Engineering Better Decision Making

Improving Decisions Through Behavioral Economic Engineering

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2017

vorgelegt
von
Dipl.-Ing. (FH) Tobias Stangl, M.Sc.
aus
Wörth an der Donau

Referent:	Prof. Ulrich W. Thonemann, Ph. D.
Korreferent:	Prof. Dr. Dirk Sliwka
Tag der Promotion:	06.12.2017

Contents

List of Figures

List of Tables

List of Abbreviations

BBC	buyback contract
BRET	Bomb Risk Elicitation Task
CLER	Cologne Laboratory of Economic Research
CRT	Cognitive Reflection Test
DOS	days of supply
ECU	experimental currency units
ITR	inventory turn rat
LBOE	Laboratory for Behavioral Operations and Economics
OECD	Organisation for Economic Co-operation and Development
OLS	ordinary least squares
ORSEE	Online Recruitment System for Economic Experiments
PIAAC	Programme for the International Assessment of Adult Competencies
RSC	revenue sharing contract
SD	standard deviation
SL	service level
SLC	service level contract
WP	wholesale price
WPC	wholesale price contract

Chapter 1

Introduction

> "*Even the most analytical thinkers are predictably irrational; the really smart ones acknowledge and address their irrationalities.*"
>
> Dan Ariely
> Behavioral Economist

1.1 Motivation

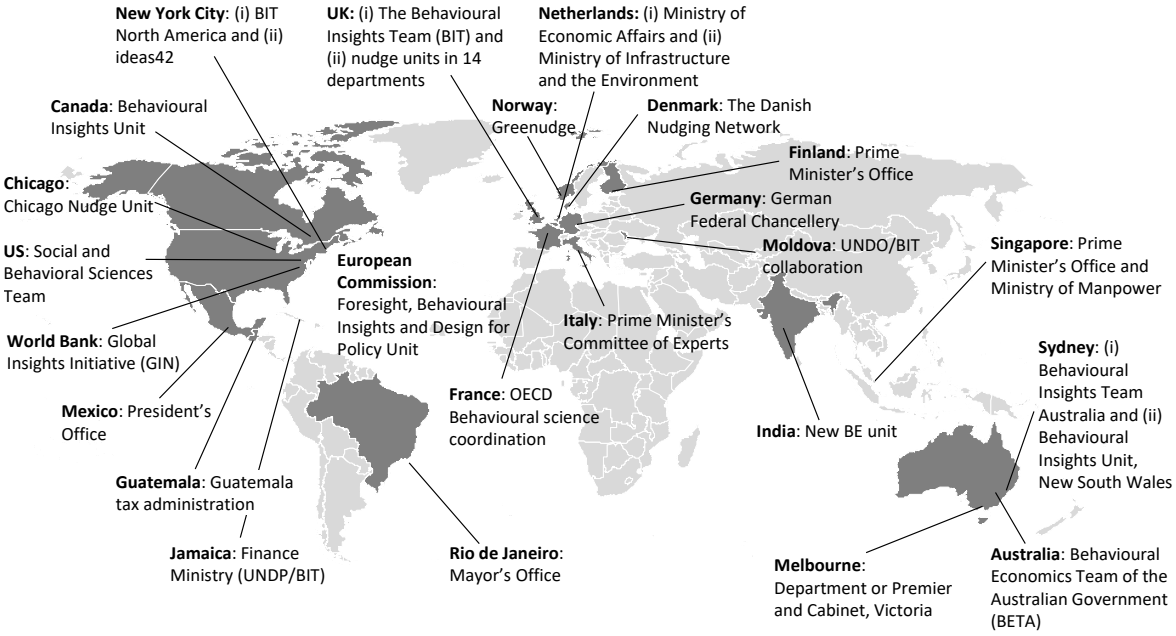
Standard economic theory relies on the neoclassical assumption that individuals are fully rational and purely self-interested. It assumes that decision makers undertake optimal decisions that maximize their own profits. As early as 1955, Simon criticized these assumptions and argues that human decision makers have limited cognitive capabilities and are prone to decision biases. He proposes satisficing as a more accurate way to model their behavior and refers to this approach as *bounded rationality* (Simon 1955, 1957). Simon (1959) recommends that economists should incorporate psychological evidence on individual behavior and that theories of decision making should be grounded in an empirically founded theory of choice that take the cognitive processes into account. The idea of bounded rationality of human decision makers began to impact economics as the first controlled laboratory experiments in decision making started to take place in the early 1960's. Pioneering work of Vernon Smith (1962, 1976) laid the ground for publishing research in experimental economics (Charness and Halladay 2017). Ever since, the field of experimental economics has seen exponential

growth every decade (Roth 1995). At the same time the field of behavioral economics has also grown in popularity and has developed new behavioral theory and models to explain the gaps between established economic theory and experimental results (Bendoly et al. 2006). Once an academic niche, behavioral and experimental economics have gained momentum in the late 1980's and early 1990's (Charness and Halladay 2017), approximately ten years after Herbert Simon received the Nobel Prize for Economics in 1978 "for his pioneering research into the decision-making process within economic organizations" (The Royal Swedish Academy of Sciences 1978).

Another broad approach to address bounded rationality started in the late 1960's by the seminal work of Tversky and Kahneman (1971, 1974), now known as the *heuristics and biases approach* (Kahneman and Frederick 2002). Tversky and Kahneman (1971) propose that when humans face complex decisions they often apply simple decision heuristics, such as the representativeness heuristic, the availability heuristic, and the anchoring and adjustment heuristic. These decision heuristics are simple rules of thumb that lead to systematic biases in judgment and decision making. Various researchers have contributed to that research stream and identified a large number of heuristics and biases in human decision making. Behavioral economists like Richard Thaler have started to incorporate behavioral insights into economic science. In 2002, Daniel Kahneman and Vernon Smith co-received the Nobel Prize in Economics "for integrating insights from psychology into economic science" and for establishing "laboratory experiments as a vital tool in empirical economic analysis", respectively (The Royal Swedish Academy of Sciences 2002). They played a central role in establishing behavioral and experimental economics as part of mainstream economics.

In the new millennium, both behavioral and experimental economics become useful tools to design government policies. Before policy makers only rarely used psychological insights or relied on theoretical predictions (Charness and Halladay 2017). After Thaler and Sunstein (2008) published their influential book *Nudge* in 2008, behavioral economists have become increasingly influential as policy advisors. Many governments have formed groups composed of behavioral and experimental economists to incorporate insights from academic research in behavioral science into the design of more effective policy solutions. In 2010 the British

Figure 1.1 Nudge Units Around the World (adapted from Chen et al. 2017)



government was the first to establish its Behavioural Insights Team[1], commonly referred to as the "Nudge Unit". In subsequent years, several governments around the world followed the British government and formed their own nudge units, including Canada with its Behavioral Insights Unit[2], Germany with a team in the Federal Chancellery (Deutscher Bundestag 2015), and the United States, which established its Social and Behavioral Sciences Team by President Obama's executive order in 2015[3]. Other initiatives around the world can be seen in Figure 1.1. In 2017, Richard Thaler received the Nobel Prize in Economics "for his contributions to behavioural economics" (The Royal Swedish Academy of Sciences 2017).

Incorporating behavioral science insights has not been limited to economics, but has taken place in other fields as well, such as accounting, finance, marketing, law, and, more recently, strategy. Operations management was perhaps the last field of management studies to embrace behavioral insights (Loch and Wu 2007). Before 2000, the field was dominated by formal mathematical models and human behavior received little attention. This has changed rapidly after the seminal paper of Schweitzer and Cachon (2000), who initiated the research stream

[1] http://www.behaviouralinsights.co.uk/
[2] http://bi.dpc.nsw.gov.au/
[3] https://sbst.gov/

of behavioral operations. They used laboratory experiments to analyze ordering behavior in the newsvendor setting. The order quantities of their subjects deviated substantially from profit maximizing order quantities and exhibited what has come to be known as the "pull-to-center" effect, because observed average order quantities are consistently between profit maximizing order quantities and mean demand. This effect was mainly attributed to anchoring and insufficient adjustment (Slovic and Lichtenstein 1971, Tversky and Kahneman 1974). Since the seminal paper of Schweitzer and Cachon (2000), the pull-to-center effect has proven robust in more than 20 experimental studies (Zhang and Siemsen 2016), holding under various demand distributions (Benzion et al. 2008, 2010) and observed and unobserved lost sales (Rudi and Drake 2014) and persists with experience and training (Bolton and Katok 2008, Bolton et al. 2012) and decision frequency (Bolton and Katok 2008, Bostian et al. 2008, Lurie and Swaminathan 2009).

Bostian et al. (2008) estimated an adaptive learning model (Camerer and Ho 1999) that tracks the observed data patterns in their newsvendor experiment. Other proposed explanations of the pull-to-center effect include bounded rationality (Su 2008), cognitive reflection (Moritz et al. 2013), ex post inventory error minimization (Schweitzer and Cachon 2000, Ho et al. 2010, Kremer et al. 2014), overconfidence of decision makers (Ren and Croson 2013), and impulse balance behavior (Ockenfels and Selten 2014, 2015). All of these theories offer explanations for why learning is insufficient to move ordering fully away from the anchor at mean demand to the optimal order quantity.

Behavioral operations has become a rapidly growing field. In the last decade, scholars have incorporated insights from psychology and behavioral economics in sub-areas, such as contracting (Katok et al. 2008, Katok and Wu 2009, Becker-Peth et al. 2013, Wu and Chen 2014), information sharing and forecasting (Özer et al. 2011, Kremer et al. 2011, Moritz et al. 2014, Kremer et al. 2016, Bolton and Katok 2017), project management (Bendoly and Swink 2007, Sting et al. 2015, Kagan et al. 2017), procurement (Engelbrecht-Wiggans et al. 2007, Davis et al. 2011, Elmaghraby et al. 2012, Haruvy and Katok 2013, Davis et al. 2014), queuing (Batt and Terwiesch 2015, Kremer and Debo 2016, Shunko et al. 2017, Yu et al. 2017), and revenue management (Bearden et al. 2008, Bendoly 2013, Kocabiyikoglu

et al. 2015). Hopp (2004) speculates that behavioral factors could be the source of the next paradigm shift within operations management. Like behavioral and experimental economics enhanced and broadened the field of economics, behavioral operations enhances and broadens the field of operations management.

1.2 Behavioral Economic Engineering

Bolton and Ockenfels (2012) defined economic engineering as "the science of designing real-world institutions and mechanisms that align individual incentives and behavior with the underlying goals". Economic mechanisms play an important role in operations management. For example, the design of procurement auctions, supply contracts, or incentive schemes directly affect prices, order quantities, and the motivation and satisfaction of employees. Mechanisms matter because they affect decision making. Incorporating insights from psychology and behavioral economics into the design of mechanisms is important because actual decision making deviates from standard economic theory in two ways:

1. Deviations from rationality (bounded rationality)

2. Deviations from selfishness (other regarding preferences)

In this dissertation we will focus on bounded rationality of human decision makers. Human decision makers are often guided by simple rules of thumb that are easy to apply but lead to deviations from normative predictions. These biases are often systematic, making humans not just boundedly rational but *predictably irrational* (Ariely 2010). These systematic deviations allow us to design mechanisms based on behavioral models with the goal of engineering better decision making.

To design effective mechanisms it is important to understand the psychological underpinnings that contribute to biases. In behavioral science, the ancient idea that cognitive processes can be partitioned into two main families has become standard under the so-called "dual process theory of mind" (Epstein 1994, Sloman 1996). There is a rich body of literature on how cognitive processes can be defined (see Stanovich and West 2000 and Evans 2008 for an overview), with the common notion that one process is more *intuitive* while the other process

Table 1.1 Two Cognitive Systems

System 1	System 2
Intuitive	Reflective
Fast	Slow
Automatic	Controlled
Effortless	Effortful
Associative	Deductive
Unconscious	Self-aware
Skilled	Rule-following

is more *reflective*. Stanovich and West (2000) and Kahneman and Frederick (2002) refer to these two cognitive processes as System 1 and System 2. Characteristics commonly attributed to the two systems are listed in Table 1.1. The operations of System 1 are intuitive, fast, automatic, and effortless. System 2 refers to operations, which are reflective, slow, controlled, and effortful. If an individual faces a judgment problem, System 1 quickly proposes an intuitive answer to the problem and System 2 can endorse or override this proposal. Although System 1 often proposes correct answers, these fast and effortless answers come with a cost: systematic biases (Arkes 1991).

The existence of systematic biases started an on-going debate about the most effective strategies for debiasing individual biases (Larrick 2004, Soll et al. 2015). Klayman and Brown (1993) suggest to group debiasing techniques into two general approaches: "modify the decision maker" or "modify the environment". The first approach assumes that the environment is more or less fixed, and, therefore, the best debiasing approach is to provide decision makers with a combination of education, training, and tools to help overcome their cognitive limitations (Soll et al. 2015). These approaches have already been tested successfully in field of behavioral operations. For example, Bolton et al. (2012) use task training to improve performance in newsvendor decisions and Ren and Croson (2013) provide their subjects with a tool for reducing overprecision and thereby improve subjects' ordering behavior. The second approach to debiasing is to modify the environment in which a decision is made in ways that reduce the likelihood of biases (Soll et al. 2015). This approach comprises incentives and choice architecture tools (Johnson et al. 2012) designed to "nudge" decision makers toward better decisions (Thaler and Sunstein 2008, Johnson et al. 2012, Soll et al. 2015), and is part

of behavioral economic engineering.

Behavioral economic engineering applied to operations management can help firms to design mechanisms to improve the performance of their processes, given a better understanding of behavioral regularities. First efforts aim to engineer better order decisions (Bolton and Katok 2008, Lee and Siemsen 2016), procurement auctions (Engelbrecht-Wiggans et al. 2007, Davis et al. 2014), supply contracts (Becker-Peth et al. 2013, Becker-Peth and Thonemann 2016) or incentive schemes (de Vries et al. 2016, Scheele et al. 2017). The aim of this dissertation is to join these efforts and improve decision making in operations through applying behavioral economic engineering.

1.3 Outline

This section outlines the structure of this dissertation. Although sharing the overall common goal of improving decisions through behavioral economic engineering, the three main chapters of this dissertation represent independent research projects. Each chapter considers an example of how decision making can be improved through behavioral economic engineering and is written in a way that the reader should be able to understand it without having read all of the prior parts of the dissertation in detail. At the end of each chapter we provide supplemental material for each project (for example, experimental instructions, screen shots of the experimental implementations, post-experimental questionnaires, etc.).

Beyond the goal of improving decisions all research projects focus on individual decision making in an operations context investigated through a common lens: controlled experiments. We conducted 58 experimental sessions at the Cologne Laboratory of Economic Research (CLER) at the University of Cologne, six sessions at the Laboratory for Behavioral Operations and Economics (LBOE) at the University of Texas at Dallas, and one experimental study on Amazon Mechanical Turk, a crowd-sourcing marketplace run by Amazon.com, Inc. (Paolacci et al. 2010, Buhrmester et al. 2011). In addition, we collected data at three business conferences (Marcus Evans 2013, Supply Chain Academy 2013, Copperberg 2013). In total we report data from 785 subjects. In the following we give a brief overview of each chapter:

Chapter 2 analyzes how performance metrics that contain equivalent information affect human decisions.[4] We consider two such performance metrics from supply chain management, days of supply and inventory turn rate, where one is the inverse of the other. We argue that individuals' evaluation of performance is affected by the metric as opposed to solely based on the fundamental attribute. We conducted three laboratory studies in which we investigate decision making in inventory management. The first study considers alternative inventory optimizations, out of which one must be selected. The second study analyzes a decision maker who must decide on the effort to invest in optimizing inventory of a specific product. The third study corresponds to the economic order quantity model. Our behavioral models suggest that decisions are affected by the metric that is used to indicate performance and we find support for the predictions in our laboratory experiments.

Chapter 3 analyzes human decision making under service level contracts.[5] Service level contracts can be parameterized, such that they have steep expected profit functions around the expected profit-maximizing order quantity – an interesting property that other supply contracts do not offer. We argue that this property leads to improved decision making. We provide analytical models and perform a laboratory experiment to analyze ordering behavior under service level contracts and compare the performance with that under wholesale price contracts, which have flat expected profit functions. In our experiment, the efficiency that human subjects achieved under a service level contract was 97.2%, compared with an efficiency of 88.1% under a wholesale price contract.

Chapter 4 examines compliance rates (trust) for forecast guidance in a simple take-the-risk or take-the-cost decision game.[6] We analyze two ways of conveying the risk in forecasts to non-expert users: providing the probability of the uncertain event or providing an explicit advice. It turns out that low numerate subjects exhibit less trust in recommendation

[4] This chapter is based on the paper by Stangl and Thonemann (2017) and benefited from comments of two anonymous referees and the editors of *Manufacturing & Service Operations Management*, seminar participants at the University of Cologne and the University of Texas at Dallas, and participants at the *2013 INFORMS Annual Meeting* and the *9th Annual Behavioral Operations Conference*.

[5] This chapter is based on the paper by Bolton et al. (2017) and benefited from seminar participants at the University of Cologne and the University of Texas at Dallas, and participants at the *2014 INFORMS Annual Meeting*, the *POMS 26th Annual Conference*, and the *2016 INFORMS Annual Meeting*.

[6] This chapter is joint work with Gary E. Bolton and Elena Katok and benefited from seminar participants at the University of Cologne, UC Riverside, and the Erasmus University Rotterdam and participants at the *12th Annual Behavioral Operations Conference*.

forecasts than high numerate subjects. While we find a positive relationship between subjects' numeracy and trust in probability forecasts, this relationship is overshadowed by the fact that even high numerate subjects use the probabilities inefficiently. Forecast guidance that blends probabilities and recommendations, in a way designed to offset the major behavioral shortcomings we observe, can improve compliance.

Chapter 5 summarizes the key results, offers concluding remarks on the contribution of this dissertation and provides a general outlook to future research in the field of behavioral operations.

1.4 Contribution

This dissertation adds to the emerging field of behavioral operations. The contributions of each research project are described in detail at the beginning of the respective chapter. The overall contribution to the field is mainly twofold:

Identifying behavioral regularities: We analyze behavioral regularities in an operations context. In our three research projects we compare different performance metrics, supply contracts, or forms of forecast guidance, respectively. Within each experimental study, the treatments share the same normative solution. Thus, fully rational decision makers should be unaffected by which of the metrics, supply contracts, or forms of forecast guidance is used. This common design feature allows us to analyze how human decision makers react to different mechanisms and to identify behavioral regularities.

Engineering better decision making: Based on the behavioral regularities that we identify in our experimental studies we propose and test mechanisms to improve performance of human decision making. We demonstrate that the design of performance metrics, supply contracts, or forecast guidance can significantly improve decisions. We also study the interaction of individual differences (for example, numeracy) with the proposed mechanism. This allows us to make recommendations of how to tailor the design to different groups of individuals.

Chapter 2

Equivalent Inventory Metrics: A Behavioral Perspective

We analyze how performance metrics that contain equivalent information affect actual decisions. We consider two such performance metrics from supply chain management, days of supply and inventory turn rate, where one is the inverse of the other. We argue that individuals' assessment of performance is also affected by the metric as opposed to solely based on the inventory value that actually matters. We perform three laboratory studies and analyze how decisions are affected by the metric used to indicate inventory performance. The first study considers alternative inventory optimizations, out of which one must be selected. The second study analyzes a decision maker who must decide on the effort to invest in optimizing inventory of a specific product. The third study corresponds to the economic order quantity model. Our behavioral models suggest that decisions are affected by the metric that is used to indicate performance and we find support for the predictions in laboratory experiments with human subjects: Under the inventory turn rate metric, individuals over-value inventory reductions. Compared to decisions under the days of supply metric, they choose worse inventory optimization options, invest more effort optimizing inventory of specific products, and choose higher ordering cost.

2.1 Introduction

Performance metrics are used to quantitatively assess the performance of organizations, functions, projects, and individuals. Important decisions are based on performance metrics, such as investment selections, budget allocations, and employee rewards. There exists a rich body of literature that provides guidance for choosing appropriate metrics (Parmenter 2010, Eckerson 2011). However, multiple metrics that contain the same information are often available, and it is unclear which one should be preferred. We refer to such metrics as *equivalent metrics*. Fully rational decision makers are unaffected by which of the equivalent metrics is used, but the decisions of actual human decision makers can be affected.

We consider equivalent metrics, where one metric is the inverse of the other. Such metrics are widely used in management. The overall performance of a company can be measured by the earnings yield and its inverse, the price-to-earnings ratio; a project can be evaluated by the payback period and its inverse, the return on investment; sales efficiency can be measured by cost per acquisition and its inverse, the acquisitions per dollar spent; employee retention can be evaluated by the employee turnover rate and its inverse, the employee retention time; and the number of calls in call centers can be measured by the incoming call rate and its inverse, the inter-arrival time.

In supply chain management, performance metrics and their inverses are used in many areas (Table 2.1). Our focus is on inventory management, which is one of the central areas of supply chain management. Inventory is part of a company's working capital and an important driver of financial performance. In inventory management, the equivalent performance metrics *days of supply* and *inventory turn rate* are commonly used (Caplice and Sheffi 1994, Hausman 2004). Days of supply measures the average duration that products are held in inventory

Table 2.1 Examples of Equivalent Metrics Used in Supply Chain Management

Area	Time based	Rate based
Inventory	Days of supply (90 days)	Inventory turn rate (4/year)
Warehousing	Picking time (30 sec/unit)	Picking rate (120 units/hr)
Production	Production time (1 min/unit)	Production rate (60 units/hr)
Reliability	Mean time between failures (10 years)	Failure rate (10%/year)

and is usually specified in terms of days. Its inverse, the inventory turn rate, measures the frequency at which the inventory stock is replenished or turned over, and is usually specified as an annual rate. An inventory system with 90 days of supply, for instance, has an inventory turn rate of 4 per year.

Days of supply and inventory turn rate are both popular in practice. In recent surveys that we conducted at three supply chain management conferences (Copperberg 2013, Marcus Evans 2013, Supply Chain Academy 2013), we asked 51 managers of manufacturing companies about the performance metrics used at their companies: 31% of the participants reported that they use days of supply, but not inventory turn rate, 31% that they use inventory turn rate, but not days of supply, and 28% that they use both metrics, while 10% used other metrics or did not provide answers. Similar results were reported by Harrison and New (2002) and Cohen et al. (2007). In informal interviews, we could not identify a consistent pattern in the rationales for choosing one metric versus the other.[1]

We are not the first to analyze the effect of equivalent metrics on decision making. Larrick and Soll (2008) analyzed how fuel efficiency metrics affect people's evaluations of fuel consumption. They conducted an experiment in which subjects had to reduce the fuel consumption of a car fleet. Subjects could increase the fuel efficiency of a fleet from 15 to 19 miles per gallon or, alternatively, of another car fleet from 34 to 44 miles per gallon. Only 25% of the participants chose the first, correct, option, that reduced fuel consumption more than twice as much as the second option. In another treatment, in which fuel efficiency was expressed in terms of the equivalent metric of gallons per 100 miles, 64% of the subjects chose the correct option. Although fuel consumption optimization and inventory optimization are decision problems from different fields, the problem structures are very similar. The setup of our first experiment is very closely related to the miles per gallon illusion experiments of Larrick and Soll (2008), and as we will show, we find similar decision biases.

The effect of metrics on decision making can be explained by attribute substitution: When

[1] To obtain insights into when these metrics are used, we consulted Knut Alicke, Master Expert of Supply Chain Management at McKinsey & Company (personal communication, May 13, 2016). Alicke has worked as a supply chain consultant with over 50 companies from various industries. He confirmed the initial insight from our surveys that there is no consistent pattern in the usage of the two metrics. While some of the companies he worked with use inventory turn rate, others use days of supply, and some use both. Alicke reported that in companies that use both metrics, inventory turn rate tends to be used frequently for financial reporting. Overall, he could not identify a clear relationship between hierarchical level or functional area and the use of the metrics that is consistent across companies.

confronted with a difficult question, people instead answer an easier question and are often actually unaware of the substitution (Kahneman and Frederick 2002), particularly if the relationships involved are non-linear (Sterman 2002). Individuals do not necessarily make decisions that optimize the fundamental attribute; instead, they optimize media that are more readily available (Hsee et al. 2003). We use attribute substitution to model the effect of metrics on inventory valuations. We show that decision makers tend to optimize the values of the metric that is used to measure inventory performance, as opposed to the fundamental attribute that is actually relevant, the inventory value. The relationship between the days of supply metric and the inventory value is linear, and decision makers who substitute inventory value by days of supply correctly value inventory changes. The relationship between the inventory turn rate metric and the inventory value is convex, and decision makers who substitute inventory value by the inventory turn rate over-value inventory changes.

We use a deductive approach for analyzing the effect of inventory metrics on decision making. We perform three laboratory studies that are based on different problem types, namely choice, effort, and cost optimization problems. Many inventory optimization problems are of these types, and we analyze one in each of our three studies.

In Study 1, we adapt Larrick and Soll's (2008) miles per gallon experiment to analyze an operations management context. We consider a choice problem, in which a decision maker must select an inventory optimization option from a set of options. Inventory managers face this type of decision when they decide on the business units, locations, or processes to optimize or when they prioritize such optimizations. We hypothesize that people more frequently decide optimally under the days of supply metric, which is proportional to inventory value, than under the inventory turn rate metric, which is convex in inventory value (Hypothesis 2.1). The results of our experiments support the hypothesis. In our main experiment, 89.3% of the choices are optimal under the days of supply metric, compared with 42.4% under the inventory turn rate metric.

Our behavioral models build on attribute substitution. Applied to our setting, attribute substitution suggests that decisions are affected by the value of the metric and not only by the inventory value, which would be optimal. We argue that the extent to which people rely

on the value of the metric depends on their cognitive reflection, which we quantify by the cognitive reflection test (CRT; Frederick 2005) score. We hypothesize that people with high CRT scores rely less on the metric and make better decisions than people with low CRT scores (Hypothesis 2.2). We analyze the CRT scores of the subjects of our main experiment and find support for the hypothesis. Subjects with high CRT scores solve up to 31.1% more problems optimally than those with low CRT scores.

In Study 2, we consider an effort problem, in which a decision maker determines the effort to invest in reducing inventory of a product. Similar problems occur, for instance, in lean manufacturing environments when employees are working on continuous optimization. Because the inventory turn rate is convex increasing in inventory reduction, it indicates a greater than actual effect of effort on inventory reductions, and we hypothesize that individuals invest more effort under the inventory turn rate metric than under the days of supply metric (Hypothesis 2.3). In our laboratory experiment, we find support for the hypothesis. Individuals invest on average 28.0% more effort under the inventory turn rate metric than under the days of supply metric.

In Study 3, we consider a cost optimization problem, in which a decision maker must determine the resources to allocate to ordering. Allocating resources to ordering is costly, but reduces inventory. The problem is closely related to the economic order quantity model, one of the standard models in inventory management. We hypothesize that individuals allocate more resources to ordering under the inventory turn rate than under the days of supply metric (Hypothesis 2.4), and our experimental results support the hypothesis. Subjects choose on average a 69.5% higher ordering cost under the inventory turn rate metric than under the days of supply metric.

