5-2021

Integrating Systems, Processes, and Human Judgment: Three Essays on Value Creation with Supply Chain Analytics

Rebekah Inez Brau
University of Arkansas, Fayetteville

Integrating Systems, Processes, and Human Judgment:
Three Essays on Value Creation with Supply Chain Analytics

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Business Administration

by

Rebekah Inez Brau
Brigham Young University
Bachelor of Science in Management, 2016
Brigham Young University
Master of Science in Instructional Psychology and Technology, 2017

May 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____
John Aloysius, Ph.D.
Dissertation Chair

_____
Christian Hofer, Ph.D.
Committee Member

_____
Enno Siemsen, Ph.D.
Ex-Officio Committee Member

_____
Nada Sanders, Ph.D.
Ex-Officio Committee Member

_____
Brent Williams, Ph.D.
Committee Member

ABSTRACT

Big-data, analytics, automation, and machine learning are changing the role of managers in supply chain and operations functions. Extant research indicates that effective value creation by analytics is achieved through careful attention to three components: technology, people, and processes. As such, the purpose of my dissertation is to improve the integration of people in current supply chain and operations management systems with ever-changing technologies and processes. I focus on one critical function—demand planning—since it is increasingly influenced by predictive analytics yet can still require human judgment. My dissertation is comprised of three essays as follow.

In Essay 1, I implement an experiment to compare existing methods of human-machine integration with two new machine learning methods. The machine learning methods of integration are classified as supervised machine learning and allow the machine learning algorithm to utilize human judgment inputs to train the model. The findings suggest that while human judgment provides a significant benefit, not all methods of integration are equal. The results indicate that the two new machine learning methods of integration that I propose are the most effective forms of integration vis-à-vis other methods commonly used in practice and studied in the academic literature.

Essay 2 consists of a field study at a large, multinational firm, testing the two machine learning methods of integration introduced in Essay 1—interactive machine learning (IML) and human guided machine learning (HGML). Analyzing the results of over three million datapoints across five product categories reveals that demand forecasts using an appropriate process to integrate machine learning and human judgment—IML and HGML—provide a significant benefit to demand planning.

In Essay 3, I study the behavioral mechanisms driving analytics use. Utilizing a multi-theoretical lens combining Adaptive Character of Thought (ACT) theory and dual process theory, I develop, and test, interventions embedded in training aimed at changing behavior and improving performance. The experiment tests the trainings on two types of demand planning tasks: 1) forecasting using IML (classified as a high-level processing task) and 2) forecasting using HGML (classified as a low-level processing task). Results of an experiment reveal declarative knowledge alone changes behavior and improves predictive performance for low-level processing tasks. In contrast, high-level processing tasks benefit from analytical thinking paired with declarative knowledge in training.

This dissertation contributes to the literature on behavioral supply chain and operations management, specifically with reference to human judgment and predictive supply chain analytics. The three overarching contributions are: 1) A comprehensive study testing the efficacy of integrating human judgment and model-based analytics when humans have contextual information unavailable to the model; 2) A first empirical behavioral study of the integration of human judgment with machine learning, that introduces HGML and IML to the behavioral operations and supply chain literature; and 3) An investigation of behavioral interventions that improve predictive analytic processes, using both theories of ACT and dual process theory. In practice, my findings provide operations and supply chain managers with guidance as to when and how human judgment should be integrated with analytics.

ACKNOLWEDGMENTS

I am deeply appreciative for the support and guidance of my committee throughout the process of completing this dissertation. I genuinely respect and admire each committee member: Dr. John Aloysius, Dr. Christian Hofer, Dr. Nada Sanders, Dr. Enno Siemsen, and Dr. Brent Williams. Their time and input throughout this process has been invaluable. I appreciate their willingness to serve on my committee.

I thank the many mentors I have had throughout the 22 years of being a student. I am extremely grateful to the faculty, staff, and fellow students in the Walton College of Business (including the Supply Chain Management Research Center) at the University of Arkansas for the education, support, and resources I received throughout the doctoral program. I can think of no better environment to have earned a PhD. I extend my sincere thanks to Tommaso Batistoni for his time spent programming the applications for all three essays, as well as the assistance he provided on miscellaneous programming throughout the data collection. I am appreciative to Heber Brau for his programming expertise and assistance.

Lastly, I thank my family and loved ones: my father, Dr. Jim Brau, for teaching me from a young age to love learning and create knowledge through rigorous research and hard work; my mother, Michelle Brau, for teaching me the power of gratitude, positive-thinking, and selfless service; my gamma, Dotty May Brau Lavallee, for teaching me to work hard (and play hard) and showing me that women can be whatever they set their mind to; my brothers, Jameson, Brigham, and Heber, for the calls, laughs, and visits; my husband, Kamiko, for his unwavering love, patience, support, and gym dates; Jake for the FaceTime's, visits, and encouragement; Dotty May for her unconditional love, smiles, and snuggles; and my closest friends for the philosophical conversations and fun distractions from stress.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

I. INTRODUCTION

The practice of supply chain management continues to experience disruption with the proliferation of big-data analytics, automation, and machine learning. As firms face the excitement of new opportunities related to analytics, they are also confronted with challenges. For example, a specific issue related to the adoption and use of analytics is the evolving role of supply chain managers. Namely, what is the role of the modern supply chain professional? The answer to this question is nontrivial to modern firms since the value-add offered from analytics is captured through three main components: technology (i.e., data, statistics, information technology), people, and processes (Marchand & Peppard, 2013; Mohr & Hürtgen, 2018). Thus, the focus of my dissertation is on optimally integrating the people in current supply chain and operations management with rapidly-evolving technologies and processes. I concentrate on one specific function—demand planning—as human judgment and predictive analytics are often used in conjunction (Arvan et al., 2019; Narayanan & Moritz, 2015; Seifert, Siemsen, Hadida, & Eisingerich, 2015).

To establish a baseline of current demand planning practice, I draw on the results of a survey conducted by Association of Supply Chain Management (ASCM) of which I was a collaborator. Additionally, I conducted formal interviews with 10 practitioners across the retail, transportation, and consumer-packaged goods industries.

1.1 Practical Application

1.1.1 ASCM Survey

Siemsen and Aloysius (2020) survey members of the ASCM to better understand "current supply chain practices and the degree to which firms actually employ such predictive analytics already in their processes" (p. 3). The total sample is comprised of respondents from 244 firms. A little over half of the respondents describe their job level as manager or advisor (52%). The

other half of respondents are grouped together as analysts/purchasing agents (28%), directors (14%), vice presidents (5%), consultants (1%) and one chief officer. Respondents work on average between 11-15 years in the same industry.

The participants are asked various questions regarding the use of predictive analytics in their organization. The most pertinent question related to my dissertation is: "Which of the following best describes demand forecasts in your [organization/division]?" The answer choices are: a) Based entirely on human judgment; b) Based on human judgment but altered through a statistical method; c) based on a statistical forecast, which is then judgmentally adjusted; d) Result from averaging a statistical and a judgmental forecast; and e) Made by a statistical software package without any involvement from human judgment. Answer choices b-d are examples of integrated demand planning practices—where human judgment is combined with algorithms. The results of the survey question are recorded in Figure 1.1.

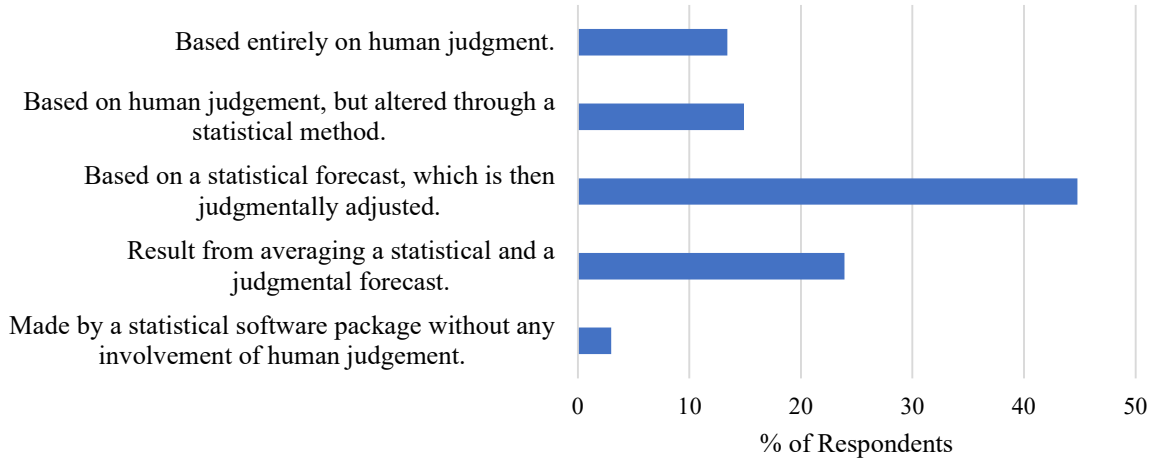


Figure 1.1. Response to the Question "Which of the following best describes demand forecasts in your [organization/division]?"

There are three findings of particular interest in Figure 1. First, the most common method of demand planning employed by respondents is, "Based on a statistical forecast, which is then judgmental adjusted," referred to through the rest of this dissertation as *judgmental adjustment*.

Second, although algorithms have been found to oftentimes be more accurate in demand planning (Dawes, Faust, & Meehl, 1989), only 3% of respondents indicate their firm relies solely on algorithms to forecast (Siemsen and Aloysius, 2020). Third, the vast majority of respondents report that their firm relies on integrated methods of forecasting (83.6%).

The results of the survey help to explain a high-level view of current practice. However, since human judgment is exercised at the individual level, I conducted formal interviews with 10 individuals to glean detailed individual thoughts and experiences on working in tandem with analytics.

1.1.2 Interviews

The interviews took place across three industries: retail, transportation, and consumer-packaged goods (CPG), with 10 individuals assigned to varying levels in their respective company. The job titles of those interviewed were: Analyst (CPG), Director of Data Science for Supply Chain (retail), two Directors of Operations (transportation), three General Managers of Operations (transportation), Associate Director Network Strategy (CPG), Senior Vice President of Operations, and Vice President of Supply Chain. Participants for the interviews were recruited through the Supply Chain Management Research Center (SCMRC) at the Sam M. Walton College of Business in the University of Arkansas.

Each of the interviews began with a set of semi-structured interview questions. The questions included: Have you worked with analytics? Describe the decision-making process (and tools you use) to reach the final forecast. How do you interact with software or statistical models? Where do you go to find information regarding trends (e.g., public press, etc.)? Do you have any pressure from upper management to change your forecasting behavior or incorporate more technology?

Two main themes emerged across industries and job levels. First, how the existing relationship between humans and analytics actually works at the company versus how it should work. Second, explaining the *how* and *why* are critical elements in training humans to work with analytics.

The first theme revealed in the interviews was that human interaction with analytics is a critical part of the current practice and future goals. For example, a General Manager of Operations (2019) said:

> *When you think about transportation, you don't necessarily think about technology. You think about getting trucks loaded and getting them to where they need to go. But [analytics and transportation are] very much intertwined…These days, you are innately connected to technology, there's not one or the other.*

One company seemed to feel as though their current practice was well aligned with their goal for future use. In fact, the same phrase emerged multiple times to describe current and future practice at the company: *"People focused, but digitally enabled"* (Director of Data Science, Supply Chain, 2020; Vice President, Supply Chain, 2019). The company elaborated on the phrase suggesting the relationship between humans and analytics involves analytics tracking, measuring, and reporting the data and humans interpreting it (Vice President, Supply Chain, 2019). One problem that was mentioned by a retailer vice president is the current incentive structure of analytics use may not be aligned appropriately:

> *What does accountability without touchability create? Frustration. So if my role is defined as a forecast analyst, my accountability is forecast accuracy. If I'm heavily leveraging AI, machine learning, to [forecast] without my input, but the outcome is bad, I'm still accountable. My bonus is impacted by that* (Vice President, Supply Chain, 2019).

The second theme that emerged from the interviews is the explanation of the *how* and *why* as two critical components in ensuring an optimal integration of analytics and humans. A Director of Data Science suggested:

> *If we want to change the way that things are done, then we have to bring [the people] along with it. You cannot just say no, this is better, and everyone is using it—that is why you have to use it. That does not help anyone…even if [managers] do not necessarily get the mathematical equation or the model, they do understand how the data is transitioning into the model or how it is kicking out a number. If you can walk them through what is happening, then it definitely helps in building the trust* (Director of Data Science, Supply Chain, 2020).

A General Manager summed up the important elements of training as communication and messaging:

> *So not only are you telling them, but this is also how it's going to be done. You have to let people know how it is going to benefit them… 'Look, the old way, it wasn't good for these reasons. And this is how the new way [analytics] is going to be easier for you. You just have to get over the learning curve and get used to it. The faster you adopt, the faster it'll be able to impact us'* (General Manager, Operations, 2019).

The General Manager continued to then explain how the communication and messaging will be different for differing audiences. An Associate Director of Network Strategy describes the same dilemma; what is the best way to effectively train multiple levels of employees:

> *How do you package a message so that it resonates with the maximum number of people and do it at a high enough level that it's sticky? Because you are going to have 5-10% of the population that wants to just roll their sleeves up and get in the data. You're going to have 5-10% that are leadership, and they're just going to want to stay at the top of the mountains, that 30,000-foot level. And then you have 80% of everybody else that will need something in between, but their background and skills coming in can sometimes be limiting* (Associate Director of Network Strategy, Transportation, 2019).

This question of how to train the maximum number of people regarding analytics use is addressed later on in my dissertation (Essay 3).

Another important idea that emerged through the interviews is that models can learn from more than just historical data—models can learn from human judgment as well. A Director of Data Science for Supply chain said this about reviewing outcomes:

> *Did the system do it better or did [humans]? It's not necessarily the system is always right or managers would always be right. Then we can learn from [the*

> *outcome] and improve [our process]. We can see what did we miss? What [was the manager] thinking because [they] gave exactly the right information. And [the outcome] turned out to be successful. So how can we grow our model, so that we can make it better later?* (Director of Data Science, Supply Chain, 2020).

This interview revealed the notion that predictive performance may be enhanced when human judgment is included as an input. Through further discussions with a company I partnered with for Essay 2, this idea became refined to suggest models could essentially capture and learn from the experiences of managers. This idea will be further discussed throughout my dissertation.

Collectively, survey and interviews can be generalized to some degree to the big picture state of current practice—humans are expected to integrate with analytics by some means. Thus, current practice can be viewed as a partnership between humans and analytics (Wilson & Daugherty, 2018). However, what is not as clear, is how humans and analytics should work together—i.e., what is the best method of integrating humans and analytics. The purpose of my dissertation is to examine the integration of people in current supply chain and operations management with rapidly-evolving technologies and processes in the context of demand planning. The following section provides an overview of the three essays that comprise my dissertation.

1.2 Structure of the Dissertation

This dissertation is comprised of three essays. The first two essays are empirical studies on the process of integrating human judgment with analytics (including machine learning models) conditioned on contextual information available to humans. My third essay develops and tests interventions to train managers on when and how to integrate their judgment with analytics. The remainder of this section provides an overview of the three essays.

In the first essay, "When Models Meet Managers: Integrating Human Judgment and Analytics," I examine the optimal integration of human judgment and analytics. I conduct a 2

(contextual information, no contextual information) x 8 (judgment forecast, judgmental adjustment, quantitative correction, combined, input to model, interactive machine learning, human guided machine learning, model forecast) laboratory experiment to compare forecasting methods of integration. The comparison includes existing methods of integration and introduces two human/machine learning methods. The experiment uses a sample of undergraduate students from a large, private American university. Results of the first study indicate that the machine learning methods are superior to existing methods.

The second essay, "Integrating Machine Learning and Human Judgment: A Study on Demand Planning in the Field," tests the two machine learning algorithms in Essay 1—interactive machine learning and human-guided machine learning—in the field. The field study is completed in partnership with a large, multinational firm. The data from the field study encompass over three million datapoints across five product categories. The analysis reveals that demand forecasts using interactive machine learning and human-guided machine learning are more accurate than the current demand planning processes used in the firm.

In the third essay, "Improving Integrated Judgmental and Analytical Processes Using Interventions Rooted in Cognitive Psychology," I develop and test interventions to train managers on when and how to integrate their judgment with analytics. Specifically, I hypothesize which trainings will help when humans are engaged in high- and low-level processing tasks. The interventions are grounded in two theories from cognitive psychology: Adaptive Character of Thought Theory (Anderson 1976, 1982, 1983, 1993) and Dual Process Theory (Kahneman, 2011; Stanovich & West, 2000; Toplak, West, & Stanovich, 2011; Toplak, West, & Stanovich, 2014). The trainings are tested in a controlled laboratory experiment. The results indicate that specific training elements which appeal to underlying cognitive mechanisms

change human behavior and predictive performance. Specifically, for high-level processing tasks, a combination of declarative knowledge (i.e., information as to how and when interference with analytics is justified), and analytical thinking training elements, decrease deviation from optimal analytics use and improve forecasting accuracy. Low-level processing tasks differ in that it is best to use only declarative knowledge in the training.

Taken collectively, these three essays offer three main contributions to the literature on behavioral supply chain management:

1. A comprehensive study of the efficacy of the various methods of integrating human judgment and model-based analytics when humans have contextual information unavailable to the model;
2. A first empirical behavioral study of the integration of human judgment with machine learning, that introduces human-guided machine learning and interactive machine learning to the behavioral operations and supply chain literature; and
3. An investigation of behavioral interventions that improve predictive analytical processes, using the multi-theoretical lens of adaptive character of thought and dual process theory.

Further, my findings provide operations and supply chain practitioners with guidance as to when and how judgment should be integrated with analytics.

1.3 References

Anderson, J.R., 1976. *Language, Memory, and Thought*. Hillsdale, New Jersey.

Anderson, J. R., 1982. Acquisition of cognitive skill. *Psychological Review*, 89: 69-403.

Anderson, J.R., 1983. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA.

Anderson, J. R., 1993. *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Dawes, R. M., Faust, D., & Meehl, P. E. 1989. Clinical versus actuarial judgment. *Science*, 243(4899): 1668-1674.

Kahneman, D., 2011. *Thinking, Fast and Slow*. Macmillan.

Marchand, D. A., & Peppard, J. 2013. Why IT fumbles analytics. *Harvard Business Review*, 91(1): 104-112

Mohr, N., & Hürtgen, H. 2018. Achieving business impact with data. *Digital McKinsey*.

Narayanan, A., & Moritz, B. B. 2015. Decision making and cognition in multi-echelon supply chains: An experimental study. *Production and Operations Management*, 24(8): 1216-1234.

Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. 2015. Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36*:* 33-45.

Siemsen, E. & Aloysius, J. 2020. Supply chains analytics and the evolving work of supply chain managers. Research report for *Association of Supply Chain Management.*

Stanovich, K.E. & West, R.F., 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5): 645-665.

Toplak, M.E., West, R.F. & Stanovich, K.E., 2011. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7): 1275.

Toplak, M.E., West, R.F. & Stanovich, K.E., 2014. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2): 147-168.

Wilson, J., & Daugherty, P. R. 2018. Collaborative Intelligence: Humans and AI are joining forces. *Harvard Business Review,* 96(4): 115-123.

II. ESSAY 1

When Managers Meet Models: Integrating Human Judgment and Analytics

ABSTRACT

Data analytics and machine learning are increasingly prevalent in emerging demand planning practice. Along with the growth in model-based demand planning, practitioners continue to employ human judgment to incorporate contextual information for increased accuracy of model forecasts. This research uses a 2 (contextual information, no contextual information) x 8 (judgment forecast, judgmental adjustment, quantitative correction, combined, input to model, interactive machine learning, human-guided machine learning, model forecast) experiment to compare existing methods of integration and two new machine learning methods. The machine learning methods of integration are classified as supervised machine learning and allow the machine learning algorithm to utilize human judgment inputs to train the model. The findings suggest that human judgment provides a significant benefit to demand planning processes. Specifically, integrated forecasts (i.e., forecasts that combine human judgment with computational analytics) can substantially improve forecast accuracy compared to non-integrated forecasts. We find however, that this improvement in accuracy is dependent on the method of integration. The two machine learning methods of integration are the most effective methods of integration vis-a-vis other methods commonly used in practice and studied in the academic literature. In keeping with theory, we give direction for further empirical testing of forecasting methods that leverage the strengths of human judgment and the strengths of models and outline potential avenues for study of implementation issues in practice.

Keywords:

Demand Planning, Machine Learning, Human Judgment, Behavioral Experiment

2.1. Introduction

Although still in its infancy, machine learning will be a game changer in supply chain. —SupplyChainToday.com

The practice of supply chain management continues to experience growing pains with the advent of industry 4.0, addressing the increasing advancement of analytics, automation, and machine learning (Waller & Fawcett, 2013). For example, a specific issue related to the adoption and use of technology-enabled tools is the evolving role of supply chain managers. One may ask, that if previous supply chain functions such as downstream demand predictions and upstream interactions are now managed by a combination of big data analytics and enterprise resource planning, what is the role of the modern supply chain professional? Practice and academic literature demonstrate that human judgment remains an essential part of operations, but understanding when and how judgment should be integrated, becomes significant in efficient supply chain management. According to extant literature, one function where human judgment continues to be critical is demand planning (Arvan et al., 2019; Narayanan & Moritz, 2015; Seifert, Siemsen, Hadida, & Eisingerich, 2015).

In the rapidly changing environment of retail supply chains, forecasting accuracy can be a decisive factor that influences the success or failure of a firm. However, despite the considerable interest and public discourse around data analytics and machine learning, modern managers continue to rely primarily on judgment (Siemsen & Aloysius, 2020). According to a recent Association of Supply Chain Management research report (Siemsen & Aloysius, 2020), only 3% of the surveyed firms rely on strictly statistical models. In contrast, 13.4% of respondents use human judgment, independent of any statistical model, when demand planning. The result most relevant to this research is that 83.6% of respondents indicated that they rely on some form of integrated forecasts (Siemsen & Aloysius, 2020). This apparent disconnect between actual

current practice (i.e., primarily human judgment) and the emerging practice (i.e., primarily models) motivates our research as we attempt to explicate how firms should *appropriately* integrate judgmental and model-based forecasts to improve their practice. Thus, our primary research question is how can we best leverage human judgment in model-based demand planning processes?

In order to explain why it may be desirable to integrate judgment and analytical models, we invoke Moravec's Paradox (Moravec, 1988; Sanders & Wood, 2020). The paradox posits that higher-level reasoning tasks (e.g., playing chess, intelligence tests) are easier to replicate with algorithms and artificial intelligence than with lower-level tasks that humans take for granted (e.g., paying attention to things that are interesting, almost anything to do with perception such as face and voice recognition, judging motivation). Although the paradox applies to multiple facets related to humans and models, we refer to prediction via human judgment and statistical models. This insight from Moravec's Paradox helps explain the relative strengths of human judgment and models in demand planning. Blattberg and Hoch (1990) summarize the complementarity of human judgment and models in demand planning (Table 2.1) by describing "*When models (humans) are weak, humans (models) are strong*" (p. 889-890). Humans thrive in the face of special events where flexibility, subjective evaluation, and contextual information are necessary (Goodwin, 2002). Contextual information refers to any information outside of the model (e.g., experience, product knowledge, actions of competitors, and personnel strikes). In contrast, models excel in stable environments where trend detection, optimal weighting of evidence, and systematic integration allow for accurate forecasts (Lawrence, Goodwin, O'Connor, & Önkal, 2006; Sanders & Ritzman, 1992).

Table 2.1. Summary of Blattberg and Hoch (1990) Strengths of Humans and Models.

Humans strengths	Models strengths
Diagnose and predict special events	Unbiased
Evaluate and incorporate subjective factors	Immune to social pressures
Flexible in adapting to changing conditions	Do not get tired, bored, or emotional
Recognize and interpret abnormal cases	Optimally weight evidence

It is possible to draw insight from the domain of chess when in 1997, IBM's Deep Blue supercomputer beat the world's best human chess player, Garry Kasparov, for the first time – a most significant event for artificial intelligence (Sanders & Wood, 2020). Kasparov, in an introspective book on the event, proposed Kasparov's Law which claims that ordinary human judgment and models integrated with the right process are better than a strong model alone:

> *A clever process beat superior knowledge and superior technology. It didn't render knowledge and technology obsolete, of course, but it illustrated the power of efficiency and coordination to dramatically improve results. I represented my conclusion like this: weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process* (Kasparov, 2017).

The emphasis lies in the process. When discussing demand planning practice, the process refers to the method of integration. Extant literature on methods of integrating judgment forecasts and model forecasts fall into four broad categories: judgmental adjustment, quantitative correction, combination, and input to model.

The most common method of integration used in the workplace is judgmental adjustment (Croson & Donohue, 1999; Lee & Siemsen, 2016; Siemsen & Aloysius, 2020; Tokar, Aloysius, Waller & Williams, 2014). Judgmental adjustment is typically what companies refer to as "using technology in the workplace" (Wilson & Daughtery, 2018). The integration occurs as humans

receive output from a model and then adjust the forecast according to their intuition or contextual information.

A second method of integration is quantitative correction (Fildes, 1991; Goodwin, Önkal, & Lawrence, 2011; Theil, 1971). Quantitative correction relies on separating bias in the human judgmental forecast into distinct components. The model is then capable of observing systematic weakness in the human's behavior. For example, humans often overestimate when they anticipate positive demand shocks (Tokar et al., 2014). The quantitative correction would recognize the systematic overestimate and then correct the forecast by the appropriate amount.

A third method of integration is via Blattberg and Hoch (1990) who illustrate how an equal-weighted average of human intuition and statistical modeling improves predictive accuracy by a substantial degree. The integration of judgment and model occurs through mechanical computation of the average of the two forecasts. To date, while this method has been shown to be advantageous in increasing forecast accuracy, it is rarely used in practice (Siemsen & Aloysius, 2020).

Lastly, Green and Armstrong (2013) encourage the use of automated quantitative methods such as extrapolation, quantitative analogies, rule-based forecasting, econometric methods, and index methods to integrate human judgment as an input to the model. With judgment as an input to the model, humans define parameters for the model using their intuition and the model then produces the forecast. It has been suggested that this method should be the most accurate, however few have empirically tested the claim (Sanders & Ritzman, 2004).

During the course of this research study, we engaged in various interviews with practicing supply chain managers, demand planners, and analysts to further understand the future of demand planning in supply chain functions. One common theme which emerged from the

interviews was the balance of control (i.e., touchability) and accountability. Most interviewees exhibited concern regarding a future trend towards automation and machine learning, often characterizing such practices as a black box (see also Petropoulos, Kourentzes, Nikopoulos, & Siemsen, 2018). We aim to clarify machine learning for both supply chain literature and practice by discussing two general types of machine learning: unsupervised and supervised. Unsupervised machine learning uses unstructured data to find patterns and draw inferences often independent of human input. Supervised machine learning, on the other hand, uses a dataset structured by a human (i.e., training data) to develop a model. As such, supervised machine learning offers a rich field of exploration in demand planning by rendering clarity to the black box of machine learning.

Against this backdrop, this paper has the following objectives:

1) Empirically test a comprehensive comparison of methods of integration during demand shocks based on human access to contextual information; and
2) Develop a method of integration to simultaneously use the strengths of human judgment and models when forecasting shocks in supply chains.

We aspire to advance the literature on supply chain and operations demand planning and provide practitioners with valuable insights. Specifically, we contribute by providing evidence which suggests that the underlying issues relating to human behavior differ depending on the way judgment is integrated with a model-based forecast (Perera, Hurley, Fahimnia, & Reisi, 2019). Thus, the results of this experiment yield insight into how best human judgment and statistical models may be integrated to improve forecast accuracy.

We find that human judgment provides significant accuracy benefits to demand planning. We show that integrated forecasts (i.e., forecasts that combine human judgment with

computational analytics) can substantially improve forecast accuracy compared to non-integrated forecasts. In addition, we find that this improvement in accuracy is dependent on the method of integration. We find that human-guided machine learning and interactive machine learning are the most effective methods of integration compared with other methods commonly used in practice and studied in the academic literature.

The remainder of this chapter proceeds as follows. Section 2.2 contains the literature review, which details judgment biases in demand planning and methods of integration. The hypothesis motivation and development are presented in Section 2.3, specifically pertaining to comprehensive comparison, interactive machine learning, and human-guided machine learning. Section 2.4 presents the methodology, describing the experimental design, program and task, and pilot tests. Data collection processes and the sample are also included in Section 2.4. The empirical results are reported and discussed in Section 2.5. Section 2.6 concludes the first essay.

