical Association, 57*(1), 21-26.

Vegso, S., Cantley, L., Slade, M., Taiwo, O., Sircar, K., Rabinowitz, P., Fiellin, M., Russi, M., Cullen, M. (2007). Extended work hours and risk of acute occupational injury: A case-crossover study of workers in manufacturing. *American Journal of Industrial Medicine, 50*, 597-603.

Williams, M. N., Grajales, C. A., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation, 18*(11), 1-14.

Wong, K., Chan, A. H., & Ngan, S. C. (2019). The effect of long working hours and overtime on occupational health: A meta-analysis of evidence from 1998 to 2018. *International Journal of Environmental Research and Public Health, 16*(12), 1-22.

Wuensch, K. L. (1994). Evaluating the relative seriousness of type I versus type II errors in classical hypothesis testing. In B. Brown (Ed.), *Disseminations of the International Statistical Applications Institute: Vol 1* (pp. 76-79). ACG Press.