In all of our studies, decisions are affected by the inventory metrics, and it seems reasonable to expect that decisions are also affected by the metrics in other inventory settings. Given the significances and magnitudes of the effects that we observe, our research suggests that management should pay close attention to choosing the right metric. In many instances, the days of supply metric will probably be the preferred choice because it is proportional to inventory value. However, there might exist situations in which the inventory turn rate

metric could be the preferred choice. Its convexity in inventory reduction could be utilized, for instance, to motivate people to reduce inventory in situations in which the inventory level is already low.

This chapter is organized as follows. Section 2.2 outlines the relevant literature. Section 2.3 presents our behavioral valuation model on which we base our hypotheses. Sections 2.4 through 2.6 contain our experimental studies. Section 2.7 concludes and discusses the managerial implications of our findings.

2.2 Related Literature

Two streams of literature are related to our research: the literature on proxy attributes and the presentation of information and the literature on dual process theory and cognitive reflection. We will review both streams below.

Proxy Attributes and the Presentation of Information. A metric such as days of supply or inventory turn rate can be considered a *proxy attribute* – an indirect and often more available measure of a more *fundamental attribute*, in our case inventory value (Keeney and Raiffa 1976). Proxy attributes are widely used in intuitive judgment and in many quantitative analyses in operations management. Kahneman and Frederick (2002) referred to the tendency to focus on a proxy attribute rather than assessing the fundamental attribute as *attribute substitution*. They revisited the earlier studies on heuristic judgment (Tversky and Kahneman 1974, 1983, Kahneman et al. 1982) and proposed a model of judgment heuristics in which the reduction of complex tasks to simpler operations is achieved by attribute substitution. Consistent with our findings, Fischer et al. (1987) found that decision makers who were presented with a proxy attribute did not necessarily translate it into the fundamental attribute and gave the proxy attribute more weight than they should have.

Keeney and Raiffa (1976) and Kahneman and Frederick (2002) argued that the use of proxy attributes can lead to systematic biases, if the computation of the fundamental attribute places a cognitive burden on the cognitive processes of decision makers. Managers often need to make quick decisions and intuitively consider the relationship between the proxy

and the fundamental attribute. Especially if the relationship is complex or probabilistic, they tend to rely on simplifying heuristics (Fischer et al. 1987). One of the major sources of complexity is non-linearity (Sterman 2002). Prior research has identified various situations in which people reason poorly if the relationship between the proxy and the fundamental attribute is non-linear. Perhaps the most noted misperception in this domain is *the MPG illusion* (Larrick and Soll 2008), whereby people rely on miles per gallon as a linear indicator of fuel efficiency. Larrick and Soll (2008) analyzed how people value fuel consumption. In a treatment in which they indicated fuel efficiency by the miles per gallon metric, people tended to over-value efficiency improvements of cars that are already efficient. In another treatment, in which fuel efficiency was expressed in terms of the equivalent metric gallons per 100 miles, which is linear in fuel efficiency, people were *nudged* toward better decisions (Thaler and Sunstein 2008, Johnson et al. 2012).

Svenson (1970, 2008) analyzed how people estimate time savings from increased driving speeds. He found that typical estimates are based on the differences in driving speeds instead of the actual time savings. Hsee et al. (2003) conducted an experiment in which they compared the effort of participants who were offered rewards that directly related to effort with that of participants who first received points that were later converted into rewards. Although the relationship between effort and reward was the same in all treatments, effort levels differed and depended on the number of points received. Another study showing the insensitivity to fundamental attributes is by Kagel et al. (1996), who analyzed how decisions are affected by proxy attributes. They conducted an experiment in which participants bargained over chips with different exchange rates. They found that participants' perceptions of fairness were more focused on the number of the chips than on the value of the chips. In another study, Soll et al. (2013) found that people expect that monthly credit payments have a roughly linear relationship with the payback period and therefore underestimate the payback period when monthly payments barely cover interest. In general, people often reason poorly about accumulation problems (see Cronin et al. 2009 and the references therein).

Even absent non-linearities, the presentation of seemingly equivalent information can affect decision making. Denes-Raj and Epstein (1994) conducted a study in which participants had

the chance to win a prize by drawing a red jelly bean from an urn. Their participants often preferred to draw from an urn containing a greater absolute number but a smaller proportion of red beans (for example, 7 in 100) than from an urn with fewer red beans but a better chance to win (for example, 1 in 10). In a series of studies, Slovic et al. (2000) analyzed the different reactions to risk presented as probability and risk presented as frequency. Experienced forensic psychologists and psychiatrists were asked to rate the likelihood that a patient would commit an act of violence. Clinicians who were given another expert's assessment of a patient's risk of violence framed in terms of relative frequency rated patients as more dangerous than those who were shown the "equivalent" risk assessment expressed as a probability. Similar results were obtained by Yamagishi (1997), whose participants perceived a disease that kills 1,286 people out of every 10,000 to be more dangerous than one that kills 24.14% of the population.

We build on the research on proxy attributes (Keeney and Raiffa 1976, Fischer et al. 1987) and attribute substitution (Kahneman and Frederick 2002) to model the effect of inventory metrics on inventory valuations. Consistent with medium maximization (Hsee et al. 2003), we show that individuals are affected by the value of the metrics and do not rely exclusively on the values of the fundamental attribute, such that metrics that contain the same information, but are framed differently, result in different decisions. The design of our Study 1 is similar to that used by Larrick and Soll (2008), and Study 2 is related to the experimental design of Hsee et al. (2003). Study 3 is not directly related to previous work in behavioral decision making but considers a setting that is closely related to one of the standard models of inventory management, that is, the economic order quantity model (Harris 1990, Erlenkotter 1990).

The results of our experiments are in line with the decision biases discussed above. Under the inventory turn rate metric, individuals over-value inventory reductions and, compared with decisions under the days of supply metric, they choose worse inventory optimization options, invest more effort optimizing inventory of specific products, and choose a higher ordering cost.

The literature has shown that people are not equally prone to such decision biases and that cognitive reflection can explain some of the variation in decision outcomes. Therefore, we next review the literature on dual process theory and cognitive reflection.

Dual Process Theory and Cognitive Reflection. In dual process theory, cognitive processes are partitioned into two qualitatively different but inter-operating types of thinking style systems (Epstein 1994, Sloman 1996). There is a rich body of literature on how cognitive processes can be defined (see Stanovich and West (2000) and Evans (2008) for an overview), with the common notion that one process is more intuitive and the other process more rational than the other. Stanovich and West (2000) and Kahneman and Frederick (2002) refer to the cognitive processes as System 1 and System 2. System 1 is intuitive, fast, automatic, and effortless, while System 2 is reflective, slow, rational, and effortful. If an individual faces a problem, System 1 generates suggestions for System 2. System 2 can endorse or override these suggestions.

Frederick (2005) proposes the CRT to measure the extent to which a person uses System 2. The CRT consists of three questions, such as "A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?". The intuitive answers are wrong – in the example, the intuitive answer is \$0.10, but the correct answer is \$0.05. The extent to which individuals choose the non-intuitive answers is measured by the CRT score, which corresponds to the number of correct answers on the test. The CRT score indicates how likely an individual is to reflect on an answer, that is, to use System 2 to override an incorrect intuitive System 1 suggestion as opposed to endorsing it. The objective nature of the CRT makes it an attractive candidate for understanding decision biases in our experiment (Oechssler et al. 2009, Toplak et al. 2011). It is brief, easy to administer, unambiguous, and widely used in laboratory experiments. Moreover, it has been proven to be a reliable predictor for task performance in the behavioral operations management literature. Individuals with high cognitive reflection have higher forecasting performance (Moritz et al. 2014) and perform better in newsvendor-type decisions (Moritz et al. 2013), and this effect is robust to controlling for intelligence. Narayanan and Moritz (2015) showed that cognitive reflection also provides a powerful predictor for the bull-whip effect and supply chain performance in a multi-echelon setting.

We build on previous research on proxy attributes to analyze the tendency of decision makers to rely on inventory metrics as a proxy for the more fundamental attribute, inventory

value. Prior research has analyzed the effect of cognitive reflection on decision making and shown that individuals with high cognitive reflection are less prone to decision biases than those with low cognitive reflection (Oechssler et al. 2009, Toplak et al. 2011). Moritz et al. (2013) were the first to analyze cognitive reflection in an inventory management context. Like us, they use the CRT score (Frederick 2005) as a proxy for cognitive reflection. In addition, we use decision time and calculator use as proxies for cognitive reflection. We find that individuals with high cognitive reflection make on average better decisions and than those with low cognitive reflection. Moreover, we show that the performance metric can moderate the effect that cognitive reflection has on task performance.

2.3 Behavioral Valuation Model

We are interested in understanding how inventory decisions are affected by the metrics that are used to indicate inventory performance. The fundamental measure of inventory performance is the inventory value. It quantifies the capital that is tied up in inventory, and profit-maximizing ("rational") individuals rely on it in their decision making. If the metrics days of supply or inventory turn rate are used to indicate inventory performance, rational individuals determine the corresponding inventory values and base their decisions on them. Cognitive science research indicates that not all decision makers use this approach and that some base their decisions on the metrics.

2.3.1 Inventory Performance Metrics

The value of the capital that is tied up in inventory, the *inventory value*, is the fundamental measure of inventory performance. For a product with unit cost c and inventory level I, the inventory value is

$$M = cI. \tag{2.1}$$

To evaluate the efficiency of inventory usage over time or to compare inventory across companies, locations, or products, the performance metrics days of supply and inventory turn rate are commonly used (Hausman 2004). The *days of supply* metric relates the inventory

value to the cost of goods sold. For a demand rate of d, the cost of goods sold is cd and the days of supply metric is defined as

$$T = \frac{M}{cd} = \frac{I}{d}, \tag{2.2}$$

where we use the variable T to indicate that it is a time-based measure. The days of supply metric measures the average duration that products are held in inventory, and a lower value indicates higher performance.

The *inventory turn rate* metric relates the cost of goods sold to the inventory value and is defined as

$$R = \frac{c\,d}{M} = \frac{d}{I}, \tag{2.3}$$

where we use the variable R to indicate that it is a rate-based measure. The inventory turn rate metric measures the frequency at which the inventory stock is replenished, and a higher value indicates higher performance. Because the days of supply metric is the inverse of the inventory turn rate metric, the two metrics are equivalent and a rational individual makes the same decisions under either metric.

2.3.2 Valuation of Inventory Reductions by Metrics

One of the key tasks of supply chain managers is to identify and implement improvements that reduce inventory. Inventory reduction can be achieved, for instance, by reducing supply lead times, automating order processing, or improving demand forecasting accuracy (see, for example, Cachon and Terwiesch 2013). Such activities require effort or financial investments, and to determine which of them to pursue, the value of the inventory reductions that they achieve must be determined. We denote the initial inventory level by I_0 and the inventory level after the reduction by I_1. An inventory level reduction of $I_0 - I_1$ reduces the inventory value by $V_M = c(I_0 - I_1)$.

Because days of supply is the reciprocal of the inventory turn rate ($T = 1/R$), the metrics are equivalent and rational individuals make the same decisions under either metric. However, if inventory performance is measured by the days of supply or inventory turn rate metric, we expect that some individuals will not invest the cognitive effort required to compute the

inventory value and will instead value inventory based on the metric. We next analyze how such individuals value inventory reductions.

Days of Supply. If an individual uses the days of supply metric as a proxy and substitute for inventory value, then the value assigned to a reduction in the days of supply metric from $T_0 = I_0/d$ to $T_1 = I_1/d$ is

$$V_T = t(T_0 - T_1), \tag{2.4}$$

where the parameter t is the value that an individual associates with a unit decrease in the days of supply metric. Following Larrick and Soll (2008), we use linear relationships between the proxy measure and the valuation. To express V_T as a function of the inventory levels, we replace T_0 with I_0/d and T_1 with I_1/d and obtain

$$V_T = \frac{t}{d}(I_0 - I_1), \tag{2.5}$$

which is the value that an individual relying on Equation (2.4) assigns to an inventory level reduction from I_0 to I_1.

Inventory Turn Rate. If an individual uses the inventory turn rate metric as a proxy and substitute for inventory value, then the value assigned to an increase in the inventory turn rate metric from $R_0 = d/I_0$ to $R_1 = d/I_1$ is

$$V_R = r(R_1 - R_0), \tag{2.6}$$

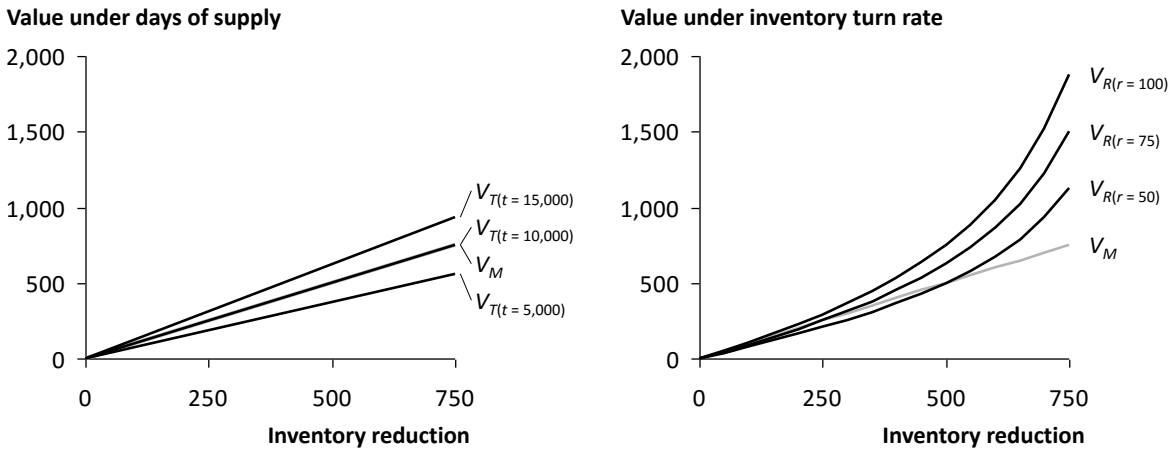
where the parameter r is the value associated with a unit increase in the inventory turn rate metric and we again use a linear relationship between the proxy measure and the valuation.

To express V_R as a function of the inventory levels, we replace R_0 with d/I_0 and R_1 with d/I_1 and obtain

$$V_R = rd\left(\frac{1}{I_1} - \frac{1}{I_0}\right), \tag{2.7}$$

which is the value that an individual relying on Equation (2.6) assigns to an inventory level reduction from I_0 to I_1. The function is strictly convex increasing in the inventory reduction,

Figure 2.1 Valuation of Inventory Reductions by Days of Supply and Inventory Turn Rate ($c = 1$, $d = 10{,}000$, $I_0 = 5{,}000$)



whereas the optimal valuation is linear increasing in it. Therefore, there does not exist a constant value for r for which the valuation is correct over a range of inventory reductions.

To optimally evaluate inventory reductions, a decision maker must use the function $r = cI_0I_1/d$. Rational decision makers use this function, but those who rely on the substitution heuristic and postulate a linear relationship between the inventory turn rate metric and the valuation use a value of r that is independent of the inventory reduction.

Figure 2.1 provides an example to illustrate how inventory changes are valued by subjects relying on the metrics. The left graph shows the valuation under the days of supply metric. The gray line indicates the optimal valuation (V_M), which is the same valuation as that under the days of supply metric with an optimal parameter value for $t = cd$ ($V_{T(t=10,000)}$). If changes in days of supply are over-valued ($t = 15,000$) or under-valued ($t = 5,000$), the days of supply valuation differs from the optimal valuation, but both depend linearly on the inventory reduction. This implies, for instance, that the value assigned to an inventory reduction is independent of the initial inventory level, which is optimal.

The right graph shows the valuation under the inventory turn rate metric. The valuation is convex increasing in the inventory reduction, which implies that for any fixed value of r, the value assigned to a given inventory reduction depends on the initial inventory level. This implies that sufficiently large inventory reductions, that is, inventory reductions that are greater than those where V_M and V_R intersect, are over-valued.

2.4 Study 1: Effect of Performance Metrics on Investment Decisions

A common management task is selecting investments from a set of investment options with different returns. Managers must decide, for instance, which of several business units, locations, or processes to optimize. We consider such a problem, in which a decision maker must determine which of multiple inventory optimization options to choose. The effect of the optimization options is indicated by the days of supply or inventory turn rate metric.

2.4.1 Behavioral Investment Models

Consider two alternative inventory optimization options for two products A and B. The initial inventory level of product A is I_0^A, and an investment in inventory optimization reduces it to I_1^A. The initial inventory level of product B is I_0^B, and an investment in inventory optimization reduces it to I_1^B. To keep the model parsimonious, we consider products with the same unit costs, demand rates, and investment costs but with different initial inventory levels and different inventory reductions.

A rational decision maker values the optimization options based on their effect on the *inventory value* and chooses Option A if

$$V_M^A = c(I_0^A - I_1^A) > c(I_0^B - I_1^B) = V_M^B \tag{2.8}$$

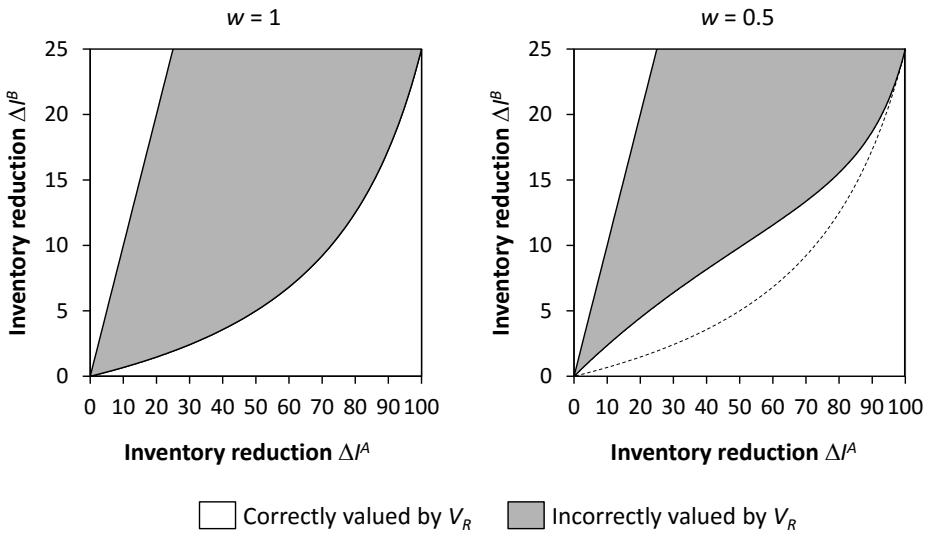
and Option B otherwise.

If the optimization options are valued by the *days of supply* metric, Option A is chosen if

$$V_T^A = \frac{t}{d}(I_0^A - I_1^A) > \frac{t}{d}(I_0^B - I_1^B) = V_T^B \tag{2.9}$$

and Option B is chosen otherwise. The only difference between the valuations by Equations (2.8) and (2.9) is that the inventory reductions in Equation (2.9) are scaled by a factor t/d. Because the factor is the same for both options, the decisions are the same under both valuations and optimal choices are made under the days of supply metric.

Figure 2.2 Effect of Initial Inventory and Inventory Reduction on Valuation under Inventory Turn Rate Metric



If the optimization options are valued by the *inventory turn rate* metric, Option A is chosen if

$$V_R^A = rd\left(\frac{1}{I_1^A} - \frac{1}{I_0^A}\right) > rd\left(\frac{1}{I_1^B} - \frac{1}{I_0^B}\right) = V_R^B \tag{2.10}$$

or, equivalently, if

$$(I_0^A - I_1^A)I_0^B I_1^B > (I_0^B - I_1^B)I_0^A I_1^A. \tag{2.11}$$

Otherwise, Option B is chosen.

The choices under the inventory turn rate metric are not always optimal because the inventory turn rate over-values inventory reductions if the initial inventory is small or the inventory reduction is large. The left graph in Figure 2.2 depicts an example with initial inventory levels $I_0^A = 100$ and $I_0^B = 25$. The gray area indicates where decision makers who rely on the inventory turn rate metric make the wrong decisions.

Thus far, we have discussed how people decide if they rely exclusively on one of the metrics or on the inventory value. Hsee et al. (2003) demonstrated that individuals do not necessarily rely exclusively on the proxy or the fundamental attribute but that their valuations are affected by the values of the proxy and the fundamental attribute, if both values are available.

Translated to our setting, this finding suggests that individuals who compute the inventory value based on the metric do not necessarily rely exclusively on the inventory value but also on the value of the metric.

Such behavior can be modeled using a weighted average of the value of the metric and the inventory value. We denote the weight that a decision maker places on the metric by w $(0 \leq w \leq 1)$ and the weight he or she places on the inventory value by $(1 - w)$. The extreme cases of $w = 0$ and $w = 1$ correspond to decision makers who rely on inventory value or the value of the metric only, respectively. Values of w strictly above zero and strictly below one model the combined approach suggested by Hsee et al. (2003). It can be shown that for all strictly positive weights w, some decisions are incorrect under the inventory turn rate metric. The right graph in Figure 2.2 shows an example for $w = 0.5$. The gray area is decreasing in w, but it always exists for weights $w > 0$, such that some decisions are not made optimally under the inventory turn rate metric. Under the days of supply metric, individuals are not prone to such decision biases, and we hypothesize the following:

Hypothesis 2.1. *Optimal investment decisions are made more frequently under the days of supply metric than under the inventory turn rate metric.*

2.4.2 Effect of Individual Thinking Styles on Decisions

To gain a better understanding of the drivers behind the potential heterogeneity of the decisions, we draw from the theory of cognitive science. We use dual process theory, which has already been successfully applied to understand heterogeneity in forecasting and decision making in the newsvendor problem (Moritz et al. 2013, 2014).

We follow Stanovich and West (2000) and Kahneman and Frederick (2002) and refer to the cognitive processes as System 1 and System 2. System 1 is intuitive, fast, automatic, and effortless, while System 2 is reflective, slow, rational, and effortful (see Section 2.2 for details). If an individual faces a problem, System 1 generates suggestions for System 2. System 2 can endorse or override these suggestions. In our problem, the option that increases the metric the most can be considered the intuitive suggestion because the metric is the proxy attribute that is directly available to the decision maker. If System 2 endorses the suggestion in the

inventory turn rate treatment, the wrong decision can be made. If System 2 is alerted and overrides an incorrect intuitive suggestion, it becomes more likely that the optimal decision is made.

Frederick (2005) proposes the CRT to measure the extent to which an individual uses System 2. The higher an individual's tendency to override an incorrect intuitive response of System 1, the higher the probability that the problem is solved optimally. Individuals with high CRT scores do not only override System 1 more frequently than individuals with low CRT scores, but they are also generally more likely to make optimal choices. Frederick (2005) and Toplak et al. (2011) found that the CRT overlaps with intelligence and cognitive ability. Therefore, we expect that individuals with high CRT scores make better decisions than individuals with low CRT scores under both metrics but that the effect is stronger under the inventory turn rate metric, where the suggestion of System 1 must be overridden, which leads to the following hypotheses:

Hypothesis 2.2.

(a) *Under the days of supply metric, individuals with high CRT scores make optimal investment decisions more frequently than individuals with low CRT scores.*

(b) *Under the inventory turn rate metric, individuals with high CRT scores make optimal investment decisions more frequently than individuals with low CRT scores.*

(c) *The effect of the CRT score on the frequency of optimal choices is stronger under the inventory turn rate metric than under the days of supply metric.*

2.4.3 Investment Experiment

We conducted a laboratory experiment in which human subjects had to decide between two inventory optimization options, where one option reduced inventory more than the other. In the experiment, we used two treatments, a days of supply treatment and an inventory turn rate treatment, that differed only in how the performance of the inventory system was measured.

Table 2.2 Treatments of Investment Experiment

	Days of supply				Inventory turns rate			
	Option A		Option B		Option A		Option B	
Problem	Initial	Optimized	Initial	Optimized	Initial	Optimized	Initial	Optimized
1	60	30	20	10	6	12	18	36
2	24	15	8	5	15	24	45	72
3	120	72	60	36	3	5	6	10

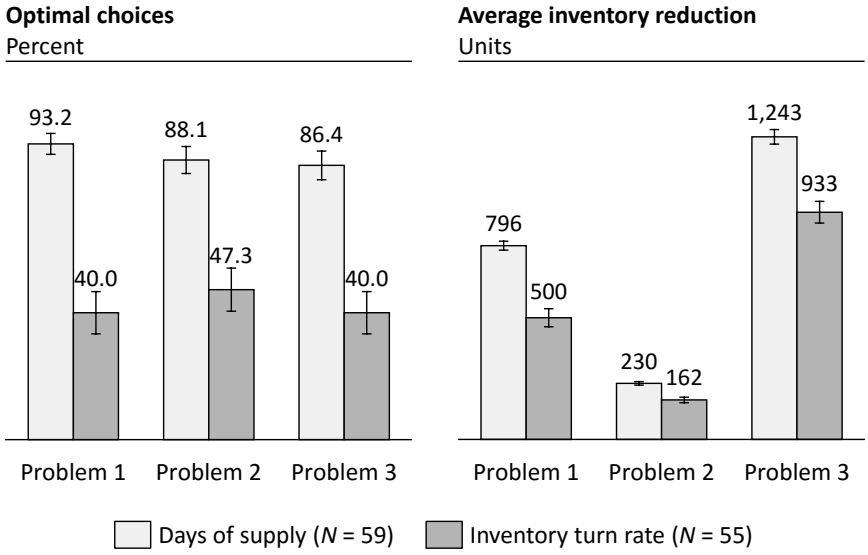
All sessions followed the same protocol. At the beginning of the experiment, subjects were randomly assigned to one of the two treatments and received the corresponding instructions (Supplementary Material 2.A). The written instructions explained how the performance metrics are computed and provided an example. Subjects were informed that they had to manage a warehouse with two products with the same unit costs and annual demand rates of 10,000 units but with different initial days of supply or inventory turn rates. Subjects were informed that they could optimize the inventory of one of the products and that they would receive a payment of 10 experimental currency units (ECUs) for each unit of inventory reduction. They were also informed that the exchange rate would be 1 euro per 3,000 ECUs. After reading the instructions, subjects could ask questions that the instructor answered privately.

During the experiment, subjects made the three investment decisions shown in Table 2.2. The problems were designed to achieve variation in the absolute values of the optimal inventory reduction and in the difference in the inventory reduction achieved under the two options. The problems were presented sequentially, and the sequence was randomized. Participants could make decisions at their own pace and were informed that the experiment was not time restricted.

After they had made their investment decisions, subjects took the CRT, stated whether they already knew the questions, and completed a post-experimental questionnaire, in which we asked questions regarding participants' attitudes and preferences, as well as general questions about the experiment. We also collected demographic data.