2.2. Related Literature

2.2.1 Judgment Biases in Demand Planning

Judgment biases refer to the systematic deviations observed in the outcomes of human judgments (Bendoly et al., 2010). Judgment biases are often predicated upon heuristics, or "principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (Tversky & Kahneman, 1974: 1124). Examples of heuristics used in judgments are representativeness, availability, and adjustment/anchoring (Tversky & Kahneman, 1974).

Representativeness occurs when judgments are based on the probability of an unknown factor which is assumed to be the same as known factors due to similarities. Representativeness heuristic is often manifest in demand planning when humans make decisions about the

systematic variation in a trend-line based on their interpretation of how they judge a random pattern should appear (Eggleton, 1982).

Availability heuristic describes a human's judgment of probability based on, "the ease with which instances or occurrences can be brought to mind" (Tversky & Kahneman, 1974: 1127). An example of availability is if a manager forecasts the next line of product based on her knowledge of the last line of product, even if the products are not similar in nature. In the context of demand planning, Seifert et al. (2015) offer guidance for when humans have access to historical and contextual information for both relying on judgment forecasting alone as well as when relying on an integration of judgment and model forecasts. When using an integration of judgment and models, they suggest using a decision support system to limit human access to historical information and instead lead the humans to focus on contextual information (Seifert et al., 2015).

Lastly, adjustment/anchoring heuristics surface when judgments are biased towards an initial value (Tversky & Kahneman, 1974). For example, anchoring heuristic manifests itself in demand planning through forecasts or adjustments that originate on a predetermined target based upon historical data instead of from an unbiased starting point. The confluence of judgment biases, heuristics, and demand planning provide an experiment rich for investigation (Harvey, 2007). Kremer, Moritz, & Siemsen (2011) confirm judgment biases in demand planning through a behavioral experiment. They find evidence that humans overreact to errors in stable environments and underreact to errors in unstable environments (Kremer et al., 2011). In another study, Kremer, Siemsen, & Thomas (2015) focus on two additional judgment biases: random judgment error and tunnel vision (i.e., when humans focus on one hypothesis at a time rather than comparing multiple hypotheses simultaneously). Interestingly, Tong and Feiler (2017) used

an operations model to derive the biases exhibited in the demand planning literature. For the reader desiring a more thorough review of the effect of such heuristics and judgment biases on demand planning, see Carter, Kaufmann, and Michel (2007).

Finally, Blattberg and Hoch (1990) provide guidance on how to optimally treat for both judgment and decision bias. Using a framework to isolate managerial intuition in the demand planning process, they test five separate business demand planning database modeling scenarios. Through their analyses, they demonstrate that, "a combination of model and manager always outperforms either of these decision inputs in isolation (Blattberg & Hoch, 1990: 887)." Extending the work of Blattberg and Hoch (1990), the next section discusses four methods of integrating human judgment and model forecasts found in the literature that attempt to increase forecast accuracy.

2.2.2 Common Methods of Integration in Existing Literature

2.2.2.1 Judgmental adjustment

We define judgmental adjustment (Figure 2.1) as any adjustment made to the output of a model forecast (Arvan, Fahimnia, Reisi, & Siemsen, 2019). It is not uncommon for human forecasters to adjust models when the forecasters feel the model does not incorporate all of the relevant contextual data (Boulaksil & Franses, 2009). The consensus of past literature and practice seems to indicate that judgmental adjustment is the most widespread method of integration (Arvan et al., 2019; Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Perera et al., 2019). In fact, even with the advances in technology, recent reports continue to find that judgmental adjustment remains the most common method of integration (Siemsen & Aloysius, 2020). As evidence, a survey of 247 firms, finds that 44.8% of firms state that they use

judgmental adjustment, which doubles the frequency of the next closest method (Siemsen & Aloysius, 2020).

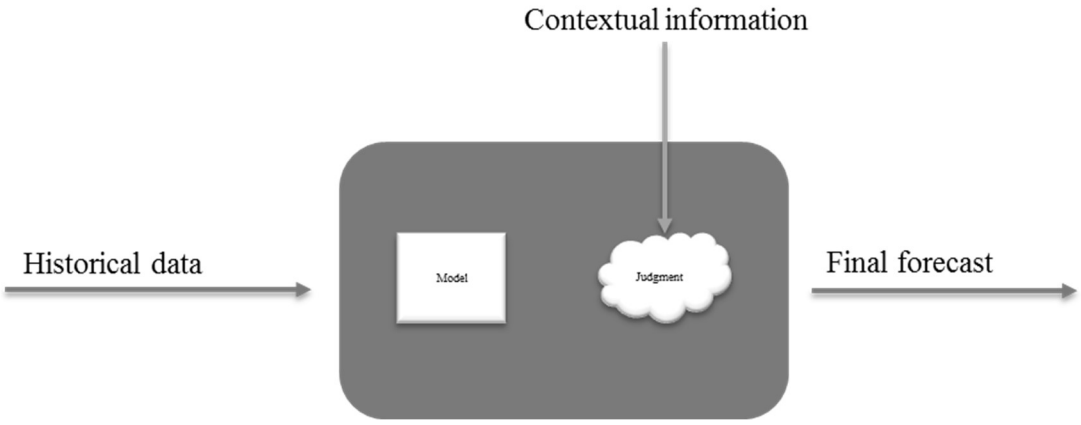


Figure 2.1. Judgmental Adjustment.

Since judgmental adjustment relies heavily on human judgment to drive the process of integration, and judgment virtually always suffers from biases, researchers have sought for demand planning best practices when engaging in judgmental adjustment (e.g., Fildes et al., 2009; Petropoulos, Fildes, & Goodwin, 2016). A common thread in the judgmental adjustment literature argues that adjustments should be made only by experts (e.g., Arvan et al., 2019), as higher levels of expertise lead to improved forecast accuracy (Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2017). Fildes et al. (2009) show that judgmental adjustments generally improve accuracy when: 1) adjustments are relatively larger in nature and 2) adjustments are negatively correlated to the forecast. Consistent with Fildes et al. (2009), Petropoulos et al. (2016) argue that after large deviations in accuracy, experts' judgmental adjustments should be paired with a specific bias correction strategy.

2.2.2.2. Quantitative correction

The second method of integration in our study is defined as an automated system that monitors judgmental forecasts and uses any detected (i.e., computed) bias in the forecasts to

adjust the next period forecast (Figure 2.2). Figure 2.2 shows the initial forecast, quantified by a human, which is corrected by the model to account for any systematic bias.

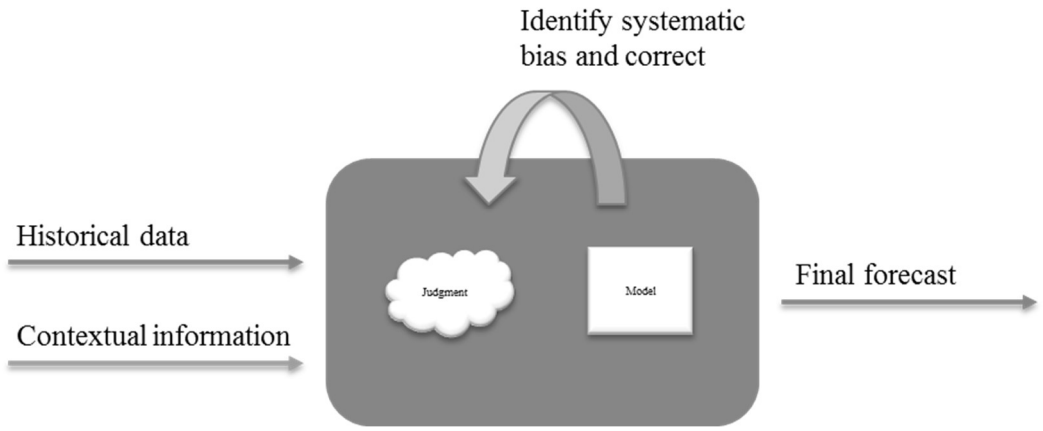


Figure 2.2. Quantitative Correction.

Research regarding quantitative correction is limited (Arvan et al., 2019). However, some existing studies demonstrate large improvements in forecasting accuracy (e.g., Elgers, Low, & Murray, 1995; Fildes, 1991; Goodwin, 2002; Sanders & Ritzman, 2004). Quantitative correction operates by accessing systematic biases in human judgment through a decomposition of the mean standard error (MSE) into unique sources of error. The most common quantitative correction is Theil's correction (Goodwin et al., 2011). Theil's correction separates MSE into two types of bias: mean and regression (Theil, 1971). The mean bias represents the historic tendency of a human to overestimate or underestimate when forecasting. The second bias, regression, refers to a systematic inability to detect patterns in the trendline (Theil, 1971). Theil demonstrates that a regression of actuals on the forecast removes both the mean and regression biases from past forecasts, so it is proposed that his method should remove the same biases (assuming they are systematic) from future forecasts (Goodwin, 2002). Studies have since verified Theil's correction and recommend its use for better accuracy in demand planning (see, Elgers et al., 1995; Goodwin, 1997; Goodwin, 2000).

A second common method of quantitative correction is proposed in Fildes (1991). This form of correction can be used when contextual information is available to the forecaster but may not be explicitly included in the model (Fildes, 1991). The Fildes method defines four determinants of forecast errors including, "an inadequate weighting in the forecast of the economic determinants of output; an implicit causal model which is misspecified, inaccurate forecasts of the determinants of output; and random shocks" (Fildes, 1991: 605). Fildes' model contains a combination of bootstrapping and expectation formation models. He uses a regression model with a series of lagged variables to correct for the error (Fildes, 1991). Elgers et al. (1995) verify the Fildes (1991) method showing an increase in forecast accuracy when used. Although research indicates quantitative correction may aid in increasing forecast accuracy, it is currently not a popular method in practice (Siemsen & Aloysius, 2020).

2.2.2.3. Combination of forecasts

The third method in our study is the mechanical combination of judgment forecasts and model-based forecasts. In the combined method, we refer to combination as the equal-weighted average of a judgment forecast and a model-based forecast, each completed independently as shown in Figure 2.3 (Blattberg & Hoch, 1990). *Combination* is used to refer to the specific method of averaged forecasts. We use *integration* in the broader sense to refer to a forecast that features both human judgment and model forecasts.

The logic for combination is explicitly rooted in the complementarity of judgment and model strengths. Blattberg and Hoch (1990) recommend using an equal-weighted average combination due to three advantages: simplicity, palatability, and accuracy:

(a) simplicity-managers do not need to understand or develop models, so the natural organizational separation of modelers and managers can continue;

(b) palatability-managers retain a considerable amount of control over the decision-making process; and

(c) accuracy, a combination of model and manager will be more accurate than the individual decision inputs (Blattberg & Hoch, 1990: 898).

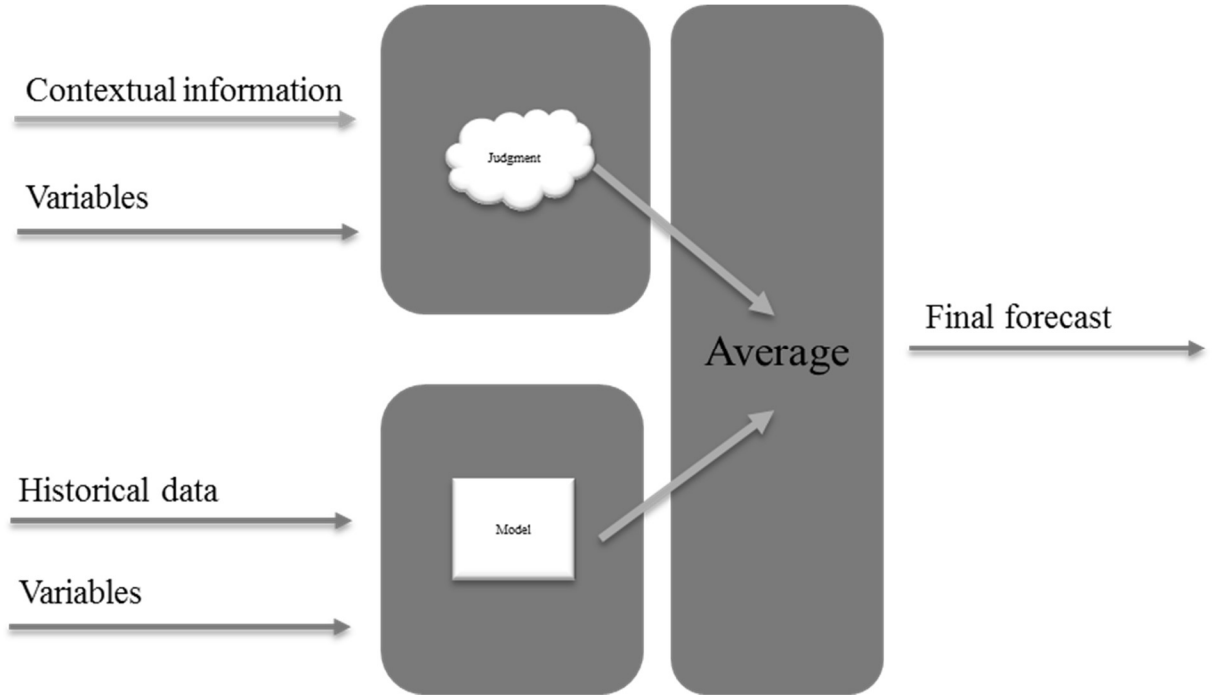


Figure 2.3. Combined Forecast.

Studies since Blattberg and Hoch (1990) test the validity of the equal-weighted average as well as circumstances when it is optimally applied. For example, Franses and Legerstee (2013) use a case study to confirm the accuracy of the Blattberg-Hoch method of combination by concluding combination is more accurate than either the judgment forecast or model-based forecast in isolation (Franses & Legerstee, 2013). In addition, Sanders and Ritzman (1995) assert that a combination of forecasts is most useful when humans have access to contextual information and when data variability is high.

2.2.2.4 Input to model

The last method of integration we test is human judgment as an input to model. This method first operationalizes judgment, typically through an algorithmic approach, and then inputs the estimated judgment into a forecasting model (Figure 2.4). There are multiple approaches to operationalize judgment into models such as: selection of variables, model specification, parameter estimation, and data analysis (Bunn & Wright, 1990). Sanders and Ritzman (2004) argue that input to model offers the most objective, unbiased integration of human judgment and model forecasts.

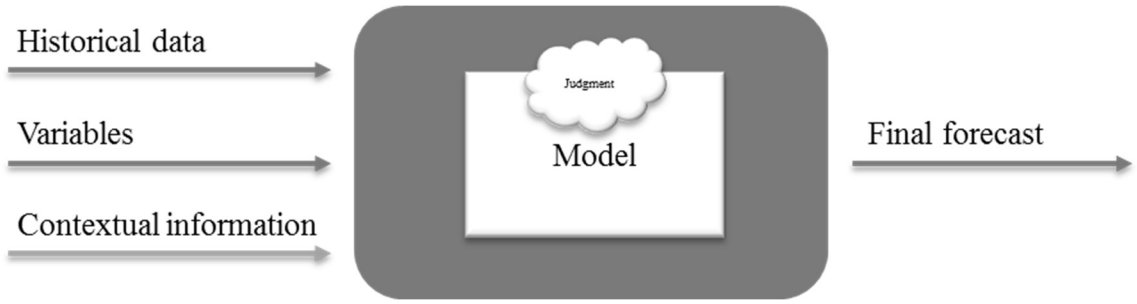


Figure 2.4. Input to Model.

Although articles such as Sanders and Ritzman (2004) suggest that input to model can be the most effective method of integration, research has largely overlooked empirical testing of input to model (Arvan et al., 2019). However, the research we do have lends some evidence that input to model can improve forecast accuracy conditioned on certain circumstances. For example, rule-based forecasting (RBF), which relies on expert human judgment and historical time series to develop rules that are then used as inputs to a model (Collopy & Armstrong, 1992), has been shown useful for complex forecasting scenarios (Adya & Lusk, 2016). Additionally, Baecke, De Baets, and Vanderheyden (2017) demonstrate that judgment, included as a predictive variable in the model during high volatility, can improve forecast accuracy. In at least three studies, evidence suggests that judgmental selection of models can outperform automated

selection of variables in regard to systematic variability such as trend and seasonality (De Baets & Harvey, 2020; Petropoulos et al., 2018; Petropoulos & Siemsen, 2021). For example, Petropoulos and Siemsen (2021) develop and test a new statistical model selection process based on insights from human judgment that provides meaningful improvements in forecast performance.

It seems that input to model is still under-tested, under-developed, and displays mixed empirical results. For example, Collopy and Armstrong (1992) demonstrate that RBF is not helpful during times of zero trend, little uncertainty, and little domain expertise. Selection of variables for the model has also been shown to rely on forecaster experience (Petropoulos et al., 2018). Hence, a primary difficulty with input to model is a supposition (and requirement) that the human forecaster has expertise and high technical knowledge (Sanders & Ritzman, 2004).

2.3. Theory

Having introduced the theoretical underpinnings established by prior literature, we now test the various methods for forecast integration. Although methods of integration previously have been empirically tested in the literature, few have focused on input to model (IM), and, to the best our knowledge, none have offered a complete comparison of all methods as we do (Arvan et al., 2019; Perera et al., 2019). Additionally, the search continues for the best process of integration to capture the strengths of human judgment and models.

2.3.1 Hypothesis 1

Judgmental adjustments are the most common method of integration (Fildes et al., 2009; Franses & Legerstee, 2010; Siemsen & Aloysius, 2020). Extant literature provides experimental evidence that humans often make unnecessary judgmental adjustments to forecasts, even when there is not enough evidence to make an informed decision (Lawrence et al., 2006; Fildes et al.,

2009). Thus, judgmental adjustments are frequently biased due to heuristics (De Baets & Harvey, 2018), judgment biases (Kremer et al., 2011), risk aversion (Kahneman & Tversky, 1979), bracing (Tokar et al., 2014), decision speed (Moritz et al., 2014), and other individual characteristics (Eroglu & Croxton, 2010). Quantitative correction allows for adjustment to combat systematic bias and is often paired with judgmental adjustment (Petropoulos et. al, 2016). However, Sanders and Ritzman (2004) suggest quantitative correction can be problematic when used during high variability times and for special events. Additionally, Green and Armstrong (2013) offer precise language regarding factors to avoid when integrating forecasts (e.g., overly complex methods, methods that do not use domain knowledge, and unstructured revisions) which are often present in a quantitative correction of judgmental forecasts (Sanders & Ritzman, 2004).

Prior pairwise comparisons among the methods of integration have mixed results (Arvan et al., 2019). For example, combinations of forecasts have resulted in greater forecast accuracy than judgmental adjustments (Franses & Legerstee, 2010; Petropoulos et al., 2018). However, judgmental adjustments have been shown to outperform combination when adjustments are made by experts (Alvarado-Valencia et al., 2017). Blattberg and Hoch (1990) identify the limitations of their method of combination by stating, "Until more is known about how to build better models, the [equal-weighted average] decision heuristic is a nonoptimal but pragmatic solution" (p. 898). Surprisingly, combination has remained a common method of integration despite research which offers insight into how to build and select better models (Petropoulos et al., 2018). One approach to build a superior model is through human input (Baecke et al., 2017; Bunn & Wright, 1990).

Unfortunately, the most promising method of integration (Sanders & Ritzman, 2004), IM, has the least empirical testing (Arvan et al., 2019). In one of the few empirical studies, Nakano and Oji (2010) offer evidence through a case study of improving forecast accuracy through the use of IM (i.e., input to model). Non-empirically tested articles (e.g., Arvan et al., 2019; Perera, et al., 2019; Sanders & Ritzman, 2004) claim that judgment should be most effective when used as an input to a model, thus allowing the model to perform the integration. We expect IM will perform better than the other methods of integration. Formally:

Hypothesis H_1. *Input to model (IM) results in improved forecast accuracy when compared to the previously discussed three other methods of integration (i.e., judgmental adjustment, quantitative correction, and combination).*

2.3.2 Hypothesis 2

To uncover the best process of integration, we return to the need for building better models (Blattberg & Hoch, 1990). One computer-centered technique that may allow for better model building is machine learning (ML). ML refers to models which learn from data without being explicitly programmed (Samuel, 1959). In other words, ML is, "the machine's ability to keep improving its performance without humans having to explain exactly how to accomplish all the tasks it's given" (Brynjolfsson & Mcafee, 2017: 2). Although ML has received considerable attention in practitioner journals (Brynjolfsson & Mcafee, 2017) along with receiving calls for research in academic journals (e.g., Bertsimas & Kallus, 2019; Feng & Shanthikumar, 2018; Pagell, Flynn, & Fugate, 2019), ML has rarely been investigated in the demand planning literature (Carbonneau, Laframboise, & Vahidov, 2008; Fildes et al., 2009; Makridakas, Spilotis, & Assimakopoulos, 2018a; Shahrabi, Mousavi, & Heydar, 2009) and continues to have relatively

little empirical testing beyond simulation (Feng et al., 2018; Makridrakas & Hibon, 2000; Makridakas Spilotis, & Assimakopoulos, 2018b; Perera et al., 2019).

ML is a broad umbrella term that can refer to many types of models, algorithms, and processes (Michie, 1968), but it typically follows seven steps (Guo, 2017). The first two steps involve gathering and preparing the data. The third step is choosing a model. Although many complex ML techniques exist (see Makridakas et al., 2018a), supervised ML is a type of practical application on behavioral data (Brynjolfsson & Mcafee, 2017). One kind of supervised ML is linear regression (Bishop, 2006). Within the context of demand planning, Carbonneau et al., (2008) provide evidence of no significant difference between complex ML (e.g., neural networks, recurrent neural networks, support vector machines) and multivariate linear regression ML. Additionally, the findings from forecasting competition M4 indicate that simpler methods are more accurate than the complex ML models such as neural networks and support vector regression (Makridakas et al., 2018b). Finally, practitioners may be more amenable to incorporating simpler models that they understand, such as regression. As such, and appealing to Occam's Razor (e.g., Myung & Pitt, 1997), the model we use in this research is multivariate linear regression (Bishop, 2006). The fourth step in ML is training the model. This training is often the primary association with ML (Guo, 2017). During the training step, the model conducts iterations of mapping a set of inputs to a set of outputs (Brynjolfsson & Mcafee, 2017). Once the training is complete, the fifth step involves evaluation of the model, preferably with data that was not used for training, to test accuracy. Following the evaluation, the model can be tuned (sixth step), and lastly, used to predict (seventh step).

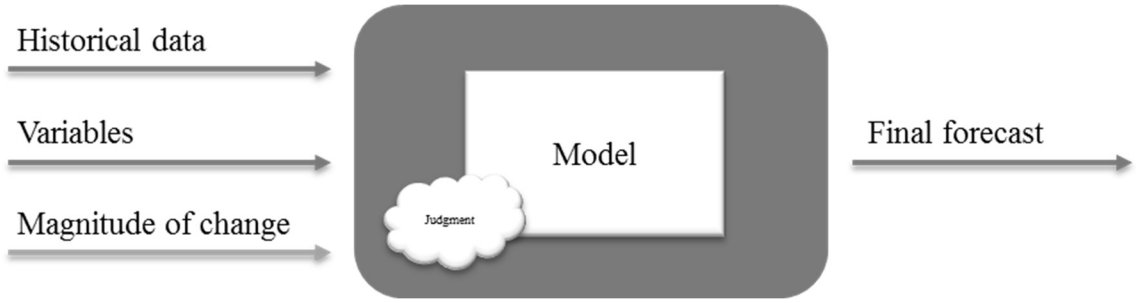


Figure 2.5. Interactive Machine Learning.

One type of ML which offers a more "hands on" approach is interactive machine learning (IML) (Amershi, Cakmak, Knox, & Kulesza, 2014). IML refers to rapid iterations of humans providing a model with corrective input (Ware, Frank, Holmes, Hall, & Witten, 2001; Fails & Olsen, 2003). The iterations between human and model in IML allow for synergy between the strengths of human judgment and the ML model. The trained model observes the data and identifies systematic trends, along with providing the optimal weights for the evidence. The human uses judgment to identify a quantity related to expected special events and introduces it as a predictive variable in the model (Figure 2.5). Research on IML has been limited; however Holzinger et al. (2019) offer evidence of a positive interaction between humans and ML. To the best of our knowledge, we are the first to experimentally study the behavioral issues involved with interactive machine learning used in demand planning. We expect IML to lead to more accurate forecasts through the iterative integration and interaction of human judgment and models. Specifically:

Hypothesis H_2. *Interactive machine learning (IML) results in improved forecast accuracy when compared to the previous four methods of integration (i.e., judgmental adjustment, quantitative correction, combination, and input to model).*

2.3.3 Hypothesis 3

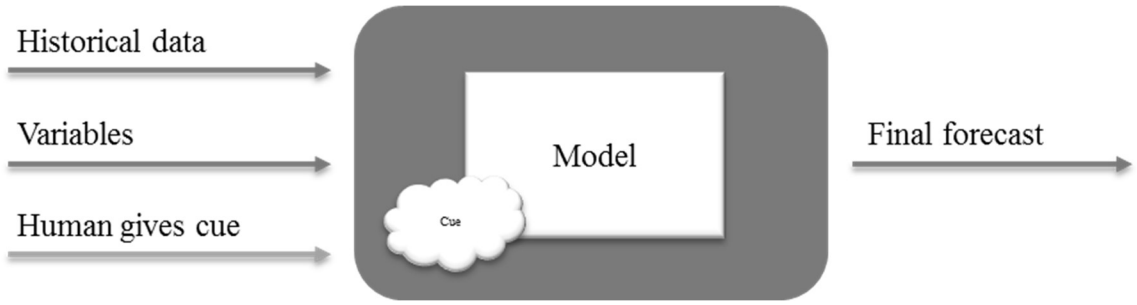


Figure 2.6. Human-Guided Machine Learning.

For our last hypothesis, we narrow the scope of IML even further and introduce the term *human-guided machine learning* (HGML). HGML relies on human forecasters to teach the model in real-time through providing information. HGML differentiates itself from IML by requiring the human to share the knowledge of contextual information through a cue (e.g., indication to the model that it is a period with a special event), rather than a quantity (Figure 2.6). According to Fildes, Goodwin, & Önkal (2019), humans often misinterpret contextual information, which leads to inaccurate estimations. As such, we propose HGML as a method of forecasting that does not rely on human judgment to quantify the magnitude of a special event. Instead, HGML relies on a model calculation of the magnitude of special events based on similar periods in the historical data. HGML is an emerging concept which has been rarely investigated (Amershi et al., 2011; Gil et al., 2019). As such, we provide the first empirical behavioral study on HGML in demand planning.

Additionally, we are the first study to empirically compare IML and HGML. The key difference between HGML and IML is that in HGML, human judgment simply recognizes special events, then hands the baton to the model which quantifies the estimated lift through specifying a variable. The HGML arrangement allows for both the human and model to use their respective strengths to complement one another (Blattberg & Hoch, 1990). Thus, we hypothesize

that HGML will be the optimal process to maximize the value of contextual information and to minimize biases. Formally:

Hypothesis H_3. *Human-guided machine learning (HGML) results in improved forecast accuracy when compared with the previous five other methods of integration (i.e., judgmental adjustment, quantitative correction, combination, input to model, and interactive machine learning).*

2.4 Design and Implementation

As the goal of this paper is to better understand an integration of human judgment and model forecasts, incentivized experiments offer an ideal platform due to the ability to control the environment. Mir, Aloysius and Eckerd (2017) provide a summary of advantages and disadvantages of using incentivized experiments. Specific to our study, the advantages of using a behavioral experiment include: the ability to control the timing and means of communication of the cue regarding an external event, consistency in the interface used for the task, and control of the integration method. We conduct our pilot tests using Amazon Mechanical Turk (MTurk) Prime members to capture behavioral regularities which can then be extended to a large population (Smith, 1982). After making adjustments from our pilot tests, we use a large sample of college students at a US university to conduct the actual experiment.

2.4.1 Experimental Design

We use a 2 (i.e., no contextual information, contextual information) X 6 (i.e., judgment forecast, judgmental adjustment, quantitative correction, input to model, interactive machine learning, human-guided machine learning) experiment with human subjects. As we subsequently explain, we also design a post-data collection manipulation of the data, resulting in a 2 (no contextual information, yes contextual information) X 8 (judgment forecast, judgmental

adjustment, quantitative correction, combined, input to model, interactive machine learning, human-guided machine learning, model forecast) design.

We define judgmental forecast (JF) as the participant's prediction using only historical data. The judgmental adjustment (JA) participant is shown the historical data as well as the model's recommendation for the next period and offers an adjustment. All JA and JF conditions are based completely on human judgment, so we do not include equations. The quantitative correction (QC) condition observes systematic bias, $Bias = Actual\ demand - Forecasted\ demand$, from the participant's forecast, $JudgmentForecast$, and adjusts accordingly (Equation 1).