A total of 114 students from the faculty of Management, Economics and Social Sciences of

Figure 2.3 Results of Investment Decision Experiment



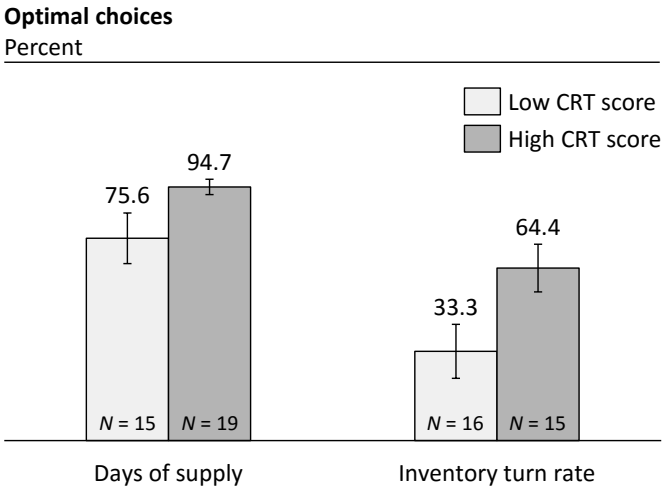
Note. Error bars indicate one standard error.

the University of Cologne were recruited via the Online Recruitment System for Economic Experiments (ORSEE, Greiner 2004). The experiment was conduced in six sessions. In total, 59 subjects were assigned to the days of supply treatment and 55 subjects to the inventory turn rate treatment. The sessions lasted 45 minutes on average and were programmed and conducted with the software z-Tree (Fischbacher 2007). The average payment was 9.29 euros, including a participation fee of 2.50 euros.

2.4.4 Results

The fractions of optimal choices are shown in the left graphs in Figure 2.3. Averaged over all problems, 89.3% of the decisions were optimal in the days of supply treatment and 42.4% were optimal in the inventory turn rate treatment. The difference in the aggregate fraction is significant (Wilcoxon test, one sided, $p < 0.001$), as are the differences in the fractions for the individual problems ($\chi^2\,(1,\, N = 114)$, $p < 0.001$ for all problems). We conclude that optimal investment decisions are made more frequently under the days of supply than under the inventory turn rate metric, which provides support for Hypothesis 2.1.

Figure 2.4 Effect of CRT Score on the Fraction of Optimal Choices



Note. Error bars indicate one standard error.

The right graphs in Figure 2.3 depict the inventory reductions that were achieved under both metrics. The inventory reductions are related to the optimal choices but are also affected by the magnitudes of the inventory reductions stipulated in the problems. In the days of supply treatment, the average total inventory reduction was 2,269 units and significantly greater than the reduction of 1,595 units in the inventory turn rate treatment (Wilcoxon test, one-sided, $p < 0.001$).

We next analyze the effect of the CRT scores on decision making. Of the 114 subjects, 49 stated that they already knew the CRT questions before the experiment, and we exclude them from the analyses. Following Oechssler et al. (2009) and Hoppe and Kusterer (2011), we pool the CRT scores of the remaining 65 subjects into a low CRT score group (CRT scores of 0 or 1) and high CRT score group (CRT scores of 2 or 3).

Figure 2.4 shows the results. Under both metrics, subjects with high CRT scores more frequently decided optimally than did those with low CRT scores. Compared with the group with low CRT scores, the group with high CRT scores solved 19.1% more problems optimally in the days of supply treatment and 31.1% more problems optimally in the inventory turn rate treatment. Both differences are significant (Wilcoxon test, one-sided, $p = 0.040$ and $p = 0.017$, respectively), which provides support for Hypotheses 2.2(a) and 2.2(b).

The results also indicate that the effect of the CRT score on performance is higher under the inventory turn rate metric than under the days of supply metric. We test the significance of the differences using a fractional logit model (Papke and Wooldridge 1996) with the fraction of optimal decisions as the dependent variable (see Table 2.3). *Metric* equals one in the days of supply treatment and zero in the inventory turn rate treatment. *CRT* equals one if the subject belongs to the high CRT score group and zero otherwise. The regression analysis yields a significant effect for the metric and the CRT group but a non-significant effect for the interaction of CRT group and metric.

We used the CRT score as an indicator of System 2 thinking in our analyses. To analyze the robustness of the results, we considered alternative indicators for System 2 thinking, that is, decision time (Dane and Pratt 2007) and calculator use (Rosenboim et al. 2013). We replaced the variable *CRT* in the regressions with the variables *Time* (average time that a subject required to reach a decision) and *Calculator* (number of decisions in which subjects used a calculator). The results of the corresponding regressions are shown in Table 2.3.

For decision time, the results are similar to those for CRT score, but the significance levels are higher. For calculator use, the results are also similar and more pronounced. Using a calculator does not significantly improve decisions in the days of supply treatment ($p = 0.721$), but using one does in the inventory turn rate treatment ($p < 0.001$). Note that the interaction term is also significant ($p = 0.050$), indicating that calculator use has significantly lower value in the days of supply than in the inventory turn rate treatment.

Overall, the results show that optimal investment decisions are made more frequently in the days of supply than in the inventory turn rate treatment. The results also indicate that System 2 thinking (operationalized by high CRT score, long decision time, and calculator use) is beneficial, in particular if the inventory turn rate is used.

We conducted the experiments in a controlled laboratory environment at the University of Cologne. The subjects were pre-experienced students from the faculty of Management, Economics and Social Sciences with an average age of 23.6 years and little or no work experience. Given the background of the students, it is unlikely that they had experience in making investment decisions, such as those that they made in the experiment. To analyze

Table 2.3 Effect of System 2 Thinking on Performance in Investment Decisions

	CRT score			Decision time			Calculator use		
Variable	DOS only	ITR only	Full sample	DOS only	ITR only	Full sample	DOS only	ITR only	Full sample
Metric			1.822**			3.169***			3.252***
			(0.667)			(0.640)			(0.504)
CRT	1.762**	1.288**	1.288**						
	(0.758)	(0.587)	(0.582)						
CRT × *Metric*			0.474						
			(0.952)						
Time				0.017	0.030***	0.030***			
				(0.011)	(0.011)	(0.011)			
Time × *Metric*						−0.013			
						(0.015)			
Calculator							0.104	0.799***	0.799***
							(0.292)	(0.203)	(0.202)
Calculator × *Metric*									−0.694**
									(0.354)
Constant	1.128***	−0.693	−0.693	1.675***	−1.493***	−1.493***	2.058***	−1.193***	−1.193***
	(0.503)	(0.447)	(0.443)	(0.458)	(0.451)	(0.449)	(0.403)	(0.306)	(0.305)
Log-likelihood	−10.75	−17.14	−27.90	−17.34	−29.32	−46.66	−17.72	−27.87	−45.58
Observations	34	31	65	59	55	114	59	55	114

Notes. Fractional logit regression. Robust standard errors in parentheses.

* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01, two-tailed.

DOS = days of supply, ITR = inventory turn rate.

Table 2.4 Inventory Investment Decisions of Managers ($N = 51$)

	Berlin		Munich		Stockholm		Total	
Metric	Subjects	Optimal	Subjects	Optimal	Subjects	Optimal	Subjects	Optimal
Days of supply	5	2 (40.0%)	9	9 (100%)	9	6 (66.7%)	23	17(73.9%)
Inventory turn rate	3	0 (0%)	10	7 (70.0%)	15	6 (40.0%)	28	13(46.4%)

whether individuals with experience in investment decisions are also subject to the decision biases we observed with students, we conducted an additional experiment with actual supply chain managers.

2.4.5 Experiment With Managers

We identified three business conferences that targeted managers at the vice presidential level and above and addressed inventory optimization: the Inventory Optimization Workshop in Berlin (Marcus Evans 2013), the Supply Chain Executive Academy in Munich (Supply Chain Academy 2013), and the Spare Parts Business Platform in Stockholm (Copperberg 2013).

At the conferences, we distributed questionnaires and asked the participants to consider a warehouse where the inventory of three products can be optimized, but budget restrictions allow them to optimize the inventory of only one product (Supplementary Material 2.B). The products had the same unit costs and demand rates. In the days of supply treatment, days of supply could be reduced from (A) 120 to 90 days, (B) 36 to 18 days, or (C) 15 to 9 days. In the inventory turn rate treatment, we provided the corresponding inventory turn rates that could be increased from (A) 3 to 4 turns per year, (B) 10 to 20 turns per year, or (C) 24 to 40 turns per year. In both treatments, A is the optimal choice.

The results of the experiment are shown in Table 2.4. At all conferences, the managers performed better under the days of supply metric than under the inventory turn rate metric. Under the days of supply metric, 73.9% of the decisions were optimal, a fraction that is significantly higher than the fraction of 46.6% optimal decisions under the inventory turn rate metric (Wilcoxon test, one-sided, $p = 0.025$).

There is substantial heterogeneity in the results across conferences, which could be attributed to the relatively small sample sizes per conference, the different backgrounds of the participants,

or the different topics covered at the conferences before the experiment. To control for such factors in the analysis, we conducted a logistic regression analysis with the metric as the independent variable and the binary decision as the dependent variable, using fixed effects for the conferences. The regression shows a significant effect of the metric on the decision (odds ratio = 0.178, $p = 0.024$). The results of this experiment indicate that the investment decisions of supply chain managers are affected by the equivalent metric used and provide additional support for Hypothesis 2.1.

2.5 Study 2: Effect of Performance Metrics on Effort

Companies continuously seek to increase their operational efficiency and reduce inventory levels by, for instance, implementing lean management practices and continuous process improvements (Chen et al. 2005, 2007, Alan et al. 2014). Such activities require effort. We analyze how the effort that people invest is affected by the performance metric used. Because different performance metrics assign different values to the effect of effort, the choice of metric can influence employee motivation and effort.

2.5.1 Behavioral Effort Model

Consider an individual who must determine the effort to invest in inventory optimization. We denote the effort cost function by $E(a)$ and assume that the function is convex and increasing in the effort level a. The effort that the decision maker invests determines the inventory level. We denote the inventory level function by $I(a)$ and the initial inventory level by $I(0)$. The function is strictly convex decreasing in effort and converges to a positive inventory level as effort goes to infinity. This functional form models the standard setting, in which more beneficial improvements are implemented before less beneficial ones. The monetary value of the inventory reduction associated with effort level a is $V_M(a) = c(I(0) - I(a))$.

The *days of supply* metric has the following value at effort level a:

$$V_T(a) = \frac{t}{d}\,(I(0) - I(a)) - E(a). \tag{2.12}$$

A decision maker who places weight $0 \leq w \leq 1$ on the metric and weight $(1-w)$ on the inventory value assigns the value

$$\widetilde{V}_T(a) = w\frac{t}{d}\left(I(0) - I(a)\right) + (1-w)c\left(I(0) - I(a)\right) - E(a) \tag{2.13}$$

to effort level a.

The *inventory turn rate* metric has the following value at effort level a:

$$V_R(a) = rd\left(\frac{1}{I(a)} - \frac{1}{I(0)}\right) - E(a). \tag{2.14}$$

A decision maker who places weight $0 \leq w \leq 1$ on the metric and weight $(1-w)$ on the inventory value assigns the value

$$\widetilde{V}_R(a) = wrd\left(\frac{1}{I(a)} - \frac{1}{I(0)}\right) + (1-w)c\left(I(0) - I(a)\right) - E(a) \tag{2.15}$$

to effort level a.

We are interested in comparing the optimal effort levels under the days of supply and inventory turn rate metrics, which requires specifying the parameters t and r. For our analyses, we use the parameter values at the initial effort level, that is, $t = cd$ and $r = cI^2(0)/d$. All results still hold if the parameters are determined at any effort level between zero and the optimal effort level under the days of supply metric.

The function $\widetilde{V}_T(a)$ is convex in the effort level, and the optimal effort level under the days of supply metric solves the first-order condition $-cI'(a_T^*) = E'(a_T^*)$. At this effort level, the first derivative of the function $\widetilde{V}_R(a)$ is

$$\widetilde{V}_R'(a_T^*) = \left(1 - \frac{I^2(0)}{I^2(a_T^*)}\right) wcI'(a_T^*). \tag{2.16}$$

$\widetilde{V}_R'(a_T^*)$ is non-negative for $0 \leq w \leq 1$ and strictly positive for $0 < w \leq 1$, which implies that the optimal effort level under the inventory turn rate metric is higher than the optimal effort level under the days of supply metric, that is, for individuals who place some weight on the metric. We hypothesize the following:

Hypothesis 2.3. *The average effort is higher under the inventory turn rate metric than under the days of supply metric.*

2.5.2 Effort Experiment

We analyzed the effect of the metrics on effort in a laboratory experiment, in which human subjects invested real effort to reduce inventory. The experiment used two treatments, a days of supply treatment and an inventory turn rate treatment, that differed only in how the performance of the inventory system was indicated.

All experimental sessions followed the same protocol. Subjects received written instructions about the experiment that explained how the performance metrics are computed and provided examples (Supplementary Material 2.C). Subjects were informed that they had to manage the inventory of a single product with an annual demand rate of 10,000 units and an initial average inventory level of 5,000 units. They were told the initial values of the performance metrics, that is, the initial value of the days of supply metric of 180 days or the initial value of the inventory turn rate metric of 2 per year. Subjects were informed that they could invest effort to reduce inventory and would receive a payment of 10 ECUs for each unit of inventory reduction. They were also informed that the exchange rate would be 1 euro per 5,000 ECUs.

We decided to design a real effort experiment because real work better captures fatigue, boredom, excitement and other affectations not present in monetary effort designs (see, for example, van Dijk et al. 2001, Carpenter et al. 2010). The effort task required subjects to position sliders on a computer screen using the computer mouse (Gill and Prowse 2012). The instructions stated the relationship between effort, measured by the number of sliders moved correctly from the initial position of 0 to the target position of 50, and average inventory:

$$\text{Average inventory level} = \frac{5,000 \text{ units}}{1 + 0.1 \cdot \text{Number of sliders positioned correctly}}. \tag{2.17}$$

Before the actual experiment, subjects completed a quiz consisting of five questions to ensure that they understood the effect of their effort on the average inventory level and the performance metric. The first question concerned the functional relationship between the

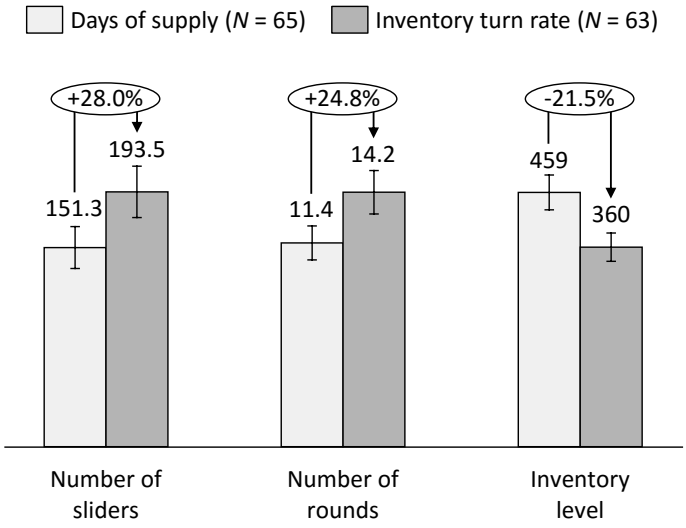
inventory level and the performance metric. The second question asked for the initial average inventory level (5,000 units) and the third question for the initial value of the performance metric (180 days or 2 per year). The fourth question asked for the inventory level after the first ten sliders were positioned correctly (2,500 units). The fifth question asked for the corresponding value of the performance metric (90 days or 4 per year).

If all five questions were answered correctly on the first attempt, subjects received 1,000 ECUs. If they needed two attempts, they received 500 ECUs. If they needed more than two attempts, they did not receive any compensation for completing the quiz. Subjects could not continue without having answered all five questions correctly; 113 subjects needed one attempt, five subjects needed two attempts, and ten subjects needed more than two attempts.

After the quiz, the actual experiment started. The experiment was played in rounds. At the beginning of a round, a screen with 48 sliders appeared, all set at an initial value of zero (see Supplementary Material 2.C for a screenshot). Subjects had two minutes to position up to 48 sliders and were informed of the time remaining in each round. In the experiment, the maximum number of sliders that a subject positioned correctly in a round was 28. After a slider was positioned correctly, the performance metric was updated. After each round, subjects could decide whether they wanted to stay for another round or to terminate the experiment. Subjects were informed that they could play as many rounds as they wished. We had to terminate the experiment for one subject in the inventory turn rate treatment after 50 rounds (nearly 120 minutes total) to avoid overlap with subjects of the subsequent session.

A total of 128 students from the faculty of Management, Economics and Social Sciences of the University of Cologne were recruited via the online recruiting system ORSEE (Greiner 2004). We ran 48 sessions and invited three students per session. To avoid having subjects who terminated the experiment affecting the effort decisions of other subjects, we placed the subjects into individual rooms, such that they could not observe one another. Subjects arrived at the instructor's office and were randomly assigned to treatments and rooms; 65 subjects were assigned to the days of supply treatment and 63 to the inventory turn rate treatment. The average compensation was 9.62 euros.

Figure 2.5 Effect of Metric on Average Invested Effort and Average Final Inventory Value



Note. Error bars indicate one standard error.

2.5.3 Results

Figure 2.5 summarizes the results. It shows that subjects, on average, invested more effort in the inventory turn rate treatment than in the days of supply treatment: They moved significantly more sliders (Wilcoxon test, one-sided, $p = 0.011$) and played significantly more rounds (Wilcoxon test, one-sided, $p = 0.025$), which provides support for Hypothesis 2.3. The figure also depicts the average final inventory level under both metrics. In the inventory turn rate treatment, the final inventory level was significantly lower than in the days of supply treatment (Wilcoxon test, one-sided, $p = 0.011$).

The results can be explained by the non-linear relationship between the inventory turn rate metric and the inventory value, which leads individuals to overestimate the impact of the effort they invested in inventory reduction. Therefore, it is more likely that an individual exerts higher effort and achieves lower inventory levels under the inventory turn rate than under the days of supply metric.

2.6 Study 3: Effect of Performance Metrics on Inventory Decisions

The two fundamental inventory models used in supply chain management are the economic order quantity model (Harris 1990, Erlenkotter 1990) and the newsvendor model (Arrow et al. 1951). A large body of literature exists that analyzes variations and extensions of these models (see, for example, Zipkin 2000). More recently, the behavioral aspects of inventory management have been addressed (Schweitzer and Cachon 2000, Bolton et al. 2012, and the references therein). The focus of this stream of research has been on analyzing decision biases and human preferences in the newsvendor setting. Our interest is in understanding the effect of inventory metrics on decision making, and we will use the simpler economic order quantity model for our analyses.

2.6.1 Behavioral Inventory Model

The economic order quantity model considers ordering and inventory holding costs. Each time an order is placed, a fixed order cost of K is charged. Orders are delivered instantaneously and placed in inventory, where they are held at an inventory holding cost per unit of h. The demand rate d is deterministic and constant, and all demand is filled from inventory.

The classical economic order quantity model considers an operational perspective and analyzes how optimal order quantities of individual products can be determined. The decision variable is typically order quantity, which is the key decision variable for inventory planners, who are in charge of placing orders with suppliers. Inventory managers who are responsible for managing larger organizational entities, such as departments, typically use aggregated metrics to assess inventory performance (Harrison and New 2002, Cohen et al. 2007). Their focus is on managing the budget, and they are often confronted with monetary decisions.

We consider such a monetary decision problem, in which a decision maker must determine the ordering cost k and the resulting inventory is indicated by an inventory metric. With ordering cost k, $n = k/K$ orders can be placed per year, which results in an average inventory level of $d/2n = dK/2k$. The higher the ordering cost is, the higher the ordering frequency and the lower the average inventory level. The total costs are $k + hdK/2k$, and the optimal ordering cost is $k^* = \sqrt{hdK/2}$.

If inventory is valued only by the *days of supply* metric, the value associated with ordering cost k is

$$V_T(k) = \frac{t}{d}\left(I(K) - \frac{dK}{2k}\right), \tag{2.18}$$

where $I(K)$ denotes the initial inventory level. A decision maker who places weight $0 \leq w \leq 1$ on the metric and weight $(1-w)$ on the inventory value and who initializes t at $k = K$ assigns the value

$$\widetilde{V}_T(k) = wh\left(I(K) - \frac{dK}{2k}\right) + (1-w)h\left(I(K) - \frac{dK}{2k}\right) - k = h\left(I(K) - \frac{dK}{2k}\right) - k \tag{2.19}$$

to ordering cost k.

Under the *inventory turn rate metric*, the value is

$$V_R(k) = rd\left(\frac{2k}{dK} - \frac{1}{I(K)}\right) - k. \tag{2.20}$$

A decision maker who places weight $0 \leq w \leq 1$ on the metric and weight $(1-w)$ on the inventory value and who initializes r at $k = K$ assigns the value

$$\widetilde{V}_R(k) = whI^2(K)\left(\frac{2k}{dK} - \frac{1}{I(K)}\right) + (1-w)h\left(I(K) - \frac{dK}{2k}\right) - k, \tag{2.21}$$

to ordering cost k.

To compare the optimal ordering cost under the two metrics, we first analyze the optimal ordering cost under the days of supply metric. The function $\widetilde{V}_T(k)$ is concave in ordering cost k, and the first-order condition yields an optimal ordering cost of $k_T^* = \sqrt{hdK/2}$. The function $\widetilde{V}_R(k)$ is also concave in ordering cost k. At the optimal ordering cost k_T^*, the first derivative of the function $\widetilde{V}_R(k)$ is

$$\widetilde{V}_R(k) = w\frac{dh - 2K}{2K}. \tag{2.22}$$

The first derivative at the optimal ordering cost is strictly positive for situations in which it is more expensive to hold the annual demand in inventory than placing one order $(dh/2 > K)$

and where individuals place some weight on the metric ($w > 0$). The first condition holds for most real inventory systems, and the second condition holds if individuals rely, to some extent, on the metric. Thus, we hypothesize the following:

Hypothesis 2.4. *Average ordering costs are greater under the inventory turn rate metric than under the days of supply metric.*

2.6.2 Inventory Experiment

We conducted a laboratory experiment in which subjects had to determine the ordering costs of three products with different inventory holding costs per unit. The experiment used two treatments, a days of supply treatment and an inventory turn rate treatment, that differed only in how the inventory level was indicated.

All experimental sessions followed the same protocol. Upon entering the lab, subjects received written instructions (Supplementary Material 2.D). Subjects were informed that they had to manage the inventory of three products with identical demand rates of 10,000 units per product but with different inventory holding costs per unit. Subjects had to determine the ordering costs of the products. They were informed that the higher that they set the ordering cost, the lower the average inventory level and the corresponding inventory holding cost for this product would be. For holding inventory, subjects incurred costs of h ECUs per average unit on hand. The inventory holding costs per unit h were 10, 15, and 20. The minimum ordering cost was 100 ECUs. Subjects received an endowment of 45,000 ECUs from which the total costs were deducted. The exchange rate was 1 euro per 5,000 ECUs.

Before the actual experiment, subjects completed a computerized quiz consisting of seven questions, separated into three parts, to ensure that they understood the relationship among ordering cost, inventory holding cost, and the performance metric to which they were exposed. The first part concerned the functional relationships among ordering cost, performance metrics, and inventory holding cost. In the second part, we asked subjects to determine the average inventory level for two values of the performance metric. In the third part, subjects had to determine the effects of decreases in the inventory level on the inventory holding cost for three different inventory holding costs per unit. If all questions within a part were answered

correctly on the first attempt, subjects received 2,000 ECUs. If they needed two attempts, they received 1,000 ECUs. If they needed more than two attempts, they did not receive any compensation for answering the corresponding part. Subjects could not continue without having answered all seven questions correctly. For the first part, 111 subjects needed one attempt, 13 subjects needed two attempts, and one subject needed more than two attempts. The corresponding frequencies for the second and third part are 115, 8, and 2 and 93, 28, and 4, respectively.

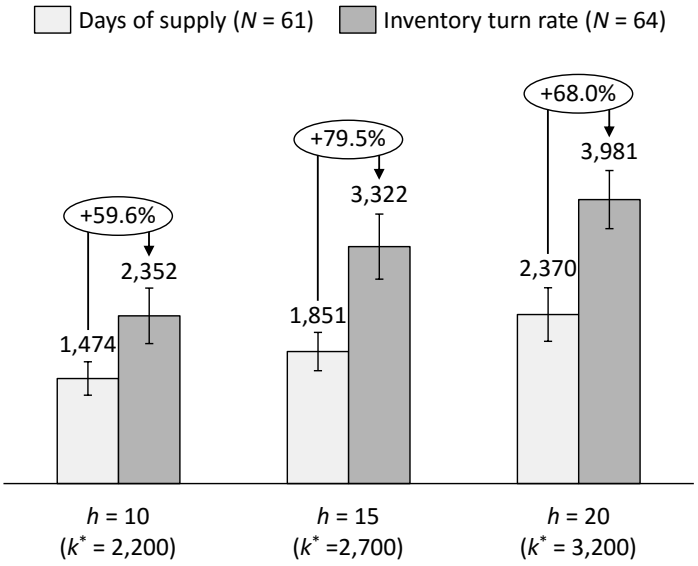
The experiment was implemented and conducted with the software z-tree (Fischbacher 2007). After they had made their investment decisions, subjects completed a post-experimental questionnaire, and we collected demographic data.

A total of 125 students of from the faculty of Management, Economics and Social Sciences of the University of Cologne were recruited via the online recruiting system ORSEE (Greiner 2004), and each subject participated in one of four sessions. Upon entering the laboratory, each subject was randomly assigned to one of two treatments, which resulted in 61 subjects for the days of supply treatment and 64 subjects for the inventory turn rate treatment. The sessions lasted 40 minutes on average, and the average payment was 8.44 euros, including a participation fee of four euros.

2.6.3 Results

The average ordering cost under the different inventory holding costs per unit are shown in Figure 2.6. For all products, ordering costs were higher under the inventory turn rate metric than under the days of supply metric. On average, the total ordering cost was 9,654 ECUs under the inventory turn rate metric and 69.5% higher than that of 5,695 ECUs under the days of supply metric. The difference in total ordering cost is significant (Wilcoxon test, one-sided, $p < 0.001$), as are the differences for the individual products (Wilcoxon test, one-sided, $p = 0.004$ for $h = 10$ and $p < 0.001$ for $h = 15$ and $h = 20$). We conclude that subjects set a higher ordering cost, on average, under the inventory turn rate metric than under the days of supply metric, which provides support for Hypothesis 2.4.

Figure 2.6 Effect of Metric on Average Ordering Cost



Note. Error bars indicate one standard error.

Figure 2.6 also indicates that ordering costs tend to be below optimality under the days of supply metric and above optimality under the inventory turn rate metric. Our model did not predict the below-optimal ordering cost that we observed in the experiment under the days of supply metric. However, this effect can be explained by prospect theory (Kahneman and Tversky 1979) and mental accounting (Thaler 1985), if the ordering cost were framed as a loss. Such losses typically loom larger than equivalent gains, which would explain the below-optimal order quantities in the days of supply treatment. Assuming that mental accounts and loss aversions are the same in both treatments, the behavioral models continue to predict the difference in ordering cost as stated in Hypothesis 4.

Because the cost function is steeper below than above the optimal solution, a given deviation from the optimal ordering cost is more costly below than above the optimum. This explains why the average total costs of 29,493 ECUs under the inventory turn rate metric are significantly below those of 42,555 ECUs under the days of supply metric (Wilcoxon test, two-sided, $p = 0.009$).

2.7 Discussion and Managerial Implications

We analyzed the effect of performance metrics on decision making and considered two equivalent metrics, days of supply and inventory turn rate. The relationship between days of supply and inventory value is linear, such that valuations that are based on the days of supply metric are proportional to those that are based on inventory value. The relationship between inventory turn rate and inventory value is convex, such that inventory reductions that are evaluated based on this metric are over-valued. We hypothesized that people decide differently under the two metrics and found support for our hypotheses in three laboratory experiments.

In the first experiment, we considered investment decisions and showed that most decisions are correct under the days of supply metric and incorrect under the inventory turn rate metric. To better understand the heterogeneity in the decisions of the inventory turn rate treatment, we applied dual process theory and found that individuals with high cognitive reflection more frequently decide optimally than those with low cognitive reflection. In the second experiment, we analyzed effort decisions and showed that individuals invest more effort under the inventory turn rate metric than under the days of supply metric. In the third experiment, we analyzed inventory decisions and showed that individuals choose a higher ordering cost under the inventory turn rate metric than under the days of supply metric.

In some situations, one metric is clearly superior to the other metric. In the investment decision study, for instance, individuals made better decisions under the days of supply metric than under the inventory turn rate metric, and thus, the days of supply metric is the superior choice. If this metric cannot be used, for instance, because corporate guidelines require using the inventory turn rate metric, decision making can still be improved by activating System 2 thinking of the decision makers. This can be supported, for instance, by reducing the emotional and cognitive load, by avoiding time pressure, and by avoiding multi-tasking during decision making.

In other situations, it is less clear which metric should be used. In our real effort study, individuals invested more effort under the inventory turn rate metric than under the days of supply metric. In situations, in which an employee is solely responsible for a dedicated

activity, such as managing raw material, work in process inventory, or finished goods inventory, the inventory turn rate metric can motivate the employee to continuously reduce inventory. However, employees investing more effort in inventory optimization have less capacity to invest in alternative activities. Employees with broader responsibilities who must also determine the activities in which to invest effort might be misguided by the inventory turn rate metric and focus on reducing inventory in areas where inventory is already low instead of areas with substantial inventory reduction potential or in other valuable activities. Furthermore, it may demotivate employees when they realize that the increase they perceived in the metric does not lead to a similar payment. If an employee is compensated based on changes in the inventory value, he or she might consider the change in the metric as a reference point (Bell 1985, Loomes and Sugden 1986, Kőszegi and Rabin 2006) for the size of a bonus payment. If bonus payments fall short of such expectation-based reference points, work satisfaction and performance might suffer (Ockenfels et al. 2015).

In our inventory management experiment, individuals chose higher order cost under the inventory turn rate metric than under the days of supply metric, which was predicted by the behavioral model. However, our model did not predict the below-optimal ordering cost that we observed in the experiment under the days of supply metric. This effect can be explained by prospect theory (Kahneman and Tversky 1979) and mental accounting (Thaler 1985), if the ordering cost were framed as a loss. Better understanding such behavioral biases seems a promising area for future research.

Our research suggests various other areas for future research. We focused on inventory management, but we expect that our insights are generalizable. In engineering, reliability can be measured by the time between failures and the failure rate, and in warehousing, performance can be measured by the picking time and the picking rate. We expect that investment decisions will be more frequently optimal under the time than under the rate metrics in both settings but that the motivation to continuously invest effort in optimizing individual areas would be higher under the rate than under the time metrics. Similar examples exist in other supply chain areas and other business functions, and it would be interesting to analyze how approaches similar to ours can be applied to them.

We considered equivalent metrics, where one was the inverse of the other. Many equivalent metrics have this property, but there are other equivalent metrics. For instance, in operations management, service performance can be measured by the fraction of filled demand or the fraction of lost sales. Similarly, equipment performance can be measured by the uptime and the downtime. One metric frames performance as gains, while the other frames it as losses, which might affect how people value the outcomes (Kahneman and Tversky 1979, Tversky and Kahneman 1981). Analyzing such effects offers interesting opportunities that we leave to future research.

Beyond the decision biases that are rooted in solid theory, biases that have not received much attention can have considerable effects on valuations. Green (2014), for instance, reported on a failed new product introduction by the A&W restaurant chain that introduced a new third pounder hamburger to rival the McDonald's Quarter Pounder. The A&W burger had more meat, was preferred in taste tests, and was less expensive, but did not sell well. Customer focus groups revealed the reason: "Why, [customers] asked the researchers, should they pay the same amount for a third of a pound of meat as they did for a quarter-pound of meat at McDonald's." As the example illustrates, it is important to understand how metrics affect valuation.

Supplementary Materials

The following instructions are translated from German. We present the instructions for the inventory turn rate treatments. In the days of supply treatments, the instructions differ from those of the inventory turn rate treatment only in the metric used to measure inventory performance.

2.A Instructions Investment Experiment

Welcome and thank you for participating in this experiment. Please do not talk to each other from now on, turn off your mobile phones, and put away all your personal belongings.

We ask you to read all instructions carefully. If you have any questions, feel free to raise your hand. The experimenter will then come to you and answer your questions in private. Moreover, after reading the instructions you will have the chance to ask questions in case anything remained unclear. All decisions are made anonymously and will be treated confidentially.

You can earn money in this experiment. How much you will earn depends on your decisions. Your earnings in the course of this experiment are expressed in a virtual unit of currency – the experimental currency unit (ECU). At the end of the experiment, you will receive 1 euro per 3,000 ECUs earned during this experiment. In addition, you will receive a show-up fee of 2.50 euros.

Introduction

The inventory turn rate metric is a measure commonly used in warehousing. It is defined as the annual demand rate divided by the average inventory level. The inventory turn rate thus indicates how many times per year the average inventory level of a product is completely depleted and replenished.

Example:

A company sells 10,000 units per year of a product. The average inventory level is 5,000 units. What is the inventory turn rate?

$$\text{Inventory turn rate} = \frac{\text{Annual demand rate}}{\text{Average inventory level}} = \frac{10,000 \text{ units/year}}{5,000 \text{ units}} = 2/\text{year}$$

At constant demand rate, an increase in the average inventory level causes a reduction in the inventory turn rate.
At constant demand rate, a reduction in the average inventory level causes an increase in the inventory turn rate.

Task description

You are in charge of a warehouse, and you will be evaluated on the basis of the average inventory level. Your warehouse contains two products featuring different inventory turn rates. From each product, 10,000 units are sold per year. The unit holding costs are the same for both products.

In each round, you can optimize the inventory management for one of the two products and thus reduce the average inventory level of this product. You will receive a bonus for each unit you reduce your average inventory level. There are no costs for the optimization itself.

You will know the current inventory turn rates of both products and how the inventory turn rates will change after the optimization. In each round, it is your task to select one of the two products for which you want to optimize inventory management.

Experimental protocol

The sequence of the experiment is as follows:

I. Decisions: You will decide in three independent rounds for which product you want to optimize the inventory management. You will receive a bonus for each unit you reduce your average inventory level.

II. Questions: You will answer three short questions.

III. Questionnaire: You will answer general questions regarding your attitudes and preferences.

IV. Questionnaire: Finally, you will answer general questions regarding the experiment and your person.

Payment

Your payment depends on the inventory reduction achieved over all three rounds. For each unit you reduce the average inventory level, you will receive 10 ECUs. At the end of the experiment, you will receive 1 euro per 3,000 ECUs that you have earned during the experiment. In addition, you will receive a show-up fee of 2.50 euros.

2.B Instructions Validation Experiment With Managers

Definition

$$\text{Inventory turn rate} = \frac{\text{Annual demand rate}}{\text{Average inventory level}}$$

Situation

You are in charge of a warehouse and you have discovered room for inventory optimization for products A, B, and C. Unfortunately, your budget restrictions allow just one optimization. You know the current inventory turns and how they will change after investing in inventory optimization.

Product		A	B	C
Annual demand rate (units)		10,000	10,000	10,000
Unit cost (€)		500	500	500
Inventory turn rate	Current situation	3	10	24
	After optimization	4	20	40

You are evaluated by average inventory value. Which product would you invest in?

At your company, which of the following metrics is used to measure inventory performance?

- □ Inventory turn rate
- □ Days of supply
- □ Both
- □ Other (please specify):

2.C Instructions Effort Experiment

Welcome and thank you for participating in this experiment. Please turn off your mobile phone, and put away all your personal belongings. We ask you to read all instructions carefully. All decisions are made anonymously and will be treated confidentially.

You can earn money in this experiment. How much you will earn depends on your decisions and your exerted effort. Your earnings in the course of this experiment are expressed in a virtual unit of currency – the experimental currency unit (ECU). At the end of the experiment, you will receive 1 euro per 5,000 ECUs earned during this experiment.

Introduction

The inventory turn rate metric is a measure commonly used in warehousing. It is defined as the annual demand rate divided by the average inventory level. The inventory turn rate thus indicates how many times per year the average inventory level of a product is completely depleted and replenished.

Example:

A company sells 10,000 units per year of a product. The average inventory level is 5,000 units. What is the inventory turn rate?

$$\text{Inventory turn rate} = \frac{\text{Annual demand rate}}{\text{Average inventory level}} = \frac{10,000 \text{ units/year}}{5,000 \text{ units}} = 2/\text{year}$$

At constant demand rate, an increase in the average inventory level causes a reduction in the inventory turn rate.

At constant demand rate, a reduction in the average inventory level causes an increase in the inventory turn rate.

Situation

You are in charge of a warehouse with a single product, and you will be evaluated on the basis of the average inventory level. Currently, your warehouse contains on average 5,000 units of this product. 10,000 units are sold per year. Therefore, the initial inventory turn rate of your warehouse is 2 per year.

Depending on your effort, you can now optimize your inventory management and increase your inventory turn rate. You will receive a bonus of 10 ECUs for each unit you reduce your average inventory level.

Task description

In this experiment your effort will be simulated by moving sliders. The sliders are initially positioned at "0" (see Figure 1 (a)). By using the mouse, you can position the slider at any integer value between "0" and "100". The more sliders you correctly position at the target position "50" (see Figure 1 (b)), the more you can reduce your average inventory level. You can adjust each slider an unlimited number of times. In each round, you have 120 seconds to do so.

(a) Initial position at "0" (b) Target position at "50"

Figure 1 Initial and target position of a slider

The average inventory level depends on the number of sliders positioned correctly as follows:

$$\text{Average inventory level} = \frac{5,000 \text{ units}}{1 + 0.1 \cdot \text{Number of sliders positioned correctly}}$$

The inventory turn rate is calculated accordingly:

$$\text{Inventory turn rate} = \frac{\text{Annual demand rate}}{\text{Average inventory level}} = \frac{10,000 \text{ units/year}}{\text{Average inventory level}}$$

Please note that the demand rate stays constant over all rounds.

Sequence of a round

In each round, the sequence is identical. Each round begins with an input screen with 48 sliders (see Figure 2). By positioning the sliders (moving them to the target position of "50"), you can reduce the average inventory level. For this task, you have 120 seconds per round. Within this time, you can freely decide how many sliders you want to position. In the upper part of the input window, you can track how the inventory turn rate changes, once you have positioned a slider correctly.

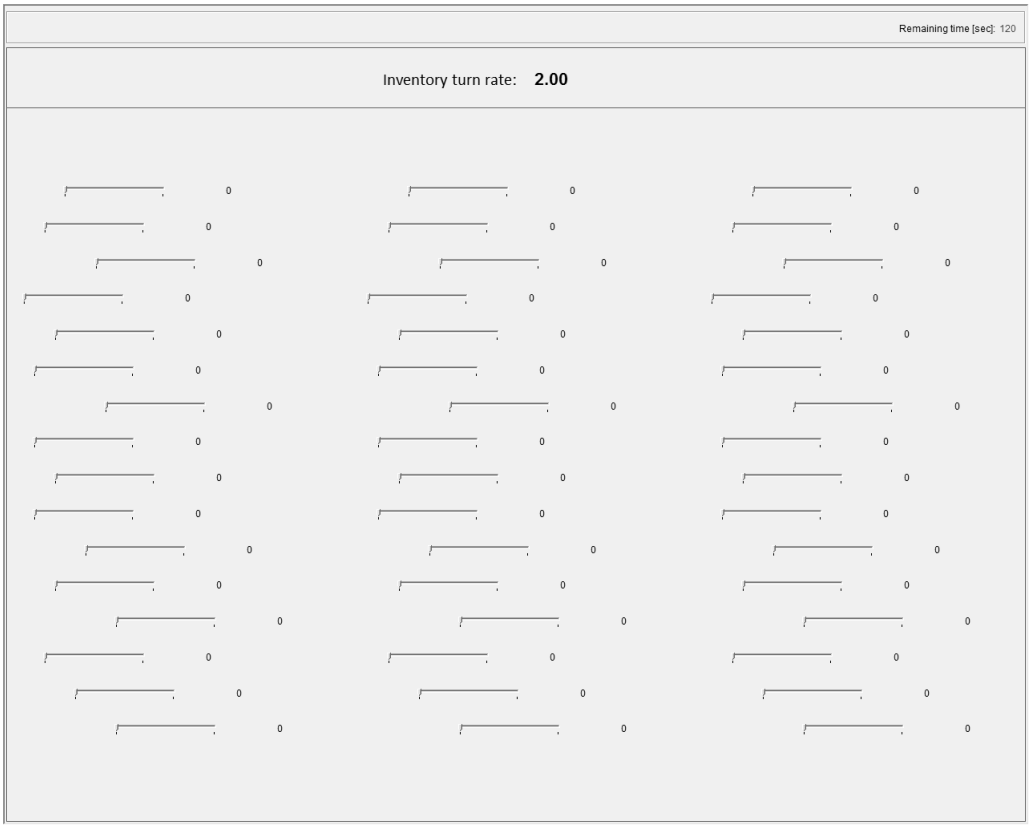


Figure 2 Input screen

At the end of each round, on the result screen, you will be informed of the extent to which you were able to increase the inventory turn rate of your warehouse. Once you press "continue", the input screen (Figure 2) appears again and the next round starts.

Please note that you will start the next round with the inventory turn rate you have achieved in the previous round. This means that you can continuously reduce your average inventory level over all rounds.

It is up to you how many rounds you exert effort. If you do not want to exert any more effort, please press "terminate experiment" on the result screen. You will then immediately receive your payment for the inventory reduction you achieved until then and are free to leave.

Experimental protocol

The sequence of the experiment is as follows:

I. Comprehension questions: First, you will answer some comprehension questions. You must answer all questions correctly to reach the next stage of the experiment. You will receive a bonus if you can answer all questions correctly on the first or second attempt.

II. Effort task: You can exert effort and thus reduce the average inventory level. It is up to you how many rounds to exert effort.

III. Questionnaire: Finally, you will answer general questions regarding the experiment and your person.

Payment

Your payment depends on the achieved inventory reduction over all rounds. For each unit you reduce the average inventory level, you will receive 10 ECUs. At the end of the experiment, you will receive 1 euro per 5,000 ECUs that you have earned during the experiment.

2.D Instructions Inventory Decision Experiment

Welcome and thank you for participating in this experiment. Please do not talk to each other from now on, turn off your mobile phones, and put away all your personal belongings.

We ask you to read all instructions carefully. If you have any questions, feel free to raise your hand. The experimenter will then come to you and answer your questions in private. Moreover, after reading the instructions, you will have the chance to ask questions in case anything remained unclear. All decisions are made anonymously and will be treated confidentially.

You can earn money in this experiment. How much you will earn depends on your decisions. Your earnings in the course of this experiment are expressed in a virtual unit of currency – the experimental currency unit (ECU). At the end of the experiment, you will receive 1 euro per 5,000 ECUs earned during this experiment. In addition, you will receive a show-up fee of 4 euros.

Introduction

The inventory turn rate metric is a measure commonly used in warehousing. It is defined as the annual demand rate divided by the average inventory level. The inventory turn rate thus indicates how many times per year the average inventory level of a product is completely depleted and replenished.

Example:

A company sells 10,000 units per year of a product. The average inventory level is 5,000 units. What is the inventory turn rate?

$$\text{Inventory turn rate} = \frac{\text{Annual demand rate}}{\text{Average inventory level}} = \frac{10,000 \text{ units/year}}{5,000 \text{ units}} = 2/\text{year}$$

At constant demand rate, an increase in the average inventory level causes a reduction in the inventory turn rate.

At constant demand rate, a reduction in the average inventory level causes an increase in the inventory turn rate.

Situation

You are in charge of a warehouse with three different products, and you will be evaluated on the basis of total annual cost. From each product, 10,000 units are sold per year.

You have to decide how much you want to invest per year in the order processing of each product. The more you invest in the order processing of a product, the higher the inventory turn rate and the lower the average inventory level, as well as the corresponding holding cost for this product.

For holding inventory, you incur costs of h ECUs per average unit on hand. The unit holding cost parameter h varies from product to product and will be displayed on the input screen.

You can adjust your decisions an unlimited number of times and display the corresponding turn rates before you submit your decisions.

Payment

The annual cost per product can be broken down as follows:

$$\text{Annual cost per product} = \text{Investment in order processing} + h \cdot \text{Average inventory level}$$

The total annual costs are made up of the sum of the annual costs per product. In addition, you will receive an endowment of 45,000 ECUs. Your profit will be calculated as follows:

$$\text{Profit} = 45{,}000 \text{ ECUs} - \text{Total annual cost}$$

At the end of the experiment, you will receive 1 euro per 5,000 ECUs that you have earned during the experiment. In addition, you will receive a show-up fee of 4 euros.

Experimental protocol

The sequence of the experiment is as follows:

I. Comprehension questions: You will answer some comprehension questions.

II. Decision: You will decide how much to invest in order processing per year.

III. Questions: You will answer eight short questions.

IV. Questionnaire: You will answer general questions regarding your attitudes and preferences.

V. Questionnaire: You will answer general questions regarding the experiment and your person.

Chapter 3

Decision Making Under Service Level Contracts: An Experimental Analysis

The ordering behavior of human decision makers under stochastic demand has been analyzed for various supply contracts. A consistent finding is that people place orders that both deviate from expected profit-maximizing quantities and exhibit high variability. We consider service level contracts, commonly used in practice but receiving little attention in the behavioral operations literature. Service level contracts have an interesting property that other supply contracts do not offer. They can be parameterized, such that they have steep expected profit functions around the expected profit-maximizing order quantity. We provide analytical models and use a laboratory experiment to analyze ordering behavior under service level contracts and compare the performance with that under wholesale price contracts, which have flat expected profit functions. Our results indicate that properly designed service level contracts can incentivize people to place close-to-optimal order quantities with low variability, resulting in high efficiency. In our experiment, the efficiency that human subjects achieved under a service level contract was 97.2%, compared with an efficiency of 88.1% under a wholesale price contract.

3.1 Introduction

We consider a standard setting in the supply chain literature, whereby a retailer orders products from a supplier to fill her customers' stochastic demand (see Cachon 2003 for an overview). A good deal of this literature has focused on analyzing expected profit-maximizing decision makers and whether various contracts incentivize first-best order quantities, that is, order quantities that maximize expected channel profit. Among the contracts commonly considered are the wholesale price contract (Arrow et al. 1951, Lariviere and Porteus 2001), in which the retailer purchases products from the supplier at a unit wholesale price and bears the full risk of excess inventory; the buyback contract (Pasternack 1985), in which the retailer can return excess inventory to the supplier at a unit buyback price; and the revenue sharing contract (Cachon and Lariviere 2005), in which the revenues of the retailer are shared with the supplier.

A more recent stream of literature examines the ordering behavior of human decision makers. This behavioral operations stream of research was initiated by Schweitzer and Cachon (2000). They used laboratory experiments to analyze ordering behavior under a wholesale price contract. The order quantities of their subjects deviated substantially from expected profit-maximizing quantities and exhibited what has come to be known as the "pull-to-center" effect, because observed average order quantities are regularly between expected profit-maximizing quantities and mean demand. The pull-to-center effect has proven robust, holding under various demand distributions (Benzion et al. 2008, 2010) and observed and unobserved lost sales (Rudi and Drake 2014) and persists with experience and training (Bolton and Katok 2008, Bolton et al. 2012) and decision frequency (Bolton and Katok 2008, Bostian et al. 2008, Lurie and Swaminathan 2009). The pull-to-center effect is not unique to the wholesale price contract, having also been observed under buyback and revenue sharing contracts (Katok and Wu 2009, Becker-Peth et al. 2013, Becker-Peth and Thonemann 2016).

One explanation for the pull-to-center effect, first put forward by Schweitzer and Cachon (2000), is anchoring and insufficient adjustment, a learning heuristic by which people first make decisions based on an initial anchor and tend to underweight additional information, thus leading to insufficient adjustment and biasing subsequent decisions toward the initial anchor

(Tversky and Kahneman 1974, Slovic and Lichtenstein 1971). Observed ordering behavior reported in the behavioral operations literature is consistent with mean demand serving as an anchor. Ordering behavior exhibits significant individual heterogeneity (Moritz et al. 2013), and thus, it is unlikely that any single explanation will fit with all or most peoples' ordering behavior (Katok 2011). Nevertheless, some models perform well at capturing aggregate trends in the data. Bostian et al. (2008) estimated the parameters of an adaptive learning model (Camerer and Ho 1999) that tracks the observed data patterns in their newsvendor experiment. Other explanations of the pull-to-center effect include bounded rationality (Su 2008), ex post inventory error minimization (Schweitzer and Cachon 2000, Ho et al. 2010, Kremer et al. 2014), overconfidence of decision makers (Ren and Croson 2013), and impulse balance behavior (Ockenfels and Selten 2014, 2015). Although diverse in their approach, all of these explanations share a broad theme in that all offer explanations for why learning is insufficient to move ordering fully away from the anchor at mean demand to the optimal order quantity.

Most supply contracts analyzed in the behavioral operations literature have relatively flat expected profit functions. For the wholesale price contract analyzed in Schweitzer and Cachon (2000), for example, order quantities that deviate by 10% from the expected profit-maximizing quantities achieve expected profits that deviate by only approximately 1% from the maximum expected profit. Buyback and revenue sharing contracts exhibit similar low sensitivities. Bolton and Katok (2008) referred to this as the flat-maximum problem. Based on the findings from Harrison (1989), who showed that increasing expected payoff differences between bidding strategies in first price auctions improves learning and performance, Bolton and Katok (2008) hypothesized that a steeper expected profit function might improve learning and reduce the pull-to-center effect. Some learning theories, such as Bostian et al.'s (2008) model, also predict faster learning when the expected profit function is steeper.

In this chapter, we analyze behavior under service level contracts; these specify the fraction of demand that a retailer is obligated to fill and the penalty that must be paid by the retailer if the realized service level is below the specified level. We show that by properly parameterizing this type of contract, a steep expected profit function can be achieved. In

addition, the stipulated service level might serve as an alternative anchor to average demand. We investigate both of these potential effects on ordering behavior.

Service level contracts are commonly used in practice (Thonemann et al. 2003, Chen and Thomas 2016). Analytical models have been developed that show how such contracts can be parameterized to incentivize first-best order quantities for expected profit-maximizing decision makers (Sieke et al. 2013 and the references therein). However, the behavior of human decision makers under service level contracts has received relatively little attention, with the notable exceptions of Katok et al. (2008) and Davis (2015). Katok et al. (2008) analyzed the effect of the length of the review period on order decisions under a service level contract and found that the order quantities of human decision makers increase in the length of the review period. Davis (2015) analyzed human decision making under a contract type that is similar to the service level contract that we consider. He analyzed a pull setting, under which a supplier fills the demand of the retailer's customers. In one of his experiments, the retailer uses a service level contract, specifying a wholesale price and a bonus payment that the retailer pays to the supplier if the supplier achieves an exogenously given service level. Davis found that the retailer sets contract parameters suboptimally and that performance under the service level contract is considerably below optimality but better than under a wholesale price contract.

In this chapter, we provide a detailed analysis of human decision making behavior under service level contracts. We use an analytical model to determine the contract parameters that incentivize first-best order quantities for expected profit-maximizing retailers and conduct a laboratory experiment with these contracts. We analyze performance under service level contracts and show that they can achieve a high degree of efficiency if they are parameterized properly, that is, if contract parameters are chosen to achieve steep expected profit functions. Then, average order quantities are closer to the expected profit-maximizing quantity and have lower variability than under a wholesale price contract that has a flatter expected profit function. In our experiment, the efficiency under the service level contract is 97.2%, compared with an efficiency of 88.1% under the wholesale price contract. We also designed an experiment to analyze whether the performance improvement can be attributed to the steepness of the

expected profit function or the potential anchoring effect of the service level. The results indicate that performance can primarily be explained by the steepness of the expected profit function whereas the anchoring effect of the service level does not seem to have a significant effect on performance.

3.2 Theoretical Analysis of the Service Level Contract

We consider a standard supply chain setting with a single supplier and a single retailer (Pasternack 1985, Lariviere and Porteus 2001, Cachon and Lariviere 2005). The retailer chooses order quantity q and places it with the supplier. When determining the order quantity, the retailer knows the distribution $F(D)$ of demand D but not the demand realization d. For our analyses, we assume that the demand density $f(D)$ is logconcave and has strictly positive support on its entire domain. Most distribution functions commonly used in inventory management have this property (Rosling 2002) and it simplifies our theoretical analyses. The supplier produces order quantity q and delivers it to the retailer at the unit wholesale price w. The retailer sells the minimum of the order quantity q and demand d to customers at unit revenue r. Excess inventory has no salvage value, and excess demand is lost. We refer to the order quantity that maximizes the retailer's expected profit as the optimal order quantity and next show how it can be determined for wholesale price and service level contracts.

A *wholesale price contract* has a single parameter, the unit wholesale price w. For order quantity q and demand realization d, the retailer's profit is

$$\pi_{WP}(w, q, d) = r \ \min(q, d) - wq. \tag{3.1}$$

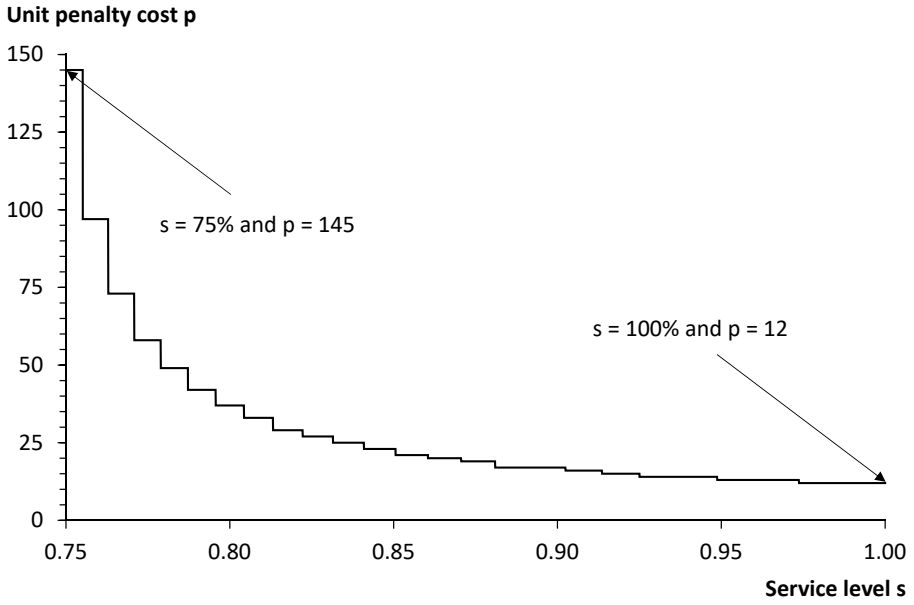
The optimal order quantity is (Arrow et al. 1951)

$$q^*_{WP}(w) = F^{-1}\left(\frac{r - w}{r}\right) \tag{3.2}$$

and we denote the optimal expected profit by $\Pi^*_{WP}(w) = E\left[\pi_{WP}(w, q^*_{WP}(w), d)\right]$.