(1) $QC_{it} = JudgmentForecast_{it} \pm \sum_{i=2}^{n} \frac{Bias_{it-1}}{n-1}$

where: i = participant,

t = period, and

n = number of periods.

The input to model (IM) prompts the participant to define parameter $Magnitude$, meaning the estimated magnitude of the change to demand, to input into the model (Equation 2). Interactive machine learning (IML) differentiates from IM in that magnitude of the change to demand, $Magnitude$, is included in the model as a predictive variable (Equation 3). Human-guided machine learning (HGML) uses a cue, Cue, to indicate whether the following period includes a special event (Equation 4):

(2) $IM_{it} = \beta_0 + \beta_1 Period_{it} + Magnitude_{it}$,

(3) $IML_{it} = \beta_0 + \beta_1 Period_{it} + \beta_2 Magnitude_{it}$, and

(4) $HGML_{it} = \beta_0 + \beta_1 Period_{it} + \beta_2 Cue_{it}$,

The post-data collection computations for combined forecast (CF) and model forecast (MF) are represented in Equations 5-6 respectively.

(5) $CF_{it} = \frac{(\beta_0+\beta_1 Period_{it})+(JudgmentForecast_{it})}{2}$,

(6) $MF_{it} = \beta_0 + \beta_1 Period_{it}$.

Lastly, we computed optimal interactive machine learning (O-IML) and optimal human-guided machine learning (O-HGML). O-IML is calculated as $Magnitude$ equal to the mean of the shock distribution ($\mu = 41$) for the shocked periods and zero for the non-shocked periods. Similarly, O-HGML is calculated as Cue equal to one for the shocked periods and zero for the non-shocked periods.

All of the conditions with contextual information also include a cue prior to each demand shock, showing a randomly-generated probability distribution of the shock ($\mu = 41$, SD = 4.32). The no contextual information conditions are given no cues throughout the task, but the demand shocks still occur. For detailed information on the forecasting task including the data generation process, contextual information, and shock distribution, please see Appendix A1.

The dependent variable of interest is mean absolute error (MAE). MAE is an appropriate measure since we are comparing forecast accuracy across the same scale of data (Hyndman & Koehler, 2006). Thus, the equation used to calculate MAE is:

(7) MAE = $\frac{\sum_{i=1}^{n} |Actual\ demand_{it} - Forecast_{it}|}{n}$.

2.4.2. Program and Task

We code the application to test our design using oTree, an open-source Python framework used to create interactive behavioral economic experiments (Chen, Schonger, & Wickens, 2016). When participants initiate the program, they are randomly assigned to a method of integration. The demand planning task begins with two training periods: a period with no

shocks to demand and a period with a shock and contextual information. Following the training periods, the participant is told the actual task begins (as per Appendix A). The participant begins the actual task with 10 periods of historical data, so the participant is familiarized with the trend line pre-forecast. The participant follows the program to forecast for 30 periods (they observe 40 periods of total data including the 10 periods of historical data). Beginning in period 11-40, a series of shocks from the randomly-generated probability distribution results in upward shifts to demand. Throughout the task participants are shown a graph of actual demand as well as forecasts. Appendix A1 also includes details on the shocks and screenshots from the program.

2.4.3. Pilot Tests

Prior to collecting our data, we conducted two pilot tests on Amazon MTurk. The first pilot test was unincentivized and illuminated our application contained some potentially unclear instructions for input to model building (IM). During the first pilot, IM asked for two inputs, a value of expected change in the quantity ordered and indication of whether the quantity should be increased or decreased. As a result of the first pilot test, we reworded the instructions to explain the two inputs more clearly. We conducted a second MTurk pilot test, this time incentivized. The input to model continued to be confusing to some participants despite the clarifications, so for the actual data collection we changed the input to model to only require one input (the value of expected change).

2.4.4. Data Collection and Sample

Following the pilot tests, we conducted our experiment using undergraduate students at a large American private university. We collected responses from 459 subjects. We eliminated 12 participants who clearly did not understand the task. The mean of the distribution of shocks that participants were given was 41, and these 12 subjects had a MAE > 100. We used a lottery

performance-based incentive (Wakker, 2007). The performance-based payment was included to increase validity following the logic from the induced-value theory in behavioral economics (Smith, 1976). The premise of induced-value theory is that, "paying subjects based on their performance in the game causes them to wish to perform better because better performance results in making more money" (Katok, 2011: 12). By using this means, experimental validity is improved (Katok, 2011). Upon beginning the task, participants were told five completed tasks would be selected at random to win the amount they earned. All participants were shown the amount they could possibly win. The payoff is calculated by:

(8) $\text{Total Payout} = \$1 + [\$9 - (0.20 * |MAE|)]$

The average bonus for our collected sample is $4.44 out of the possible $9. Following the post-data collection calculation of combined forecast, optimal interactive machine learning, optimal human-guided machine learning, and model forecast conditions (426 calculated responses), result in a total of 873 observations for 30 periods. The first 10 forecast periods allow for learning effects and training the model. To conduct our analysis, we use only the last 20 shocked periods of the demand planning task (periods 21-40 with shocks) aggregated to the participant-level to achieve independence of units. Thus, the number of observations in the sample is 873.

2.5. Results

Descriptive statistics are shown in Table 2.2. Following the Shapiro-Wilk Test for normality, we find evidence of non-normality ($W= 0.013$, $Z = 15.784$, $p < .001$). Therefore, to test our hypotheses, we conduct a Kruskal-Wallis H test (Conover, 1999). The Kruskal-Wallis H-test is a non-parametric substitute for the standard F-test used when conditions of normality are violated. Since the distribution of our data is non-normal, Kruskal-Wallis is an appropriate

method to test the differences between methods of integration. Thus, the null and alternative hypotheses for the Kruskal-Wallis H-test in the context of our analysis are:

H_0: The distribution of MAE is the same across categories of method of integration.

H_a: At least one of the distributions of MAE differs across categories of method of integration.

Testing the data of Table 2.2, the results of the Kruskal-Wallis H-test are significant (H = 255.14, df= 14, p < 0.001) indicating that at least one of the distributions of MAE differs across methods of integration.

Table 2.2. Descriptive Statistics of Mean Absolute Error (MAE).

	Mean Absolute Error					
	No contextual information			Contextual information		
Method of Forecasting	Mean	SE	n	Mean	SE	n
Judgment forecast (JF)	22.75	1.89	55	15.21	1.20	47
Judgmental adjustment (JA)	20.34	1.71	54	13.70	0.76	45
Quantitative correction (QC)	21.17	1.26	61	20.45	2.09	67
Combined forecasts (CF)	19.35	0.96	55	14.39	0.67	48
Input to model (IM)	23.56	9.17	61	18.45	1.83	57
Interactive machine learning (IML)	13.44	0.60	69	10.72	0.59	61
Human guided machine learning (HGML)	11.90	0.20	62	10.82	0.52	69
Model forecast (MF)	11.90	0.20	62			

Table 2.3 summarizes the pairwise and comparisons between methods of integration. $\boldsymbol{H_1}$ (i.e., IM results in improved forecast accuracy when compared to the three other methods of integration, the first three test rows of Table 2.3) is not supported since the differences between IM and the other three methods of integration are not significant (*IM-JA* = 4.75, p = 0.86; *IM-QC* = -2.00, p = 0.58; *IM-CF* = 4.06, p = 0.88). $\boldsymbol{H_2}$ (i.e., IML results in improved forecast accuracy when compared to the four prior methods of integration) is supported. The differences between IML and the prior four methods of integration are negative and significant (*IML-JA* =-2.98, p <

0.001; *IML-QC* = -9.73, p > 0.001*; IML-CF* = -3.67, p < 0.001; *IML-IM* = -7.73, p < 0.001). Lastly, $\boldsymbol{H_3}$ is receives mixed results. The hypothesis holds for the comparisons between HGML and the first four methods of integration, but not from IML (*HGML-JA* = -2.87, p < 0.001; *HGML-QC* = -9.63, p < 0.001; *HGML-CF* = -3.57, p < 0.001; *HGML-IM* = -7.63, p < 0.001). Thus, both IML and HGML are superior to the other methods, but they are not statistically different between themselves.

Table 2.3. Pairwise Comparisons of MAE Between Methods of Integration with Contextual Information.

		Mean Absolute Error		
Sample 1-Sample 2	Difference	Std. Error	Sig.	Hypotheses
IM-JA	4.75	26.28	0.86	H1
IM-QC	-2.00	23.75	0.58	H1
IM-CF	4.06	25.82	0.88	H1
IML-JA	-2.98	25.90	0.00	H2
IML-QC	-9.73	23.32	0.00	H2
IML-CF	-3.67	25.43	0.00	H2
IML-IM	-7.73	24.28	0.00	H2
HGML-JA	-2.87	25.25	0.00	H3
HGML-QC	-9.63	22.60	0.00	H3
HGML-CF	-3.57	24.77	0.00	H3
HGML-IM	-7.63	23.59	0.00	H3
HGML-IML	0.10	23.16	0.66	H3

.Notes. Abbreviations of methods of integration: judgmental adjustment (JA), quantitative correction (QC), combined forecast (CF), input to model (IM), interactive machine learning (IML), human-guided machine learning (HGML). Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Next, to better understand human's use of contextual information, we next compare the no contextual information conditions (NCI) and contextual information (CI) conditions across methods of forecasting. Table 2.4 summarizes the pairwise comparisons between NCI and CI methods of integration.

Table 2.4. Pairwise Comparisons of MAE by Method of Integration (Contextual Information – No Contextual Information).

	Mean Absolute Error			
Sample 1-Sample 2	Difference	Std. Error	Sig.	Percent Change
Judgment forecast (JF)	-7.54	50.09	0.00	-50%
Judgmental adjustment (JA)	-6.64	50.90	0.00	-49%
Quantitative correction (QC)	-0.72	44.63	0.00	-4%
Combined forecast (CF)	-4.96	50.90	0.00	-34%
Input to model (IM)	-5.11	46.45	0.00	-28%
Interactive machine learning (IML)	-2.72	44.32	0.02	-25%
Human guided machine learning (HGML)	-1.08	44.13	0.69	-0.10

Notes. Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

We conduct the same procedure used for our main analysis above using Mean Error as the dependent variable to provide greater insight into the biases associated with each method. Mean Error is calculated as:

(9) $\text{ME} = \frac{\sum_{i=1}^{n} Actual\ demand_{it} - Forecast_{it}|}{n}$.

Table 2.5 summarizes the descriptive statistics. The ME reveals that in all methods of forecasting except input to model the forecasts are negative, meaning the forecasts are too low.

Table 2.5. Descriptive Statistics of Mean Error.

	Mean Error			
	No contextual information		Contextual information	
Methods of Forecasting	Mean	SE	Mean	SE
Judgment forecast (JF)	-22.52	2.11	-11.52	1.55
Judgmental adjustment (JA)	-21.32	1.92	-5.56	1.85
Quantitative correction (QC)	-20.60	1.57	-13.28	2.52
Combined forecasts (CF)	-20.62	1.06	-15.07	0.77
Input to model (IM)	14.40	10.45	11.77	2.68
Interactive machine learning (IML)	-13.73	0.49	-10.96	0.79
Human guided machine learning (HGML)	-12.95	0.22	-11.34	0.65
Model forecast (MF)	-12.95	0.22		

Using a Kruskal-Wallis H-Test, we compare the ME between methods of integration with contextual information (as conducted in our main analysis). The results of the Kruskal-Wallis H-Test are significant ($H = 255.14$, df= 14, $p < 0.001$) indicating that at least one of the distributions of ME differs across methods of integration.

Table 2.6. Pairwise Comparisons of ME Between Methods of Integration with Contextual Information.

		Mean Error	
Sample 1-Sample 2	Difference	Std. Error	Sig.
IM-JA	17.33	26.28	0.00
IM-QC	25.05	23.75	0.00
IM-CF	26.84	27.34	0.00
IML-JA	-5.40	25.90	0.14
IML-QC	2.32	23.32	0.80
IML-CF	4.11	25.43	0.02
IML-IM	-22.73	24.28	0.00
HGML-JA	-5.78	25.25	0.04
HGML-QC	1.94	22.60	0.72
HGML-CF	3.73	24.77	0.05
HGML-IM	-23.11	23.59	0.00
HGML-IML	-0.38	23.16	0.54

Notes. Abbreviations of methods of integration: judgmental adjustment (JA), quantitative correction (QC), combined forecast (CF), input to model (IM), interactive machine learning (IML), human-guided machine learning (HGML). Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

The pairwise comparisons (Table 2.6) reveal that the difference between IM and the other existing methods of integration is actually much larger than the MAE showed. In fact, IM has a significantly higher ME than the other methods of integration (*IM-JA* = 17.33, $p < 0.001$; *IM-QC* = -25.05, $p < 0.001$; *IM-CF* = 26.84, $p < 0.001$).

Lastly, we conduct the comparison of O-IML and O-HGML with the six methods of integration included in our main analysis. The descriptive statistics of each are respectively O-IML: Mean = 3.32, SE = 0.07, n = 71; O-HGML: Mean = 3.34, SE = 0.07, n = 65. The results of

a Kruskal-Wallis H-Test analyzing the methods of integration with contextual information only are significant ($H = 295.67$, df= 9, $p < 0.001$).

2.6. Discussion and Conclusion

The purpose of our research is to identify the most accurate method of integrating forecasts by focusing on the process of integrating the strengths of human judgment and models. As such, we propose and test three hypotheses. Hypothesis $\boldsymbol{H_1}$ focuses on the four extant methods of integration found in the literature which hypothesize that input to model (IM) would be the most accurate. Surprisingly, contrary to prior literature, we find IM to be less accurate when compared to judgmental adjustment (JA) and combined forecast (CF). Most of the participants randomly-assigned to IM provide large, estimated changes to demand even when they had no contextual information for such predictions (M = 10, SD = 17.62). The additional human-input estimated change to demand introduces more error which the model cannot correct for, thus resulting in lower forecast accuracy.

Hypothesis $\boldsymbol{H_2}$ investigates interactive machine learning (IML) which is like input to model (IM) in that the human estimates a change to demand. However, IML differs from IM in that the estimated change to demand is weighted by the model. Humans still provide large, estimated changes to demand, but the model removes much of the introduced error through weighting the variable lower. Therefore, IML results in improved forecast accuracy when compared to IM. The results from the first two hypotheses suggest that humans are not accurate when quantifying an expected change to demand. We argue the inaccuracy could partially be due to judgment biases which remains unchecked prior to entering the model (Bendoly et al., 2010; Fildes et al., 2009; Kremer et al., 2011).

Lastly, Hypothesis $\boldsymbol{H_3}$ tests human-guided machine learning (HGML) compared to the other methods of integration. HGML removes the function of the human to quantify the change to demand. As predicted, HGML reduces error and results in improved accuracy. However, the comparison between IML and HGML reveal that although the methods are superior to the other methods, they are not different from each other. Thus, we find that the methods of forecasting that use machine learning, and that employ the strengths of human judgment and model, forecast in the most precise manner.

Our study provides implications for research and practice. First, our research is, to the best of our knowledge, the first to test the ability of humans to leverage contextual information when comparing across methods of integration. Although humans do not use contextual information optimally, we can see there is an improvement in the methods of forecasting that have contextual information over those that do not. Second, we provide further insight into human interaction with model prediction via machine learning. We build on past literature regarding information processing in demand planning (e.g., Moritz et al., 2014) and find that humans struggle to judge the magnitude of demand shocks. Third, we develop a new machine learning method of integration within the context of demand planning: human-guided machine learning (HGML). We provide empirical support for HGML by showing an 72% improvement of forecast accuracy compared to existing methods of forecasting when used optimally.

Although our research offers new insights into the process of utilizing the strengths of both humans and models in demand planning, our current study relies on human judgment rather than expertise. Since demand planning experience has some evidence of improving accuracy (Önkal et al., 2003), we encourage and plan future empirical testing of input to model (IM), interactive machine learning (IML), and human-guided machine learning (HGML) with experts

such as managers and demand planners. Another avenue for future research is to investigate the combination of multiple experts' human judgment rather than relying on one individual (Larrick & Soll, 2006).

In conclusion, our results suggest contextual information leveraged by humans can enhance forecasting accuracy. However, what matters most is the method used to integrate the contextual information with the model. The results of our testing indicate that human-guided machine learning (HGML) and interactive machine learning (IML) are the most effective methods of integrating model and human-judgment forecasts. We encourage more investigation into the benefits of machine learning both for theory and practice. Managers seeking to improve forecast accuracy may find that asking employees to share the contextual information with the model, rather than directly adjusting the forecast, can help eliminate bias. Another strength of human-guided machine learning is the ability to secure knowledge that may previously have resided only in the minds of individuals. Human-guided machine learning enables the model to learn continually learn from forecaster experience. Additionally, when designing forecasting support systems (FSS), our results show the importance of using the correct method that employs the strengths of users and algorithms. Thus, we encourage practitioners to consider implementing human-guided machine learning as it has potential to improve predictive performance.

2.7. References

Adya, M., & Lusk, E. J. 2016. Development and validation of a rule-based time series complexity scoring technique to support design of adaptive forecasting DSS. *Decision Support Systems*, 83: 70-82.

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. 2017. Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting,* 33(1): 298-313.

Amershi, S., Lee, B., Kapoor, A., Mahajan, R., & Christian, B. 2011. Human-guided machine learning for fast and accurate network alarm triage. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine,* 35(4): 105-120.

Arvan, M., Fahimnia, G., Reisi, M. & Siemsen, E. 2019. Integrating human judgment into quantitative forecasting methods: A review. *Omega,* 86: 237-252.

Baecke, P., De Baets, S., & Vanderheyden, K. 2017. Investigating the added value of integrating human judgment into statistical demand forecasting systems. *International Journal of Production Economics*, 191: 85-96.

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

Bertsimas, D. & Kallus, N. 2019. From predictive to prescriptive analytics. *Management Science,* Articles in Advance.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blattberg, R. C. & Hoch, S. J. 1990. Database models and managerial intuition: 50% model+ 50% manager. *Management Science,* 36(8): 887-899.

Bolger, F. & Harvey, N. 1993. Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology Section A*, 46(4): 779-811.

Boulaksil, Y. & Franses, P. H. 2009. Experts' stated behavior. *Interfaces*, 39(2): 168-171.

Brynjolfsson, E. & Mcafee, A. 2017. The business of artificial intelligence. *Harvard Business Review*.

Bunn, D. & Wright, G. 1991. Interaction of judgmental and statistical forecasting methods: issues & analysis. *Management Science,* 37(5): 501-518.

Carbonneau, R., Laframboise, K., & Vahidov, R. 2008. Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3): 1140-1154.

Carter, C. R., Kaufmann, L., & Michel, A. 2007. Behavioral supply management: a taxonomy of judgment and decision-making biases. *International Journal of Physical Distribution and Logistics Management,* 37:631–69.

Chen, D. L., Schonger, M., & Wickens, C. 2016. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9: 88-97.

Collopy, F. & Armstrong, J. S. 1992. Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, 38(10): 1394-1414.

Conover, W.J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York, NY: Wiley Publishers

Croson, R. & Donohue, K. 2006. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science*, 52(3): 323-336.

De Baets, S. & Harvey, N. 2018. Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support. *International Journal of Forecasting*, 34(2): 163-180.

De Baets, S. & Harvey, N. 2020. Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research,* In press.

Donohue, K., Katok, E., & Leider, S. (Eds.). 2019. *The Handbook of Behavioral Operations*. John Wiley & Sons.

Eggleton, I. R. 1982. Intuitive time-series extrapolation. *Journal of Accounting Research*, 20(1): 68-102

Elgers, P. T., Lo, M. H., & Murray, D. 1995. Note on adjustments to analysts' earnings forecasts based upon systematic cross-sectional components of prior-period errors. *Management Science*, 41(8): 1392-1396.

Eroglu, C. & Croxton, K. L. 2010. Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26(1): 116-133.

Fails, J. A., & Olsen Jr, D. R. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*: 39-45.

Feng, Q., & Shanthikumar, J. G. 2018. How research in production and operations management may evolve in the era of big data. *Production and Operations Management*, 27(9): 1670-1684.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. 2016. Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & Service Operations Management*, 18(1): 69-88.

Fiebrink, R.; Cook, P. R.; and Trueman, D. 2011. Human Model Evaluation in Interactive Supervised Learning. In *Proceedings of the Conference on Human Factors in Computing Systems*:147–156. New York: Association for Computing Machinery.

Fildes, R. 1991. Efficient use of information in the formation of subjective industry forecasts. *Journal of Forecasting,*10: 597-617.

Fildes, R., P. Goodwin, M. Lawrence, & K. Nikolopoulos. 2009. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting,* 25: 3–23.

Fildes, R., & Goodwin, P. 2013. Forecasting support systems: What we know, what we need to know. *International Journal of Forecasting*, 2(29): 290-294.

Fildes, R., Goodwin, P., & Önkal, D. 2019. Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, 35(1): 144-156.

Franses, P. H., & Legerstee, R. 2010. Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting,* 29(3): 331-340.

Franses, P. H., & Legerstee, R. 2013. Do statistical forecasting models for SKU-level data benefit from including past expert knowledge? *International Journal of Forecasting,* 29(1): 80-87.

Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S., Yang, Q., & Jahanshad, N. 2019. Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces:* 614-624.

Goodwin, P. 2002. Integrating management judgment and statistical methods to improve short-term forecasts. *Omega,* 30(2): 127-135.

Goodwin, P., & Wright, G. 1994. Heuristics, biases and improvement strategies in judgmental time series forecasting. *Omega*, 22(6): 553-568.

Goodwin, P., Önkal, D., & Lawrence, M. 2011. Improving the role of judgment in economic forecasting. *Oxford Handbook of Economic Forecasting.* Oxford University Press: 163-192.

Green, K. C. & Armstrong, J. S. 2013. Demand forecasting: Evidence-based methods. In Thomas, C., & Shughart, W. (Eds.) *The Oxford handbook of evidence-based management*. Oxford University Press.

Guo, Y. 2017. The seven steps of machine learning. https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e. Accessed January 6, 2020.

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

Harvey, N. 2007. Use of heuristics: Insights from forecasting research. *Thinking & Reasoning*, 13(1): 5-24.

Holzinger, A. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop?. *Brain Informatics*, 3(2): 119-131.

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C. M., & Palade, V. 2019. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49(7): 2401-2414.

Hyndman, R. J., & Koehler, A. B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4): 679-688.

Jordan, M. I., & Mitchell, T. M. 2015. Machine Learning: Trends, Perspectives, and Prospects. *Science*, *349*(6245): 255-260.

Kahneman, D., & Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2): 263-292.

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

Kasparov, G. 2017. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. PublicAffairs.

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

Kremer, M., Siemsen, E., & Thomas, D. J. 2015. The sum and its parts: Judgmental hierarchical forecasting. *Management Science*, 62(9): 2745-2764.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. 2006. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3): 493-518.

Lee, Y. S. & Siemsen, E. 2016. Task decomposition and newsvendor decision making. *Management Science,* 63(10): 3226-3245.

Lee, Y. S., Seo, Y. W., & Siemsen, E. 2018. Running behavioral operations experiments using Amazon's Mechanical Turk. *Production and Operations Management,* 27(5): 973-989.

Levitt, S. D., & List, J. A. 2009. Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review*, 53(1): 1-18.

Lim, J. S. & O'Connor, M. 1995. Judgmental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making,* 8(3): 149-168.

Lobo, G. J. & Nair, R. D. 1990. Combining judgmental and statistical forecasts: an application to earnings forecasts. *Decision Sciences,* 21(2): 446-460.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. 2018. On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, 64: 19-40.

Larrick, R. P., & Soll, J. B. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, 52(1): 111-127.

Makridakis, S. & Winkler, R. L. 1983. Averages of forecasts: Some empirical results. *Management Science,* 29(9): 987-996.

Makridakis, S., & Hibon, M. 2000. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4): 451-476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. 2018a. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS One*, 13(3): 1-26.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. 2018b. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4): 802-808.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54-74.

Michie, D. 1968. “Memo” functions and machine learning. *Nature,* 218(5136): 19-22.

Mitchell, T. (1997). Introduction to machine learning. *Machine Learning*, 7: 2-5.

Mir, S., Aloysius, J. A., & Eckerd, S. 2017. Understanding supplier switching behavior: The role of psychological contracts in a competitive setting. *Journal of Supply Chain Management*, 53(3): 3-18.

Moritz, B., Siemsen, E., & Kremer, M. 2014. Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management*, 23(7): 1146-1160.

Moravec, H. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

Myung, I.J. and Pitt, M.A., 1997. Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1): 79-95.

Nakano, M. and Oji, N. 2010. The transition from a judgmental to an integrative method in demand forecasting: A case study of a Japanese company. *International Journal of Operations and Production Management,* 32(4): 386-397

Narayanan, A., & Moritz, B. B. 2015. Decision making and cognition in multi-echelon supply chains: an experimental study. *Production and Operations Management*, 24(8): 1216-1234.

Önkal, D., Yates, J. F., Simga-Mugan, C., & Öztin, Ş. 2003. Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes*, 91(2): 169-185.

Pagell M., Flynn, B., & Fugate, B. 2019. Call for papers for the 2020 emerging discourse incubator emerging approaches for developing SCM theory. *Journal of Supply Chain Management*, 55(4): 129-131.

Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. 2019. The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research*, 274(2): 574-600.

Petropoulos, F., Fildes, R., & Goodwin, P. 2016. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, 249(3): 842-852.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. 2018. Judgmental selection of forecasting models. *Journal of Operations Management,* 60: 34-46.

Petropoulos, F. & Siemsen, E. 2021. Forecast selection and representativeness. Working paper.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3): 210-229.

Sanders, N. R., & Ritzman, L. P. 1995. Bringing judgment into combination forecasts. *Journal of Operations Management*, 13(4): 311-321.

Sanders, N. R., & Ritzman, L. P. 2004. Integrating judgmental and quantitative forecasts: methodologies for pooling marketing and operations information. *International Journal of Operations and Production Management,* 24(5): 514-529.

Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. 2015. Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36*:* 33-45.

Siemsen, E. & Aloysius, J. 2020. Supply chains analytics and the evolving work of supply chain managers. Research report for *Association of Supply Chain Management.*

Shahrabi, J., Mousavi, S. S., & Heydar, M. 2009. Supply chain demand forecasting: a comparison of machine learning techniques and traditional methods. *Journal of Applied Sciences*, 9(3): 521-527.

Sheridan, T. B. 1995. Human centered automation: Oxymoron or common sense? In *IEEE International Conference on Systems, Man and Cybernetics*. Intelligent Systems for the 21st Century, 1: 823-828

Smith, V. L. 1982. Microeconomic systems as an experimental science. *The American Economic Review*, 72(5): 923-955.

Theil, H. 1971. *Applied Economic Forecasting*. North-Holland Publishing Company, Amsterdam.

Tokar, T., Aloysius, J. A., Waller, M., & Williams, B. 2014. Bracing for demand shocks: An experimental investigation. *Journal of Operations Management,* 32(4): 205-216.

Tong, J., & Feiler, D. 2017. A behavioral model of forecasting: Naive statistics on mental samples. *Management Science*, 63(11): 3609-3627.

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

Wakker, P. P. 2007. Message to referees who want to embark on yet another discussion of the random-lottery incentive system for individual choice. URL: http://people.few.eur. nl/wakker/miscella/debates/randomlinc. htm, Access Date, 3 June 2020.

Waller, M. A., & Fawcett, S. E. 2013. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2): 77-84.

Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, I. H. 2001. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3): 281-292.

Wilson, J., & Daugherty, P. R. 2018. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review,* 96(4): 115-123.

2.8. Appendix A1. Information about Behavioral Experiment and Screenshots

2.8.1. Data Generation Process for Demand Planning Task

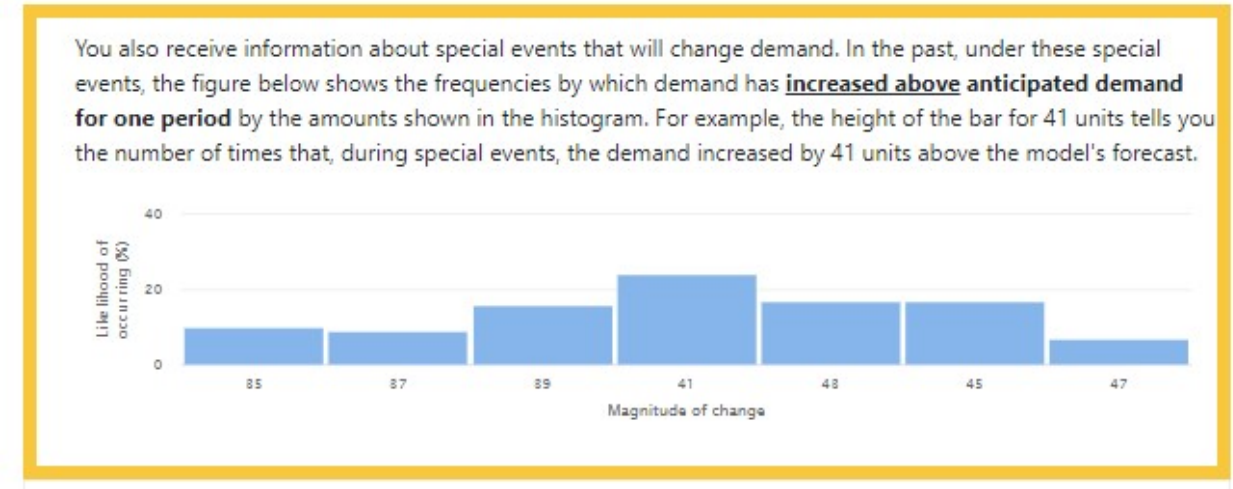
The historical data follow the trend used for actual demand:

$$ActualDemand_{it} = \beta_0 + \beta_1 Period_{it} + \varepsilon_{it}$$

where: i = observation
t = period
β_0 = 100
β_1 = 5
ε_{it} = X ~ U(-3, 3).