A *service level contract* specifies the fraction of demand that the retailer is obligated to fill

Figure 3.1 Combinations of Service Levels and Unit Penalty Costs Incentivizing Optimal Order Quantities of 75



Note. Demand is discrete and uniformly distributed between 1 and 100, and p is restricted to integers.

and the financial consequences of failing to do so. The fraction of demand that must be filled is referred to as the service level s. For a demand realization of d units, the retailer must fill at least sd units. If the retailer ordered fewer than sd units, a unit penalty cost of p is charged for each unit difference between sd and q. If the retailer ordered at least sd units, no penalty is charged. For order quantity q and demand realization d, the retailer's profit is

$$\pi_{SL}(s,p,w,q,d) = r\ \min(q,d) - wq - p\,(sd-q)^{+}\,. \tag{3.3}$$

The retailer's optimal order quantity $q^{*}_{SL}(s,p,w)$ can be determined by solving (Sieke et al. 2013)

$$w - r\,(1-F(q)) - p\left(1-F\left(\frac{q}{s}\right)\right) = 0 \tag{3.4}$$

and we denote the optimal expected profit by $\Pi^{*}_{SL}(s,p,w) = E\left[\pi_{SL}(s,p,w,q^{*}_{SL}(s,p,w),d)\right]$.

The service level contract has three parameters and two degrees of freedom. Observe from Equation (3.4) that a given optimal order quantity can be achieved by different combinations

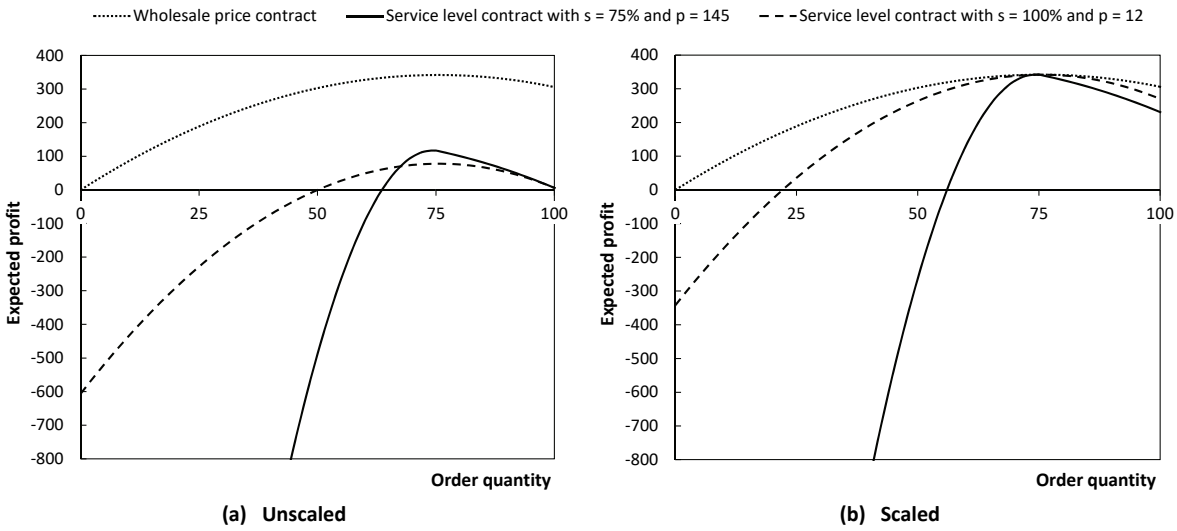
of the contract parameters. For instance, consider a supplier with a unit production cost of $c = 3$, a unit revenue of $r = 12$, and uniformly distributed customer demand between 1 and 100. The expected supply chain profit-maximizing order quantity is $q^*_{SC} = F^{-1}(0.75) = 75$ units. For a service level contract with unit wholesale price $w = 6$, Figure 3.1 depicts the combinations of service level s and unit penalty cost p for a retailer's optimal order quantity of $q^*_{SL}(s, p, w) = 75$ units; for example, service level $s = 75\%$ and unit penalty cost $p = 145$, service level $s = 100\%$ and unit penalty cost $p = 12$, or any combination of service level and unit penalty cost on the curve.

Note that in the case of uniformly distributed demand between 1 and 100, the same demand distribution as used in our experiment, and a unit wholesale price of $w = 6$, $s = 75\%$ is the lowest service level that yields a profit-maximizing order quantity of 75. For instance, imposing a service level of 65% as opposed to 75%, will not get the optimal order above 65 independent of the penalty p, because for order quantities of 65 or above the expected penalty payment will be zero and hence there is no incentive to exceed an order quantity of 65.

Although the order quantity that maximizes expected profit is the same for all combinations of s and p on the curve, the expected profit functions are different. For the extreme cases, that is, for $s = 75\%$ and $p = 145$ and for $s = 100\%$ and $p = 12$, the expected profit functions are depicted in Figure 3.2(a). The expected profit function of the wholesale price contract, which we will use as a benchmark, is also shown. The graphs show that the contracts have the same optimal order quantities but different optimal expected profits. For our analyses, it will prove useful to scale the contracts, such that they have the same optimal expected profits. This can be achieved by adding $\Pi^*_{WP}(w) - \Pi^*_{SL}(s, p, w)$ to the profit function $\pi_{SL}(s, p, w, q, d)$ of Equation (3.3). The results are shown in Figure 3.2(b).

The graphs in Figure 3.2(b) indicate that the service level contract with a low service level and a high unit penalty cost ($s = 75\%,\ p = 145$) has a steeper expected profit function than that with a high service level and a low unit penalty cost ($s = 100\%,\ p = 12$) and that both service level contracts have steeper expected profit functions than the wholesale price contract. The latter observation will be important for deriving our hypotheses, and the following proposition states that it holds in general (all proofs are provided in the Section 3.7):

Figure 3.2 Expected Profit Functions for Different Contracts with Optimal Order Quantity of 75



Proposition 3.1. *For all wholesale price contracts with* $0 < w < r$*, there exists a service level contract with the same optimal order quantity and a steeper expected profit function.*

The steepness of the expected profit function is affected by the service level s and the unit penalty cost p. The higher the service level, the higher is the expected number of units for which the penalty cost must be paid and the lower the unit penalty cost can be, which results in a flatter expected profit function (see Figure 3.2). The following proposition states the effect of the service level on the steepness of the expected profit function:

Proposition 3.2. *For a given optimal order quantity* q^**, the steepness of the expected profit function of the service level contract is decreasing in the service level* s*.*

The proposition implies that the contract with the lowest service level and highest unit penalty cost, in our example the contract with $s = 75\%$ and $p = 145$, has the steepest expected profit function. Many of our analyses will rely on this contract. For notational convenience, we will refer to it as the *steep service level contract* and to the contract with the flattest expected profit function, that is, the contract with $s = 100\%$ and $p = 12$, as the *flat service level contract*.

Under a steep expected profit function, it is more costly to deviate with average orders from the optimal quantity than under a flat expected profit function. A similar effect holds for the

variability of order quantities, which we measure by within-subject standard deviations of quantities because within-subject variability represents the extent to which a subject adjusts his or her order quantities. Between-subject variability reflects the extent to which subjects differ in their level behavior (Rudi and Drake 2014).

For concave expected profit functions, the marginal profit loss that is incurred by deviating from the optimal order quantity is increasing in the distance between the order quantity and the optimal order quantity. This implies that given order quantity $\bar{q} \leq q^*$, ordering $\bar{q} - \Delta q$ instead of $\bar{q}$ reduces the expected profit by more than ordering $\bar{q} + \Delta q$ increases the expected profit. If multiple orders are placed and they exhibit variability, the variability is more costly the more concave the expected profit function is. The following proposition addresses this implication for our contracts:

Proposition 3.3. *For a given optimal order quantity q^* and a given average order quantity $\bar{q}$, order variability is more costly under a service level contract than under a wholesale price contract.*

Both the steepness and concavity of the service level contract's expected profit function are decreasing in the service level s, which results in the following proposition:

Proposition 3.4. *For a given optimal order quantity q^* and a given average order quantity $\bar{q} \leq q^*$, the costliness of order variability is decreasing in the service level s.*

3.3 Development of Hypotheses

In supply contracting experiments, actual order quantities deviate significantly from the expected profit-maximizing quantities and exhibit substantial variability (see Section 3.1 for references). For example, the expected profits under a wholesale price contract in the baseline treatments of Bolton et al. (2012) are 13.3% below optimality. In their experiment, approximately one-half of the performance gap can be attributed to deviations of actual average orders from order quantities and one-half to order variability. Other studies have reported similar results (for example Rudi and Drake 2014), which indicates that two issues

must be addressed to achieve efficient ordering behavior: average orders must be close to expected profit-maximizing quantities and must exhibit low variability.

Bostian et al. (2008) hypothesized that suboptimal ordering behavior can be attributed to the flatness of the expected profit function:

> The flatness of the expected profit function in the neighborhood of $[q^*]$ implies a low average payoff penalty for choosing an order quantity that is merely close to the optimum. As a result, subjects may not have an economic incentive to be very circumspect in their decisions, and so lazy decision making could possibly explain the pull-to-center effect. (p. 593)

Bolton and Katok (2008) offered similar arguments.

We are interested in measuring how the steepness of the expected profit function affects ordering profits. To compare the steepnesses of different contracts, we use the sensitivity of the expected profit function with respect to deviations of the order quantity from the optimum.

Similar to Bostian et al. (2008), we define an anchor factor α to quantify the deviation of actual average orders $\bar{q}$ from the optimal order quantity q^* toward the mean demand μ:

$$\bar{q} = \alpha\mu + (1-\alpha)q^*. \tag{3.5}$$

Using the anchor factor α, which reflects the degree of the pull-to-center effect, we define sensitivity:

$$\epsilon_\alpha = 1 - \frac{\Pi\left(\alpha\mu + (1-\alpha)q^*\right)}{\Pi(q^*)}. \tag{3.6}$$

Sensitivity $\epsilon_{40\%}$, for instance, is the proportion of optimum expected profit that is lost if a weight of 40% is placed on mean demand. For the wholesale price contract in Figure 3.2(b), the sensitivity is $\epsilon_{40\%} = 1.9\%$. The flatness of the expected profit function is not unique to the wholesale price contract but can be observed under other supply contracts that have been analyzed in the supply chain literature. Table 3.1 provides empirical estimates for the anchor factor based on Equation (3.5) and the sensitivities of typical supply contracts.

Table 3.1 Steepnesses of Selected Contracts Analyzed in the Literature

			Contract parameters						
			Retail price	Wholesale price	Buyback price	Revenue share		Sensitivity	
Contract	Authors	Demand	r	w	b	λ	α	$\epsilon_{40\%}$	$\epsilon_{60\%}$
WPC	Schweitzer and Cachon 2000	U(1,300)	12	3	-	-	60%	1.8%	4.1%
	Bolton and Katok 2008	U(0,100)	12	3	-	-	56%	1.9%	4.1%
	Bostian et al. 2008	U(1,100)	4	1	-	-	36%	1.9%	4.2%
	Bolton et al. 2012[a]	U(1,100)	12	3	-	-	89%	1.9%	4.2%
	Rudi and Drake 2014[b]	N(1000,400^2)	12	3	-	-	70%	0.8%	1.8%
BBC	Ren and Croson 2013	N(100,30^2)	10	4	2	-	93%	0.6%	1.3%
	Katok and Wu 2009	U(0,100)	12	9	8	-	78%	1.9%	4.1%
		U(50,150)	12	9	8	-	37%	0.8%	1.8%
RSC	Katok and Wu 2009	U(0,100)	12	1	-	1/3	44%	1.9%	4.1%
		U(50,150)	12	1	-	1/3	89%	0.8%	1.8%

Notes. WPC = wholesale price contract, BBC = buyback contract, RSC = revenue sharing contract.
[a]Pooled data from managers and students in Phase 2 of the basic treatments.
[b]Data of the uncensored treatment.

The contracts in Table 3.1 all have low sensitivities, such that deviating from the expected profit-maximizing order quantity has a small effect on expected profit. Under service level contracts, such low sensitivities can be avoided, which makes deviating from the expected profit-maximizing order quantity more costly. The service level contracts used in our experiment have sensitivities of up to $\epsilon_{40\%} = 25.8\%$ and $\epsilon_{60\%} = 61.0\%$. Therefore, we expect lower deviations from the optimal order quantity under a service level contract than under contracts with low sensitivity. Proposition 3.1 states that for all wholesale price contracts, we can design a service level contract with the same optimal order quantity but a steeper expected profit function. We refer to such contracts as *properly designed service level contracts* and state the hypotheses for the experiment as follows:

Hypothesis 3.1. *Under a properly designed service level contract, average order quantities are closer to the optimal order quantity than under a wholesale price contract.*

From Proposition 3.3, we know that order variability is more costly under a properly designed service level contract than under a wholesale price contract. We therefore hypothesize the following:

Hypothesis 3.2. *Under a properly designed service level contract, orders are less variable than under a wholesale price contract.*

The expected behavior stated in Hypotheses 3.1 and 3.2 has an immediate consequence for expected profits. The closer order quantities are to the optimal order quantity and the less variability they exhibit, the higher is the expected supply chain profit. A standardized measure of supply chain profit is supply chain efficiency, that is, the expected supply chain profit achieved divided by the expected supply chain profit from the optimal order, and we hypothesize the following:

Hypothesis 3.3. *Under a properly designed service level contract, expected supply chain efficiency is higher than under a wholesale price contract.*

The above hypotheses concern performance differences between service level contracts and wholesale price contracts. For a given optimal order quantity, there exists a set of service level contracts with different combinations of contract parameters (Figure 3.1). Although these service level contracts have the same expected profit-maximizing order quantity, their expected profit functions differ in steepness (Proposition 3.2) and concavity (Proposition 3.4). We argued above that anchoring on mean demand and order variability are decreasing in the steepness and concavity of the expected profit function. For a given optimal order quantity, the steep service level contract has the steepest and the flat service level contract has the flattest expected profit function, and we hypothesize the following:

Hypothesis 3.4. *Under a steep service level contract,*

(a) *average order quantities are closer to the optimal order quantity,*

(b) *order quantities are less variable, and*

(c) *supply chain efficiency is higher*

than under a flat service level contract.

Table 3.2 Treatments Used in Laboratory Experiment

	Fixed payment	Wholesale price	Retail price	Service level	Penalty cost	Optimal order	Sensitivity	
Treatment	E	w	r	s	p	q^*	$\epsilon_{40\%}$	$\epsilon_{60\%}$
1. WPC	–	3	12	–	–	75	1.9%	4.2%
2. $SLC_{75\%,\,145}$	225	6	12	75%	145	75	25.8%	61.0%
3. $SLC_{100\%,\,12}$	264	6	12	100%	12	75	3.9%	8.4%
4. $SLC_{75\%,\,6}$	205	6	12	75%	6	60	0.6%	1.3%

WPC = wholesale price contract, SLC = service level contract.

3.4 Experiment

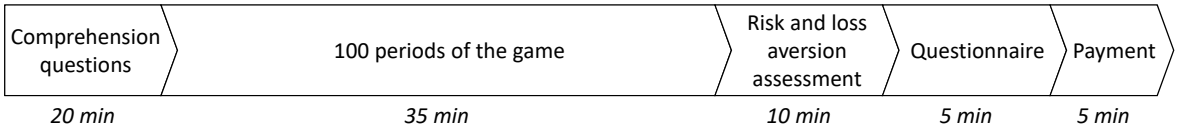
We use a laboratory experiment to analyze human decision making under service level contracts. Our experiment has four treatments, one wholesale price contract treatment, which serves as a benchmark, and three service level contract treatments.

3.4.1 Design

Table 3.2 provides an overview of the four treatments of our laboratory experiment. All treatments used discrete uniformly distributed demand between 1 and 100 and a retail price of r = 12 francs. The fixed payments ensured that ordering the optimal quantities would yield the same expected profit of 342 francs in all treatments. All payouts were expressed in laboratory francs. Subjects were informed that francs would be converted into cash at an exchange rate of 3,000 francs to the dollar at the end of the experiment.

We chose a high-profit condition such that inventory is optimally stocked above average demand ($q^* > \mu$) because this setting offers greater possible gains from coordination, and thus coordinating contracts, such as service level contracts, are more likely to be observed in practice (Katok and Wu 2009, Wu and Chen 2014). Furthermore, service levels below 50% are uncommon in practice (Gruen et al. 2002). For the wholesale price contract of Treatment 1, we used a wholesale price of w = 3 francs, resulting in an optimal order quantity of q^* = 75 units. The wholesale price of the service level contracts was w = 6 francs. In Treatment 2, we used the steep service level contract, and in Treatment 3, we used the flat service level contract (see Figure 3.1).

Figure 3.3 Experimental Protocol



Treatments 1 to 3 are sufficient to test our hypotheses. The hypotheses are based on arguments regarding differences in the steepnesses of the expected profit functions and do not address the possibility that the service level exhibits an anchoring effect. To analyze whether such an anchoring effect exists, we included Treatment 4, in which we used a service level contract with the same service level as in Treatment 2 but with a lower unit penalty cost.

3.4.2 Protocol

All sessions were conducted at the Laboratory for Laboratory for Behavioral Operations and Economics at the University of Texas at Dallas and followed the experimental protocol in Figure 3.3. The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

Upon entering the laboratory, subjects were randomly assigned to a private computer terminal and given time to read the instructions. After they had read the instructions, subjects could ask questions that were answered privately. During the experiment, communication between subjects was prohibited and none was observed.

Before the actual experiment started, subjects completed a computerized quiz with 11 (wholesale price contract treatment) or 17 questions (service level contract treatments). The quiz comprised three sections. In the first and second sections of the service level contract treatments, subjects had to determine the purchase cost, the number of units sold, the revenue, the service level, the number of units short of the target, the penalty cost, and the profit for two examples that were identical across treatments. In the wholesale price contract treatment, questions regarding the service level, the number of units short of the target, and the penalty cost were excluded. The third section contained general questions about the experiment. The questions and statistics on the answers are provided in Supplementary Material 3.C. If all

questions of a section were answered correctly on the first attempt, subjects received 1,000 francs. If they needed a second attempt, they received 500 francs. If they needed more than two attempts, they did not receive any compensation for the section. Subjects could continue only after they had correctly answered all questions in a section. We used this approach to ensure that subjects had a good understanding of the cost accounting and profit calculation for the particular contract addressed in their treatment.

At the beginning of each period, subjects were reminded of all contract parameters. After each period, they were shown a detailed breakdown of the profit calculation. After the main experiment, all subjects completed two additional tasks (see Supplementary Material 3.D for details). The first task was a computerized version of the risk elicitation task introduced by Holt and Laury (2002). The second task was the computerized loss aversion measurement task of Gächter et al. (2010), which was adapted from an earlier protocol of Fehr and Goette (2007). Subjects earned francs depending on their decisions and the outcome of the risky lotteries.

Finally, subjects answered some general questions, provided demographic data (see Supplementary Material 3.E), and were paid, in private, their total individual earnings. The total earnings were based on the performance on the quiz, the profits achieved over the 100 periods of the main experiment, and the two lotteries that we used to elicit subjects' risk and loss aversion. The sessions lasted on average approximately 75 minutes. Actual average earnings, including a $5 participation fee, were $17.24.

3.4.3 Subjects

A total of 116 subjects participated in six sessions of the experiment. In each session, subjects were randomly assigned to one of the four treatments. Each subject participated in exactly one session, and cash was the only incentive offered. Subjects were students recruited through an online recruitment system from the subject pool of the University of Texas at Dallas. The majority of our subjects were graduate students (81%), and the rest were undergraduates (3% freshmen or sophomores and 16% juniors or seniors). The average age was 23.6 years (SD = 2.0 years).

Table 3.3 Summary Statistics

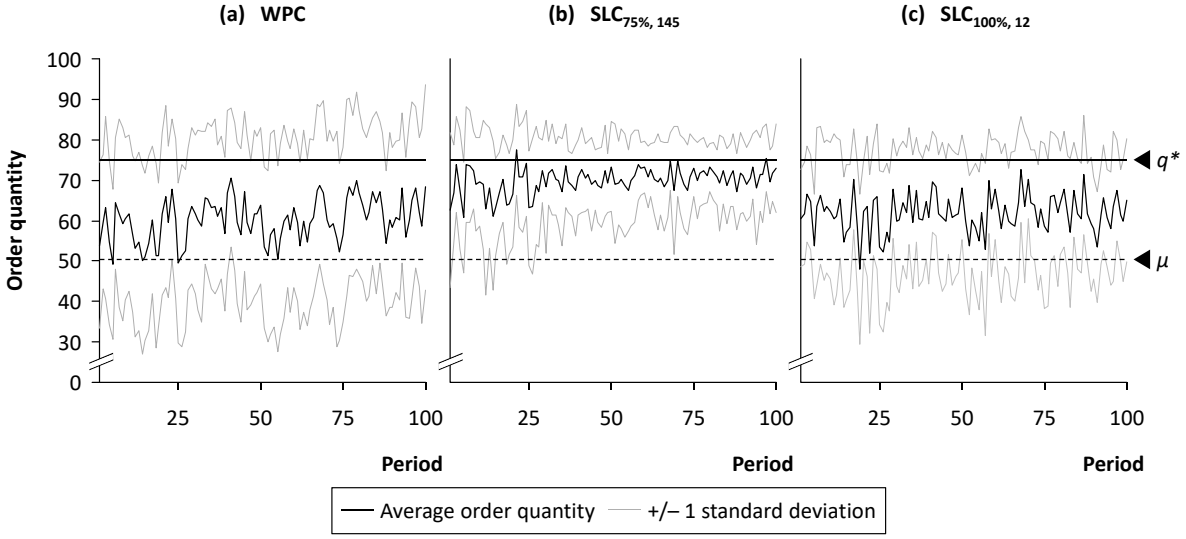
	WPC	SLC1	SLC2	SLC3
	N = 30	N = 28	N = 30	N = 28
Subjects' average order quantities				
Normative	75	75	75	60
Mean	60.16	69.93	61.79	54.37
Median	60.59	70.31	60.65	53.76
Std. deviation	12.5	5.59	6.05	8.56
Within-subject standard deviation				
Normative	0	0	0	0
Mean	15.51	8.99	13.88	16.18
Median	15.53	8.78	14.22	15.62
Std. deviation	7.33	4.51	5.72	5.47
Expected supply chain efficiency				
Normative	100%	100%	100%	100%
Mean	88.15%	97.19%	92.19%	89.19%
Median	92.07%	98.05%	91.67%	89.57%
Std. deviation	13.42%	2.72%	3.82%	5.17%

WPC = wholesale price contract, SLC = service level contract.

3.5 Results

We will first test the hypotheses concerning the higher performance of service level contracts compared to wholesale price contracts. Then, we will test the hypotheses concerning the effect of the service level contract parameters on performance, and finally, we will analyze a potential anchoring effect of the service level. Unless otherwise stated, we use the Wilcoxon signed-rank test for one-sample tests and the Mann-Whitney test for two-sample tests. All p-values we report below are two-tailed. Summary statistics are provided in Table 3.3. For all comparisons below, we tested for differences in subjects' risk (Holt and Laury 2002) and loss aversion (Gächter et al. 2010) across treatments and could not find any significant differences (all $p > 0.1$). Thus, there is no evidence that a difference in risk or loss aversion across treatments drives our results.

Figure 3.4 Average Order Quantities by Period Under Wholesale Price Contract and Service Level Contracts



3.5.1 Service Level Contracts Versus Wholesale Price Contract

Our first set of analyses compares performance under service level contracts with that under a wholesale price contract. The hypotheses state that average order quantities are closer to optimal quantities (Hypothesis 3.1), that they have lower variability (Hypothesis 3.2), and that they result in higher supply chain efficiency (Hypothesis 3.3) under a properly designed service level contract than under a wholesale price contract. We will first compare ordering behavior under the wholesale price contract with that under the steep service level contract of Treatment 2, which has a 40%-sensitivity that is over ten times that of the wholesale price contract, and with that under the flat service level contract of Treatment 3, which has a 40%-sensitivity that is approximately twice that of the wholesale price contract.

3.5.1.1 Average Order Quantities

Figure 3.4 depicts the average order quantities per period under (a) the wholesale price contract, (b) the steep service level contract, and (c) the flat service level contract. Under the *steep service level contract*, average orders are closer to the optimal order quantity than are those under the wholesale price contract. The average order quantities are 5.1 units

below optimality under the service level contract versus 14.8 units under the wholesale price contract. This difference is significant ($p < 0.001$), which provides support for Hypothesis 3.1.

Under the *flat service level contract*, average order quantities are slightly above the average order quantities under the wholesale price contract, but we do not observe a similar magnitude in the difference to that observed under the steep service level contract. Average order quantities are 13.2 units versus 14.8 units below optimality for the flat service level contract and the wholesale price contract, respectively. The difference is small (1.6 units) and not significant ($p = 0.965$) and thus only provides directional support for Hypothesis 3.1.

3.5.1.2 Order Variability

From Table 3.3, we see that the within-subject standard deviation of order quantities under the *steep service level contract* is 8.99 and lower than that under the wholesale price contract (15.51). The difference in within-subject standard deviation is significant ($t(56) = 4.04$, $p < 0.001$), providing support for Hypothesis 3.2.

Under the *flat service level contract*, we also observe less order variability than under the wholesale price contract. However, the within-subject standard deviation difference is small and not significant ($t(58) = 0.96$, $p = 0.342$), providing only directional support for Hypothesis 3.2.

3.5.1.3 Supply Chain Efficiency

The above analysis shows that under properly designed service level contracts, average orders are closer to the optimal quantities and have lower variability than under a wholesale price contract, which should result in higher supply chain efficiency. To compute supply chain efficiency, we must specify the production costs of the supplier. Without loss of generality, we set them equal to the wholesale price in the wholesale price contract, that is, $w = 3$.

Under the *steep service level contract*, supply chain efficiency is 97.2% and significantly higher than that under the wholesale price contract, 88.1% ($p < 0.001$), providing support for Hypothesis 3.3. Under the *flat service level contract*, supply chain efficiency is 92.2% and

also higher than that under the wholesale price contract, but the difference is not significant ($p = 0.399$), again providing only directional support.

The results of our experiment provide directional support for Hypotheses 3.1, 3.2, and 3.3; that is, all experimental results are in the directions stated in these hypotheses. For the steep service level contract, all differences are highly significant ($p < 0.001$ for all comparisons between the steep service level contract and the wholesale price contract). For the flat service level contract, the differences are not significant. We conclude that properly designed service level contracts tend to outperform wholesale price contracts but that it is important to design a supply contract with a steep expected profit function to realize the performance potential that this contract type offers.

3.5.2 Steep Versus Flat Service Level Contract

The above analyses indicated that the steepness of the expected profit function affects ordering behavior. Hypothesis 3.4 states the performance differences between steep and flat service level contracts with respect to (a) average order quantities, (b) order variability, and (c) supply chain efficiency, and we next formally test this hypothesis.

Average orders under the steep service level contract are significantly above those under the flat service level contract ($p < 0.001$), which provides support for Hypothesis 3.4(a). From Table 3.3, we see that a steeper expected profit function leads to less order variability among service level contracts. The within-subject standard deviation is significantly lower under the steep than under the flat service level contract ($t(58) = 3.60$, $p < 0.001$), providing support for Hypothesis 3.4(b). We also find support for Hypothesis 3.4(c). The efficiency under the steep service level contract is significantly higher than the efficiency under the flat service level contract ($p < 0.001$).

3.5.3 Service Level Anchor

We can explain the superior performance of the steep service level contract compared with the flat service level contract and the wholesale price contract by the steepnesses of the expected profit functions. However, there exists another factor that can potentially affect behavior,

that is, the stipulated service level. One might argue that a service level provides another anchor in addition to mean demand and therefore affects order quantities.

If the service level served as an anchor, increasing the service level and maintaining the optimal order quantity would increase average orders. The comparison of the steep service level contract with service level 75% and the flat service level contract with service level 100% shows that the service level contract with the higher service level has smaller average order quantities. However, it has also a flatter expected profit function, and we cannot exclude the possibility that we observed superposed effects: A flatter expected profit function reduces order quantities, and a higher service level anchor increases them.

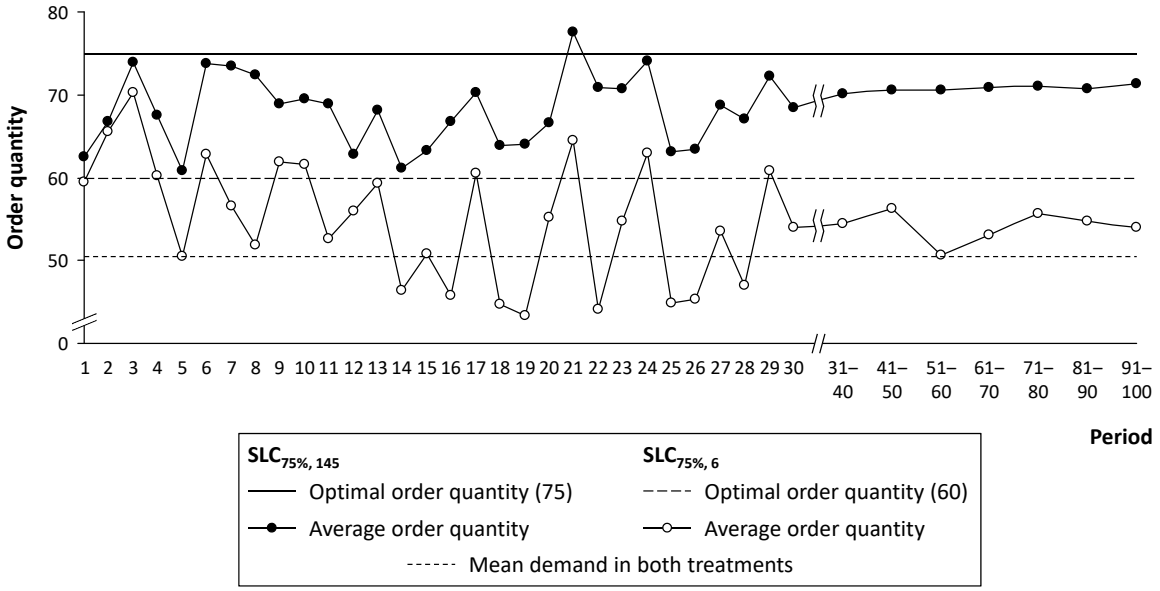
Ideally, we would design a service level contract with the same expected profit function and optimal order quantity but with different service levels. Unfortunately, this is not possible. When we vary the service level, we must change the unit penalty (see Figure 3.1) and thus the expected profit function to maintain the same optimal order quantity. However, we can use the results of Treatment 4 to obtain an indication of whether people anchor on the service level.

In Treatment 4, we used the same service level of 75% as in Treatment 2 but used a unit penalty cost of $p = 6$ instead of $p = 145$. The contract in Treatment 4 has a flatter expected profit function than the steep service level contract from Treatment 2, and its optimal order quantity is 15 units smaller (60 as opposed to 75 units).

If the service level served as an anchor, the difference in average orders should be smaller than the difference in the optimal order quantity, but this is not the case. As Table 3.3 indicates, average orders in Treatment 4 are 54.4 units, that is, 69.9 units – 54.4 units = 15.5 units below the average orders under the service level contract from Treatment 2. Because we held the service level constant, this change must be attributed to the change in the expected profit function. Because average orders differ by approximately the same quantity as the optimal order quantities, we have another indication that the flatter expected profit function, and not the service level anchor, explains behavior under service level contracts.

Figure 3.5 shows the average per period order quantities for Treatments 2 and 4. We observe that they start at approximately the same level and then diverge over 30 rounds

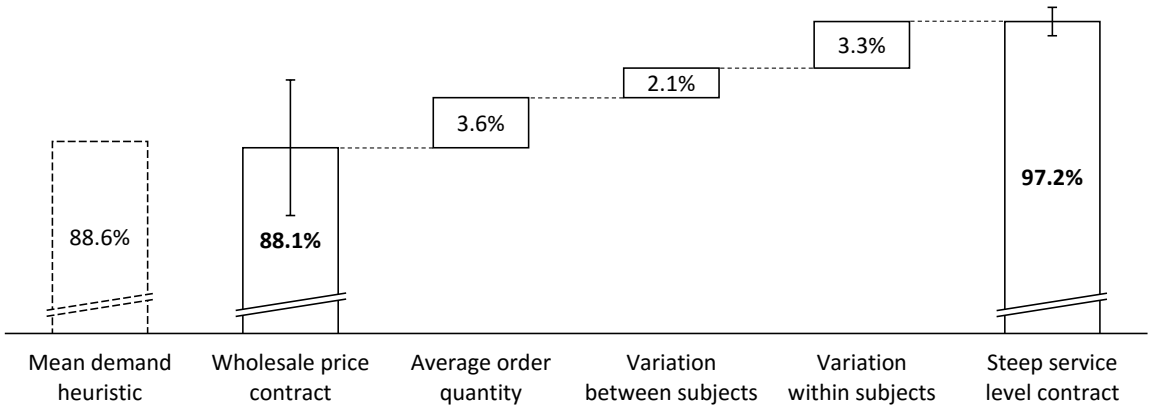
Figure 3.5 Average Order Quantities by Period in Treatments 2 and 4



before they level out. Average order quantities in the first period of the treatments do not significantly differ (average order quantities of 62.6 and 59.5 in Treatments 2 and 4, respectively, $p = 0.404$). Fitting a simple trend line to the data from the first 30 periods of Treatment 4, we find a significant order decrease of 0.361 units per period (standard error = 0.146, OLS two-tailed $p = 0.019$), which is significantly different from that of Treatment 2 (OLS two-tailed $p = 0.029$), in which we do not observe a significant trend over the fist 30 periods (OLS two-tailed $p = 0.780$). The results suggest that subjects might initially anchor on the stipulated service level and then adjust toward their final decision over time.

We note that neither the comparison of Treatments 2 and 3 nor the comparison of Treatments 2 and 4 can exclude the possibility that a service level anchoring effect exists that is superposed by the effect that steepness of the expected profit function has on ordering. However, they indicate that if an anchoring effect existed, it diminished over time, and its effect size would be much smaller than the size of the expected profit function steepness effect.

Figure 3.6 Effect of Average Order Quantities and Order Variability on Efficiency



Note. Error bars indicate one standard error.

3.6 Discussion and Managerial Implications

The majority of supply contracts analyzed in the literature have flat expected profit functions. Under such contracts, moderate deviations from optimal order quantities are inexpensive, and we have argued that this flatness contributes to the pull-to-center effect. We hypothesized that a supply contract with a steep expected profit function can reduce the pull-to-center effect, and our experimental results are in line with the prediction. Under the steep service level contract, average order quantities were 6.8% below optimality, compared with 19.8% under the wholesale price contract.

We argued that the steepness of the expected profit function also affects order variability and hypothesized that order variability is lower under steep than under flat profit functions. Our experimental results provided support for the hypothesis. Under the steep service level contract, the standard deviation of order quantities was 8.99, compared with 15.51 under the wholesale price contract.

Because better average order quantities and lower order variability result in higher efficiency, we expected that efficiency would be higher under the steep service level contract than under the wholesale price contract. Our experimental results were in line with this expectation. The efficiency of the steep service level contract was 97.2%, compared with an efficiency of

88.1% under the wholesale price contract. Figure 3.6 shows how the difference in efficiency can be attributed to differences in mean order quantities and variabilities and indicates that both factors play an important role in explaining the efficiency differences that we observed.

Figure 3.6 also shows the efficiency under a mean demand heuristic, whereby the expected demand is ordered in every period. This heuristic can serve as a benchmark and contextualizes the performance that we observed. The performance under the wholesale price contract was even worse than that of the mean demand heuristic, albeit not significantly. Thus, ordering mean demand in each period would result in a similar efficiency to what subjects achieved in the lab under a wholesale price contract. In both settings, efficiency is over 11% below optimality. Under a service level contract, the gap was reduced to less than 3%, which indicates that it is important to consider aspects of human behavior when selecting supply contracts.

In addition to incentivizing average order quantities that are close to optimal order quantities, having low variability, and resulting in high efficiency, service level contracts have another useful property. They can be parameterized to incentive first-best order quantities for any desired expected profit division among suppliers and retailers. Other supply contracts with two or more contract parameters have the same property, but unlike these contracts, service level contracts can be parameterized to achieve three objectives simultaneously: incentivizing first-best order quantities, offering a steep expected profit function, and dividing expected profits arbitrarily among supply chain partners.

The service level parameters in our experiment were chosen to explore our steepness of the curve hypothesis. The controlled environment of our study enables us to show that, in principle, a service level contract can induce more optimal behavior than a mathematically comparable wholesale price contract. An interesting next step is to explore whether the kinds of parameterizations we observe in the field are sufficient to induce the same kind of favorable results. A stake-size experiment of this sort is probably best conducted at the field level, since emulating the size of field stakes, and the associated consequences to decision makers, is difficult to do outside the field environment.

3.7 Proofs

We present the proofs for continuous demand distributions and denote the wholesale price under the wholesale price contract by w_{WP} and under the service level contract by w_{SL}.

Proof of Proposition 3.1

Proof. The optimal order quantity under a wholesale price contract with $0 < w_{WP} < r$ is $q^* = F^{-1}\left(\frac{r-w_{WP}}{r}\right)$. Consider a service level contract with $w_{WP} < w_{SL} < r$, $0 < s \leq 1$, and

$$p = \frac{w_{SL} - w_{WP}}{1 - F\left(\frac{q^*}{s}\right)}, \tag{3.7}$$

Because the second derivative of the expected profit function,

$$\frac{d^2}{dq^2}\Pi_{SL}(s, w_{SL}, q) = -w_{SL}f(q) - (r - w_{SL}))\, f(q) - \frac{w_{SL} - w_{WP}}{s\left(1 - F\left(\frac{q^*}{s}\right)\right)} f(\tfrac{q}{s})$$

is negative, the expected profit function is concave in q and the retailer's optimal order quantity can be determined by solving

$$\frac{d}{dq}\Pi_{SL}(s, w_{SL}, q) = -w_{SL}F(q) + (r - w_{SL})\,(1 - F(q)) + \frac{w_{SL} - w_{WP}}{1 - F(\frac{q^*}{s})}\left(1 - F(\tfrac{q}{s})\right) = 0 \tag{3.8}$$

for q. With $q = q^*$ in Equation (3.8), we obtain

$$\frac{d}{dq}\Pi_{SL}(s, w_{SL}, q)\Big|_{q=q^*} = -w_{SL}\frac{r - w_{WP}}{r} + (r - w_{SL})\left(1 - \frac{r - w_{WP}}{r}\right) + w_{SL} - w_{WP} = 0,$$

which proves that a service level contract with unit penalty cost chosen according to Equation (3.7) has the same optimal solution as the wholesale price contract.

We next prove that the expected profit function of the service level contract is steeper than

that of the wholesale price contract. For $q < q^*$,

$$\begin{aligned}\frac{d}{dq}\Pi_{SL}(s, w_{SL}, q) &= r\,(1 - F(q)) - w_{SL}\left(1 - \frac{1 - F(\frac{q}{s})}{1 - F(\frac{q^*}{s})}\right) - w_{WP}\left(\frac{1 - F(\frac{q}{s})}{1 - F(\frac{q^*}{s})}\right) \\ &> r\,(1 - F(q)) - w_{WP}\left(1 - \frac{1 - F(\frac{q}{s})}{1 - F(\frac{q^*}{s})}\right) - w_{WP}\left(\frac{1 - F(\frac{q}{s})}{1 - F(\frac{q^*}{s})}\right) \\ &= \frac{d}{dq}\Pi_{WP}(w_{WP}, q).\end{aligned}$$

Analogously, it can be shown for $q > q^*$ that $\frac{d}{dq}\Pi_{SL}(s, w_{SL}, q) < \frac{d}{dq}\Pi_{WP}(w_{WP}, q)$. □

Proof of Proposition 3.2

Proof. Consider a service level contract with $w_{WP} < w_{SL} < r$, $0 < s \leq 1$, and p chosen according to Equation (3.7). The expected profit function has steepness $\frac{d}{dq}\Pi_{SL}(s, w_{SL}, q) = -w_{SL}F(q) + (r - w_{SL})\,(1 - F(q)) + \frac{w_{SL} - w_{WP}}{1 - F(\frac{q^*}{s})}\left(1 - F(\frac{q}{s})\right)$. To determine the effect of the service level s on the steepness, we analyze the first derivative of steepness with respect to service level:

$$\begin{aligned}\frac{d}{ds}\frac{d}{dq}\Pi_{SL}(s, w_{SL}, q) &= (w_{SL} - w_{WP})\frac{f(\frac{q}{s})\frac{q}{s^2}\left(1 - F\left(\frac{q^*}{s}\right)\right) - \left(1 - F(\frac{q}{s})\right)f\left(\frac{q^*}{s}\right)\frac{q^*}{s^2}}{\left(1 - F\left(\frac{q^*}{s}\right)\right)^2} \\ &= (w_{SL} - w_{WP})\frac{1 - F(\frac{q}{s})}{\left(1 - F\left(\frac{q^*}{s}\right)\right)s}\left(\frac{q}{s}\frac{f(\frac{q}{s})}{1 - F\left(\frac{q}{s}\right)} - \frac{q^*}{s}\frac{f(\frac{q^*}{s})}{1 - F\left(\frac{q^*}{s}\right)}\right). \qquad (3.9)\end{aligned}$$

Let $h(x) \equiv f(x)/(1 - F(x))$ denote the failure rate. All logconcave distributions have an increasing failure rate (Bagnoli and Bergstrom 2005). From Lariviere and Porteus (2001) we know that distributions with an increasing failure rate have an increasing generalized failure rate, that is, $xh(x)$ is increasing x. Therefore $\frac{q}{s}f(\frac{q}{s})/\left(1 - F(\frac{q}{s})\right)$ is increasing in q and it follows that Equation (3.9) is negative for $q < q^*$ and positive for $q > q^*$. □

Proof of Proposition 3.3

Proof. Consider a service level contract with $w_{WP} < w_{SL} < r$, $0 < s \leq 1$, and p chosen according to Equation (3.7). To prove the proposition, we show that the expected profit function of the

service level contract is more concave than the expected profit function of the wholesale price contract:

$$\begin{aligned}\frac{d^2}{dq^2}\Pi_{SL}(s,w_{SL},q) &= -w_{SL}f(q) - (r-w_{SL})\,f(q) - \frac{w_{SL}-w_{WP}}{1-F(\frac{q^*}{s})}\frac{f(\frac{q}{s})}{s}\\ &= -rf(q) - \frac{w_{SL}-w_{WP}}{1-F(\frac{q^*}{s})}\frac{f(\frac{q}{s})}{s}\\ &< -rf(q) = \frac{d^2}{dq^2}\Pi_{WP}(w_{WP},q).\end{aligned}$$ □

Proof of Proposition 3.4

Proof. Consider a service level contract with $w_{WP} < w_{SL} < r$, $0 < s \leq 1$, and p chosen according to Equation (3.7). We show that the concavity of the expected profit function is decreasing in s, that is, that the second derivative of the expected profit function with respect to q becomes less negative as s increases. We next provide the proof for $q = q^*$ and then show that the result also holds for $q < q^*$. The concavity of expected profit function at $q = q^*$ is

$$\frac{d^2}{dq^2}\Pi_{SL}(s,w_{SL},q)\Big|_{q=q^*} = -rf(q^*) - (w_{SL}-w_{WP})\,\frac{1}{q^*}\frac{q^*}{s}\frac{f(\frac{q^*}{s})}{1-F\left(\frac{q^*}{s}\right)}. \tag{3.10}$$

The concavity of the expected profit (Equation (3.10)) is increasing in s if $\frac{q^*}{s}\frac{f\left(\frac{q^*}{s}\right)}{1-F\left(\frac{q^*}{s}\right)} = \frac{q^*}{s}h\left(\frac{q^*}{s}\right)$ is decreasing in s. $\frac{q^*}{s}h\left(\frac{q^*}{s}\right)$ is a generalized failure rate with a logconcave density function, which is decreasing in s. Thus, the right term in Equation (3.10) is decreasing in s, implying that concavity is decreasing as s increases, that is, $\frac{d}{ds}\frac{d^2}{dq^2}\Pi_{SL}(s,w_{SL},q)\Big|_{q=q^*} > 0$.

To prove that the proposition also holds for $q < q^*$, we consider the first derivative of the concavity with respect to s:

$$\begin{aligned}\frac{d}{ds}\frac{d^2}{dq^2}\Pi_{SL}(s,w_{SL},q) &= -(w_{SL}-w_{WP})\frac{-q\left(1-F(\frac{q^*}{s})\right)f'(\frac{q}{s}) + f(\frac{q}{s})\left(-s+sF\left(\frac{q^*}{s}\right)-q^*f\left(\frac{q^*}{s}\right)\right)}{s^3\left(-1+F\left(\frac{q^*}{s}\right)\right)^2}\\ &= (w_{SL}-w_{WP})\frac{f(\frac{q}{s})}{s^3\left(1-F\left(\frac{q^*}{s}\right)\right)}q\left(h\left(\frac{q^*}{s}\right)\frac{q^*}{q}+\frac{s}{q}+\frac{f'(\frac{q}{s})}{f(\frac{q}{s})}\right)\end{aligned} \tag{3.11}$$

We proved above that this derivative is positive for $q = q^*$, which implies that

$$h\left(\frac{q^*}{s}\right) + \frac{s}{q^*} + \frac{f'\left(\frac{q^*}{s}\right)}{f\left(\frac{q^*}{s}\right)} > 0$$

Logconcavity of the density function f implies that $\frac{f'\left(\frac{q}{s}\right)}{f\left(\frac{q}{s}\right)}$ is decreasing in q. Thus for $q < q^*$, the last term in Equation (3.11) is positive and the proposition also holds for $q < q^*$. □

Supplementary Materials

3.A Sample Instructions

The sample instructions below are for the wholesale price (Treatment 1) and the service level contract with $s = 75$ and $p = 145$ (Treatment 2). Instructions for the other service level contracts (Treatments 3 and 4) considered in our study are presented analogously but with different parameters.

At the end of the experiment, we asked our subjects to rate, on a 7-point Likert scale (from "strongly disagree" to "strongly agree"), how much they agree or disagree with the statement "*The instructions were clear and precise*" (see 3.E). A Kruskal-Wallis H test showed that there was no statistically significant difference in the rating across our four treatments, $\chi^2(3) = 2.845$, $p = 0.416$, with a mean rank of 6.0, 5.8, 5.4, and 5.3 for Treatments 1, 2, 3, and 4, re spectively.

Instructions for the Wholesale Price Contract

The purpose of today's session is to study how people make decisions in a particular situation. If you follow these instructions carefully and make good decisions, you could earn a considerable amount of money. If you have any questions, feel free to raise your hand and the experimenter will come to you and answer your question.

During this session, you will play a game from which you can earn money. Your earnings in this session are expressed in 'francs' with the following exchange rate:

3,000 francs = \$1

Description of the game

You are a retailer who sells a single generic product. In each period of the game, you will order the product from an external supplier at a purchase price of **3 francs** per unit and sell the product to customers at a sales price of **12 francs** per unit.

You play 100 periods with identical activities:

- At the beginning of each period, you determine the order quantity before you know what your customers will demand. You can choose your order quantity freely between 0 and 100.
- Once you have submitted your order, the computer generates a customer demand.
  To generate a customer demand, the computer draws a random number **between 1 and 100**. All customer demands between 1 and 100 are equally likely. The demand drawn for any one period is independent of the demand from earlier periods. So a small or large demand in earlier periods has no influence on whether demand is small or large in later periods.
- Demand is filled and your profit is computed. There are two different cases:
  - If customer demand is less than or equal to your order quantity, all customer demand can be filled. Your profit is:

$$\text{Profit} = 12 \cdot \text{Customer Demand} - 3 \cdot \text{Order Quantity}$$

  - If customer demand is greater than your order quantity, only customer demand up to the order quantity can be filled. Your profit is:

$$\text{Profit} = 12 \cdot \text{Order Quantity} - 3 \cdot \text{Order Quantity}$$

- If your order quantity was greater than the demand, the remaining stock is disposed of at **no cost** at the end of the period. In other words, remaining inventory is worthless and is not carried over to later periods.

Profit calculation per period

Your profit in each period is:

$$\begin{aligned} \text{Profit} &= \text{Sales Price} \cdot \text{Customer Demand Filled} && (\textit{Revenue}) \\ &- \text{Purchase Price} \cdot \text{Order Quantity} && (\textit{Order Cost}) \end{aligned}$$

Please be aware that you can also make a loss. Should you have accumulated losses after the 100 periods, these will be set against your show-up fee of \$5.

Example

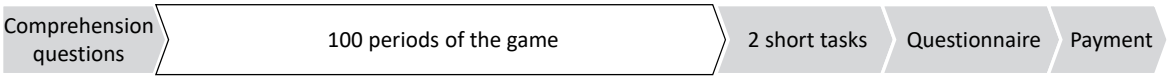
Suppose customer demand is 60 units and you ordered 80 units:
All customer demand can be filled, and your profit in this period is:

$$\text{Profit} = 12 \cdot 60 - 3 \cdot 80 = 480$$

Now, suppose customer demand is 60 units and you ordered 40 units:
Only 40 units can be filled, and your profit in this period is:

$$\text{Profit} = 12 \cdot 40 - 3 \cdot 40 = 360$$

Sequence of the experiment



Payment determination

At the end of the session, your total earnings will be converted into U.S. dollars at a rate of \$1 per 3,000 francs, added to your show-up fee of \$5, and paid to you in cash.

Instructions for the Service Level Contract

The purpose of today's session is to study how people make decisions in a particular situation. If you follow these instructions carefully and make good decisions, you could earn a considerable amount of money. If you have any questions, feel free to raise your hand and the experimenter will come to you and answer your question.

During this session, you will play a game from which you can earn money. Your earnings in this session are expressed in 'francs' with the following exchange rate:

3,000 francs = $1

Description of the game

You are a retailer who sells a single generic product. In each period of the game, you will order the product from an external supplier at a purchase price of **6 francs** per unit and sell the product to customers at a sales price of **12 francs** per unit.

You play 100 periods with identical activities:

- At the beginning of each period, you receive a fixed endowment of **225 francs**.
- You determine the order quantity before you know what your customers will demand. You can choose your order quantity freely between 0 and 100.
- Once you have submitted your order, the computer generates a customer demand. To generate a customer demand, the computer draws a random number **between 1 and 100**. All customer demands between 1 and 100 are equally likely. The demand drawn for any one period is independent of the demand from earlier periods. So a small or large demand in earlier periods has no influence on whether demand is small or large in later periods.
- Demand is filled and your fill rate is calculated. There are two different cases:

  - If customer demand is less than or equal to your order quantity, all customer demand can be filled. The fill rate is:

$$\text{Fill Rate} = \frac{\text{Customer Demand Filled}}{\text{Customer Demand}} = \frac{\text{Customer Demand}}{\text{Customer Demand}} = 100\%$$

  - If customer demand is greater than your order quantity, only customer demand up to the order quantity can be filled. The fill rate is:

$$\text{Fill Rate} = \frac{\text{Customer Demand Filled}}{\text{Order Quantity}} = \frac{\text{Order Quantity}}{\text{Customer Demand}} < 100\%$$

- The fill rate target is **75%**; that is, the target is filling at least 75% of actual period demand. An amount of **145 francs** is deducted for each unit you fall short of target.
- If your order quantity was greater than the demand, the remaining stock is disposed of at **no cost** at the end of the period. In other words, remaining inventory is worthless and is not carried over to later periods.

Profit calculation per period

Your profit in each period is:

$$\begin{aligned} \text{Profit} &= \text{Endowment} \\ &+ \text{Sales Price} \cdot \text{Customer Demand Filled} && (\textit{Revenue}) \\ &- \text{Purchase Price} \cdot \text{Order Quantity} && (\textit{Order Cost}) \\ &- \text{Deduction} \cdot \text{Units Short of Target} && (\textit{Deduction}) \end{aligned}$$

Please be aware that you can also make a loss. Should you have accumulated losses after the 100 periods, these will be set against your show-up fee of $5.