The forecasting task presents a participant with instructions to insert a prediction for the next period into an entry field. The task begins showing the historical 10 periods which follow the trendline. The task includes randomly generated shocks equally-likely ranging from 35, 37, 39, 41, 43, 45, to 47 units. The demand shock occurs during periods 12, 15, 16, 18, 19, 21, 23, 26, 27, 29, 31, 33, 34, 36, 38.

2.8.2 Cue About Contextual Information



2.8.3. Screenshot of Training Period

Instructions

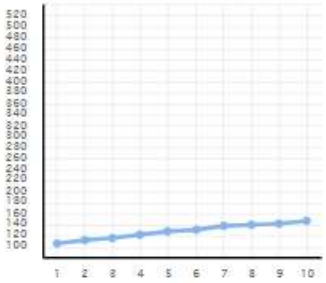
Prior to beginning the task, you will complete two training periods. Please take the training very seriously because **if you responses indicate you do not understand the task, you will not be allowed to continue in the study and will not be paid. Additionally, if you do not learn how to forecast, you will lose money from your potential bonus of $9.**

Imagine you are managing inventory for a particular item in a retail store. Your firm uses a model to make forecasts and **you may have access to the output**. You may also receive information about special events that will change demand from the anticipated level. The **information about special events that you may receive is not incorporated** into the firm's model.

Every selling period you have the task of predicting demand for the next selling period. The historical data from the ten previous selling periods is provided in the graph.

The model your firm uses forecasts demand will be **156** in Period 11.

Historical Data



What is the model forecast?

Example: You were told above the model forecasts 156, so enter 156

Please enter the **model forecast** for Period 11:

How much do you want to change the model forecast?

*Example: If you think the **change (c)** to demand will be 5 units different from the model forecast **(c = 5), you would input 5 and select the range 1 to 5**. Note: If you do not wish to change the model forecast, input 0.*

Please input **the change (c) from 156** for Period 11:

Please choose the range you expect **the change (c) from 156** for Period 11: ---------

What is your forecast?

*Example: If you think demand will be 170 **instead of 156, input 170**. Notice that your forecast will be the model forecast adjusted for the change (c) that you just input above.*

Please enter **your forecast** for Period 11:

Next

2.8.4. Screenshot of Instructions

Instructions

Now you begin the actual task. You can earn money depending partly on the accuracy of your forecasts and partly on chance. Accuracy is calculated as the absolute error, which is the absolute value of the difference between the forecast and the actual (i.e., Accuracy = | Forecast – Actual|). Your final payoff will be calculated based on the mean absolute error of your forecasts. If for example, your mean absolute error is 2 units, you will be penalized $0.40 from the maximum of $20.00 that you could have earned. **At the completion of this study, five participants will randomly be selected to win the amount earned!**

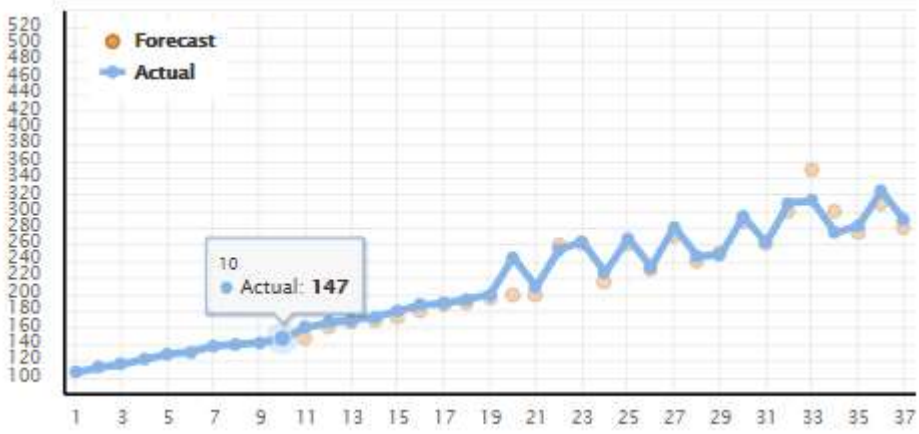
Please pay careful attention to the information that you are given and think carefully about your responses.

Imagine you are managing inventory for a particular item in a retail store. Your firm uses a model to make forecasts and **you may have access to the output**. You may also receive information about special events that will change demand from the anticipated level. The **information about special events that you may receive is not incorporated** into the firm's model.

Every selling period you have the task of predicting demand for the next selling period. The historical data from the ten previous selling periods is provided in the graph.

2.8.5. Screenshot of Trend Feedback

The orange points on the graph indicate the participant's input, and the blue line represents the actual demand.



III. ESSAY 2

Integrating Machine Learning and Human Judgment:

A Study on Demand Planning in the Field

ABSTRACT

While firms use automated machine learning algorithms in their demand planning processes, human judgment continues to feature in these processes. This research examines two methods of integrating machine learning and human judgment into demand planning. We implement a field study at a large, multinational firm testing interactive machine learning (IML) in which humans estimate adjustment quantities due to special events, and human-guided machine learning (HGML) in which humans merely input information about the special events. Analyzing the results of over three million datapoints across five product categories reveals that demand forecasts using IML and HGML are more accurate than the current demand planning process used in the firm. These findings suggest that using an appropriate process to integrate machine learning and human judgment—IML and HGML—provides a significant benefit to demand planning. Furthermore, the paper provides insights on how to implement IML and HGML in practice.

Keywords:

Machine Learning, Behavioral Operations, Human Judgment, Demand Planning

A clever process beat superior knowledge and superior technology. It didn't render knowledge and technology obsolete, of course, but it illustrated the power of efficiency and coordination to dramatically improve results. I represented my conclusion like this: weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process (Kasparov, 2017, from Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins, p. 246).

3.1 Introduction

Over the past several decades, practitioners and academics alike have recommended the implementation of machine learning (ML) into existing demand planning practices (Ferreira, Lee, and Simchi-Levi, 2016; Kremer, Moritz, & Siemsen, 2011; Makridakis & Winkler, 1983; Makridakis, Spiliotis, & Assimakopoulos, 2020). ML—"the machine's ability to keep improving its performance without humans having to explain exactly how to accomplish all the tasks it's given" (Brynjolfsson & Mcafee, 2017, p. 2)—offers more sophisticated models than previously available, as well as the ability to process large amounts of data quickly (Carbonneau, Laframboise, & Vahidov, 2008; Ferreira, Lee, and Simchi-Levi, 2016). However, despite the considerable interest and public discourse around automation and the usefulness of ML in demand planning, modern managers continue to rely primarily on judgment (Siemsen & Aloysius, 2020). The apparent disconnect between emerging practice (i.e., primarily models) and actual current practice (i.e., primarily human judgment) motivates this research as we attempt to explicate how to conjunctively balance the strengths of ML and human judgment.

Since the purpose of this paper is to examine the integration of machine learning and human judgment in demand planning, we first discuss the strengths of each. The strengths of models lie in predicting systematic variability (Blattberg & Hoch, 1990). Therefore, the more stable (less volatile), the time series, the more accurate models' predictions (Schubert, 2012). This is difficult because the demand patterns of many products and services contain elements of

non-systematic variability. Fortunately, the strengths of humans lie in predicting non-systematic variability (Seifert, Siemsen, Hadida, & Eisingerich, 2015), specifically when the variability is due to special events. We use the term special events to describe any event that causes a significant change to demand. Humans thrive in the face of special events where flexibility, subjective evaluation, and contextual information are necessary (Arvan et al., 2019; Blattberg & Hoch, 1990). According to Seifert et al., 2015, contextual information refers to any non-historical information that is relevant to a special event (p. 34). Examples of contextual information include abnormal weather or seasonality; surprise actions of competitors; and rapid changes in political climates. As an illustration, a firm that produces canned goods and distributes these through several retailers in a geographical region, may receive a hurricane alert for that region with the impending event that will increase demand for the product. For readers seeking a more comprehensive review of contextual information, see Arvan et al., 2019. Although humans may improve predictions by offering insight into non-systematic variability, solely relying on human judgment for predictions often results in a performance loss due to biases (Kremer, et al., 2011). Blattberg and Hoch (1990) demonstrate that judgment biases can be ameliorated through integrating models and judgment rather than using either in isolation (p. 887).

It is possible to draw insight from the domain of chess when in 1997, IBM's Deep Blue supercomputer beat the world's best human chess player, Garry Kasparov, for the first time – a most significant event for artificial intelligence (Sanders & Wood, 2019). Kasparov, in an introspective book on the event, proposed Kasparov's Law which claims that ordinary human judgment and models integrated with the right process are better than a strong model alone (see quote at beginning of article). Kasparov's conclusion emphasizes the importance of the process.

This research documents two processes that are designed to capture the strengths of supervised learning systems and human judgment simultaneously: interactive machine learning (IML) and human-guided machine learning (HGML).

ML is often separated into two paradigms: supervised learning and unsupervised learning. Supervised learning systems rely on a labeled dataset of inputs x and outputs y. Through iterations, the system creates a "learned mapping $f(x)$" which is then used to produce a prediction y^* for each x^* (Jordan & Mitchell, 2015, p. 257). Common examples of mapping f include linear regression, decision trees, neural networks, and support vector machines (Bishop, 2006; Jordan & Mitchell, 2015). Linear regression is considered one of the simplest supervised ML algorithms when used iteratively (Goodfellow et al., 2016; Hastie et al., 2009; and Murphy, 2012).

In contrast, unsupervised learning systems involve analysis of an unlabeled dataset, thus requiring the system to gain its own experience without human guidance (Jordan & Mitchell, 2015, p. 258). Unsupervised learning is not as common as supervised learning, but examples include clustering algorithms (e.g., K-mean, and Gaussian mixture models) and dimension reduction methods (Bishop, 2006; Jordan & Mitchell, 2015). Other types of ML include reinforcement learning, semi-supervised learning, and discriminative learning (see Jordan & Mitchell, 2015 for an in-depth review of ML).

The two processes discussed in this paper are based on an extension of supervised learning systems. Namely, once the model is sufficiently trained, the model continues to learn from new features which are guided by the human. Therefore, a combination of trend detection (model) and identification of special events (human judgment) are used to create a map between demand variability and causal events. The difference between IML and HGML lies in the

amount of information the human includes in the features shared with the model. IML allows the human to provide an estimated change to demand in units due to the special event and HGML allows the human to indicate whether there is a special event.

This research contributes to academia and practice. Our research illustrates the importance of the process of integrating ML and human judgment. We introduce two processes for integrating ML and human judgment—IML and HGML—to be used in demand planning. We offer evidence of an improvement in the demand forecast accuracy over the current demand planning process used by a firm when IML and HGML are used. Our study offers a solution to the current disconnect between research encouraging the use of ML (Arvan et al., 2019; Kremer et al., 2011; Makridakis & Winkler, 1983; Makridakis et al., 2020) and practice relying on human adjustments (Siemsen & Aloysius, 2020).

3.2 Related Literature

This paper explores processes of how best to integrate ML and human judgment in the context of demand planning. The concept of integrating machines and humans originated in the engineering "human-in-the-loop machine control loop" (Fung et al., 1992, p. 1). *Human-in-the-loop* referred to an automated machine outfitted with an operatable part (e.g., a gripper of a robotic manipulator) which a human operator was able to manually control given instances of "unscheduled tasks and unpredictable disturbances" (Fung et al., 1992, p.7). Human-in-the-loop (HITL) has since been adopted by artificial intelligence (AI) researchers to describe human interaction with AI-based automation (Holzinger, 2016; Sheridan, 1995; Zanzotto, 2019). HITL research has covered a wide range of algorithms and contexts (Amershi, Cakmak, Knox, & Kulesza, 2014; Gil, Honaker, Gupta, Ma, D'Orazio, Garijo, Gadewar, & Yang, 2019). Two

human-in-the-loop ML integration processes that are most relevant to this research, although loosely-defined, are *interactive machine learning* and *human-guided machine learning*.

IML (Figure 3.1) is often used to describe any form of interaction between the ML model and user (Amershi et al., 2014). The term *interactive machine learning* is first introduced by Fails and Olsen Jr. (2003) as a process that uses rapid iterations between the model and user thus initiating a collaboration between the machine intelligence and human intelligence. Fails and Olsen Jr. (2003) illustrate the notion of IML using an image-processing system Crayons. The Crayons system begins like a supervised ML system with inputs of human-labeled training data (i.e., classified images). However, once the image is classified in the system, output from the system is displayed, and the user can refine the classifier further or accept the classifier. The iteration between the model and human throughout the training process allows the user to input new data to the training dataset, thus quickly solving problems in the model as they arise. Holzinger et al. (2019) provide a more concentrated definition of IML as a process that, "integrates the human into the algorithmic loop" (p. 2402). Using an optimization experiment, Holzinger et al. (2019) offer evidence that incorporating human knowledge positively influences machine learning.

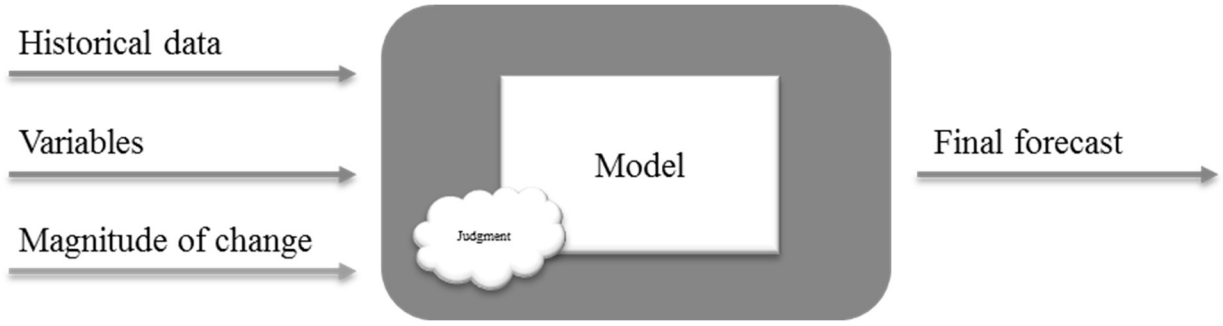


Figure 3.1. Interactive Machine Learning.

Another process in the literature, HGML (Figure 3.2), focuses on a user acting as a guide to a fully automated ML system (Gil et al., 2019). The main difference between IML and HGML is users providing input (IML) versus users providing guidance (HGML). HGML is an emerging concept which has been rarely investigated (Amershi et al., 2011; Gil et al., 2019). However, we believe the concept has great potential in the context of demand planning where humans often misinterpret contextual information (Fildes, Goodwin, & Önkal, 2019).

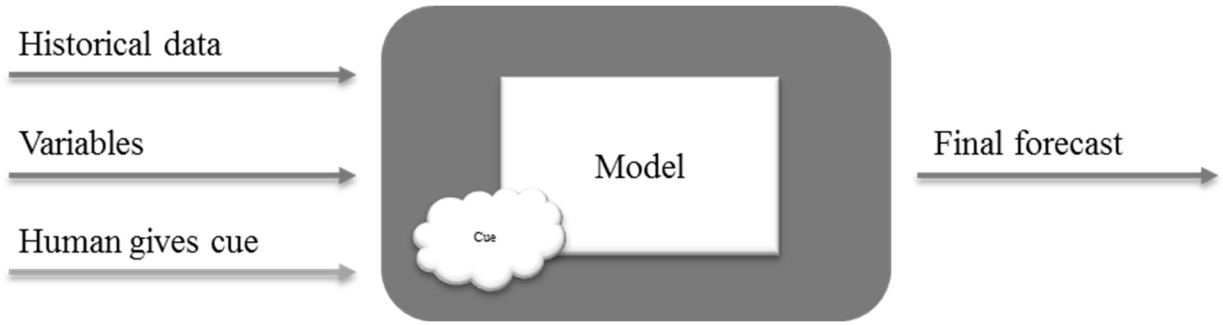


Figure 3.2. Human-Guided Machine Learning.

Extant research on integrating models and human judgment in demand planning has primarily focused on humans adjusting statistical models (Arvan et al., 2019; Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009), quantitative correction of judgment forecasts (Fildes, 1991; Theil, 1971), equal-weighted average of statistical model and judgment forecasts (Blattberg & Hoch, 1990), and judgmental selection of statistical models (Bunn & Wright, 1990; Petropoulos, Kourentzes, Nikolopoulos, & Siemsen, 2018). The few studies that discuss ML methods in demand planning rely solely on automated ML (i.e., no human interaction) and offer mixed results of the usefulness (Carbonneau et al., 2008; Makridakis et al., 2018).

In summary, IML and HGML offer unique processes of integrating ML models and human judgment in demand planning. To the best of our knowledge, we are the first to study the behavioral issues involved with IML used in demand planning. We also provide the first

empirical behavioral study on HGML in demand planning. Additionally, we are the first study to empirically compare IML and HGML.

3.3 Theory

3.3.1 Hypothesis 1

Judgmental adjustments (JA)—any adjustment made to the output of a system forecast--are the most common method of integration (Arvan et al., 2019; Siemsen & Aloysius, 2020). This is because it is not uncommon for human forecasters to feel the system does not incorporate all of the relevant contextual data (Boulaksil & Franses, 2009). One difficulty associated with JA is the reliance on human judgment to drive the process of integration since judgment virtually always suffers from biases. Examples of biases include risk aversion (Kahneman & Tversky, 1979), judgment biases (Kremer et al., 2011), bracing (Tokar et al., 2014), decision speed (Moritz, Siemsen, & Kremer, 2014), and other individual characteristics (Eroglu & Croxton, 2010). Researchers have frequently sought for best practices when engaging in JA (e.g., Fildes et al., 2009; Petropoulos, Fildes, & Goodwin, 2016). A common thread in the JA literature argues that adjustments should be made only by experts (e.g., Arvan et al., 2019), as higher levels of expertise lead to improved forecast accuracy (Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2017).

Past studies have focused on the comparison of systems relying on statistical models and JA. To the best of our knowledge, we are one of the first studies to compare ML and JA. We expect since ML systems often have greater ability to process more data, ML systems are even more accurate than statistical model systems. Regardless of the accuracy of the ML system, there is still inability of the ML system to access contextual information the human forecaster has

access to. Therefore, drawing from existing literature, we predict JA made by experts will increase accuracy over the ML system:

***Hypothesis* H_1.** *Judgmental adjustment (JA) results in improved forecast accuracy when compared to the machine learning system (ML).*

3.3.2 Hypothesis 2

One common grievance with JA is that humans often make unnecessary judgmental adjustments to forecasts, even when there is not enough evidence to make an informed decision (Fildes et al., 2009; Lawrence, Goodwin, O'Connor, & Önkal, 2006). Prior literature summarizes adjustment reasons into four overarching types: 1) contextual information outside the model, 2) social pressures, 3) gain control or sense of ownership (i.e., lack of trust in the system), and 4) misinterpretation of data (Petropoulos et al., 2016). An attractive feature of IML and HGML is the ability to assign weights to the adjustment based on the usefulness of past adjustments. IML and HGML are thus able to ameliorate biases and weight un-useful adjustments accordingly.

Research on IML and HGML has been limited; however Holzinger et al. (2019) offer evidence of a positive interaction between humans and ML. To the best of our knowledge, we are the first to study the behavioral issues involved with IML and HGML used in demand planning. We expect IML and HGML to lead to more accurate forecasts through the iterative integration and interaction of the ML system and human judgment. Specifically:

***Hypothesis* H_2.** *Interactive machine learning (IML) and human-guided machine learning (HGML) result in improved forecast accuracy when compared judgmental adjustment (JA).*

3.3.3. Hypothesis 3

According to Fildes et al. (2019), humans often misinterpret contextual information, which leads to inaccurate estimations. The design of HGML acknowledges the human tendency to misinterpret contextual information and therefore removes the human's role of estimating change to demand. Instead, HGML relies on a model calculation of the magnitude of special events based on similar periods in the historical data. The concept of HGML aligns best with the strengths of ML (systematic variability) and human judgment (non-systematic variability).

HGML is an emerging concept which has been rarely investigated (Amershi et al., 2011; Gil et al., 2019). As such, we provide the first empirical behavioral study on HGML in demand planning. We hypothesize that HGML will be the optimal process to maximize the value of contextual information and to minimize biases. Formally:

***Hypothesis* $\mathbf{H_3}$.** *Human-guided machine learning (HGML) results in improved forecast accuracy when compared to interactive machine learning (IML).*

3.4 Design and Implementation

The study is conducted using a natural field study to test the hypothesized relationships between the various processes of integrating ML and human judgment. Natural field studies are an attractive method to study behavior since the researchers maintain control (internal validity) while retaining realism (external validity) without the subjects knowing they are being studied (Harrison & List, 2004; Levitt & List, 2009). We conducted the field study in a large, multinational firm over a period of 30 weeks.

The existing demand planning process in the firm relies on a sophisticated ML system as well as human demand planners. The ML system the firm uses relies on unsupervised machine learning algorithms such as gradient-boosting optimization, Gaussian processes, and hierarchical

modeling. Forecasts are prepared eight weeks in advance using a four-to-six week forecast horizon. Each week, demand planners are tasked with reviewing the ML system demand forecast at the category-SKU-store level and have the option to adjust the ML system forecast (as is common practice in most firms, see Siemsen & Aloysius, 2020). If the demand planners adjust the ML system demand forecast, the adjustment also requires a reason to be assigned to the adjustment.

Although adjustments may be made due to a combination of multiple types, the firm labeled the 12 commonly-used adjustment reasons across the five product categories included in our sample as adjustments made due to contextual information outside the model (Type 1 in Petropoulos et al., 2018). Table 1 lists the 12 adjustment reasons and the corresponding number of adjustments attributed to each reason sorted according to the firm's list. The specific reasons for adjustment have been removed in some cases to preserve anonymity. Table 3.1 illustrates how frequently human demand planners adjust the ML system demand forecast. The predominant reason of adjustment is Promotion 4, a promotion unique to the firm, where adjustments are made 12.88% of the time.

Table 3.1. Number of Adjustments and Demand Forecasts per Reason at the Category-SKU-Store Level.

Adjustment Reason	N Adjustments	N Demand Forecasts	Frequency of Adjustments
Weather	984	369,180	0.27%
Summer vacations	16,244	183,328	8.86%
Demand shifts between store locations	5,080	480,200	1.06%
Phantom inventory	32,000	369,180	8.67%
Store-specific events	10,444	364,996	2.86%
Promotion 1	940	285,612	0.33%
Promotion 2	4,728	400,816	1.18%
Promotion 3	1,284	266,896	0.48%
Promotion 4	85,448	663,528	12.88%
Promotion 5	532	111,020	0.48%
Model 1	932	183,328	0.51%
Model 2	2,192	31,636	6.93%

The field study contains four treatment groups: 1) the existing method of integration—judgmental adjustment (JA)—used by the firm, 2) IML, 3) HGML, and 4) ML forecast with no human adjustment as the control group.

The JA treatment follows the existing demand planning process where the human demand planner observes the forecast from the ML system and chooses whether to adjust the ML system forecast.

For both IML and HGML models we use a multivariate linear regression algorithm. We choose to implement IML and HGML using multivariate linear regression for two main reasons: 1) existing literature does not show a significant difference between multivariate linear regression and more complex ML algorithms (Carbonneau et al., 2008), and 2) the simpler the algorithm, the easier for the company to implement.

The IML treatment requires 10 weeks of training where the IML model (see Equation 1) uses the actual sales as targets and the intercept (β_0), ML system forecast (SYS_FCST), and amount of adjustment ($QTY\_ADJUSTED$) per reason for adjustment ($REASONk$) as features in the training dataset:

(1) $ActualDemand_{it} = \beta_0 + \beta_1 SYS\_FCST_{it} + \beta_k REASONk_{it}$

where i = category-SKU-store-level; *t* = week; *SYS_FCST* = ML system forecast (no human adjustment); REASON = quantity of adjustment for adjustment reason k.

Once the IML model is trained, using the same symbology, the forecast is calculated as:

(2) $IML_{FCST_{it}} = \beta_0 + \beta_1 SYS\_FCST_{it} + \beta_k REASONk_{it}$

Similarly, the HGML model is trained using the first 10 weeks of data where the targets were actual sales and the features included the intercept (β_0), ML system forecast absent of

human adjustment (SYS_FCST), and the amount of adjustment ($QTY\_ADJUSTED$) per reason for adjustment ($REASONk$):

(3) $ActualDemand_{it} = \beta_0 + \beta_1 SYS\_FCST_{it} + \beta_k REASONk_{it}$

where REASON variables $= \begin{cases} 0 \; if \; QTY\_ADJUSTED = 0 \\ 1 \; if \; QTY\_ADJUSTED \neq 0 \end{cases}$

Once the HGML model is trained, the forecast is calculated as:

(4) $HGML_{FCSTit} = \beta_0 + \beta_1 SYS\_FCST_{it} + \beta_k REASONk_{it}$

Lastly, the control treatment is the ML system forecast absent of human adjustment.

At the category-SKU-store-level, our dataset contains 3,709,720 weekly demand forecasts ranging from February-September 2020. Table 3.2 summarizes the dataset by number of SKUs and stores per category.

Table 3.2. Number of SKUs and Stores per Category.

Category	N SKUs	N Stores
1	18	213
2	25	485
3	15	265
4	21	485
5	51	354

Once the mean absolute error (MAE) is computed at the category-SKU-store-level for each week, the MAE is aggregated to the category-reason level to achieve independence, resulting in a final dataset with 340,772 demand forecast observations. Each treatment group has 85,193 observations.

3.5 Results

We initially test the MAEs of the category-reason level for normality following the Kolmogorov-Smirnov Test and find evidence of non-normality ($D = 0.354$, $p < 0.001$). Therefore, to test our hypotheses, we conduct a Kruskal-Wallis H test (Conover, 1999) which is

a non-parametric substitute for the standard F-test used when conditions of normality are violated. The null and alternative hypotheses for the Kruskal-Wallis H-test in the context of our analysis are:

H_0: The distribution of MAE is the same across categories of treatment.

H_a: At least one of the distributions of MAE differs across categories of treatment.

The results of the Kruskal-Wallis H-test are statistically significant ($H = 707.56$, $p < 0.001$) indicating that at least one of the distributions of MAE differs across treatments. In other words, the processes used to integrate ML and human judgment are not synonymous.

Table 3.3. Kruskal-Wallis H-Test Results for Hypothesis H_1: Judgmental Adjustment (JA) Compared to Machine Learning System (ML) Demand Forecast.

Adjustment Reason	MAE_{JA}	Between-treatment comparison	MAE_{ML}
Weather	6.61	$p < 0.01$	**6.04**
Summer vacations	2.97	$p < 0.01$	**2.79**
Demand shifts between store locations	5.42	$p < 0.01$	**5.01**
Phantom inventory	6.61	$p < 0.01$	**6.04**
Store-specific events	7.13	$p < 0.01$	**6.55**
Promotion 1	8.45	$p < 0.01$	**7.74**
Promotion 2	7.03	$p < 0.01$	**6.45**
Promotion 3	2.67	$p < 0.01$	**2.48**
Promotion 4	5.42	$p < 0.01$	**5.01**
Promotion 5	5.97	$p = 0.18$	5.58
Model 1	2.97	$p < 0.01$	**2.79**
Model 2	8.46	$p = 0.11$	7.85

Hypothesis **H_1** predicts that JA is more accurate than the ML system forecast because of JA's ability to include contextual information that the ML system does not have access to. However, the results do not support H_1 since out of the 12 adjustment reasons, 10 were more accurate when using the ML system (see Table 3.3). This finding suggests JA fails to effectively include the contextual information.