Example

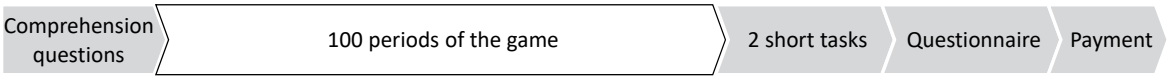
Suppose customer demand is 60 units and you ordered 80 units:
All customer demand can be filled, and your fill rate = 60/60 = 100%. Because your fill rate is above 75%, you do not incur a deduction. Your profit in this period is:

$$\text{Profit} = 225 + 12 \cdot 60 - 6 \cdot 80 = 465$$

Now, suppose customer demand is 60 units and you ordered 40 units:
Only 40 units can be filled, and your fill rate = 40/60 = 66.7%. 45 units, 5 units more, would have been required to achieve a fill rate of 75%. You incur a deduction of 145 francs for each of the 5 units you are short of target. Your profit in this period is:

$$\text{Profit} = 225 + 12 \cdot 40 - 6 \cdot 40 - 145 \cdot 5 = -260$$

Sequence of the experiment

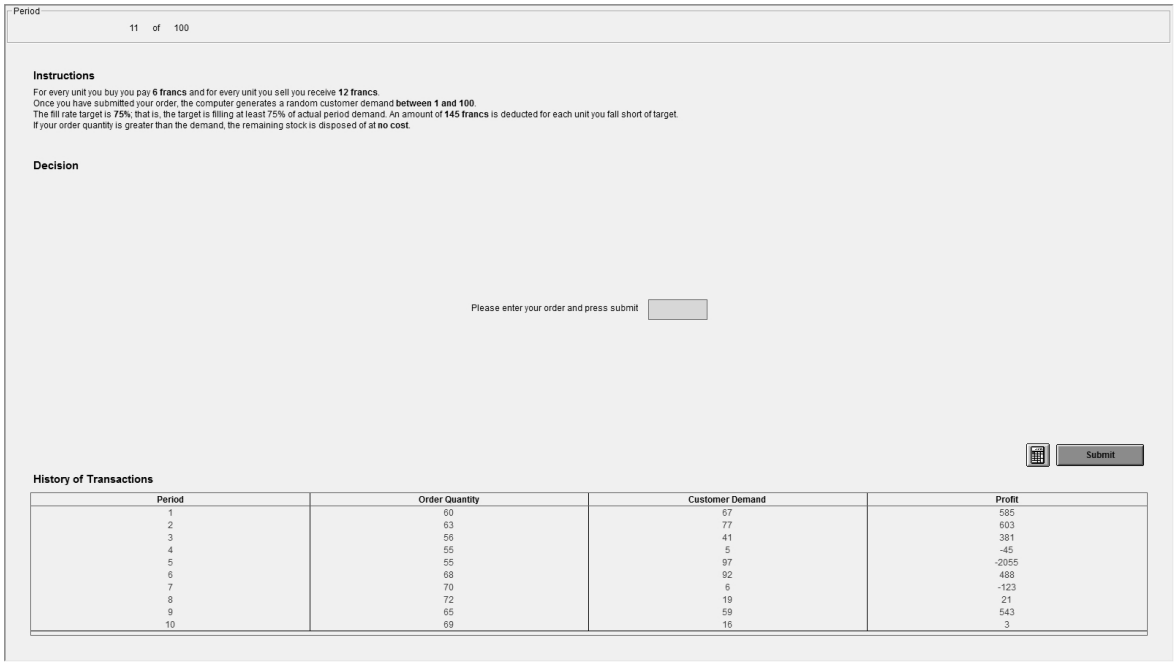


Payment determination

At the end of the session, your total earnings will be converted into U.S. dollars at a rate of $1 per 3,000 francs, added to your show-up fee of $5, and paid to you in cash.

3.B Decision and Result Screens

Decision Screen



Period

11 of 100

Instructions

For every unit you buy you pay **6 francs** and for every unit you sell you receive **12 francs**.
Once you have submitted your order, the computer generates a random customer demand **between 1 and 100**.
The fill rate target is **75%**; that is, the target is filling at least 75% of actual period demand. An amount of **145 francs** is deducted for each unit you fall short of target.
If your order quantity is greater than the demand, the remaining stock is disposed of at **no cost**.

Decision

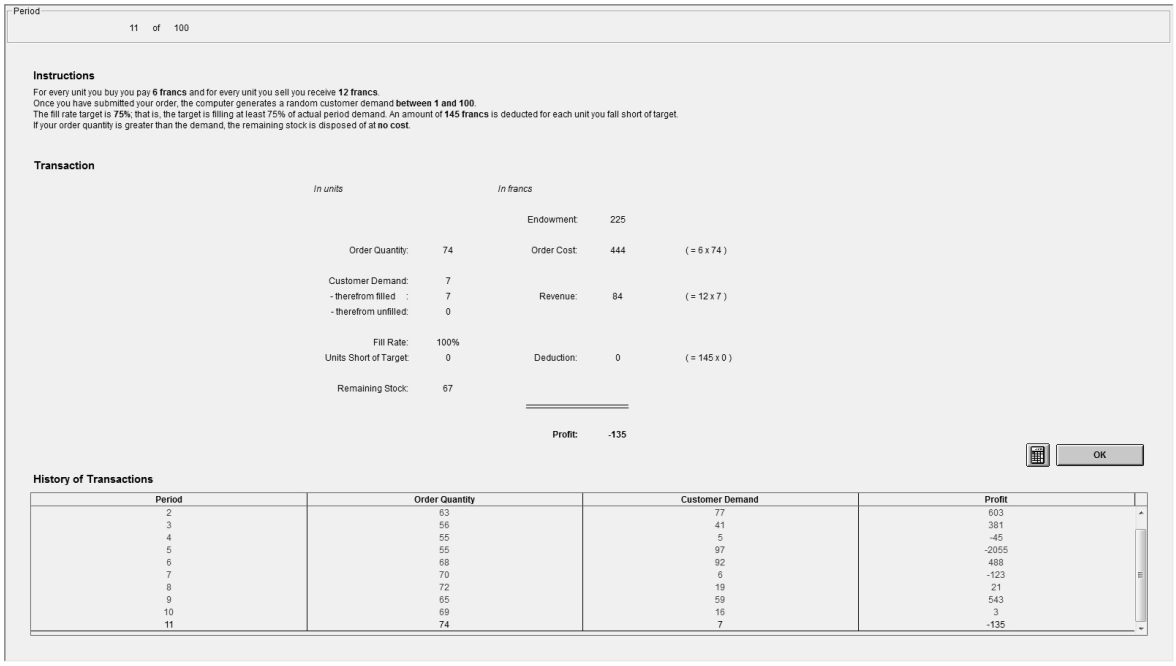
Please enter your order and press submit

Submit

History of Transactions

Period	Order Quantity	Customer Demand	Profit
1	60	67	585
2	63	77	603
3	56	41	381
4	55	5	-45
5	55	97	-2055
6	68	92	488
7	70	6	-123
8	72	19	21
9	65	59	543
10	69	16	3

Result Screen



Period

11 of 100

Instructions

For every unit you buy you pay **6 francs** and for every unit you sell you receive **12 francs**.
Once you have submitted your order, the computer generates a random customer demand **between 1 and 100**.
The fill rate target is **75%**; that is, the target is filling at least 75% of actual period demand. An amount of **145 francs** is deducted for each unit you fall short of target.
If your order quantity is greater than the demand, the remaining stock is disposed of at **no cost**.

Transaction

In units		*In francs*		
		Endowment:	225	
Order Quantity:	74	Order Cost:	444	(= 6 x 74)
Customer Demand:	7			
- therefrom filled :	7	Revenue:	84	(= 12 x 7)
- therefrom unfilled:	0			
Fill Rate:	100%			
Units Short of Target:	0	Deduction:	0	(= 145 x 0)
Remaining Stock:	67			
		Profit:	**-135**	

OK

History of Transactions

Period	Order Quantity	Customer Demand	Profit
2	63	77	603
3	56	41	381
4	55	5	-45
5	55	97	-2055
6	68	92	488
7	70	6	-123
8	72	19	21
9	65	59	543
10	69	16	3
11	**74**	**7**	**-135**

3.C Comprehension Questions

In this section, we provide the questions and answers from the computerized quiz that our subjects had to pass before they could continue with the main part of the experiment. The sample quiz below contains the questions and answers for the service level contract with $s = 75$ and $p = 145$. Note that the correct answers are indicated with dots (⊙). The questions from the other treatments in our study are presented analogously, except for the differences outlined in Section 3.4.2.

Section 1

Suppose you ordered **60 units** and the computer generated a customer demand of **90 units**.

(1) What are your *Order Costs* in this period?

- ⊙ 360 francs
- ○ 540 francs
- ○ 720 francs

(2) How many units can you sell in this period?

- ○ 30
- ⊙ 60
- ○ 90

(3) What is your *Revenue* in this period?

- ○ 360 francs
- ⊙ 720 francs
- ○ 1080 francs

(4) What is your *Fill Rate* in this period?

- ○ 33.3%
- ⊙ 66.7%
- ○ 100%

(5) How many units more (if any) would have been required to achieve a fill rate of 75%?

- ○ 0
- ⊙ 8
- ○ 30

(6) What is your *Deduction* in this period?

- ○ 0 francs
- ⊙ 1160 francs
- ○ 4350 francs

(7) What is your *Profit* in this period? (Note that your endowment is 225 francs.)

- ⊙ -575 francs
- ○ -215 francs
- ○ 585 francs

Section 2

Suppose you ordered **60 units** and the computer generated a customer demand of **30 units**.

(1) What are your *Order Costs* in this period?

- ○ 180 francs
- ⊙ 360 francs
- ○ 720 francs

(2) How many units can you sell in this period?

- ⊙ 30
- ○ 60
- ○ 90

(3) What is your *Revenue* in this period?

- ○ 180 francs
- ⊙ 360 francs
- ○ 720 francs

(4) What is your *Fill Rate* in this period?

- ○ 33.3%
- ○ 66.7%
- ⊙ 100%

(5) How many units more (if any) would have been required to achieve a fill rate of 75%?

- ⊙ 0
- ○ 8
- ○ 30

(6) What is your *Deduction* in this period?

- ⊙ 0 francs
- ○ 1160 francs
- ○ 4350 francs

(7) What is your *Profit* in this period? (Note that your endowment is 225 francs.)

- ○ -135 francs
- ⊙ 225 francs
- ○ 585 francs

Section 3

(1) All 100 periods are payoff relevant.

- ⊙ True
- ○ False

(2) You can choose your order quantity freely between 0 and 100.

- ⊙ True
- ○ False

(3) Depending on your order quantity and customer demand, you can also make losses.

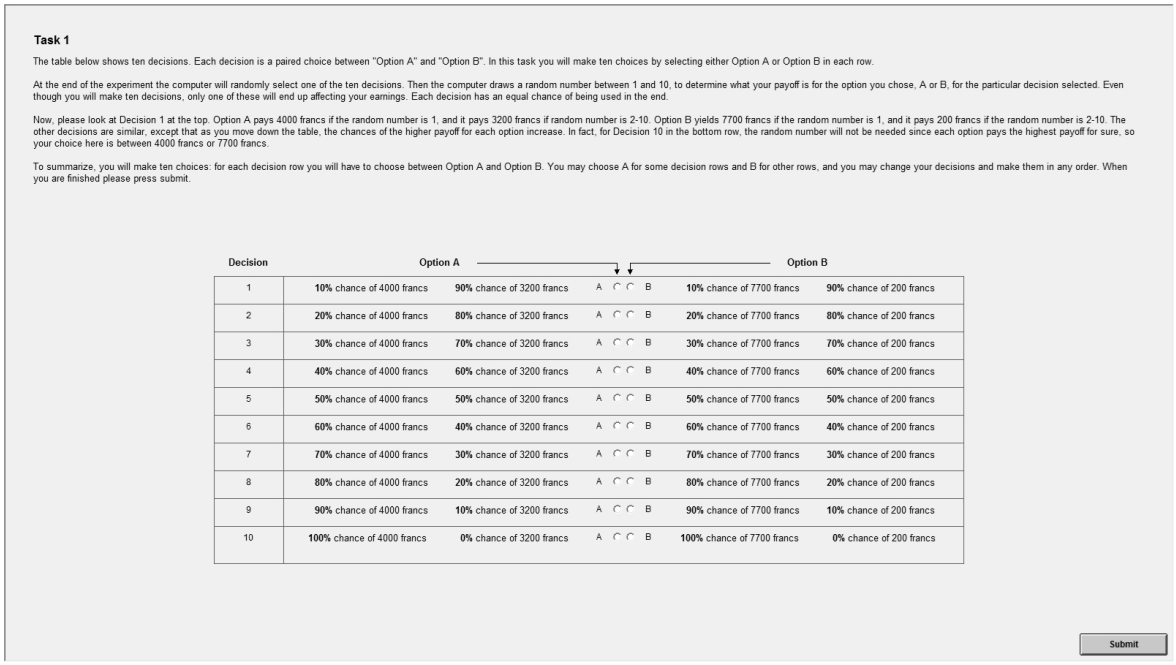
- ⊙ True
- ○ False

Figure 3.7 Proportion of Required Attempts to Pass a Section of the Quiz by Treatment

Figure 3.7 shows the proportion of subjects who required one, two, or more attempts to pass a specific section of the quiz. Comparing the average number of attempts within a section across our four treatments, we did not find any significant differences, except for the first section, in which subjects in the $\text{SLC}_{75\%,\,145}$ treatment required on average 1.9 more attempts to correctly answer all questions than did subjects in the WPC treatment ($p = 0.017$). This is not surprising because the first section of the WPC treatment had three fewer questions and required less calculation effort.

3.D Additional Tasks

Task 1 – Holt and Laury's (2002) Risk Elicitation Task



Task 1

The table below shows ten decisions. Each decision is a paired choice between "Option A" and "Option B". In this task you will make ten choices by selecting either Option A or Option B in each row.

At the end of the experiment the computer will randomly select one of the ten decisions. Then the computer draws a random number between 1 and 10, to determine what your payoff is for the option you chose, A or B, for the particular decision selected. Even though you will make ten decisions, only one of these will end up affecting your earnings. Each decision has an equal chance of being used in the end.

Now, please look at Decision 1 at the top. Option A pays 4000 francs if the random number is 1, and it pays 3200 francs if random number is 2-10. Option B yields 7700 francs if the random number is 1, and it pays 200 francs if the random number is 2-10. The other decisions are similar, except that as you move down the table, the chances of the higher payoff for each option increase. In fact, for Decision 10 in the bottom row, the random number will not be needed since each option pays the highest payoff for sure, so your choice here is between 4000 francs or 7700 francs.

To summarize, you will make ten choices: for each decision row you will have to choose between Option A and Option B. You may choose A for some decision rows and B for other rows, and you may change your decisions and make them in any order. When you are finished please press submit.

Decision	Option A			Option B	
1	**10%** chance of 4000 francs	**90%** chance of 3200 francs	A ○ ○ B	**10%** chance of 7700 francs	**90%** chance of 200 francs
2	**20%** chance of 4000 francs	**80%** chance of 3200 francs	A ○ ○ B	**20%** chance of 7700 francs	**80%** chance of 200 francs
3	**30%** chance of 4000 francs	**70%** chance of 3200 francs	A ○ ○ B	**30%** chance of 7700 francs	**70%** chance of 200 francs
4	**40%** chance of 4000 francs	**60%** chance of 3200 francs	A ○ ○ B	**40%** chance of 7700 francs	**60%** chance of 200 francs
5	**50%** chance of 4000 francs	**50%** chance of 3200 francs	A ○ ○ B	**50%** chance of 7700 francs	**50%** chance of 200 francs
6	**60%** chance of 4000 francs	**40%** chance of 3200 francs	A ○ ○ B	**60%** chance of 7700 francs	**40%** chance of 200 francs
7	**70%** chance of 4000 francs	**30%** chance of 3200 francs	A ○ ○ B	**70%** chance of 7700 francs	**30%** chance of 200 francs
8	**80%** chance of 4000 francs	**20%** chance of 3200 francs	A ○ ○ B	**80%** chance of 7700 francs	**20%** chance of 200 francs
9	**90%** chance of 4000 francs	**10%** chance of 3200 francs	A ○ ○ B	**90%** chance of 7700 francs	**10%** chance of 200 francs
10	**100%** chance of 4000 francs	**0%** chance of 3200 francs	A ○ ○ B	**100%** chance of 7700 francs	**0%** chance of 200 francs

Submit

Task 2 – Gächter et al.'s (2010) Lottery Choice Task to Assess Loss Aversion

Task 2

The table below shows seven different gambles. In this task you will make seven choices by selecting for each of the gambles whether you want to play it ("Accept") or not ("Reject").

Each of the seven gambles has two possible outcomes with 50% chance of occurring. Your earnings for this task will be determined by: (1) whether you accept or reject a gamble and (2) which of the two possible outcomes occur.

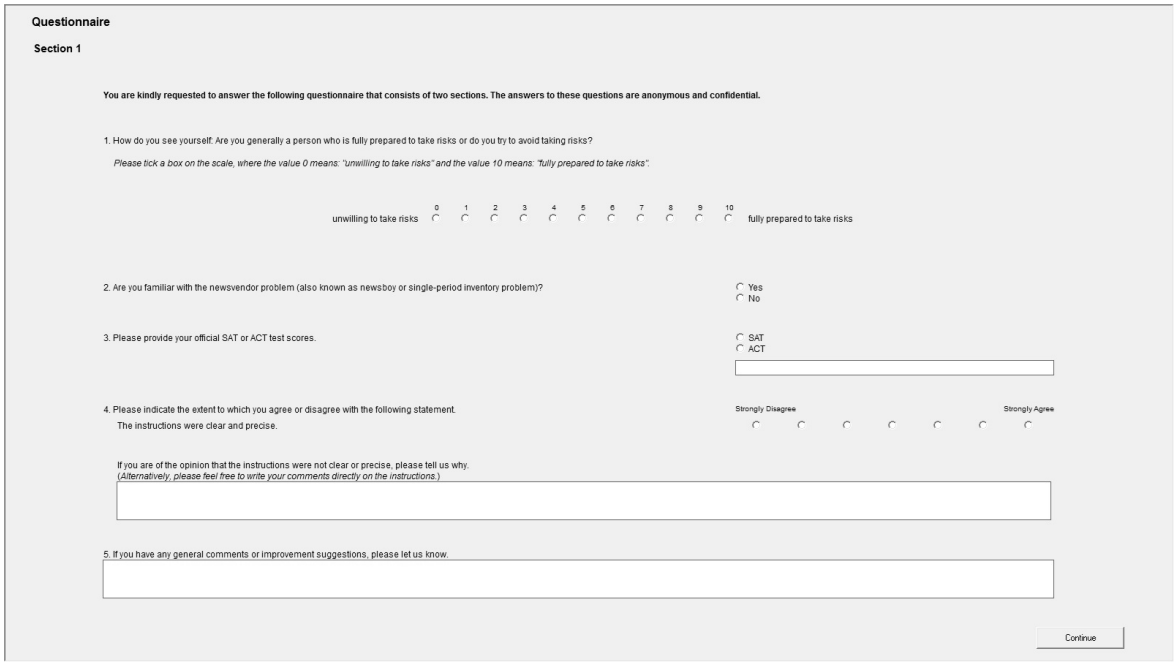
At the end of the experiment the computer will randomly select one of the seven gambles and will randomly select one of the two possible outcomes for the selected gamble. For example, let's assume the computer randomly selects gamble 1. If you have chosen to accept gamble 1, you lose 1000 francs with 50% chance and you win 6000 francs with 50% chance. If you have chosen to reject gamble 1, you do not win or lose any money.

Please select for each gamble whether you want to accept or reject it. When you are finished please press submit.

Gamble	Accept	Reject
1. With 50% chance, you **lose 1000 francs**. With 50% chance, you **win 6000 francs**.	○	○
2. With 50% chance, you **lose 2000 francs**. With 50% chance, you **win 6000 francs**.	○	○
3. With 50% chance, you **lose 3000 francs**. With 50% chance, you **win 6000 francs**.	○	○
4. With 50% chance, you **lose 4000 francs**. With 50% chance, you **win 6000 francs**.	○	○
5. With 50% chance, you **lose 5000 francs**. With 50% chance, you **win 6000 francs**.	○	○
6. With 50% chance, you **lose 6000 francs**. With 50% chance, you **win 6000 francs**.	○	○
7. With 50% chance, you **lose 7000 francs**. With 50% chance, you **win 6000 francs**.	○	○

Submit

3.E Post-Experimental Questionnaire



Questionnaire

Section 1

You are kindly requested to answer the following questionnaire that consists of two sections. The answers to these questions are anonymous and confidential.

1. How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

Please tick a box on the scale, where the value 0 means: "unwilling to take risks" and the value 10 means: "fully prepared to take risks".

unwilling to take risks 0 1 2 3 4 5 6 7 8 9 10 fully prepared to take risks

2. Are you familiar with the newsvendor problem (also known as newsboy or single-period inventory problem)?
- Yes
- No

3. Please provide your official SAT or ACT test scores.
- SAT
- ACT

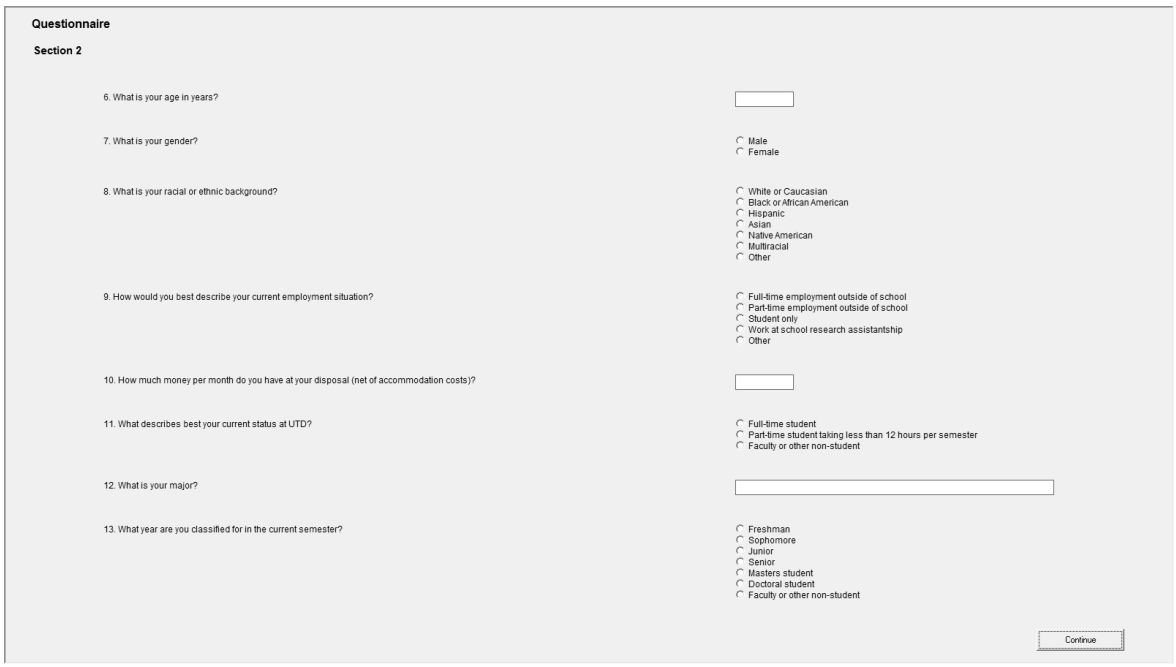
4. Please indicate the extent to which you agree or disagree with the following statement.
The instructions were clear and precise.

Strongly Disagree ... Strongly Agree

If you are of the opinion that the instructions were not clear or precise, please tell us why.
(Alternatively, please feel free to write your comments directly on the instructions.)

5. If you have any general comments or improvement suggestions, please let us know.

Continue



Questionnaire

Section 2

6. What is your age in years?

7. What is your gender?
- Male
- Female

8. What is your racial or ethnic background?
- White or Caucasian
- Black or African American
- Hispanic
- Asian
- Native American
- Multiracial
- Other

9. How would you best describe your current employment situation?
- Full-time employment outside of school
- Part-time employment outside of school
- Student only
- Work at school research assistantship
- Other

10. How much money per month do you have at your disposal (net of accommodation costs)?

11. What describes best your current status at UTD?
- Full-time student
- Part-time student taking less than 12 hours per semester
- Faculty or other non-student

12. What is your major?

13. What year are you classified for in the current semester?
- Freshman
- Sophomore
- Junior
- Senior
- Masters student
- Doctoral student
- Faculty or other non-student

Continue

Chapter 4

Trusting the Forecast: The Role of Numeracy

Relying on a scientific forecast to make decisions is an act of trust. Conventionally, forecast guidance that includes uncertainty measures is generally thought to be for quantitatively sophisticated decision makers, while firmer forecast guidance that omits uncertainty measures is more easily and generally understood. In a controlled study of decision making in a simple take-the-risk or take-the-cost decision game, we examine compliance rates (trust) for forecast guidance provided as probabilities as well as recommendations. Most strikingly, and contrary to our initial expectation, low numerate subjects exhibit less trust in recommendation forecasts than do high numerates. While we find a positive relationship between subjects' numeracy and trust in probability forecasts, this relationship is overshadowed by the fact that even high numerate subjects use the probabilities inefficiently. Forecast guidance that blends probabilities and recommendations, in a way designed to offset the major behavioral shortcomings we observe, improves compliance; especially for high numerates. We argue that improving low numerate individuals' trust in forecasting will require a new approach.

4.1 Introduction

Many important business and personal decisions have to be made under risk and uncertainty. Typically, experts use scientific models to derive forecasts, while non-expert users employ these forecasts to make the decisions. With the growing availability of big data and predictive analytics, more and more decisions are guided by expert models, ranging from improving emergency responsiveness (Green and Kolesar 2004, Pinker 2007), to supply base diversification, to talent management (Arellano et al. 2017). The forecasts for these specific applications are inherently uncertain. For the forecast users, the value of scientific forecasts rests largely with a reduction of uncertainty over future events, thereby enabling better decision making today (Weber 1994, Fox and Tversky 1998).

Definitions of trust vary but one commonly accepted definition across multiple disciplines is that "trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another" (Rousseau et al. 1998, p. 395). For most people, relying on an expert forecast for decision guidance is an act of trust (Burgman 2016). Forecasts are by nature uncertain, and most forecast users have only a partial understanding of the workings and expected accuracy of the scientific models that underlie the forecast. To the extent that forecast users misunderstand the underlying uncertainty, or fail to use this information properly, they are vulnerable to decision mistakes that, rightly or wrongly, may get attributed to a faulty forecast.

Diversity in user numeracy complicates forecast guidance on forecast uncertainty, a clear measure of which is, in decision theory, critical to good decision making. Forecast uncertainty is difficult to convey effectively in words because different people interpret words like 'likely' differently (Bryant and Norman 1980, Beyth-Marom 1982, Wallsten 1986, Karelitz and Budescu 2004). Conventionally, forecast guidance that includes quantitative measures of uncertainty is thought to be for numerically sophisticated decision makers, while firmer forecast guidance that omits uncertainty measures, such as point forecasts or recommendations of what decision to take, is more easily and generally understood. Regarding trust there is a potential trade-off here: The measure of uncertainty is, in decision theory, sufficient to make optimal decisions but may also be prone to misunderstanding or misuse particularly among

numerically less sophisticated users (Schwartz et al. 1997). Firmer guidance that omits or de-emphasizes uncertainty – point forecasts or recommendations for example – is easier to incorporate into decision making (assuming users deem the information credible) but may also raise questions of reliability as users observe forecast errors (Bliss et al. 1995, Meyer and Bitan 2002).

Here, we examine this trade-off in the context of a cost-loss game, in which a decision maker chooses whether to take the risk of a loss or take a cost to avoid the risk (Bilham 1922, Thompson 1952). Many of the important risk decisions people face have a cost-loss structure. Examples include whether to evacuate in the face of an impending hurricane, how to invest for retirement, whether to elect a preventative medical procedure, and how to vote on climate change policies. For each of these examples a forecast is typically available to aid in the decision.

Making an informed decision in a cost-loss situation requires an evaluation of how likely it is that the loss occurs. Providing the relevant probabilities is a straightforward way of conveying this information to the forecast-users. Behavioral studies of cost-loss games find that providing numerical forecast information of risk can improve the quality of decisions relative to providing a point forecast only (Roulston et al. 2006). Work on categorical recommendations find a cry wolf effect when the recommendation turns out to be a false alarm (Meyer and Bitan 2002, Roulston and Smith 2004, Bolton and Katok 2017). However, it is likely that providing the relevant probabilities is only meaningful to the extent that people have the ability to process basic probability and numerical concepts, a construct called numeracy (Schwartz et al. 1997).