Hypothesis **H_2** predicts JA will be less accurate than IML and HGML. Hypothesis **H_2** is supported for all 12 adjustment reasons (see Table 3.4). This finding suggests the value of contextual information outside the ML system can be captured more effectively using IML or HGML rather than JA.

Table 3.4. Kruskal-Wallis H-Test Results for Hypothesis H_2: Judgmental Adjustment (JA) Compared to Interactive Machine Learning (IML) and Human-Guided Machine Learning. (HGML).

Adjustment Reason	MAE_{IML}	Between-treatment comparison	MAE_{JA}	Between-treatment comparison	MAE_{HGML}
Weather	**5.83**	$p < 0.01$	6.61	$p < 0.01$	**6.10**
Summer vacations	3.38	$p = 0.04$	2.97	$p = 0.02$	**2.83**
Demand shifts between store locations	**4.89**	$p < 0.01$	5.42	$p < 0.01$	**4.89**
Phantom inventory	**6.04**	$p < 0.01$	6.61	$p < 0.01$	**5.89**
Store-specific events	**6.42**	$p < 0.01$	7.13	$p < 0.01$	**6.37**
Promotion 1	**7.48**	$p < 0.01$	8.45	$p < 0.01$	**7.47**
Promotion 2	**6.24**	$p< 0.01$	7.03	$p< 0.01$	**6.24**
Promotion 3	**2.47**	$p< 0.01$	2.67	$p< 0.01$	**2.47**
Promotion 4	8.76	$p< 0.01$	5.42	$p< 0.01$	**4.94**
Promotion 5	**5.48**	$p< 0.01$	5.97	$p< 0.01$	**5.49**
Model 1	**2.80**	$p< 0.01$	2.97	$p< 0.01$	**2.80**
Model 2	**7.88**	$p = 0.01$	8.46	$p< 0.01$	**7.77**

Hypothesis **H_3** predicts IML will be less accurate than HGML. H_3 is not supported overall as 11 of the 12 difference are not significant (see Table 3.5). The non-difference suggests there is not a statistically significant difference between the MAE of IML and HGML. However, there is one special event—Promotion 4—where there is a significant difference between IML and HGML. This finding suggests there may be reasons where HGML is more accurate, however there needs to be more tests before a conclusion can be made.

Although our research is focused on the integration of ML and human judgment, we also include a comparison between the ML system—absent of human adjustment—and IML, HGML (see Table 3.6). Interestingly, there is never an adjustment reason where the ML system is

significantly better than either IML or HGML. This finding indicates there is value in the contextual information that can increase predictive accuracy.

Table 3.5. Kruskal-Wallis H-Test Results for Hypothesis H_3: Interactive Machine Learning (IML) Compared to Human-Guided Machine Learning (HGML).

Adjustment Reason	MAE_{IML}	Between treatments	MAE_{HGML}
Weather	5.83	p = 0.15	6.10
Summer vacations	3.38	p = 0.71	2.83
Demand shifts between store locations	4.89	p = 0.99	4.89
Phantom inventory	6.04	p = 0.55	5.89
Store-specific events	6.42	p = 0.87	6.37
Promotion 1	7.48	p = 0.98	7.47
Promotion 2	6.24	p = 0.99	6.24
Promotion 3	2.47	p = 0.97	2.47
Promotion 4	8.76	p = 0.05	**4.94**
Promotion 5	5.48	p = 0.96	5.49
Model 1	2.80	p = 0.97	2.80
Model 2	7.88	p = 0.70	7.77

Table 3.6. Kruskal-Wallis H-Test Results for Machine Learning System (ML) Compared to Interactive Machine Learning (IML) and Human-Guided Machine Learning (HGML).

Adjustment Reason	MAE_{IML}	Between-treatment comparison	MAE_{ML}	Between-treatment comparison	MAE_{HGML}
Weather	**5.83**	p < 0.01	6.04	p = 0.09	6.10
Summer vacations	3.38	p = 0.33	2.79	p = 0.54	2.83
Demand shifts between store locations	**4.89**	p < 0.01	5.01	p < 0.01	**4.89**
Phantom inventory	6.04	p = 0.03	6.04	p = 0.01	**5.89**
Store-specific events	**6.42**	p = 0.04	6.55	p = 0.03	**6.37**
Promotion 1	**7.48**	p < 0.01	7.74	p < 0.01	**7.47**
Promotion 2	**6.24**	p < 0.01	6.45	p < 0.01	**6.24**
Promotion 3	2.47	p = 0.42	2.48	p = 0.40	2.47
Promotion 4	8.76	p = 0.78	5.01	p = 0.03	**4.94**
Promotion 5	5.48	p = 0.08	5.58	p = 0.08	5.49
Model 1	2.80	p = 0.94	2.79	p = 0.91	2.80
Model 2	7.88	p = 0.38	7.85	p = 0.20	7.77

3.6 Discussion and Conclusion

The purpose of our research is to identify the most accurate process of integrating ML and human judgment in demand planning. As such, we propose and test three hypotheses.

Hypothesis $\mathbf{H_1}$ focuses on the comparison between accuracy of demand forecasts made by the existing demand planning process of the firm, JA, and the ML system. We hypothesized JA is more accurate than the ML system. The results indicate out of the 12 adjustment reasons, 10 were more accurate when using the ML system, and the other two reasons had no significant difference between JA and the ML system. Therefore, Hypothesis $\mathbf{H_1}$ is not supported. There are various ways this finding may be explained according to existing literature. First, human judgment relies on various heuristics that introduce bias into the demand forecast (Eroglu & Croxton, 2010; Kahneman & Tversky, 1979; Kremer et al., 2011; Moritz et al., 2014; Tokar et al., 2014; Tversky & Kahneman, 1974). Therefore, the biases introduced may outweigh the value of the contextual information. Second, humans frequently misinterpret contextual information (Fildes et al., 2019). Third, the existing literature base primarily focuses on statistical systems (Arvan et al., 2019), and our study uses a sophisticated ML system. Due to the intricacies of the ML system in place, it may be better to rely on the ML system forecast, absent of human judgment, rather than allow the biases introduced in JA to damage the demand forecast.

Hypothesis $\mathbf{H_2}$ predicts that JA will be less accurate compared to IML and HGML. The MAE for all 12 adjustment reasons indicate JA is significantly higher (less accurate) than IML and HGML, so Hypothesis $\mathbf{H_2}$ is supported. This finding suggests that contextual information does have value, however, the value is not captured when human judgment is not weighted according to systematic biases. This finding contributes to the literature by offering two

processes to effectively integrate ML and human judgment that effectively capture the strengths of ML and the value of contextual information.

Lastly, Hypothesis **H_3** predicts HGML will be more accurate than IML because HGML removes the role of estimating the change to demand caused by the special event from human judgment. Humans often misinterpret contextual information (Fildes et al., 2019) and adjust demand forecasts when no adjustment should be made (Lawrence et al., 2006; Fildes et al., 2009). However, our results do not support Hypothesis **H_3** since 11 out of the 12 reasons for adjustment are not significantly different. The results do indicate that for 1 out of the 12 reasons for adjustment, HGML is significantly more accurate. This finding is interesting since it suggests there are certain events that are best left to the model to quantify. Further research is needed to uncover which events should use purely HGML. Our comparison between IML and HGML is the first, to the best of our knowledge, and introduces a promising new avenue for future research.

Our study provides implications for research and practice. Our research contributes to the literature on integrating human judgment in the demand planning process. We illustrate the importance of the process of integrating ML and human judgment. Namely, we offer evidence of two processes for integrating ML and human judgment—IML and HGML—that improve accuracy whereas the current process for integrating human judgment, JA, decreases accuracy. IML and HGML continue to allow human demand forecasters some control over the demand planning outcome, as suggested by Dietvorst, Simmons, and Massey (2016), while controlling for biases that inevitably are present with human judgment.

Our results suggest that managers seeking to improve forecast accuracy may find that asking employees to share the contextual information with the model, rather than directly

adjusting the forecast, can help eliminate bias. Another strength of IML and HGML is the ability to secure knowledge that may previously have resided only in the minds of individuals. IML and HGML enable the model to continually learn from forecaster experience. Firms implementing IML and HGML need not replace the current system used for demand planning. Instead, IML and HGML use the system forecast as an integral part and strategically weight the human adjustment. In fact, the current demand planning process can look the same to the manager performing the adjustments. This is valuable since it removes much of the complexity involved in change management. Additionally, when designing forecasting support systems (FSS), our results show the importance of using the correct method that employs the strengths of users and algorithms. Thus, in order to capture the value of contextual information outside the model, we encourage practitioners consider implementing IML and HGML as they have potential to improve predictive performance.

This research has limitations. First, although the natural field study is conducted in a large, multinational firm, it is still within one firm. Running a natural field study in multiple firms and comparing between firms would be a useful avenue for future research. Additionally, the timeframe during which we ran the field study was brief (between 20 and 30 weeks). Future studies could use a period of longer than 30 weeks to allow for more adjustments. IML and HGML learn based on the adjustments, thus the more data available, the more accurate the models.

In conclusion, our research indicates that using an appropriate process of integrating machine learning and human judgment—IML and HGML—provides a significant benefit to demand planning. Despite the movement towards automated demand forecasts and reliance on sophisticated ML methods, our research suggests human judgment continues to add value when

integrated appropriately. We encourage further research into IML and HGML in various demand planning scenarios.

3.7 References

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. 2017. Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting*, 33(1): 298-313.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. 2014. Power to the People: The role of humans in interactive machine learning. *AI Magazine*, 35(4): 105-120.

Arvan, M., Fahimnia, G., Reisi, M. & Siemsen, E. 2019. Integrating human judgment into quantitative forecasting methods: A review. *Omega*, 86: 237-252.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blattberg, R. C. & Hoch, S. J. 1990. Database models and managerial intuition: 50% model+ 50% manager. *Management Science*, 36(8): 887-899.

Boulaksil, Y. & Franses, P. H. 2009. Experts' stated behavior. *Interfaces*, 39(2): 168-171.

Brynjolfsson, E. & Mcafee, A. 2017. The business of artificial intelligence. *Harvard Business Review*: 1-20.

Bunn, D. & Wright, G. 1991. Interaction of judgmental and statistical forecasting methods: Issues & analysis. *Management Science*, 37(5): 501-518.

Carbonneau, R., Laframboise, K., & Vahidov, R. 2008. Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3): 1140-1154.

Conover, W.J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York, NY: Wiley Publishers

De Baets, S. & Harvey, N. 2020. Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research*, 284(3): 882-895.

Dietvorst, B. J., Simmons, J. P., & Massey, C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3): 1155-1170.

Eroglu, C. & Croxton, K. L. 2010. Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26(1): 116-133.

Fails, J. A., & Olsen Jr, D. R. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*: 39-45.

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

Fiebrink, R.; Cook, P. R.; and Trueman, D. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the Conference on Human Factors in Computing Systems*:147–156. New York: Association for Computing Machinery.

Fildes, R. 1991. Efficient use of information in the formation of subjective industry forecasts. *Journal of Forecasting*, 10: 597-617.

Fildes, R., P. Goodwin, M. Lawrence, & K. Nikolopoulos. 2009. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25: 3–23.

Fildes, R., Goodwin, P., & Önkal, D. 2019. Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, 35(1): 144-156.

Franses, P. H., & Legerstee, R. 2010. Do experts' adjustments on model-based sku-level forecasts improve forecast quality? *Journal of Forecasting*, 29(3): 331-340.

Fung, P. T., Norgate, G., Dilts, T. A., Jones, A. S., & Ravindran, R. 1992. *U.S. Patent No. 5,116,180*. Washington, DC: U.S. Patent and Trademark Office: 1-12.

Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S., Yang, Q., & Jahanshad, N. 2019. Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*: 614-624.

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

Holzinger, A. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop?. *Brain Informatics*, 3(2): 119-131.

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C. M., & Palade, V. 2019. Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence*, 49(7): 2401-2414.

Jordan, M. I., & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245): 255-260.

Kahneman, D., & Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2): 263-292.

Kasparov, G. 2017. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. PublicAffairs.

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

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. 2006. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3): 493-518.

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

Makridakis, S. & Winkler, R. L. 1983. Averages of forecasts: Some empirical results. *Management Science*, 29(9): 987-996.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. 2018. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS One*, 13(3): 1-26.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. 2020. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54-74.

Mitchell, T. (1997). Introduction to machine learning. *Machine Learning*, 7: 2-5.

Moritz, B., Siemsen, E., & Kremer, M. 2014. Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management*, 23(7): 1146-1160.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. 2018. Judgmental selection of forecasting models. *Journal of Operations Management*, 60: 34-46.

Petropoulos, F., Fildes, R., & Goodwin, P. 2016. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour?. *European Journal of Operational Research*, 249(3): 842-852.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3): 210-229.

Sanders, N. R., & Wood, J. D. 2019. *The Humachine: Humankind, Machines, and the Future of Enterprise*. Routledge.

Schubert, S. 2012. Forecastability: A new method for benchmarking and driving improvement. *Foresight: The International Journal of Applied Forecasting*, 26: 7-15.

Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. 2015. Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36: 33-45.

Sheridan, T. B. 1995. Human centered automation: Oxymoron or common sense?. In *IEEE International Conference on Systems, Man and Cybernetics*. Intelligent Systems for the 21st Century, 1: 823-828

Siemsen, E. & Aloysius, J. 2020. Supply chain analytics and the evolving work of supply chain managers. Research report for *Association of Supply Chain Management*.

Theil, H. 1971. *Applied Economic Forecasting*. North-Holland Publishing Company, Amsterdam.

Tokar, T., Aloysius, J. A., Waller, M., & Williams, B. 2014. Bracing for demand shocks: An experimental investigation. *Journal of Operations Management*, 32(4): 205-216.

Tong, J., & Feiler, D. 2017. A behavioral model of forecasting: Naive statistics on mental samples. *Management Science*, 63(11): 3609-3627

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

Zanzotto, F. M. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64: 243-252.

IV. ESSAY 3

Improving Integrated Judgmental and Analytical Processes

Using Interventions Rooted in Cognitive Psychology

ABSTRACT

Modern supply chain managers are increasingly adopting data analytics, and practice reveals that change management is a critical part of successfully employing those analytics. The purpose of this paper is to understand the behavioral mechanisms driving analytics use and to provide actionable insights. We utilize a multi-theoretical lens using Adaptive Character of Thought (ACT) theory and dual process theory to develop hypotheses. Our research uses a laboratory to develop and test interventions on use of analytics. The laboratory experiment follows a 2 (declarative knowledge, no declarative knowledge) x 2 (analytical thinking, no analytical thinking) x 3 (interactive machine learning, human-guided machine learning, participant's choice between interactive machine learning or human-guided machine learning) design. We find that declarative knowledge training, when paired with analytical thinking, has the largest impact on improving behavior and performance for tasks that require high levels of information-processing capacity.

Keywords:

Supply Chain Analytics, Behavioral Experiment, Interventionist-based Research

4.1 Introduction

Big data analytics are a significant part of the business world and some sources suggest that analytics will soon be an essential component for business survival (Singh, 2018). However, many businesses have found that capturing the value of analytics is challenging since it requires an efficient use of three disparate components: technology (i.e., data, statistics, information technology), people, and processes (Mohr & Hürtgen, 2018). Interestingly, executives highlight that the most important part of a successful analytics adoption is to, "Place people at the heart of the initiative" (Marchand & Peppard, 2013). Despite practitioners suggesting the importance of the people component of analytics, extant research in supply chain has mostly focused on the other two components: technology and processes (Sodero, Jin & Barratt, 2019). The purpose of this research is to focus on people and their underlying behavioral mechanisms associated with supply chain analytics use. We do so through a laboratory study and an interventionist-based approach (Chandrasekaran, De Treville, & Browning, 2020; Oliva, 2019), working in tandem with practice.

Interventionist-based research is used to explore, "the role an intervention … can play in testing and developing theory" (Oliva, 2019, p. 711). Interventionist-based research is founded on the same empirical motivation as design science, which is, "submitting ideas/models/frameworks to the ultimate empirical assessment by testing their usefulness and applicability" (Oliva, 2019, p. 711). However, interventionist-based approaches depart from the focus on design propositions and instead rely on the "use of interventions to make improvements" (Oliva, 2019, p. 721). Since the use of supply chain analytics is relatively new and has relatively limited empirical studies (Srinisavan & Swink, 2018), interventionist-based

research can provide an effective framework to test theory and to iteratively solve complex problems in practice (Van de Ven, 2007).

Through an iterative collaboration with a retailer, we develop training as a part of a change management initiative for supply chain employees working with analytics for truck scheduling and dispatching. The relationship with the company began through interviews pertaining to analytics use. The company suggested they would like to collaborate to understand current behavior in regard to analytics use with the goal to improve efficacy. Throughout the iterative conversations with the company, as well as decision makers, two main topics emerged: 1) decision makers must understand the definition of optimal behavior regarding analytics use and 2) decision makers must be deliberate about their decisions. Using theories from cognitive psychology, we develop interventions focused on training decision makers on their use of analytics.

The first theory we use is Adaptive Character of Thought (ACT) theory to develop an intervention regarding the first topic: decision makers must understand the definition of optimal behavior regarding analytics use. ACT theory explains human's use of declarative and procedural knowledge in determining behavior. The second theory we use is dual process theory to develop an intervention regarding the second topic: decision makers must be deliberate about their decisions. We use dual process theory to account for individual differences in decision making.

Within this research paradigm and context, our research questions are:

1) Can declarative knowledge increase optimal behavior when using analytics?
2) Can declarative knowledge improve performance in supply chain decision making tasks?
3) Can analytical thinking increase optimal behavior when using analytics?

4) Can analytical thinking improve performance in supply chain decision making tasks?
5) Can a combination of declarative knowledge and analytical thinking be incorporated into a training program that effectively increases professional supply chain employees' optimal behavior when using analytics?

This research provides three main contributions. First, it implements a multi-theoretical lens to, "develop more robust and comprehensive explanations for empirical questions that have traditionally been addressed from…exclusive lenses" (Okhuysen & Bonardi, 2011, p. 9). ACT explains learning and behavior but does not account for individual differences. Thus, we employ dual process theory to account for heterogeneity in individual's judgments. Additionally, by combining the two theories, we find that using analytical thinking is one way to help individuals better access the common cognitive system for higher-level processing. Second, this research extends previous research by furthering the investigation on "how and why instructional training impacts…decision making" (Tokar, Aloysius, & Waller, 2012, p. 536). Third, we provide empirical evidence regarding best practice for the use of supply chain analytics.

4.2 Related Literature

4.2.1. Supply Chain Analytics

Although analytics has been a popular topic in supply chain practitioner (Brynjolfsson & McAfee, 2017; Marchand & Peppard, 2013; McAfee & Brynjolfsson, 2012) and academic (Schoenherr & Speier-Pero, 2015; Srinivasan & Swink, 2018; Waller & Fawcett, 2013) literature, studies have been "mostly anecdotal" in nature (Srinivasan & Swink, 2018). In this line of literature, both practice and academia refer to analytics using a multitude of terms (e.g., big data, big data analytics, business analytics, big data business analytics, data analytics, and supply chain analytics) which can cause confusion. In this study, we use the umbrella term

analytics and define it as, "tools, techniques, and processes that enable a firm to process, organize, visualize, and analyze data, thereby producing insights that enable data-driven operational planning, decision-making, and execution" (Srinivasan & Swink, 2018, p. 1851).

Analytics differ from other related fields (e.g., statistics, optimization, data mining) by being both quantitative and qualitative, looking at the past and future with different conditions, and approximating relationships between variables while using deductive mathematical methods (Waller & Fawcett, 2013). Analytics are often classified into three specific categories: descriptive, predictive, and prescriptive (Delen & Demirkan, 2013). Descriptive analytics aim to identify problems and opportunities within existing processes and functions. Descriptive analytics employ performance analysis using models, statistical techniques, and tools in an effort to make more accurate and efficient decisions (Delen & Demirkan, 2013). Predictive analytics, "involves the use of mathematical algorithms and programming to discover explanatory and predictive patterns within data" (Wang et al., 2016 p. 100; see also Delen & Demirkan, 2013; Waller & Fawcett, 2013). Prescriptive analytics, "involves the use of data and mathematical algorithms to determine and assess alternative decisions that involve objectives and requirements characterized by high volume and complexity, with the aim to improve business performance" (Wang et al., 2016, p. 101). Typically, descriptive analytics are used to identify what is currently occurring within the business and predictive and prescriptive analytics are used to aid in making decisions for the future (Wang et al., 2016).

Existing research suggests that supply chain adoption of analytics has significantly lagged relative to other industries (Srinivasan & Swink, 2018). However, analytics is becoming a substantial, and arguably a critical, part of the business world (Singh, 2018). As such, there is a unique opportunity to empirically study supply chain analytics as firms begin to move beyond

descriptive analytics. Additionally, most existing studies that empirically examine the use of supply chain analytics are from an organizational perspective (e.g., Bendoly, 2016; Choi, Wallace, & Wang, 2018; Srinivasan & Swink, 2018). Yet, the individual perspective is also an important piece to understanding the value creation of supply chain analytics (Mohr & Hürtgen, 2018).

For example, individual's unnecessary interference with analytics (e.g., adjustments and overrides) may lead to a dampened ability to capture the competitive edge provided by analytics (Srinisavan & Swink, 2018), decreased firm performance (Chen, Preston, & Swink, 2015; McAfee & Brynjolfsson, 2012), and failed analytics adoptions (Marchand & Pepper, 2013). Therefore, our research seeks to address the avoidable individual friction with analytics through developing interventions. In the following sections we review cognitive theories that can help explain individual behavior to provide a theoretical context for developing interventions to help supply chain individual's use of analytics.

4.2.2. Adaptive Character of Thought (ACT) Theory

Cognitive psychology holds the belief that all humans have a "common cognitive system for higher-level processing" (Anderson, 1983, p. 1). Some of the functions included in the higher-level system are memory, language, problem solving, and decision-making (Anderson, 1983). ACT posits all higher-level functions can be explained by productions. Productions are comprised of conditions, which specify *when* the production applies, and actions, which specify *what* occurs when the production applies (Anderson, 1982). Thus, "the sequence of productions that apply in a task correspond to the cognitive steps taken in performing the task" (Anderson, 1982, p. 3).

Two critical components of ACT are declarative and procedural knowledge. Declarative knowledge refers to factual information such as facts, definitions, and events which are encoded as parts (or chunks) of declarative structures in the working and long-term memory (Anderson, 1983). Procedural knowledge, on the other hand, refers to the knowledge of "how to do things" (Anderson, 1983, p. 215) and is represented as productions. There are three distinguishing features between declarative and procedural knowledge (Anderson, 1976). First, declarative knowledge is retained in an "all-or-none manner," it is either possessed or not (Anderson, 1976, p. 117). In contrast, procedural knowledge can be partially possessed (e.g., one may have a limited ability to play the piano). Second, declarative knowledge is acquired in an instance, whereas procedural knowledge is acquired gradually through performance (Anderson, 1976). In the words of Ryle (1949, p. 59):

> *Learning how or improving an ability is not like learning that or acquiring information. Truths can be imparted, procedures can only be inculcated, and while inculcation is a gradual process, imparting is relatively sudden. It makes sense to ask at what moment someone became apprised of a truth, but not to ask at what moment someone acquired a skill* (Ryle, 1949, p. 59).

Third, declarative knowledge can be communicated verbally through descriptions (Anderson, 1976). However, unlike declarative knowledge, procedural knowledge can only be measured through behavior (i.e., performance).

The ACT production system is made up of three memories: working, declarative, and production. Working memory is defined as, "declarative knowledge, permanent or temporary, in the active state" (Anderson, 1983, p. 20). Most of the processes in the ACT production system involve working memory in some way. *Encoding* refers to the process of accumulating declarative knowledge from the outside world into the working memory. The declarative knowledge is then *stored* in the declarative memory so it can be *retrieved* when prompted.

Productions (condition-action pairs), cycle through *matching* the knowledge in working memory with the conditions of the production, thus leading to *execution*. *Performance*, in turn, refers to the conversion of productions into behavior. Figure 4.1 provides an illustration of the ACT production system (Anderson, 1983, p. 19).

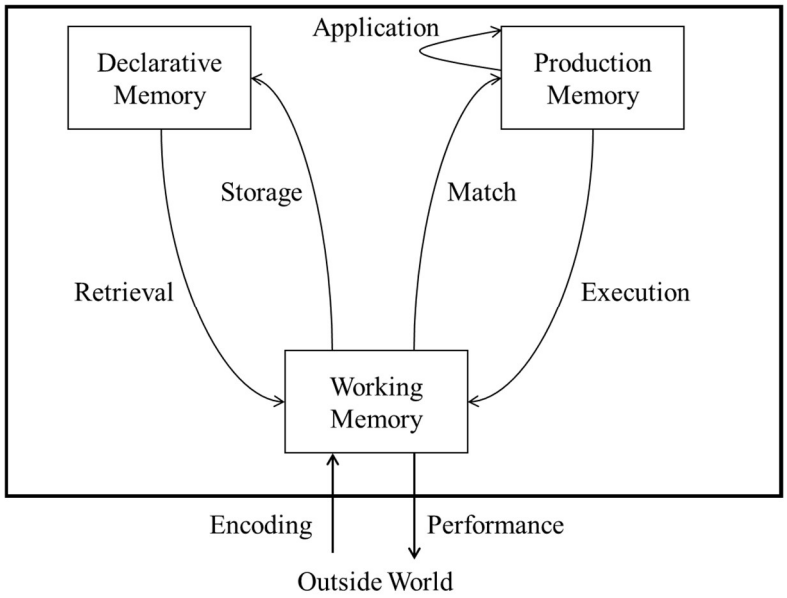


Figure 4.1. A General Framework for the ACT Production System. Based on Figure 1.2 from p. 19 of "The Architecture of Cognition" by J.R Anderson, 1983, Cambridge, MA: Harvard University Press.

ACT posits that an increase in declarative knowledge can aid individuals' acquisition of procedural knowledge (Anderson, 1976, 1982, 1983, 1993). Thus, ACT suggests that active declarative representations, which are not necessarily long-term, are critical for procedural learning (Anderson, 1993, p. 24). In turn, an increase in procedural learning results in improved cognitive functions (Anderson, 1983). In the words of Tokar, et al. (2012, p. 528), "when people have declarative knowledge of the task domain…they can better interpret that feedback, develop explanations, and make better decisions."

ACT has been used in prior literature to explain performance relating to cognitive functions such as skill acquisition (Anderson, 1982), memory and problem-solving (Anderson,

1995), and creativity (Kim & Zhong, 2017). Additionally, segments of ACT have been used in the supply chain literature. For example, declarative knowledge has been applied as training (Wu & Katok, 2006) and intervention (Tokar, et al., 2012) in inventory replenishment tasks. Interestingly, both Wu and Katok (2006) and Tokar et al. (2012) find mixed results when studying the application of various forms of declarative knowledge with intent to improve judgment and decision-making. At least a portion of the reason why results are mixed can most likely be explained by heterogeneity between individuals (Moritz, Hill, & Donohue, 2013). In the next section we discuss such heterogeneity.

4.2.3. Dual Process Theory

Heterogeneity between individuals regarding judgment and decision-making behavior in operations and supply chain management has frequently been studied using dual process theory (Choo, Nag, & Xia, 2015; Ball, Shah, & Donohue, 2018; Moritz, Hill, & Donohue, 2013; Narayanan & Moritz, 2015; Timmer & Kaufmann, 2019; Weinhardt, Hendijani, Harman, Steel, & Gonzalez, 2015). Dual process theory posits that human cognition can be split into two processes of reasoning: System 1 and System 2 (Stanovich & West, 2000) or intuitive and analytical (Weinhardt et al., 2015). System 1 (intuitive), "is characterized as automatic, largely unconscious, and relatively undemanding of computational capacity (Stanovich & West, 2000, p. 658). In contrast, System 2 (analytical), "conjoins the various characteristics that have been viewed as typified controlled processing" (Stanovich & West, 2000, p. 658) or "effortful activities" (Kahneman, 2011, p. 23). Examples of System 2 processes include focusing attention, following rules, and making deliberate choices between various options (Kahneman, 2011). Dual process theory postulates that both intuitive and analytical reasoning are concurrently engaged during judgment and decision making (Stanovich & West, 2000). Kahneman (2011, p. 24)

demonstrates the relationship between System 1 and System 2 processes through a brief anecdote:

> *"…Systems 1 and 2 are both active whenever we are awake. System 1 runs automatically, and System 2 is normally in a comfortable low-effort mode, in which only a fraction of its capacity is engaged. System 1 continuously generates suggestions for System 2: impressions, intuitions, intentions, and feelings. If endorsed by System 2, impressions and intuitions turn into beliefs, and impulses turn into voluntary actions. When all goes smoothly, which is most of the time, System 2 adopts the suggestions of System 1 with little or no modification. You generally believe your impressions and act on your desires, and that is fine—usually…System 2 is activated when an event is detected that violates the model of the world that System 1 maintains"* (Kahneman, 2011, p. 24).

One key assumption of dual process theory is that humans are "cognitive misers" (Toplak, West, & Stanovich, 2014, p. 147). The term "cognitive misers" refers to the human tendency to depend on System 1, intuitive processing whenever possible because it requires less effort (Simon, 1955; Toplak, West, & Stanovich, 2014; Tversky & Kahneman, 1974). The assumption of cognitive misers asserts that if there are several ways of solving the same problem, "people will eventually gravitate to the least demanding course of action" (Kahneman, 2011; Toplak, West, & Stanovich, 2014). Thus, humans have a natural tendency to be lazy (or efficient, depending on one's point of view) and engage in System 2 (analytical) processing as little as possible (Kahneman, 2011). As such, much of the day is spent relying on System 1 (intuition) (Kahneman, 2011). However, when the brain encounters a situation which requires attention, System 2 (analytical) is triggered through an "override function" (Toplak, West, & Stanovich, 2014, p. 147).

Although the human brain is equipped with numerous mechanisms geared to deal with varying situations, the "override function" is especially important within dual process theory (Toplak, West, & Stanovich, 2014, p. 148). The override function refers to the brain's shift between System 1 to System 2 processing, thus replacing default reasoning with more deliberate

thought (Evans & Stanovich, 2013). The switch between systems of reasoning is important in judgment and decision-making tasks as failures to override System 1 when it is wrong will greatly impact performance (Toplak, West, & Stanovich, 2014). Factors that influence the success of the override function include feeling of rightness (Thompson, 2009; Thompson, Turner, & Pennycock, 2011), confidence (Shynkarkuk & Thompson, 2006; Thompson et al., 2011), disposition to think deliberately (Stanovich, 2009, 2011), and cognitive ability (Evans & Stanovich, 2013).

Extant literature has frequently used the cognitive reflection task (CRT) to measure the individual tendency to engage the "override function" (Frederick, 2005; Kahneman & Frederick, 2002; Toplak, West, & Stanovich, 2011). The CRT consists of three questions which have a quick answer from System 1 reasoning, and a correct answer which would require an override from System 2 reasoning (Frederick, 2005). Research in many fields including operations and supply chain management (Moritz, Hill, & Donohue, 2013; Moritz, Siemsen, & Kramer, 2014; Narayanan & Moritz, 2015) have frequently used the CRT when studying human behavior (Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2011; Toplak, West, & Stanovich, 2014).

In the following section, we develop hypotheses aimed at altering individual's cognitive mechanisms underlying behavior using ACT and dual process theory. Our hypotheses employ the logic of ACT and dual process theory separately and combined to, "develop a more robust and comprehensive explanation" (Okhuysen & Bonardi, 2011, p. 9) of judgment and decision-making behavior in supply chain analytics use.

4.3. Theory

4.3.1. Hypothesis 1

Petropoulos, Fildes, and Goodwin (2016) list four reasons why individuals adjust/override model-based statistical forecasts: 1) contextual information outside the model, 2) social pressures, 3) gain control or sense of ownership (i.e., lack of trust in the system), and 4) misinterpretation of data. When conducting interviews with various firms, we found justifications from individuals pertaining to their interference with analytics that can be classified into the same four categories. Therefore, we postulate that these four reasons also explain why individuals often use analytics non-optimally. Since individuals' non-optimal use of analytics may negatively impact the value creation of supply chain analytics, we develop hypotheses regarding interventions aimed at altering the individual's cognitive mechanisms underlying the non-optimal behavior.

According to ACT, behavior relating to higher-cognitive processes, such as judgment and decision making, is continually refined through altering the productions, or procedural knowledge (Anderson, 1982). Productions are comprised of conditions and action pairs: conditions decide when a production applies, and action is what occurs when the conditions apply (Anderson, 1982). Since conditions are comprised of segments of declarative knowledge (Anderson, 1983), it follows that declarative knowledge is a key factor in determining actions and behavior. Thus, providing individuals declarative knowledge about optimal analytics use should help individuals "tune" the existing conditions in their brains relating to their behavior associated with analytics use (Anderson, 1983, p. 34). Recall the four broad reasons individuals use analytics non-optimally: contextual information, social pressures, gain control (i.e., lack of trust), and misinterpretation of data (Petropoulos et al., 2016). Through the lens of ACT, these

four reasons serve as conditions which decide when an individual should interfere with the analytics.

ACT posits that declarative knowledge is established through exposal to declarative representations and is formed whether the representations are long- or short-term (Anderson, 1993). As such, declarative knowledge embedded in training can aid in improving decision-making behavior and task performance. Prior literature offers evidence of the positive effect declarative knowledge can have in judgment and decision-making tasks. For example, in a multi-echelon inventory replenishment task, Wu and Katok (2006) find that declarative knowledge can increase performance when paired with communication. In the same context, declarative knowledge has been successfully implemented as a debiasing intervention (Tokar, Aloysius, & Waller, 2012). We propose that exposing individuals to declarative representations through training can help individuals tune their conditions, and thus their behavior (Anderson, 1983).

The first condition that individuals use when determining whether to interfere with analytics is contextual information (Petropoulos et al., 2016). Contextual information refers to any information which is outside of the analytics (e.g., weather and seasonality, actions of competitors, problems in purchasing, difficulties in managing the production process, personnel strikes, or competitor-related and environment-related rumors; Arvan et al., 2019, p. 245). Non-optimal interference with analytics based on contextual information would most likely be caused by a lack of knowledge regarding the data behind the analytics. We propose that tuning this condition through declarative representations would require explicitly informing individuals regarding the data behind the analytics. Providing greater transparency pertaining to the inputs of the analytics could help individuals know what variables are included. Individuals may then find

the contextual information that they were considering is either already included or needs to be added through their intervention.

The second condition that individuals rely upon to interfere with analytics is social pressure (Petropoulos et al., 2016). Social pressure is often due to targets set by senior managers (Fildes, Goodwin & Önkal, 2007; Lawrence, O'Connor, & Edmundson, 2000). During one of our interviews with a professional engaged in the use of supply chain analytics, she justified many of her overrides of the analytics recommendation because the change would "save the company money." However, not all targets set by senior managers, or the reasoning behind the targets, may be known to supply chain individuals. In the interview, the individual understood the short-term goal of saving the company money but did not know about the long-term cost incurred by overriding the analytics. We argue, if individuals have more knowledge regarding what specific goals of the company are, as related to analytics use, and how their behavior impacts the company holistically, they will be more likely to use analytics optimally.

The third condition that may cause individuals to use analytics non-optimally is to gain control (Petropoulos et al., 2016). A need to gain control may be caused by a lack of trust in the analytics (Önkal & Gönül, 2005; Petropoulos et al., 2016), discomfort (Arunachalam, Kumar, & Kawalek, 2018; Boudreau & Robey, 2005; Kache & Seuring, 2017; Schoenherr & Speir-Pero, 2015), or anxiety (Venkatesh, 2000). These responses may be caused by humans regarding the analytics as "black-boxes" (Önkal & Gönül, 2005). As such, many articles discussing analytics implementation emphasize the importance of change management (e.g., Singh & Del Giudice, 2019). During our interviews, one theme that emerged was the need to, "bring individuals along the process of integrating analytics into supply chain functions" by using a "glass-box, not a black-box." Declarative representations through training may allow for transparency into the

inner workings of the analytics. Additionally, studies have shown that establishing credibility (i.e., increasing trust) in technology is best done through humans, since humans are more likely to listen to humans than to technology (Önkal, Goenuel, & Lawrence, 2008; Onkal, Goodwin, Thomson, Gonul, & Pollock, 2009). Therefore, by providing declarative representations through a human (e.g., training) aimed in establishing credibility in the system, individuals may be more likely to develop declarative knowledge that enables trust in the analytics, and thus more optimal use.

The fourth, and last condition, that individuals use when non-optimally engaging with analytics is misinterpretation of data (Petropoulos et al., 2016). Misinterpretation of data can occur when individuals misread visualizations associated with analytics (Bendoly, 2016). Humans often struggle to correctly recognize patterns and distinguish trends (Blattberg & Hoch, 1990; Eggleton, 1982; Kremer, 2011; Kremer, Siemsen, & Thomas, 2015). Additionally, when engaged in judgment and decision tasks, humans rely on heuristics and biases which are often incorrect (Bendoly et al., 2010; Tversky & Kahneman, 1974). Since judgments and decisions begin with an interpretation of declarative knowledge which is continually refined, providing new declarative knowledge could aid in the formation of new, less-biased, procedures (Anderson 1982). Indeed, Tokar, Aloysius, and Waller (2012) find that providing supply chain managers with declarative knowledge through training reduces biases in an inventory replenishment task.

The tasks completed by an integration of analytics and human judgment are broad (Delen & Demirkan, 2013; Srinivasan & Swink, 2018; Waller & Fawcett, 2013). Therefore, the tasks supply chain individuals work with can be classified as both high- and low-level processing tasks. We predict declarative knowledge disseminated through training will be helpful for both high- and low-level tasks since new declarative knowledge can aid in creating better procedures

(Anderson, 1982). In summary, in the context of a supply chain individual's optimal use of analytics, we propose that declarative knowledge disseminated through training will decrease non-optimal behavior, which in turn, will increase forecasting performance for both high- and low-level processing tasks. Formally:

H1_high*. Declarative knowledge will: a) decrease deviation from optimal behavior and b) improve forecasting performance for high-level processing tasks.*

H1_low. *Declarative knowledge will: a) decrease deviation from optimal behavior and b) improve forecasting performance for low-level processing tasks.*

4.3.2. Hypothesis 2

Dual process theory posits that relying on System 1 reasoning when situations are unfamiliar, high-risk, and timely, is dangerous as, "these are the circumstances in which intuitive errors are probable" (Kahneman, 2011, p. 79). As such, during times of unfamiliarity the override function should be engaged to allow for more deliberate action. However, the override function often remains un-triggered and System 1 defaults to a quick, easy answer relying on heuristics and biases (Toplak, West, & Stanovich, 2011). The propensity and ability to override differs between individuals (Toplak, West, & Stanovich, 2014). As such, heterogeneity in individual performance on judgment and decision tasks may in part be explained by the ability to engage the override function. For example, in the forecasting literature, Moritz, et al. (2013) provide evidence that common biases in an inventory replenishment task can be explained by an individual's performance on the CRT. Additionally, Moritz, et al. (2014) find that individuals who score higher on the CRT (indicating a stronger propensity to engage the override function) have lower forecast errors.

We propose that relying on System 1 for high-level processing tasks may further fuel the four factors that drive individual non-optimal use of analytics. For example, the first two reasons that individuals use analytics non-optimally are contextual information and social pressures. If individuals use System 1 reasoning when interfering with analytics due to contextual information, the individual only has the ability to process and integrate one thing at a time (Kahneman, 2011). The limited information processing available to System 1 therefore lacks the capacity to integrate existing information provided by the analytics with the additional contextual information. In contrast, the information processing available to System 2 is greater as it can "deal with multiple distinct topics at once" (Kahneman, 2011, p. 36). As such, individuals will better process the contextual information in conjunction with other factors when deciding to interfere with the analytics, resulting in more optimal use. Similarly, since System 1 is highly susceptible to biases (Toplak, West, & Stanovich, 2011), if individuals rely on System 1 reasoning when using analytics non-optimally due to social pressures, they will likely be biased in their recollection of targets set by senior managers. However, the switch to System 2 reasoning would allow for greater deliberate thinking and the ability to follow rules more closely (Kahneman, 2011; Toplak, West, & Stanovich, 2014).

The second two reasons that individuals use analytics non-optimally are to gain control and misinterpretation of data. These two situations can be attributed to System 1 reasoning since both errors are rooted in emotions and biases (Kahneman, 2011; Toplak, West, & Stanovich, 2011). In contrast, System 2 is effectively neutral, rational, and less subjective to heuristics and biases (Toplak, West, & Stanovich, 2011). Thus, a switch to analytical thinking should mitigate biases and emotions influencing improper adjustments. Similarly, instances of non-optimal use of analytics due to misinterpretation of data are likely lessened when individuals rely on System

2 reasoning. Indeed, dual process theory suggests that System 2 thinking is better at interpreting statistical information since it is more deliberate, thus relying less on first-glance or quick intuition (Kahneman, 2011).

We do not hypothesize about the effect of analytical training on low-level processing tasks since System 1 reasoning is predominantly used for low-level processing tasks. Based on the assumption that humans are cognitive misers, once individuals resume the low-level processing task after the training and recognize the task does not require much effort, it is expected the individual will switch back to System 1 reasoning (Simon, 1955; Toplak, West, & Stanovich, 2014; Tversky & Kahneman, 1974). Therefore, following dual process theory logic, we propose that activating System 2 reasoning will help individuals think deliberately and thus their use of analytics will be closer to optimal and more accurate when engaged in high-level processing tasks.

H2. *Analytical thinking will: a) decrease deviation from optimal behavior and b)improve forecasting performance for high-level processing tasks.*

4.3.3. Hypothesis 3

According to ACT, behavior is continually refined through productions (i.e., procedural knowledge). However, procedural knowledge can use only active knowledge, or knowledge which is in the working memory (Anderson, 1983, p. 26). Therefore, it follows that the power of declarative knowledge to influence behavior by means of revising procedural knowledge can only effectively be utilized when active in the working memory. In a similar vein, dual process theory appeals to the ability of System 2 to train memory to obey new instructions (Kahneman, 2011, p. 36). Since individuals differ in memory and cognitive ability, actively triggering System 2 in individuals engaged in high-level processing tasks may ensure that declarative knowledge is

activated thus improving the effect of declarative knowledge. For individuals engaged in low-level processing tasks, there are two competing arguments. The first argument follows the same reasoning as high-level processing tasks. Utilizing System 2 processing can actively use declarative knowledge to revise procedural knowledge and create new habits (Anderson, 1983). In contrast, the second argument is based on the assumption that humans are cognitive misers. As such, we can assume individuals will remain in System 1 processing as long as the task is not demanding (Kahneman, 2011; Toplak, West, & Stanovich, 2014). Therefore, analytical thinking will not improve the behavior and performance on a task that could be completed with System 1 processing.

H3_high*. Analytical thinking will strengthen the effect of declarative knowledge to a) decrease deviation from optimal behavior and b) improve forecasting performance for high-level processing tasks.*

H3_low*. Analytical thinking will strengthen the effect of declarative knowledge to a) decrease deviation from optimal behavior and b) improve forecasting performance for low-level processing tasks.*

4.4. Design and Implementation

Demand planning is a common function in supply chain that relies on analytics. The company we worked with discussed their desire to decrease human overrides of the analytical system. Therefore, we test the hypotheses using an experiment in the demand planning context.

4.4.1. Experimental Design

We use a 2 (no declarative knowledge, declarative knowledge) x 2 (no analytical thinking, analytical thinking) x 2 (interactive machine learning, human-guided machine learning) experiment with human subjects. The trainings are described in Section 4.4.2, below. Interactive

machine learning (IML) is a method of forecasting that uses participant input of the estimated magnitude of change to demand, $Magnitude$ (Equation 1):

(1) $IML_{it} = \beta_0 + \beta_1 Period_{it} + \beta_2 Magnitude_{it}$

where: i = participant; t = period

IML is optimally used when humans input a quantity from the specified demand distribution during shocked periods only. Human-guided machine learning (HGML) uses a cue, Cue, rather than an amount, to indicate whether the following period includes a special event (Equation 2):

(2) $HGML_{it} = \beta_0 + \beta_1 Period_{it} + \beta_2 Cue_{it}$

HGML is optimally used when the participant indicates to the model that there is a special event when there is a special event. Otherwise, the participant should indicate that there are no changes.

The two dependent variables of interest are optimal behavior and mean absolute error (MAE). We calculate optimal behavior (OB) as:

(3) OB = $\frac{\#\ correct\ periods}{n} * 100$

where: n = total number of periods

Correct periods are defined as the periods that the participant correctly identifies whether or not it is a shocked period. Additionally, the measure of accuracy we use is MAE since we are comparing forecast accuracy across the same scale of data (Hyndman & Koehler, 2006). Thus, the equation used to calculate MAE is:

(4) MAE = $\frac{\sum_{i=1}^{n} |Actual\ demand_{it} - Forecast_{it}|}{n}$

4.4.2. Program and Task

We code the application to test our design using oTree, an open-source Python framework used to create interactive behavioral economic experiments (Chen, Schonger, & Wickens, 2016). The forecasting task for Essay 3 is similar to the task used in Essay 1. When participants initiate the program, they are randomly assigned to a treatment. The participant then completes two training periods (one with no shock, one with a shock) and begins the actual task. The actual task begins with 10 periods of historical data, so the participant is familiarized with the trend line pre-forecast. The actual task is a total of 40 periods (10 periods historical data and 30 periods of forecasting). Beginning in period 11-40, a series of shocks from the randomly-generated probability distribution results in upward shifts to demand. Throughout the task participants are shown a graph of actual demand as well as forecasts. (See 2.9 Appendix A and 4.7 Appendix A for details on the program and shock distribution.) Upon completing periods 11-30, the participants are shown their assigned intervention. Following the intervention, the participants complete 10 periods of forecasting. At the end of the task, participants are shown the CRT-2 (Thomson & Oppenheimer, 2016), an updated version of the CRT which is not as likely to be familiar to the participants compared to the CRT, and demographic questions (e.g., age, gender).

To test a complete factorial of the two interventions, we have four groups per method of forecasting. The first group receives only the declarative knowledge training. The declarative knowledge training provides information regarding the model forecast, when to change the model forecast, and how to be the most accurate. The participants are then asked, “when should you change the model forecast,” to ensure they understand the information. The second group does not receive the declarative knowledge training but does receive the analytical thinking

training. The analytical thinking training follows the task used in Anseel, Lievens, and Schollaert (2009) to stimulate System 2 reasoning. The participants are shown the graph documenting actual demand and the forecast so they can see forecast accuracy. The participants are then shown a feedback report with the definition of forecast accuracy, how an expert would do, and their score. Following the feedback report, participants are asked four questions: 1) What periods were you the most accurate?; 2) Why do you think you did well in those periods?; 3) What periods were you the least accurate?; and 4) Why do you think you did poorly in those periods?. The third group receives both the declarative knowledge and analytical thinking training. The fourth group acts as the control group and does not receive an intervention. Instead, the fourth group sees the question "Are you ready to continue?" and must type "Yes" in the box.

4.4.3. Pilot Tests

Prior to the data collection, we implemented a pilot test to ensure participants understood the instructions and were properly incentivized. The pilot test revealed the instructions in the training periods were not clear enough for the participants to fully understand the task. We updated the instructions to be clearer, and we also added precautions so participants could only continue to the actual task if they answered all the training questions correctly. Following the update to the instructions, we conducted a second pilot test to ensure the instructions were clear.

4.4.4. Data Collection and Sample

Following the pilot tests, we conducted the experiment using undergraduate students at a large American private university. Following Thomas (2011), university students are an appropriate context for our experiment. The students were enrolled in an introductory business class. We collected responses from 611 students. We eliminated seven participants who participated in the study twice. We used a lottery performance-based incentive (Wakker, 2007).

The performance-based payment was included to increase validity following the logic from the induced-value theory in behavioral economics (Smith, 1976). Upon beginning the task, participants were told five completed tasks would be selected at random to win the amount they earned. All participants were shown the amount they could possibly win. The payoff was calculated by:

(5) Total Payout = $1+[$20-(0.20*|MAE|)]

The average bonus for our collected sample is $17.37 out of the possible $20. The first 10 forecast periods allow for learning effects and training the model. To conduct our analysis, we use the average of the shocked periods 20-29 (pre-intervention) and the average of the 31-40 (post-intervention) for each subject. Thus, the number of observations in the sample is 1,210, or two repeated measures of 605 observations. The number of observations per treatment are all greater than 70 (Table 4.1).

Table 4.1. Sample Size per Training for Methods of Forecasting.

Interactive Machine Learning		
	Analytical thinking	None
Declarative knowledge	73	76
None	88	71
Human-guided Machine Learning		
	Analytical thinking	None
Declarative knowledge	71	74
None	80	71

4.5 Results

Table 4.2, below, summarizes descriptive statistics for each method of forecasting and training. Recall, optimal behavior (OB) is 100. Generally inspecting the means reveals that on average the behavior (OB) has increased towards optimal and the forecasting performance has improved (MAE decreased) between pre- and post-intervention training groups.

Table 4.2. Descriptive Statistics for Pre- and Post-Intervention Means of Optimal Behavior (OB) and Mean Absolute Error (MAE) for Methods of Forecasting.

			Optimal Behavior		Mean Absolute Error	
Method of forecasting	Training	Intervention	Mean	Std. dev.	Mean	Std. dev.
Interactive machine learning	Declarative knowledge	Pre	66.58	40.58	9.85	6.18
		Post	78.95	36.21	9.36	5.11
	Analytical thinking	Pre	63.41	43.34	9.69	5.64
		Post	68.41	43.71	9.54	4.79
	Both	Pre	51.23	44.09	12.52	8.50
		Post	62.74	46.11	10.48	5.19
	None	Pre	52.96	44.48	11.98	5.24
		Post	56.34	47.25	11.49	4.89
Human-guided machine learning	Declarative knowledge	Pre	62.43	34.71	10.42	5.37
		Post	86.22	26.47	8.90	4.41
	Analytical thinking	Pre	69.75	30.36	11.11	5.76
		Post	72.00	31.11	10.16	4.83
	Both	Pre	77.18	27.94	9.30	5.08
		Post	84.51	26.23	8.41	3.88
	None	Pre	75.21	32.15	9.29	5.58
		Post	75.21	31.98	9.49	5.00

We use a repeated measures general linear model (GLM) to test the between- and within-effects of the trainings for each method of forecasting. We include both dependent variables, OB and MAE, in the model. As such, training is the between-subjects factor and each participant's average pre-intervention measure of OB and MAE, and average post-intervention measure of OB and MAE, are the within-subject factors. The repeated measures GLM is grouped by method of forecasting. Table 4.3 summarizes the results of the repeated measures GLM. Since the repeated measures GLM is conducted using only two repeated measures, the assumption of sphericity is met (Field, 2013).

Table 4.3. Repeated Measures GLM for Optimal Behavior (OB) and Forecasting Performance (MAE).

			Optimal behavior (OB)		Performance (MAE)	
Method of forecasting	Factor	Type	df	F	df	F
Interactive machine learning (IML)	Intercept	B	1	718.59***	1	1290.84***
	Declarative knowledge (DK)	B	1	0.97	1	0.04
	Analytical thinking (AT)	B	1	0.23	1	0.04
	DK*AT	B	1	8.39***	1	11.58***
	Intervention	W	1	24.06***	1	7.12***
	Intervention*DK	W	1	5.55**	1	2.52
	Intervention*AT	W	1	0.01	1	1.02
	Intervention*DK*AT	W	1	0.14	1	2.54
	Error	B	304		304	
Human-guided machine learning (HGML)	Intercept	B	1	2379.15***	1	1304.79***
	Declarative knowledge (DK)	B	1	2.16	1	1.99
	Analytical thinking (AT)	B	1	0.13	1	0.17
	DK*AT	B	1	3.09*	1	3.70*
	Intervention	W	1	24.19***	1	10.65***
	Intervention*DK	W	1	18.10***	1	2.93*
	Intervention*AT	W	1	4.39**	1	0.3
	Intervention*DK*AT	W	1	7.61***	1	3.36*
	Error	B	292		292	

*p < 0.10, **p < 0.05, ***p < 0.01

4.5.1 Effectiveness of the Trainings

We first examine the test of within-subject effects from the repeated measures GLM to test the efficacy of the training using both dependent measures, behavior (OB) and forecasting performance (MAE). We address the results of OB first, followed by MAE.

The results reveal the OB between pre- and post-interventions are significantly different for both IML ($F(1, 304) = 24.06$, $p < 0.001$) and HGML ($F(1,292) = 24.19$, $p < 0.001$). To understand the effect of declarative knowledge, analytical thinking, and both trainings together, between pre- and post-intervention we focus on the interactions (Intervention*DK,

Intervention*AT, and Intervention*DK*AT). Figures 4.2A-F below show the interaction plots for within-subjects effects.

Starting first with IML, the interaction between intervention and declarative knowledge is significant (IML: $F(1,304) = 5.55$, $p = 0.02$). Since the interaction is significant, we next observe the pairwise comparisons to evaluate the differences between pre- and post-intervention (Table 4). The pairwise comparisons reveal that there is a positive and significant difference between pre- and post-intervention OB for those who received the declarative knowledge training (IML: $\overline{\text{Post} - \text{Pre}} = 12.37$, $p < 0.001$). The positive sign indicates that the OB is better post-intervention. Next, the pairwise comparison between pre- and post-intervention OB for the analytical thinking training indicates a positive and significant difference (IML: $\overline{\text{Post} - \text{Pre}} = 5$, $p = 0.06$). The training with both analytical thinking and declarative knowledge also positively and significantly differs between pre- and post-intervention (IML: $\overline{\text{Post} - \text{Pre}} = 11.51$, $p = 0.006$). Lastly, as predicted, there is no significant difference in behavior (OB) between the pre- and post-intervention groups for those who did not receive training (IML: $\overline{\text{Post} - \text{Pre}} = 3.38$, $p = 0.23$).

Turning to HGML, the interaction between intervention and declarative knowledge is significant (HGML: $F(1,292) = 18.10$, $p < 0.001$). As such, we examine the pairwise comparisons (Table 4.4). The mean difference between pre- and post-intervention behavior (OB) for those who received the declarative knowledge training is positive and significant (HGML: $\overline{\text{Post} - \text{Pre}} = 23.78$, $p < 0.001$) as well as for those who received both analytical thinking and declarative knowledge training (HGML: $\overline{\text{Post} - \text{Pre}} = 7.33$, $p = 0.03$). The difference between those who received analytical thinking is not significant (HGML: $\overline{\text{Post} - \text{Pre}} = 2.25$, $p = 0.48$) suggesting the behavior is not different after the analytical thinking training. Additionally, as

predicted, there is no difference between pre- and post-intervention behavior for those who did not receive training (HGML: $\overline{\text{Post} - \text{Pre}} = 0$, p = 1.0).

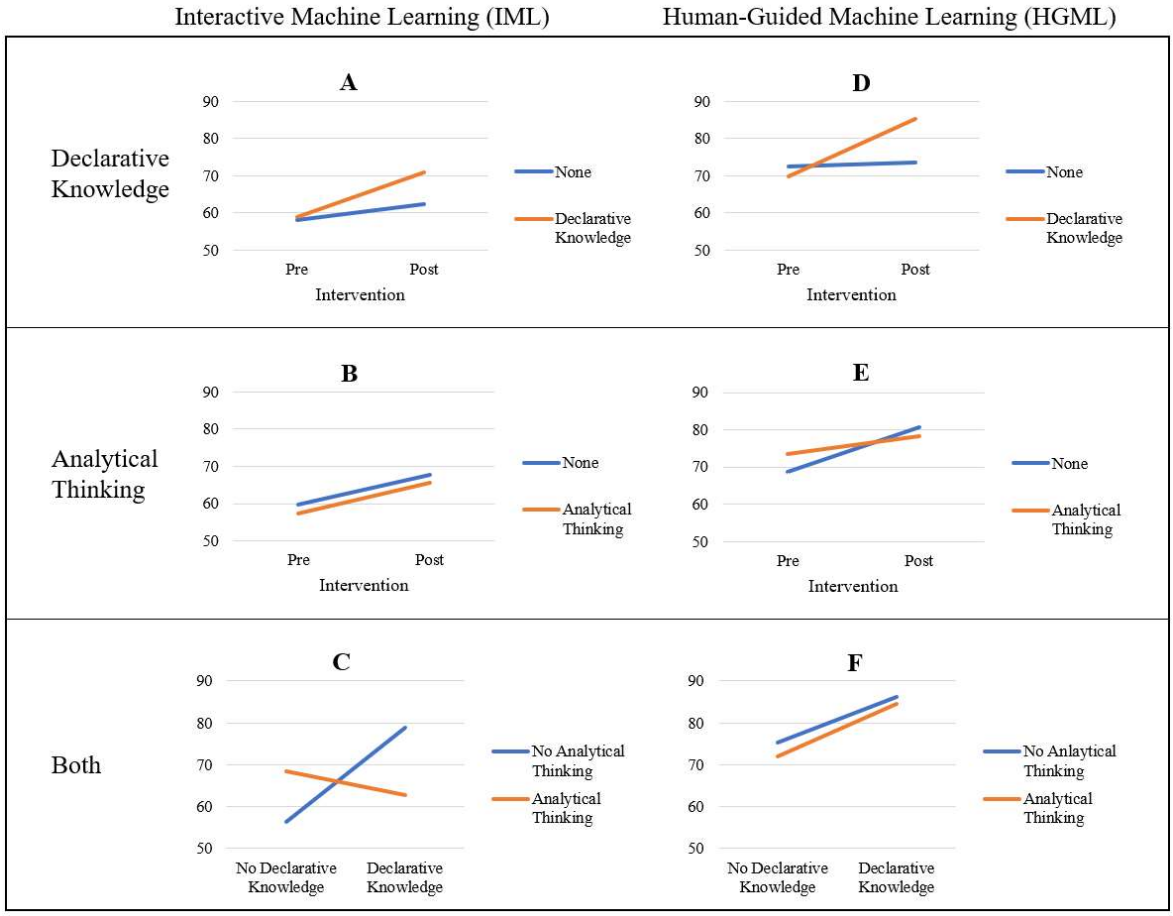


Figure 4.2. Interaction Plots of Within-Subject Effects for Optimal Behavior (OB). Note: The Y-Axis of all plots is Estimated Marginal Means of Optimal Behavior (OB).

The results reveal the MAE between pre- and post-interventions are significantly different for both IML ($F(1, 304) = 10.65$, $p < 0.001$) and HGML ($F(1,292) = 7.12$, $p = 0.01$). To understand the effect between pre- and post-intervention of declarative knowledge, analytical thinking, and both trainings jointly on forecasting performance, we test the interactions (Intervention*DK, Intervention*AT, and Intervention*DK*AT). Figures 4.3A-F below show the interaction plots for within-subjects effects.

Table 4.4. Pairwise Comparisons for Within-Subjects Effects.

Method of forecasting	Post – Pre Intervention	Optimal Behavior (OB)		Forecasting Performance (MAE)	
		Mean diff. (Post-Pre)	Std. Error	Mean diff. (Post-Pre)	Std. Error
Interactive machine learning (IML)	Declarative knowledge	12.37***	3.55	-0.49	0.60
	Analytical thinking	5.00*	2.65	-0.15	0.38
	Both	11.51***	4.07	-2.03**	0.84
	None	3.38	2.78	-0.49	0.51
Human-guided machine learning (HGML)	Declarative knowledge	23.78***	4.09	-1.51***	0.59
	Analytical thinking	2.25	3.18	-0.95**	0.44
	Both	7.33**	3.36	-0.89**	0.43
	None	0.00	2.75	0.2	0.44

*p < 0.10, **p < 0.05, ***p < 0.01

Beginning with IML, none of the interactions are statistically significant for forecasting performance (MAE). However, the pairwise comparisons indicate that there is a negative and significant difference between pre- and post-intervention MAE for subjects who received both analytical thinking and declarative knowledge training (IML: $\overline{\text{Post} - \text{Pre}}$ = -2.03, p = 0.02). The negative sign indicates a lower MAE after the analytical thinking and declarative knowledge training. The MAE was not different post-intervention for declarative knowledge alone (IML: $\overline{\text{Post} - \text{Pre}}$ = -0.49, p = 0.42), analytical thinking alone (IML: $\overline{\text{Post} - \text{Pre}}$ = -0.15, p = 0.70), or no training (-0.49, p = 0.34).

Considering HGML, the interaction between intervention and declarative knowledge is significant (HGML: F(1,292) = 2.93, p = 0.09). The pairwise comparison of pre- and post-intervention MAE for declarative knowledge reveals a negative and significant difference (HGML: $\overline{\text{Post} - \text{Pre}}$ = -1.51, p = 0.01). Additionally, analytical thinking (HGML: $\overline{\text{Post} - \text{Pre}}$ = -0.95, p = 0.03), and analytical thinking with declarative knowledge (HGML: $\overline{\text{Post} - \text{Pre}}$ = -0.89, p = 0.04), are negatively and significantly different, whereas no training is not significant

(HGML: $\overline{\text{Post} - \text{Pre}} = 0.2$, $p = 0.65$). Therefore, in all three groups, MAE improved after training.

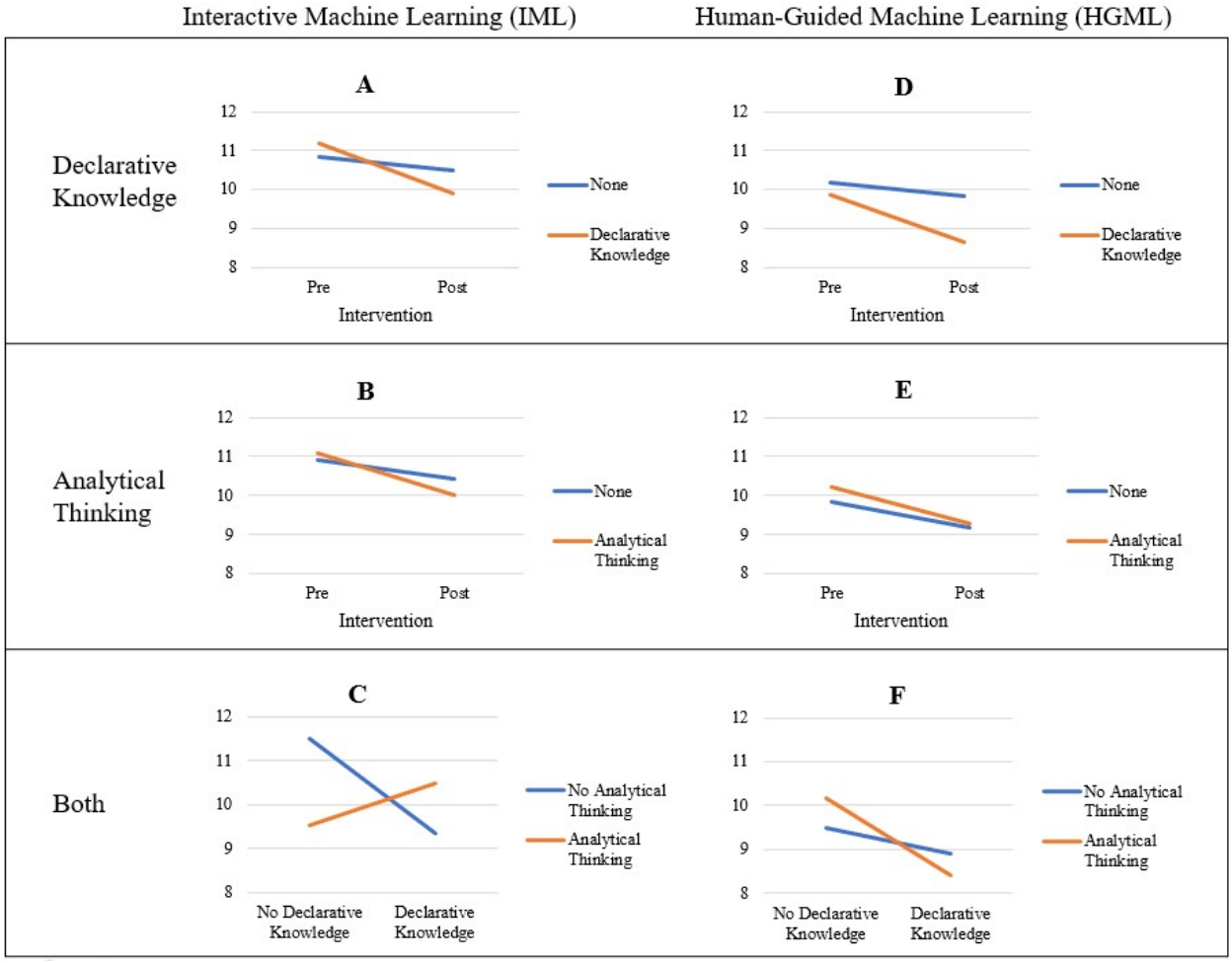


Figure 4.3. Interaction Plots of Within-Subject Effects for Forecasting Performance. (MAE). Note: The Y-Axis of all plots is Estimated Marginal Means of Mean Absolute Error (MAE).

In summary, when comparing pre- and post-intervention OB for IML, the most improvement is seen using declarative knowledge alone or declarative knowledge and analytical trainings together. When comparing pre- and post-intervention MAE for IML, the greatest improvement is experienced by using declarative knowledge and analytical thinking together. For HGML, when comparing pre- and post-intervention OB and MAE, declarative knowledge training results in the best improvement for both.

We now turn our focus to the between-treatment effects for the tests of the **H1-H3**. Interpreting the test of between-subject effects from the repeated measures GLM reveals that the OB and MAE between trainings are significantly different for the interaction between declarative knowledge and analytical thinking for both IML (OB: $F(1,304) = 8.39$, $p = 0.004$; MAE: $F(1,304) = 11.58$, $p = 0.001$) and HGML (OB: $F(1,292) = 3.09$, $p = 0.08$; MAE: $F(1,292) = 3.70$, $p = 0.06$). To gain additional insight into the differences between trainings we examine the pairwise comparisons (Table 4.5).

Table 4.5. Pairwise Comparisons of Optimal Behavior (OB) and Forecasting Performance (MAE).

		Optimal Behavior (OB)			Forecasting Performance (MAE)		
Method of forecasting	$Training_1 - Training_2$	Mean diff.	Std. error	Hypothesis	Mean diff.	Std. error	Hypothesis
Interactive machine learning	Declarative knowledge-None	8.99*	4.75	H1a	0.01	0.86	H1b
	Analytical thinking-None	1.62	4.59	H2a	0.35	0.83	H2b
	Both-None	8.13*	4.79	H3a	-1.54*	0.86	H3b
Human-guided machine learning	Declarative knowledge-None	23.78***	4.84	H1a	-1.71***	0.69	H1b
	Analytical thinking-None	2.25	4.75	H2a	-1.15*	0.68	H2b
	Both-None	7.32	4.89	H3a	-1.09	0.7	H3b

*p < 0.10, **p < 0.05, ***p < 0.01

4.5.2 Optimal Behavior

Our first hypothesis**, H1a_high**, declarative knowledge will decrease deviation from optimal behavior for high-level processing tasks, is supported (IML: $\overline{\text{Declarative knowledge} - \text{None}} = 8.99$, $p = 0.06$. Additionally, **H1a_low**, declarative knowledge will decrease deviation from optimal behavior for low-level processing tasks, is supported (HGML: $\overline{\text{Declarative knowledge} - \text{None}} = 23.78$, $p < 0.001$). **H2a**, analytical thinking will

decrease deviation from optimal for high-level processing tasks, is not supported (IML: $\overline{\text{Analytical Thinking} - \text{None}}$ = 1.62, p = 0.72). The non-significance may be explained by the within-subject effects since there is not a difference in OB after the analytical thinking training for IML. A possible theoretical explanation of this finding follows in the subsequent discussion section.

Our third group of hypotheses consider the impact of analytical thinking on declarative knowledge. **H3a_high**, analytical thinking will strengthen the effect of declarative knowledge to decrease deviation from optimal behavior for high-level processing tasks, is supported (IML: $\overline{\text{Both} - \text{None}}$ = 8.13, p = 0.09). Theory indicated that **H3a_low**, analytical thinking will strengthen the effect of declarative knowledge to improve forecasting performance for low-level processing tasks, was an empirical issue. As such, we tested **H3a_low** as an exploratory hypothesis. The results reveal analytical thinking does not strengthen the effect of declarative knowledge in low-level processing tasks (HGML: $\overline{\text{Both} - \text{None}}$ = 7.32, p = 0.14). Again, we discuss explanations and implications of this finding in the discussion section.

4.5.3 Forecasting Performance

H1b_high, declarative knowledge will improve forecasting performance for high-level processing tasks, is not supported (IML: $\overline{\text{Declarative knowledge} - \text{None}}$ = 0.01, p = 0.86). The comparison between those who received the declarative knowledge training and those who did not receive the training when using HGML results in a negative and significant difference (HGML: $\overline{\text{Declarative knowledge} - \text{None}}$ = -1.71, p = 0.01). This finding provides evidence that declarative knowledge improves forecasting performance for low-level processing tasks. Therefore **H1b_low** is supported. **H2b_high**, analytical thinking will improve forecasting

performance for high-level processing tasks, is not supported (IML: $\overline{\text{Analytical Thinking} - \text{None}}$ = 0.35, p = 0.68).

The results of testing **H3b_high**, analytical thinking will strengthen the effect of declarative knowledge for high-level processing tasks, report a negative and significant difference (IML: $\overline{\text{Both} - \text{None}}$ = -1.54, p = 0.08). Thus, since the group that received both analytical thinking and declarative knowledge trainings has a significantly lower MAE than the control group, **H3b_high** is supported. We also test our exploratory hypothesis **H3b_low.** The comparison between groups that receive both analytical thinking and declarative knowledge trainings and the control group is not significant (HGML: $\overline{\text{Both} - \text{None}}$ = -1.09, p = 0.12).

4.5.4 Robustness Check

One of the fundamental assumptions of our research is that the propensity and ability to override System 1 reasoning differs between individuals (Toplak, West, & Stanovich, 2014). In order to fully test our hypotheses regarding triggering analytical thinking through training, we collect the subjects' responses to the CRT-2 as students are less likely to be familiar with the CRT-2 questions (Thomson & Oppenheimer, 2016). We then re-estimate the repeated measure GLMs including both dependent variables (OB and MAE) with the control variables gender, age, CRT-2. The repeated measures GLM between pre- and post-interventions including the controls supports the findings in the main analysis (Table 4.6).

Table 4.6. Robustness Check.

Dependent Variable			Optimal Behavior (OB)		Forecasting Performance (MAE)	
Method of forecasting	Factor	Type	df	F	df	F
Interactive machine learning (IML	Intercept	B	1	7.201***	1	11.91***
	CRT-2	B	1	1.24	1	6.26***
	Age	B	1	0.737	1	0
	Gender	B	1	0.08	1	0.46
	Declarative Knowledge (DK)	B	1	0.66	1	0
	Analytical Thinking (AT)	B	1	0.26	1	0.02
	DK*AT	B	1	8.07***	1	10***
	Intervention	W	1	0.23	1	1.53
	Intervention*CRT-2	W	1	0.07	1	0.08
	Intervention*Age	W	1	0.02	1	2.56
	Intervention*Gender	W	1	1.6	1	1.64
	Intervention*DK	W	1	5.08**	1	2.08
	Intervention*AT	W	1	0.02	1	1
	Intervention*DK*AT	W	1	0.2	1	2.69
	Error	B	301		301	
Human-guided machine learning (HGML)	Intercept	B	1	7.98***	1	22.24***
	CRT-2	B	1	1.16	1	5.07**
	Age	B	1	0.19	1	1.61
	Gender	B	1	1.83	1	1.56
	Declarative Knowledge (DK)	B	1	2.14	1	1.95
	Analytical Thinking (AT)	B	1	0.08	1	0.29
	DK*AT	B	1	3.33*	1	4.20**
	Intervention	W	1	1.14	1	0.30
	Intervention*CRT-2	W	1	0.12	1	1.64
	Intervention*Age	W	1	0.16	1	0.10
	Intervention*Gender	W	1	2.69	1	5.66**
	Intervention*DK	W	1	17.82***	1	3.06*
	Intervention*AT	W	1	3.99**	1	0.33
	Intervention*DK*AT	W	1	8.02***	1	3.65*
	Error	B	289		289	

*p < 0.10, **p < 0.05, ***p < 0.01

4.6 Discussion and Conclusion

This study seeks to develop training strategies using theories based in cognitive psychology and to test the training for its effectiveness in decreasing the deviation from optimal behavior and improving forecasting performance. Our first group of hypotheses, H1, predicts

that declarative knowledge will decrease deviation from optimal behavior and improve forecasting performance in both high- and low-level processing tasks. Our second group of hypotheses, H2, predicts that analytical thinking will decrease deviation from optimal behavior and improve forecasting performance in high-level processing tasks. Our third group of hypotheses predicts that analytical thinking will strengthen the use of declarative knowledge to decrease deviation from optimal and improve forecasting performance in high-level processing tasks.

H1a is supported across both high- (**H1a_high**) and low-level (**H1a_low**) processing tasks showing strong support for the role of declarative knowledge in aligning behavior closer to optimal. However, the results offer evidence that the forecasting performance of only low-level processing tasks is improved by declarative training (**H1b_low**). One possible explanation for this finding is the difference between the mechanics of IML and HGML. Recall, IML requires the individual to estimate the magnitude of the change to demand due to a special event. The magnitude provided by the individual is then weighted by the IML model according to accuracy. Thus, it is possible that the IML model does not have enough time to evaluate the updated estimated magnitudes and to assign an appropriate weight. In contrast, HGML does not require an estimation of magnitude from the individual, the model calculates the magnitude based on historical data of special events. Therefore, once the individual's behavior increases optimality, the model can quickly adapt to estimate the magnitude of what has occurred in the past. We predict over a longer period of time the forecasting performance of IML would improve.

The next group of hypotheses predict that analytical training improves the overall behavior and performance for high-level processing tasks. The results do not provide evidence that analytical training improves behavior and performance for high-level processing tasks

(IML). We believe this finding can be explained by the assumption that humans are cognitive misers (Simon, 1955; Toplak, West, & Stanovich, 2014; Tversky & Kahneman, 1974). Recall that in the experiment, although IML is a high-level processing task, humans acclimate, and over time, instinctively create habits to complete tasks as efficiently as possible (Toplak, West, & Stanovich, 2014). Since humans naturally gravitate towards the processing system that requires relatively less effort—System 1—it follows that without any new information or disruption, individuals form habits and complete the forecasting task more automatically.

The third hypothesis unifies declarative knowledge and analytical thinking. Specifically, **H3** predicts that analytical thinking strengthens the effect of declarative knowledge in decreasing deviation from optimal behavior and MAE. **H3a_high** and **H3b_high** are supported for IML providing evidence that analytical thinking strengthens the effect of declarative knowledge for high-level processing tasks. This finding suggests that by actively triggering System 2 reasoning in individuals engaged in high-level processing tasks, the new information provided in the declarative knowledge training is used to revise procedural knowledge. Since the declarative knowledge training teaches new practices, and individuals engaged in a high-level processing task are likely already managing other distinct cognitive processes (i.e., estimation of magnitude), individuals are best able to handle all the processes when relying on System 2. Rather than allowing System 2 processing to remain untriggered, the analytical thinking training prompts individuals to make the switch to System 2 processing. **H3a_low** and **H3b_low** are exploratory hypotheses because theory does not support a clear prediction. We find no evidence that analytical thinking enhances the declarative knowledge training for low-level processing tasks. This finding is likely due to the tendency to stay with System 1 thinking as often as possible. Once the training is complete, and participants return to the low-level processing task,

the participants are no longer dealing with multiple processes, and can therefore revert to System 1.

The overarching findings of this research are summarized by method of forecasting in Table 4.7. For high-level processing tasks, it is best to use both declarative knowledge and analytical thinking in the training. The analytical thinking component of the training seems to encourage individuals to engage their override function, despite their particular propensity to switch from System 1 to System 2. By engaging System 2 reasoning, individuals are then better able to process multiple types of information more effectively and accurately (Kahneman, 2011; Toplak, West, & Stanovich, 2011; Toplak, West, & Stanovich, 2014). On the other hand, tasks that require less deliberate thought are better-off implementing trainings using declarative knowledge.

Table 4.7. Summary of Results for Hypotheses.

Hypotheses		High-level processing	Low-level processing
H1	a.	Supported	Supported
	b.	Not supported	Supported
H2	a.	Not supported	N/A
	b.	Not supported	N/A
H3	a.	Supported	Exploratory
	b.	Supported	Exploratory

Although our research provides important insights into the cognitive mechanics behind judgment and decision-making behavior, there are two potential limitations to consider. First, our experiment relies on a student sample rather than practitioners. Since the purpose of our study is to understand the common cognitive elements embedded in basic human judgment and decision making, student subjects are a viable (Lonati et al., 2018), and perhaps preferred (Thomas, 2011) sample. However, now that the complex cognitive underpinnings of the training have been verified in a carefully controlled environment, we encourage future research to test the training

in a natural environment with practitioners. A second potential limitation is that our experiment is a total of 30 periods (10 training, 10 pre-intervention, and 10 post-intervention). Future research could consider testing the effects of declarative knowledge and analytical thinking training over a longer span of time.

This research has three main contributions. First, we combine two theories—ACT and dual process theory—to address a possible cause of heterogeneity in individual's judgments. By using the theories together, we find evidence that training using analytical thinking is one way to help individuals better access the cognitive system for higher-level processing. Thus, declarative knowledge becomes more useful during uncertain times or complex tasks when the individual has engaged the override function and is processing the declarative knowledge with System 2 thinking. Second, we provide nuanced insight into how and why declarative knowledge impacts judgment and decision making (Tokar, Aloysius, & Waller, 2012). Specifically, we find declarative knowledge training has more of an impact on improving behavior and performance when paired with analytical thinking for tasks that require a higher information-processing capacity. Third, we provide empirical evidence regarding best practice for the use of supply chain analytics. Although human judgment is an important component that can be included in forecasting functions, there is a proper time and place.

We encourage industry to include analytical thinking training in combination with their current training (typically instructional and based on declarative knowledge). Table 4.8 contains a summary for which types of trainings should be considered when implementing analytics. Declarative knowledge trainings are important for both high- and low-level processing tasks where humans must interact with analytics. The declarative knowledge trainings should provide specific guidance as to when overriding the analytical system is appropriate. For example,

Petropoulos, Fildes, and Goodwin (2016) suggest the circumstances when it is most appropriate to override the analytical system are times when there is contextual information outside the analytical system.

Table 4.8. Suggestions for Training per Method of Forecasting.

Level of Processing Required For Task	Suggested Training
High	Rely on both declarative knowledge and analytical thinking training to decrease deviation from optimal behavior and improve performance.
Low	Rely on declarative knowledge training to decrease deviation from optimal behavior and improve performance.

In conclusion, supply chain analytics are an exciting advancement in forecasting and the best level of value creation is captured when individuals are brought along to be used conjointly with the analytics. This study provides evidence for effective elements of training—declarative knowledge and analytical thinking—that aid humans in navigating their evolving roles. The integration of analytics and human behavior is an interesting and important avenue future research will hopefully continue to explore.

4.7 References

Anderson, J.R., 1976. *Language, Memory, and Thought*. Hillsdale, New Jersey.

Anderson, J. R., 1982. Acquisition of cognitive skill. *Psychological Review*, 89: 369-403.

Anderson, J.R., 1983. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA.

Anderson, J. R., 1993. *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Anseel, F., Lievens, F. & Schollaert, E., 2009. Reflection as a strategy to enhance task performance after feedback. *Organizational Behavior and Human Decision Processes*, 110(1): 23-35.

Arunachalam, D., Kumar, N. & Kawalek, J.P., 2018. Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E: Logistics and Transportation Review*, 114: 416-436.

Arvan, M., Fahimnia, B., Reisi, M. & Siemsen, E., 2019. Integrating human judgement into quantitative forecasting methods: A review. *Omega*, 86: 237-252.

Ball, G.P., Shah, R. & Donohue, K., 2018. The decision to recall: A behavioral investigation in the medical device industry. *Journal of Operations Management*, 62: 1-15.

Bendoly, E., 2016. Fit, bias, and enacted sensemaking in data visualization: frameworks for continuous development in operations and supply chain management analytics. *Journal of Business Logistics*, 37(1): 6-17.

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

Blattberg, R.C. & Hoch, S.J., 1990. Database models and managerial intuition: 50% model+ 50% manager. *Management Science*, 36(8): 887-899.

Boudreau, M.C. & Robey, D., 2005. Enacting integrated information technology: A human agency perspective. *Organization Science*, 16(1): 3-18.

Brynjolfsson, E. & Mcafee, A.N.D.R.E.W., 2017. The business of artificial intelligence. *Harvard Business Review*: 1-20.

Chandrasekaran, A., S de Treville, & Browning, T. 2020. Editorial: Intervention-based research (IBR)—What, where, and how to use it in operations management. *Journal of Operations Management*: 1-9.

Chen, D.Q., Preston, D.S. & Swink, M., 2015. How the use of big data analytics affects value creation in supply chain management. *Journal of Management Information Systems*, 32(4): 4-39.

Chen, D.L., Schonger, M. & Wickens, C., 2016. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9: 88-97.

Choi, T.M., Wallace, S.W. & Wang, Y., 2018. Big data analytics in operations management. *Production and Operations Management*, 27(10): 1868-1883.

Choo, A.S., Nag, R. & Xia, Y., 2015. The role of executive problem solving in knowledge accumulation and manufacturing improvements. *Journal of Operations Management*, 36: 63-74.

Delen, D. & Demirkan, H., 2013. Data, information and analytics as services, *Decision Support Systems,* 55: 359–363.

Eggleton, I.R., 1982. Intuitive time-series extrapolation. *Journal of Accounting Research*: 68-102.

Evans, J.S.B. & Stanovich, K.E., 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3): 223-241.

Field, A. 2013. *Discovering Statistics Using IBM SPSS Statistics*. Sage.

Fildes, R., Goodwin, P. & Önkal, D., 2019. Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, 35(1): 144-156.

Frederick, S., 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4): 25-42.

Hyndman, R.J. & Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4): 679-688.

Kache, F. & Seuring, S., 2017. Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. *International Journal of Operations & Production Management,* 37(1):10-36.

Kahneman, D., 2011. *Thinking, Fast and Slow*. Macmillan.

Kahneman, D. & Frederick, S., 2002. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 49: 81.

Kim, Y.J. & Zhong, C.B., 2017. Ideas rise from chaos: Information structure and creativity. *Organizational Behavior and Human Decision Processes*, 138: 15-27.

Kremer, J., 2011. *Using Statistics to Plan*. Book Market.

Kremer, M., Siemsen, E. & Thomas, D.J., 2016. The sum and its parts: Judgmental hierarchical forecasting. *Management Science*, 62(9): 2745-2764.

Lawrence, M., O'Connor, M. & Edmundson, B., 2000. A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, 122(1): 151-160.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. 2018. On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, 64: 19-40.

Marchand, D.A. & Peppard, J., 2013. Why IT fumbles analytics. *Harvard Business Review*, 91(1): 104-112.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. & Barton, D., 2012. Big data: the management revolution. *Harvard Business Review*, 90(10): 60-68.

Mohr, N. and Hürtgen, H., 2018. *Achieving Business Impact with Big Data*. McKinsey, 15.

Moritz, B. B., Hill, A. V., & Donohue, K. L., 2013. Individual differences in the newsvendor problem: Behavior and cognitive reflection. *Journal of Operations Management*, 31(1-2): 72-85.

Moritz, B., Siemsen, E., & Kremer, M., 2014. Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management*, 23(7): 1146-1160.

Narayanan, A. & Moritz, B.B., 2015. Decision making and cognition in multi-echelon supply chains: An experimental study. *Production and Operations Management*, 24(8): 1216-1234.

Okhuysen, G., & Bonardi, J. P., 2011. The challenges of building theory by combining lenses. *Academy of Management Review,* 36(1): 6-11.

Oliva, R., 2019. Intervention as a research strategy. *Journal of Operations Management*, 65(7): 710-724.

Önkal, D., Gönül, M.S. & Lawrence, M., 2008. Judgmental adjustments of previously adjusted forecasts. *Decision Sciences*, 39(2): 213-238.

Onkal, D. & Gonul, M.S., 2005. Judgmental adjustment: a challenge for providers and users of forecasts. *Foresight: The International Journal of Applied Forecasting*, 1: 13-17.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S. & Pollock, A., 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, 22(4): 390-409.

Petropoulos, F., Fildes, R. & Goodwin, P., 2016. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*, 249(3): 842-852.

Ryle, G., 1949. The concept of mind Hutchinson. *London, UK*.

Schoenherr, T. & Speier-Pero, C., 2015. Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics*, 36(1): 120-132.

Shynkarkuk J.M., & Thompson V.A., 2006. Confidence and accuracy in deductive reasoning. *Memory Cognition,* 34:619–632

Simon, H.A., 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1): 99-118.

Singh, S.K. & Del Giudice, M., 2019. Big data analytics, dynamic capabilities and firm performance. *Management Decision* 57(8): 1729-1733.

Singh, T., 2018, Artificial intelligence in enterprises – businesses are waking up. *Forbes* (Oct 22).

Sodero, A., Jin, Y.H. & Barratt, M., 2019. The social process of Big Data and predictive analytics use for logistics and supply chain management. *International Journal of Physical Distribution & Logistics Management* 49(7):706-726.

Srinivasan, R. & Swink, M., 2018. An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective. *Production and Operations Management*, 27(10): 1849-1867.

Stanovich, K.E., 2009. What intelligence tests miss. *The Psychology of Rational Thought*. Yale University Press, New Haven and London.

Stanovich, K.E., 2011. *Rationality and the Reflective Mind*. Oxford University Press, New York.

Stanovich, K.E. & West, R.F., 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5): 645-665.

Thomas, R.W., 2011. When student samples make sense in logistics research. *Journal of Business Logistics*, 32(3): 287-290.

Thomson, K.S. & Oppenheimer, D.M., 2016. Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1): 99.

Thompson, V.A., 2009. Dual-process theories: a metacognitive perspective. In: Evans, Frankish K (eds) In *Two Minds: Dual Processes and Beyond*. Oxford University Press, Oxford: 171–196.

Thompson, V.A., Turner, J.A.P. and Pennycook, G., 2011. Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3): 107-140.

Timmer, S. & Kaufmann, L., 2019. Do managers' dark personality traits help firms in coping with adverse supply chain events? *Journal of Supply Chain Management*, 55(4): 67-97.

Tokar, T., Aloysius, J.A. & Waller, M.A., 2012. Supply chain inventory replenishment: The debiasing effect of declarative knowledge. *Decision Sciences*, 43(3): 525-546.

Toplak, M.E., West, R.F. & Stanovich, K.E., 2011. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7): 1275.

Toplak, M.E., West, R.F. & Stanovich, K.E., 2014. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2): 147-168.

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

Van de Ven, A.H., 2007. *Engaged Scholarship: A Guide for Organizational and Social Research*. Oxford University Press on Demand.

Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating perceived behavioral control, computer anxiety and enjoyment into the technology acceptance model. *Information Systems Research*, 11(4), 342-365.

Waller, M.A. & Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2): 77-84.

Wang, G., Gunasekaran, A., Ngai, E.W. & Papadopoulos, T., 2016. Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, 176: 98-110.

Weinhardt, J.M., Hendijani, R., Harman, J.L., Steel, P. & Gonzalez, C., 2015. How analytic reasoning style and global thinking relate to understanding stocks and flows. *Journal of Operations Management*, 39: 23-30.

Wu, D.Y. & Katok, E., 2006. Learning, communication, and the bullwhip effect. *Journal of Operations Management*, 24(6): 839-850.

4.7. Appendix A3. Information about Interventions and Screenshots

4.7.1 Declarative Knowledge Screenshot

As a reminder, At the completion of this study, five participants will randomly be selected to win the amount earned.

The following information is provided to help you make more accurate forecasts:

What is used to compute the model forecast?

- The model uses all historical data to analyze trends and patterns.
- The model has no access to information about special events.

When should you change the model forecast?

- The model specializes in providing unbiased estimates when analyzing the historical data.
- Best practice is to **only change the model forecast when you have information about a special event.** In other words, if you receive information about a special event, change the model accordingly.change the model accordingly. If you do not receive information about a special event, do not change the model.

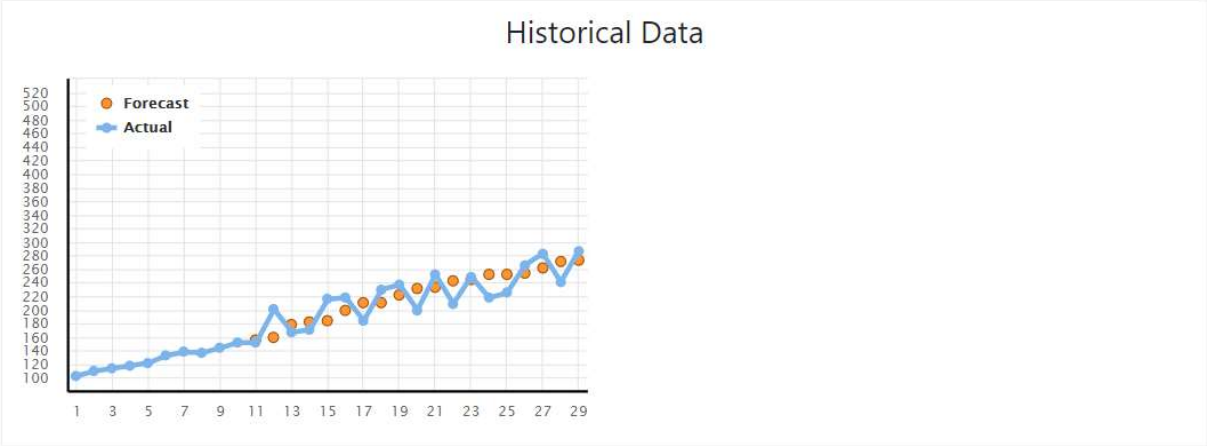
How can you be the most accurate?

- The most accurate forecast will rely on the model's unbiased estimate and your use of the information about special events which is not available to the model.
- For example: when asked for the forecast for Period 31, if you have information about a special event in Period 31, change the model. If you do not receive information about a special event in Period 31, do not change the model in Period 31.

When should you change the model forecast?

4.7.2 Analytical Thinking Screenshot

The following exercise is provided to help you make more accurate forecasts:



Forecast Accuracy

Definition: this score represents the accuracy of your forecasts. It is computed as an average of the absolute difference between the forecast and actual.

Expert: people who score low on this outcome are more accurate.

Your score: **21.33**

Please think reflectively and complete the following questions:

What periods were you the most accurate?

Why do you think you did well in those periods?

What periods were you the least accurate?

Why do you think you did poorly in those periods?

V. CONCLUSION

The purpose of this dissertation is to examine the integration of supply chain and operations management people with new technologies and processes. The first two essays provide empirical studies on the process of integrating human judgment with analytics (including machine learning models) conditioned on contextual information available only to humans. My third essay develops and tests interventions to train managers on when and how best to integrate their judgment with analytics.

Essay 1 assesses the optimal integration of human judgment and analytics in the demand planning context. The study uses a laboratory experiment to compare forecasting methods of integration. The comparison includes existing methods of integration (i.e., judgment forecast, judgmental adjustment, quantitative correction, combined, input to model, and model forecast) and introduces two machine learning methods (interactive machine learning and human guided machine learning). Essay 1 is, to best of my knowledge, the first study to test the ability of humans to leverage contextual information when comparing across methods of integration. The findings suggest that contextual information leveraged by humans can enhance forecasting accuracy. Most importantly, the method used to integrate the contextual information with the model ultimately determines how useful the contextual information is in improving forecasting performance. Results of the comparison across all methods of integration—both existing methods and my two new methods—reveal that human-guided machine learning (HGML) and interactive machine learning (IML) are the most effective methods of integrating model and human-judgment forecasts.

Essay 2 tests the two machine learning algorithms introduced in Essay 2, IML and HGML, in the field. The field study is conducted with a large, multinational firm. The data from the field study encompass over three million datapoints across five product categories. The field

study compares the accuracy of demand forecasts from the existing demand planning process used by the company (i.e., judgmental adjustment of sophisticated machine learning forecast) with demand forecasts from IML and HGML. Analysis of the results indicate that demand forecasts using IML and HGML are more accurate than the current demand planning process used by the firm. IML and HGML control for biases that inevitably are present with human judgment while continuing to allow human demand forecasters some control over the demand planning outcome, as suggested by Dietvorst, Simmons, and Massey (2016).

Essay 3 develops and tests interventions to train managers on when and how to optimally integrate their judgment with analytics. Since the tasks that require supply chain practitioners to integrate with analytics are diverse (Waller & Fawcett, 2013), I test both high- and low-level processing tasks. The interventions are grounded in two theories from cognitive psychology: Adaptive Character of Thought (ACT) Theory and Dual Process Theory. In essence, the theories used collectively offer a framework to explain heterogeneity in individual behavior when intervening with analytical systems. The results suggest undesirable human interference with analytics may be mitigated through specific training elements which appeal to the underlying cognitive mechanisms influencing behavior. Specifically, for high-level processing tasks, it is best to use a combination of declarative knowledge (i.e., information as to how and when interference with analytics is justified) and analytical thinking in the training. For low-level processing tasks it is best to use only declarative knowledge in the training.

Taken collectively, these three essays offer three encompassing theoretical contributions: 1) A comprehensive study of the efficacy of different methods of integrating human judgment and model-based analytics when humans have contextual information unavailable to the model; 2) A first empirical behavioral study of the integration of human judgment with machine

learning, that introduces HGML and IML to the behavioral operations and supply chain literature; and 3) An investigation of behavioral interventions that improves predictive analytic processes, using the combination of ACT and dual process theory.

In light of the complex nature of modern supply chains, fraught with unprecedented disruptions, contextual information is a critical resource (Arvan et al., 2019). Since contextual information is typically not available to predictive analytical systems, human judgment is oftentimes the only way that contextual information can be included in the model (Perera et al., 2019). Although existing literature has developed and studied various methods of integrating human judgment and analytical systems, there has yet to be a comprehensive comparison of methods of integration where contextual information is manipulated. Essay 1 provides this empirical confirmation of the conceptual prediction, "the key to utilizing the benefits of contextual information seems to be lying under its integration into the forecasting process" (Arvan et al., 2019, p. 245).

The concept of integrating models and humans in a manner as to capitalize on the strengths of each is not new (Blattberg & Hoch, 1990). However, methods of integration discussed in the literature have yet to consider utilizing machine learning in conjunction with human judgment. The introduction of IML and HGML to behavioral supply chain and operations management illustrates the advantages of appropriately using the strengths of both models and humans. Results from both the carefully-controlled laboratory study and the in-practice demand-planning field study reveal that IML and HGML improve predictive accuracy.

Existing research on human judgment frequently discusses heuristics and biases that oftentimes impede predictive performance. ACT and dual process theory have been used separately to explain cognitive mechanisms affecting predictive performance. However, as

illustrated in Essay 3, the theories, when taken together, provide further guidance as to approaches that improve predictive performance. Namely, the use of System 2 thinking while receiving training on a high-level processing task is useful for changing sub-optimal behavior and improving predictive performance.

Altogether, the three essays provide nuanced insight as to the role of a modern supply chain professional. One thing is for certain—human judgment remains a valuable resource when used correctly. Focusing on the demand planning context, my research reveals that human judgment is most useful when contextual information is available regarding a special event and the human judgment is used conjointly with machine learning (i.e., IML/HGML). Despite the continual changes to supply chain functions, value creation is possible when human strengths are recognized and utilized appropriately.

References

Arvan, M., Fahimnia, G., Reisi, M. & Siemsen, E. 2019. Integrating human judgment into quantitative forecasting methods: A review. *Omega,* 86: 237-252.

Blattberg, R. C. & Hoch, S. J. 1990. Database models and managerial intuition: 50% model+ 50% manager. *Management Science,* 36(8): 887-899.

Dietvorst, B. J., Simmons, J. P., & Massey, C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3): 1155-1170.

Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. 2019. The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research*, 274(2): 574-600.

Waller, M. A., & Fawcett, S. E. 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2): 77-84.

VI. APPENDIX

5.1 Appendix A. Institutional Review Board Protocol Approvals

To:	Rebekah Inez Brau
From:	Douglas James Adams, Chair IRB Committee
Date:	11/28/2017
Action:	**Exemption Granted**
Action Date:	11/28/2017
Protocol #:	1711083455
Study Title:	Supply Chains Analytics and the Evolving Work of Supply Chain Managers

The above-referenced protocol has been determined to be exempt.

If you wish to make any modifications in the approved protocol that may affect the level of risk to your participants, you must seek approval prior to implementing those changes. All modifications must provide sufficient detail to assess the impact of the change.

If you have any questions or need any assistance from the IRB, please contact the IRB Coordinator at 109 MLKG Building, 5-2208, or irb@uark.edu.

cc: John Aloysius, Investigator

UNIVERSITY OF ARKANSAS

To:	John Aloysius WCOB 475D
From:	, Chair
Date:	09/13/2018
Action:	**Exemption Granted**
Action Date:	09/13/2018
Protocol #:	1808137974
Study Title:	Supply Chains Analytics and the Evolving Work of Supply Chain Managers

The above-referenced protocol has been determined to be exempt.

If you wish to make any modifications in the approved protocol that may affect the level of risk to your participants, you must seek approval prior to implementing those changes. All modifications must provide sufficient detail to assess the impact of the change.

If you have any questions or need any assistance from the IRB, please contact the IRB Coordinator at 109 MLKG Building, 5-2208, or irb@uark.edu.

cc: Rebekah Inez Brau, Investigator

UNIVERSITY OF ARKANSAS

To:	John Aloysius WCOB 475D
From:	Douglas James Adams, Chair IRB Committee
Date:	08/05/2019
Action:	**Exemption Granted**
Action Date:	08/05/2019
Protocol #:	1902177243
Study Title:	Supply Chains Analytics and the Evolving Work of Supply Chain Managers

The above-referenced protocol has been determined to be exempt.

If you wish to make any modifications in the approved protocol that may affect the level of risk to your participants, you must seek approval prior to implementing those changes. All modifications must provide sufficient detail to assess the impact of the change.

If you have any questions or need any assistance from the IRB, please contact the IRB Coordinator at 109 MLKG Building, 5-2208, or irb@uark.edu.

cc: Bekki Brau, Investigator

VII. VITAE

REBEKAH INEZ (BEKKI) BRAU

Walton Doctoral Fellow
PhD Candidate

University of Arkansas
WCOB #438
Fayetteville, AR 72701
RBrau@walton.uark.edu
http://bekkibrau.com/
https://www.linkedin.com/in/bekkibrau

EDUCATION

Ph.D., Supply Chain Management, University of Arkansas; Fayetteville, AR
The Sam M. Walton College of Business
Dissertation: Integrating Systems, Processes, and Human Judgment: Three Essays on Value Creation with Supply Chain Analytics
Dissertation Committee (alphabetical order):
John Aloysius, Dissertation Chair, U. Arkansas, Oren Harris Chair in Logistics
Christian Hofer, U. Arkansas, Associate Professor of Supply Chain Management
Nada Sanders, Northeastern U., Distinguished Professor of Supply Chain Management
Enno Siemsen, U. Wisconsin, Proctor & Gamble Bascom Professor
Brent Williams, U. Arkansas, Garrison Endowed Chair in Supply Chain Management
Degree Awarded: May 2021

M.S., Instructional Psychology and Technology, Brigham Young University; Provo, UT
McKay School of Education
Concentration: Research and Evaluation
Capstone Project Title: Plan, Do, Study, Act (PDSA): A Supply Chain Management Learning Activity
Project Committee (alphabetical order):
Royce Kimmons, Assistant Professor of Instructional Psychology and Technology
Jason McDonald, Chair, Associate Professor of Instructional Psychology & Tech
Scott Webb, Associate Professor of Supply Chain Management
Degree Awarded: December 2017

B.S., Management, Brigham Young University; Provo, UT
Marriott School of Business
Concentration: Organizational Behavior Pre-PhD Program
Business European Study Abroad with Dr. Kristen DeTienne
Degree Awarded: June 2016

RESEARCH ACTIVITY

PEER-REVIEWED PUBLICATIONS

Barker, J. & Brau, R.I. Shipping Surcharges and LSQ: Pricing the Last Mile, *International Journal of Physical Distribution and Logistics Management*, 50(6), 2020, 667-691.

Brau, R.I., Gardner, J., McDonald, J., & Webb, S. Plan, Do, Study, Act (PDSA): A Supply Chain Management Learning Activity, *Decision Sciences Journal of Innovative Education* 17(1), 2019, 6-32 (from master's capstone project).

Brau, J.C., Brau, R.I., Rowley, T., & Swenson, M.J. An Empirical Analysis of the Success Factors in an Introductory Financial Management Class, *Journal of the Academy of Business Education* 18, 2017, 1-52 (lead article, from undergraduate program).

Brau, R.I. A Framework for Teaching the Goal of the Firm in Introductory Business Classes: Shareholder Wealth Maximization Ethicality and Classical Philosophical Paradigms, *Journal of the Utah Academy of Sciences, Arts & Letters* 93, 2016, 135-161, (from undergraduate program).

Brau, J.C., Brau, R.I, Owen, S., & Swenson, M.J. The Determinants of Student Performance in a University Marketing Class, *Business Education Innovation Journal* 8(2), 2016, 21-31, (from undergraduate program).

CONFERENCE PROCEEDINGS

Brau, J.C., Brau, R.I., Swenson, M.J. & Rowley, T., An Empirical Analysis of the Success Factors in an Introductory Business Class, *Academy of Management Proceedings* (1), 2016, 17349.

Brau, R.I. A Framework for Teaching the Goal of the Firm in Introductory Business Classes: Shareholder Wealth Maximization Ethicality and Classical Philosophical Paradigms, *Proceedings of the International Association for Business and Society* 27, 2016, 1-20.

WORK IN PROGRESS

Brau, R.I., Aloysius, J., & Siemsen, E. When Managers Meet Models: Integrating Human Judgment and Machine Learning

Brau, R.I., Aloysius, J., & Siemsen, E. Integrating Machine Learning and Human Judgment: A Study on Demand Planning in the Field

Brau, R.I., Aloysius, J., & Sanders, N. Improving of Supply Chain Performance Through Analytics: Designing Interventions Based on Cognitive Psychology

Brau, R.I., Hofer, C., & Aloysius, J. Firm's Strategic Orientation Towards Analytic Skill Acquisition and Expectations of Firm Performance

Aloysius, J., Davis, F., & Brau, R.I. Preferences for Computerized Decision Aids: The Influence of Decision Anxiety on Cognitive Mechanisms

CONFERENCE PRESENTATIONS & POSTERS (* indicates I presented)

Brau, R.I., Aloysius, J. & Siemsen, E. (August 2021) Integrating Machine Learning and Human Judgment: A Study on Demand Planning in the Field. *Academy of Management (AOM)*. Peer-reviewed presentation. (Nominated for Operations & Supply Chain Management (OSCM) Division Chan Hahn Best Paper Award.)*

Brau, R.I., Aloysius, J. & Siemsen, E. (April 2021) Integrating Machine Learning and Human Judgment: A Study on Demand Planning in the Field. *Production and Operations Management Society (POMS).* Invited presentation.*

Brau, R.I., Aloysius, J. & Siemsen, E. (November 2020) When Managers Meet Models: Integrating Human Judgment and Analytics. *Decision Sciences Institute (DSI)*. Peer-reviewed presentation.*

Brau, R.I., Aloysius, J. & Siemsen, E. (October 2020) When Managers Meet Models: Integrating Human Judgment and Analytics. *International Symposium on Forecasting (ISF)*. Invited presentation.*

Brau, R.I., Aloysius, J. & Siemsen, E. (August 2020) When Managers Meet Models: Integrating Human Judgment and Analytics. *Academy of Management (AOM)*. Peer-reviewed presentation. (Nominated for Operations & Supply Chain Management (OSCM) Division Best Student Paper Award.)*

Brau, R.I., Aloysius, J. & Siemsen, E. (April 2020) When Managers Meet Models: Integrating Human Judgment and Analytics. *Production and Operations Management Society (POMS).* Invited presentation. (Conference cancelled after acceptance due to Covid-19.)

Brau, R.I., Aloysius, J. & Siemsen, E. (September 2019) When Models Meet Managers: Optimal Integration of Statistical Model-Based and Judgmental Forecasting. *Council of Supply Chain Management Professionals* (*CSCMP)*. Peer-reviewed presentation.*

Brau, R.I., Aloysius, J. & Siemsen, E. (May 2019) When Models Meet Managers: Optimal Integration of Statistical Model-Based and Judgmental Forecasting. *Production and Operations Management Society (POMS).* Invited presentation.*

Aloysius, J., Brau, R.I. & Siemsen, E. (November 2018) When Models Meet Managers: Optimal Integration of Statistical Model-Based and Judgmental Forecasting. *Institute for Operations Research and the Management Sciences (INFORMS).* Invited presentation.

Brau, R.I. (September 2018) When Models Meet Managers: Optimal Integration of Statistical Model-Based and Judgmental Forecasting. *Council of Supply Chain Management Professionals* (*CSCMP).* (Won first prize award.) Poster.*

Barker, J. & Brau, R.I. (September 2018). Shipping Surcharges and Logistics Disruptions: Pricing Strategies to Attract and Retain Online Customers. *Council of Supply Chain Management Professionals* (*CSCMP)*, *Academic Research Symposium (ARS).* Peer-reviewed presentation.

Brau, R.I. (July 2018) When Models Meet Managers: Optimal Integration of Statistical Model-Based and Judgmental Forecasting. *Behavioral Operations Management Conference (BOM), Young Scholars Workshop.* Peer-reviewed presentation.*

Brau, R.I. (April 2017) Instructional Designer Values as Revealed Through the Use of Inscriptions. *Utah Academy of Sciences, Arts, & Letters (UASAL)*. Peer-reviewed presentation.*

Brau, R.I. (June 2016). A Framework for Teaching the Goal of the Firm in Introductory Business Classes: Shareholder Wealth Maximization Ethicality and Classical Philosophical Paradigms. *International Association of Business and Society (IABS)*. Peer-reviewed presentation.*

Brau, R.I. (March 2016). Conditions of Shareholder Wealth Maximization Ethicality under Classical Philosophical Paradigms. *Utah Academy of Science, Arts, & Letters (UASAL)*. Peer-reviewed presentation.*

Brau, J.C., Brau, R.I., Rowley, T., & Swenson, M.S. (August 2016). An Empirical Analysis of the Factors of Success in an Introductory Business Class. *Academy of Management (AOM)*. Peer-reviewed presentation.*

Brau, J.C., Brau, R.I., Rowley, T., & Swenson, M.S. (March 2016). An Empirical Analysis of the Factors of Success in a Principles of Finance Class. *Utah Academy of Science, Arts, & Letters (UASAL)*. Peer-reviewed presentation.*

Brau, J.C., Brau, R.I., Rowley, T., & Swenson, M.S. (September 2015). An Empirical Analysis of the Factors of Success in a Principles of Finance Class. *Academy of Business Education (ABE)*. (This paper won the best paper award at ABE.) Peer-reviewed presentation.

RESEARCH AWARDS

Chan Hahn Best Paper Award Nomination, Operations & Supply Chain Management (OSCM) Division, Academy of Management (AOM), Award announced August 2021.

Best Student Paper Award Finalist, Operations & Supply Chain Management (OSCM) Division, Academy of Management (AOM), August 2020.

First Place for Best Practical Application Poster, Academic Research Symposium (ARS), CSCMP Academic Research Strategies, September 2018.

Walton Doctoral Fellowship, 2017-2021.

Best Paper Award, Academy of Business Education (ABE), September 2015.

RESEARCH GRANTS

Association for Supply Chain Management, Associate Researcher
Supply Chains Analytics and the Evolving Work of Supply Chain Managers
Future of Supply Chains Grant, $41,500 (2017), $24,500 (2019)
Principals: Enno Siemsen and John Aloysius

Office of Research & Creative Activities, BYU, Primary Author
Externalities of Corporate Social Responsibility on Employee Compensation
University-wide grant competition for $1,500 of support
Mentor Professor: Dr. Nile Hatch

RESEARCH ASSISTANT EXPERIENCE

Graduate Research Assistant for Dr. John Aloysius (2017-present), U. Arkansas
Focus on behavioral operations and supply chain management.

Research Assistant for Drs. Sheli Sillito Walker & Katie Liljenquist (2014-2017), BYU
Led over a dozen studies in the MSB Behavioral Research Lab as the principal lab investigator, compiled data from the studies, and wrote literature reviews.

Research Assistant for Dr. Andrew Holmes (2013-2014), BYU
Edited scholarly articles and white papers.

RESEARCH INTERESTS

Supply Chain Analytics
Consumer Behavior in Supply Chains
Logistics & Transportation
Behavioral Operations & Supply Chain Management

RESEARCH METHODOLOGIES

Laboratory and Field Experiments
Interventionist Research
Econometric Analysis

TEACHING ACTIVITY

TEACHING EXPERIENCE

Graduate Teaching Instructor, University of Arkansas, 2018-present

Spring 2021: International Logistics, SCMT 3643 (Hybrid, solo instructor)
Section 1 size: 24 Instructor rating: TBA
Section 2 size: 24 Instructor rating: TBA
Course Description: Logistics activities in international business with special emphasis on international sourcing and distribution channels, international transportation, import and export procedures, international sale and payment terms, and documentation. Special emphasis is placed on current events and their effect on the management of operations of U.S.-based organizations. Prerequisite: ((ECON 2013 and ECON 2023), or ECON 2143) and SCMT 2103.

Fall 2019: Supply Management, SCMT 3613 (Hybrid, solo instructor)
Section 1 size: 28 Instructor rating: 4.83/5.0
Section 2 size: 28 Instructor rating: 4.85/5.0
Spring 2019: Supply Management, SCMT 3613 (Lecture, solo instructor)
Section size: 41 Instructor rating: 4.94/5.0
Fall 2018: Supply Management, SCMT 3613 (Hybrid, solo instructor)
Section size: 38 Instructor rating: 4.70/5.0
Course Description: SCMT 3613: This course covers the critical sourcing and procurement processes: strategic sourcing, source to pay, and supplier relationship management. Additionally, it covers innovative efforts to grow sourcing contribution to demand-driven supply chain integration, including sustainability, technology, and risk management. Prerequisite: ((ECON 2013 and ECON 2023) or ECON 2143)

EMBA Guest Lecturer, University of Arkansas, 2019-present

Spring 2019: Predictive Supply Chain Analytics, SCMT 5693

TEACHING DEVELOPMENT ACTIVITIES

UArk PhD Seminar on University-Level Teaching, WCOB 6111 (2018)
Many graduate courses in instructional psychology and technology, BYU (2016-2017)
Academy of Management Teaching and Learning Conference (TLC) (2016)

TEACHING INTERESTS

Supply Management, Purchasing, Procurement
Supply Chain Strategy and Logistics
Predictive Analytics

SERVICE ACTIVITY

PROFESSIONAL SERVICE

Junior Editorial Review Board, Journal of Supply Chain Management (JSCM), 2020-present
Organizing Committee, Behavioral Operations Conference (BOC), 2021
Reviewer, Transportation Research Part E., 2021
Reviewer, Journal of Business Logistics (JBL), 2021
Reviewer, Journal of Operations Management (JOM), 2020
Reviewer, Council of Supply Chain Management Professionals Conference (CSCMP), 2018
Reviewer, Academy of Management Conference (AOM), 2017
Reviewer, International Association of Business Society Conference (IABS), 2016
Student Facilitator, Teaching Ethics at Universities Conference, BYU Wheatley Institute, 2015

COMMUNITY SERVICE

Women's group leader for local church congregation, Rogers AR (2020-present)
Youth leader for local church congregation, Rogers AR (2020)
Spanish translator for local church congregation, Rogers, AR (2019-2020)
Sunday instructor for local church congregation, Bentonville, AR (2018)
Program Director for local Boys & Girls Club, Provo, UT (2017)
Women's group instructor for local church congregation, Provo, UT (2016)
Served church mission to Mexico City, Mexico (2014)
YMAD Humanitarian Service Trip, Banjar India (2012)

INDUSTRY EXPERIENCE & INTERESTS

INDUSTRY ENGAGEMENT

Supply Chain Management Research Center (SCMRC) Emerging Research Symposium, *Invited presenter* (November 2020)

Plug and Play Tech Center Leading Supply Chain Innovation Webinar Part 1: Transforming your Business Analytics: Maturity Cycle for Decision Making, *Invited presenter* (May 2020)

Walmart International Supplier Collaboration Board, *Invited presenter* (February 2020)

Supply Chain Management Research Center (SCMRC) Emerging Research Symposium, *Invited presenter* (November 2019)

INDUSTRY EXPERIENCE

Bekki Brau Photography, *Founder and Owner*, Provo, UT (Jun 2012-2017)
Photography services for graduations, dances, weddings, and other events
Contracted with local jewelry store and concert venue to provide photography

MyEducator, *Intern*, Orem, UT (Jan 2017-April 2017)
Audited instructional design for three separate electronic textbooks.

WorldStock (Overstock.com), *Consultant*, Salt Lake City, UT (Aug 2013-Dec 2013)
Created and organized consulting project to expand supply chain into South Africa
Served as team leader and company liaison, focusing on supply chain management

Citizens for Decency, *Intern*, Provo, UT (Jan 2013-Aug 2013)
Chosen as team leader of five-person team; created new chapter
Redefined project to give direction and best serve needs of organization

Women on Wall Street, *Participant*, New York, NY (July 2012)
Visited Goldman Sachs, Credit Suisse, Morgan Stanley, Citigroup, among others
Focused on women effectiveness in business

LANGUAGES AND SKILLS

Speak, read, and write Spanish (I have lived in Mexico and Puerto Rico)
Intermediate programmer in Python, (Django, oTree), JavaScript, HTML, CSS
Proficient in Stata, SPSS, R, and EViews statistical analysis software

PERSONAL

Career goal is to become a business professor at a well-respected research university
Hobbies include family activities, traveling (40 countries so far), photography, reading, working out, yoga, swimming, and meditation
Women relay team triathlon winner
Family: Married to D. Kamiko Ho'okano Adcock; Dotty May (Jack Russell Terrier)