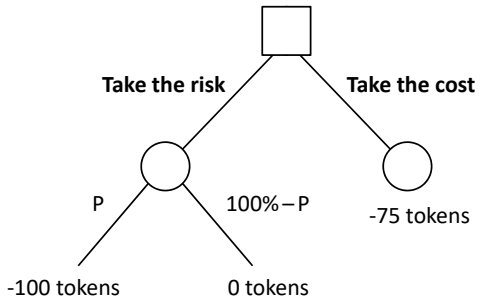
People differ substantially in their numeracy skills, and many people are innumerate (Schwartz et al. 1997, Lipkus et al. 2001, Cokely et al. 2012). The 2003 National Assessment of Adult Literacy indicates that about half the U.S. population has only very basic or below basic numeracy skills (Kutner et al. 2006). The Survey of Adult Skills (PIAAC) shows that in almost all OECD countries a sizable proportion of adults has poor numeracy skills, 23% of adults, on average (OECD 2016). The survey also reveals that across countries and economies, there is a positive correlation between numeracy skills and trust. While the causal nature of this relationship is difficult to discern, it clearly matters, because trust is the foundation of

economic behavior. Gurmankin et al. (2004) revealed that trust in numerical information increases in numeracy skills. They hypothesize that low numerate individuals will be more likely than their high numerate peers to reject information that they perceive to be inaccurate or unreliable. Furthermore, low numerate individuals are less likely to retrieve and use appropriate numerical principles, are less likely to use accuracy-enhancing System 2 forms of thinking, and are more vulnerable to System 1 cognitive errors (Kahneman 2003, Peters et al. 2006, 2007). Peters et al. (2007) found that this effect of numeracy is not due to general intelligence.

In some conditions of our experiment, forecasts are offered in the form probabilities of the loss event, a form of forecast that emphasizes the inherent uncertainty in the underlying model. In other conditions, the forecast is offered as a recommendation of whether to take the cost or take the risk, a form of forecast that provides the optimal action given the expected probability of loss while de-emphasizing the underlying model uncertainty. Both kinds of forecasts are common. We measured numeracy in each condition, to learn more about the role of numerical abilities and its interaction with the form of forecast. Numeracy is deemed important because it affects risk comprehension and the efficiency of decision making (Reyna et al. 2009). The hypothesis we test in this study is that people with higher numeracy skills are better able to utilize probability information, while lower numeracy people perform better with recommendation.

This chapter is organized as follows. In Section 4.2, we present the details of our experimental design, protocol, and sample. In Section 4.3, we present the results of our study, starting with aggregate descriptive statistics, the role-of-numeracy analysis, and following with analyses of the behavioral strength and weaknesses of probabilities and recommendations. In Section 4.3, we also present forecast guidance that blends probabilities and recommendations in a way designed to offset the major behavioral shortcomings. In Sections 4.4, we summarize our conclusions and discuss the managerial implications of our findings.

Figure 4.1 The Cost-Loss Game



4.2 Experiment

4.2.1 Design

4.2.1.1 Forecast guidance in a cost-loss game

Figure 4.1 depicts the extensive form of the cost-loss game used in the experiment. The decision maker chooses between two actions: *take the risk*, in which case she loses 100 tokens with probability P or *take the cost*, in which case she loses 75 tokens for certain.

Each subject in the experiment played this game 100 times, with the value of P varying each round. The expected cost minimizing decision depends on P; specifically, the implied choice rule is take the risk if the expected loss, $P \times 100$, is less than 75, and take the cost if greater than 75. Given the monetary stakes and the amount of game repetition, this rule should approximate the optimal decision rule for all but highly risk averse individuals.[1] We use this rule to benchmark forecast compliance; that is, we say that a decision is forecast compliant if it is optimal given the forecast information about P available to the decision maker.

The experiment manipulates the forecast information decision makers have about probability P and compare the optimality of subject decision making across manipulations. At the beginning of all conditions, subjects are told that the probability of the loss event will average to 50% over the series of cost-loss games they play. In the baseline *Neither* condition, this is

[1] The major conclusions we will draw are robust to the assumption that players are risk averse; see Bolton and Katok (2017).

all the information subjects are given. The optimal action in each round is to take the risk. In the other conditions, subjects also receive round-by-round forecast guidance.

In the *Probability* condition, forecast guidance is communicated as the actual value of P for that round. In the *Recommendation* condition, guidance takes the form of advice: "take the risk" or "take the cost" (the value of P is not provided). The recommendation follows the optimal decision rule. The Both condition, both probability and recommendation are presented together. For all three of these conditions, the optimal decision rule stipulates taking the risk if $P < 75\%$; take the cost otherwise. The round-by-round forecasts provide better information about the risk than does knowing only the average cross-rounds probability, so decision makers with round-by-round forecasts should experience fewer total losses (costs paid plus losses incurred from risks) than users in the Neither condition.

Because it is optimal when no day-to-day forecast guidance is available, take the risk is the *status quo* action. In this context, the function of the day-to-day forecast is to alert the decision maker that the status quo action should be abandoned in favor of taking the cost, the *siren action*.

4.2.1.2 Assessment of Numeracy

We assessed a subject's numeracy with seven questions and scored it as the total number of correct responses. Table 4.1 shows the numeracy questions accompanied by the percentage of subjects who responded correctly to each item.

Item 1 to 3 are taken from Schwartz et al. (1997) and the remaining items are taken from the Berlin Numeracy Test by Cokely et al. (2012). Building on the work of Lipkus et al. (2001) and Schwartz et al. (1997), Cokely et al. (2012) provide a relatively short and reliable instrument that has been proven to be the strongest single predictor of individual differences in understanding everyday risks, such as evaluating risk in numerical and non-numerical claims about consumer products and medical treatments or interpreting weather forecasts (Ghazal et al. 2014). The Berlin Numeracy Test was developed to assess numeracy of educated and highly educated samples, such as college students. Cokely et al. (2012) suggests to combine the Berlin Numeracy Test with the Schwartz et al. test when assessing

Table 4.1 Numeracy Questions Accompanied by the Percentage of Subjects Who Responded Correctly to Each Item

Item	% Correct
1. Imagine that we flip a fair coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips?	87.1
2. In the BIG BUCKS LOTTERY, the chance of winning a $10 prize is 1%. What is your best guess about how many people would win a $10 prize if 1,000 people each buy a single ticket to BIG BUCKS?	81.1
3. In ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?	56.2
4. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3, or 5)?	56.2
5. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?	38.8
6. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in a choir 100 are men. Out of the 500 inhabitants that are not in a choir 300 are men. What is the probability that a randomly drawn man is a member of the choir?	37.8
7. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red?	16.4

individuals who have lower levels of educational attainment. They tested the combined score on Amazon's Mechanical Turk online labor market and showed that it provides a fast assessment with excellent discriminability. We added two additional items to specifically test subjects' comprehension of expected values, such as "Imagine we toss a fair coin. If head comes up you win $20, if tail comes up you win $100. What is the expected payoff of this gamble?" We find a strong positive correlation between the combined score suggested by Cokely et al. (2012) and the number of correct answers to our additional items ($r = 0.552$, $p < 0.001$).

All questions were incentivized such that subjects earned ten cents for each question answered correctly.

4.2.2 Protocol

The experiment was conducted online using a self-developed Javascript software (implemented in Qualtrics). Upon accessing the experiment, subjects were randomly assigned to one of our four conditions, and this assignment remained constant throughout all rounds for each subject. The forecast information given to the subject (Neither, Probability, Recommendation,

or Both) depended on the condition and subjects were shown the corresponding on-screen instructions (Supplementary Material 4.A) at the beginning of the experiment.

After reading the instructions but before playing the cost-loss game, subjects had to pass a comprehension quiz that was the same for all conditions. The quiz comprised five questions provided in Supplementary Material 4.B. Subjects could continue only after they had correctly answered all questions on the first attempt (approximately one third were screened out). We used this approach to ensure that subjects have read and understood the instructions.

At the beginning of each of the 100 rounds of play, a subject receives an endowment of 150 tokens (to avoid "bankruptcy problems" in which subjects would owe the experimenter money; the endowment is fixed and does not change the normative analysis of the optimal action). Forecast information depending on the condition was made available to subjects before they made their decisions of whether to take the risk or take the cost. A random draw from a uniform distribution, consistent with the loss probability P for that round, determined the outcome of whether the loss occurred. Draws were independent across rounds, subjects, and conditions. After the decision, the outcome and payoff was displayed to the subjects (see screen shots of the computer interface in Supplementary Material 4.C).

After the main part of the experiment, we asked all subjects to "briefly describe how [they] have decided when to take the risk and when to take the cost". Subjects then answered the seven numeracy questions (Table 4.1) as well as our two additional items, and completed a risk elicitation task. We decided to use the Bomb Risk Elicitation Task (BRET) introduced by Crosetto and Filippin (2013) to asses subjects' risk preferences, because it requires minimal numeracy skills but still allows precise estimation of both risk aversion and risk seeking. The BRET asks subjects to choose the number of boxes they want to collect from a set of 100 boxes, one of which contains a bomb. Earnings increase with the number of boxes collected (ten cents per box) but are equal to zero if one of them contains the bomb. Thus, the number of boxes collected is a good proxy for subjects' risk appetite (for instructions and results see Supplementary Material 4.D).

At the conclusion of the experiment we asked subjects for additional demographic information, including age, gender, highest level of education completed, employment status, and

Table 4.2 Sample Demographics (N = 201)

Demographics	Percentage
Age[a]	35.0 (10.9)
Gender (female)	54.7
Highest level of education	
High school graduate	16.9
Some college	29.4
Bachelor's degree	38.8
Master's degree	12.4
Doctoral or advanced professional degree	2.5
Employment status	
Not working	19.4
Working (paid employee)	62.2
Working (self-employed)	18.4
Annual income from all sources before taxes	
$20,000 and under	28.9
$20,001 to $40,000	33.8
$40,001 to $60,000	21.9
$60,001 to $80,000	9.5
Over $80,000	6.0

Note. [a] *M* (*SD*)

own annual income from all sources before taxes (see post-experimental questionnaire in Supplementary Material 4.E).

4.2.3 Subjects

We recruited subjects on Amazon's Mechanical Turk (MTurk) online labor market (Buhrmester et al. 2011, Paolacci and Chandler 2014). We restricted participation to residents of the United States. After accepting to participate in our study, subjects were referred to an external website containing our online experiment. A total of 361 MTurk workers started our survey and were randomly assigned to one of our four conditions. Out of this 361 workers, 110 were screened out in the quiz and 50 did not complete the survey, resulting in a sample size of 201 subjects (demographic features of the sample are summarized in Table 4.2). There were no statistically significant differences neither for the failure rates in the quiz ($\chi^2(3) = 1.287$, $p = 0.732$) nor for the drop-out rates ($\chi^2(3) = 3.187$, $p = 0.364$) across conditions. Selective attrition is therefore not an important concern for our study.

Table 4.3 Summary of the Experimental Conditions

		Probability	
		Not provided	Provided
Recommendation	Not provided	*Neither* ($n = 50$)	*Probability* ($n = 60$)
	Provided	*Recommendation* ($n = 47$)	*Both* ($n = 44$)

Subjects required on average 31 minutes to complete the experiment. Actual average earnings, including a $1 participation fee, were $6.64, resulting in an average wage of more than $13 per hour, which are substantial earnings on MTurk. Table 4.3 summarizes the condition labels and sample sizes.

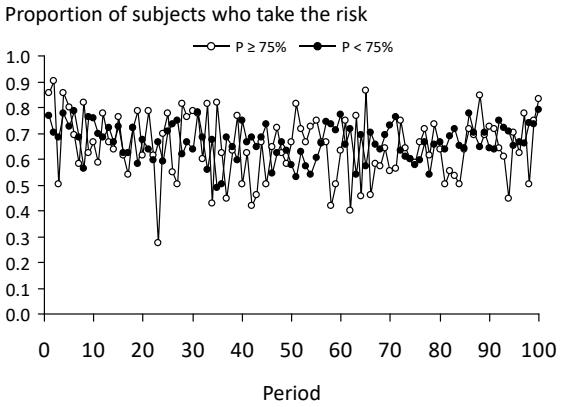
4.3 Results

4.3.1 Overall Results

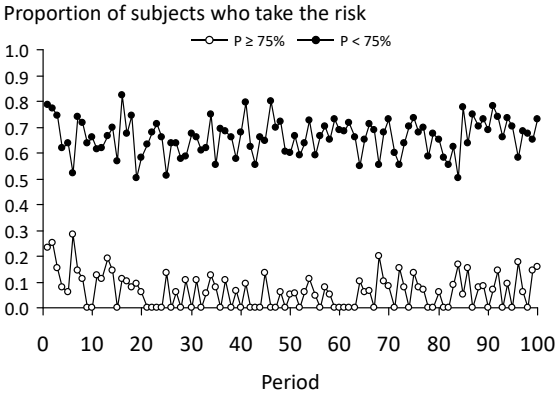
We begin by plotting the proportion of decisions to take the risk over time in Figure 4.2. Fitting a simple trend line to the data in Figure 4.2 (where each data point corresponds to the proportion of subjects who take the risk per round), we find a significant trend only if the advice is to take the cost for the Probability (Figure 4.2(b)) and the Both (Figure 4.2(d)) condition (random effects regression, two-tailed $p < 0.05$ in both cases). If the optimal action is to take the cost in the Probability condition the overall average increase in compliance is 0.04 percentage points per round (standard error = 0.016); if the optimal action is to take the cost in the Both condition, the overall average increase in compliance is 0.10 percentage points per round (standard error = 0.021).

Overall, we can conclude that there is little to no trend in our data (see Figure 4.5). Therefore, to further investigate how different information affects behavior we plot the average compliance across all rounds and subjects by condition, and broken out on whether the optimal action was to take the risk or take the cost in Figure 4.3. On the right side, the figure

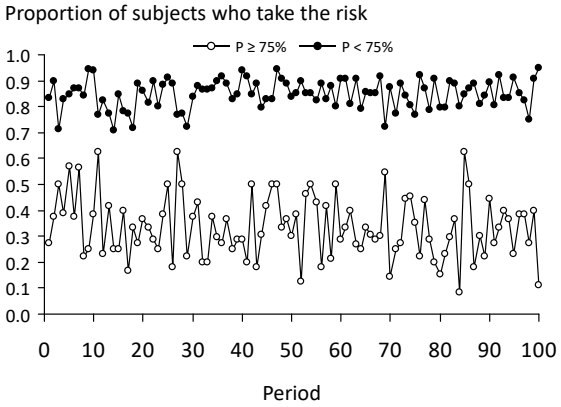
Figure 4.2 Proportion of Subjects Who Take the Risk



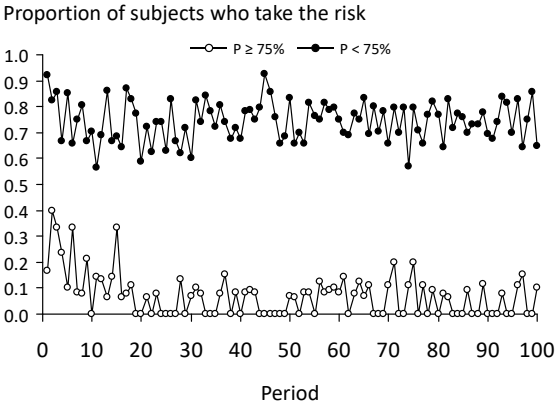
(a) Neither Recommendation Nor Probability



(b) Probability Only



(c) Recommendation Only



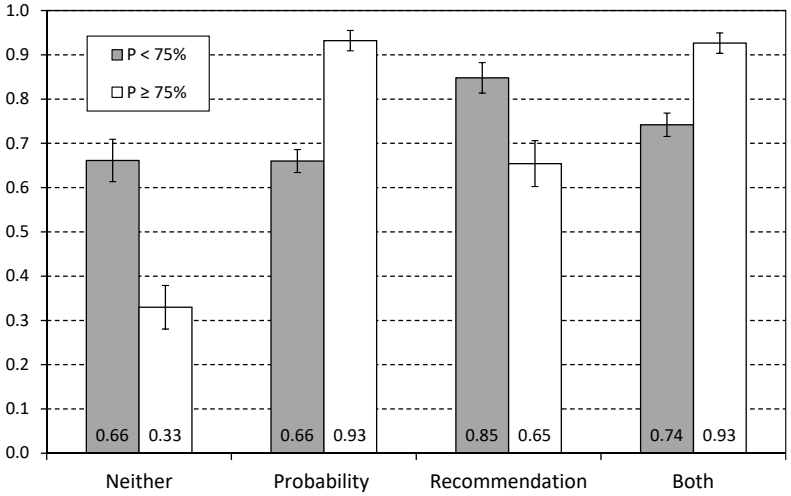
(d) Both Recommendation and Probability

also presents pairwise Wilcoxon rank-sum tests that compare forecast compliance for each pair of conditions in our study.

Absent of forecast information, subjects take the risk with a frequency of 66%. Getting a probability forecast that implies that taking the risk is optimal does not increase this frequency. Receiving a recommendation to take the risk significantly increases this frequency to 85%. Comparing the compliance in the Probability and Recommendation condition, we can conclude that the latter lead to higher compliance if the forecast implies to take the risk.

The frequency of taking the cost, absent of forecast guidance, is 33%. Both probability and recommendation forecasts significantly improve this frequency to 93% and 65%, respectively. A direct comparison of these frequencies shows that, if the forecast implies to take the cost, the

Figure 4.3 Mean Proportion of Time in Compliance With the Forecast by Treatment



Wilcoxon rank-sum tests (two-tailed, n = number of subjects)

Comparison	*p* -values
P < 75%	
H_0: Neither = Probability	0.417
H_0: Neither = Recommendation	0.015
H_0: Neither = Both	0.721
H_0: Probability = Recommendation	0.000
H_0: Probability = Both	0.040
H_0: Recommendation = Both	0.000
P ≥ 75%	
H_0: Neither = Probability	0.000
H_0: Neither = Recommendation	0.000
H_0: Neither = Both	0.000
H_0: Probability = Recommendation	0.000
H_0: Probability = Both	0.244
H_0: Recommendation = Both	0.000

Note. Error bars indicate one standard error.

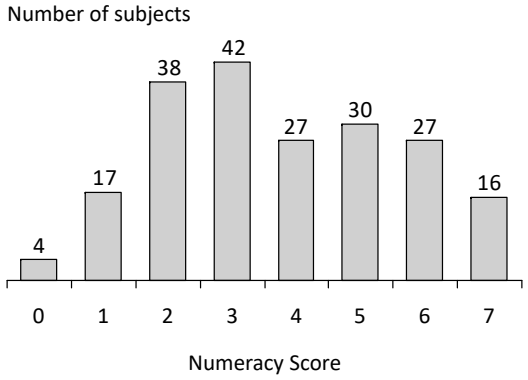
probability forecast is now more effective to increase compliance than is the recommendation forecast.

Giving subjects both probability and recommendation forecasts leads to similar compliance as probability only forecasts. We do not find a difference between Probability and Both treatment in their effectiveness to induce compliance with the siren action (take the cost). Adding verbal recommendation to the probability forecast, however, helps to improve forecast compliance (from 66% to 74%) if the forecast implies to take the risk; but still remains behind the compliance of 85% in the Recommendation condition. Overall, we can conclude that the results from Bolton and Katok (2017) replicate quite well qualitatively on MTurk.

4.3.2 The Role of Numeracy

Figure 4.4 shows the distribution of the numeracy scores in our sample. The mean numeracy score is 3.74 (SD = 1.84). A Shapiro-Wilk W test for normality indicates that we cannot reject that numeracy is normally distributed in our sample (p = 0.100), with skewness of 0.13. There was no statistically significant difference between conditions ($F(3, 197)$ = 1.78, p = 0.153).

Figure 4.4 Distribution of Numeracy Scores

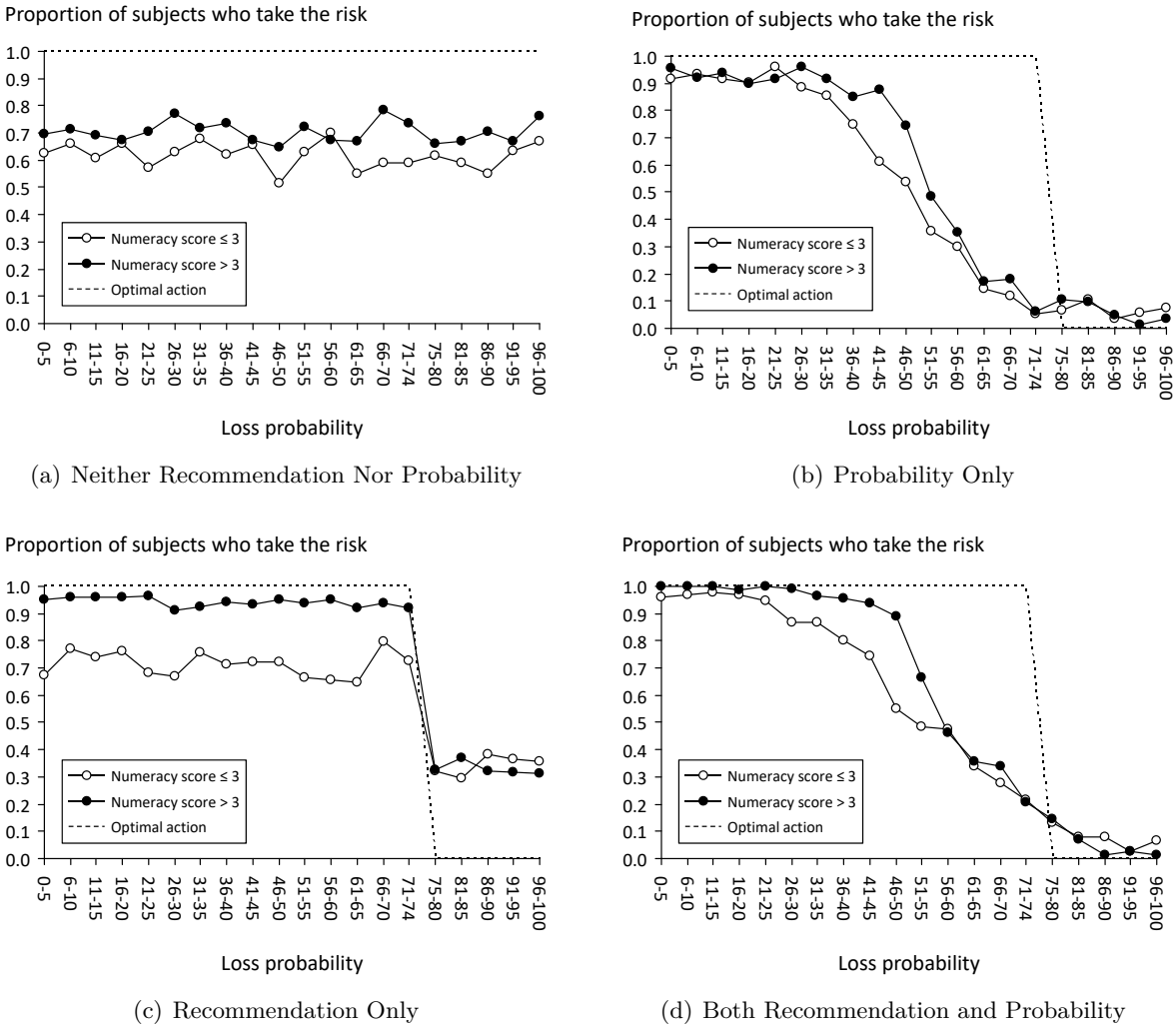


To get a sense of the role of numeracy, Figure 4.5 displays the proportion of decisions to take the risk by loss probability P (averaged over 5-percentage-point blocks). We partitioned subjects into low numerate (numeracy score 3 and below) and high numerate (numeracy score 4 and above) subjects.

Absent round-by-round forecast information, high numerate subjects seem to be somewhat more likely to take the optimal action (take the risk) than low numerate subjects (on average 70% versus 62%). If a probability forecast is provided, Figure 4.5(b), high numerate subjects are somewhat more likely to take the optimal action when the forecast implies to take the risk (69% versus 63%). When the forecast implies to take the cost there is no difference in the proportion of subjects who take the cost (93%) between high and low numerate subjects. In the Recommendation condition, Figure 4.5(c), high numerate subjects are again more likely to take the optimal action when the forecast implies to take the risk (94% versus 71%). When the forecast implies to take the cost, high numerate subjects are only slightly more likely to follow the forecast than low numerate subjects (67% versus 63%). If both probability and a verbal recommendation was provided. Figure 4.5(d), the pattern looks pretty much as the pattern in the Probability condition. However, the differences between high and low numerate subjects are more pronounced when the forecast implies to take the risk (81% versus 70%). When the forecast implies to take the cost, high numerate subjects are only slightly more likely to follow the forecast than low numerate subjects (94% versus 92%).

Overall, it appears that high numerate subjects are more likely to comply with the

Figure 4.5 Proportion of Subjects Who Take the Risk by Numeracy



(a) Neither Recommendation Nor Probability

(b) Probability Only

(c) Recommendation Only

(d) Both Recommendation and Probability

forecast than low numerate subjects. The effect of numeracy seems particularly strong in the Recommendation condition. Table 4.4 presents estimated panel logistic regression models for each condition. The dependent variable in all models is 1 when the optimal action is taken. The variable *Numeracy* takes the value of the subject's numeracy score. The estimates confirm our previous observations: There is a strong positive effect of numeracy on compliance in the Recommendation and the Both condition. Both coefficients are highly significant. In the Probability condition we find a smaller and weakly significant effect of numeracy on compliance. These effects still hold if we control for age, gender, education, and income (for correlations between variables see Supplementary Material 4.F).

Table 4.4 Effect of Numeracy on the Likelihood to Take the Optimal Action

Explanatory variable	Neither	Probability	Recommendation	Both
Numeracy	0.614**	0.090*	0.384***	0.219***
	(0.313)	(0.049)	(0.105)	(0.067)
Constant	–0.207	0.756***	0.336	0.772***
	(1.363)	(0.205)	(0.462)	(0.237)
Log-likelihood	-1752	-3275	-1933	-2126
Observations	5000	6000	4700	4400

Notes. Random-effects logistic regression. Standard errors in parentheses.
* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01, two-tailed.

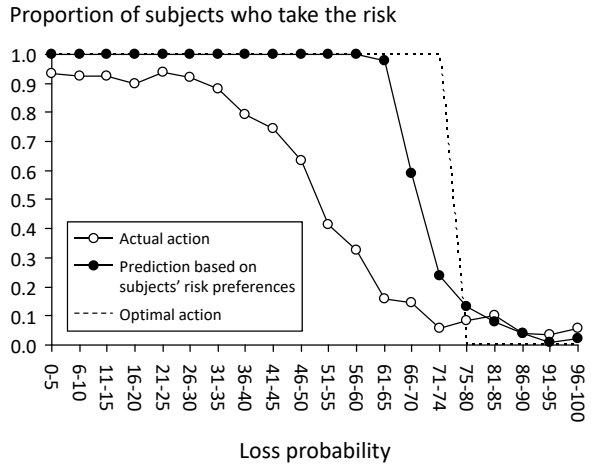
Re-estimating the regression model with the pooled data from the Probability and the Recommendation condition, adding an indicator variable for the Recommendation condition crossed with *Numeracy* we find that numeracy skills are significantly more important in the Recommendation condition than in the Probability condition ($p = 0.006$). Repeating this analysis with the pooled data from the Recommendation condition and the Both condition we find no significant difference in the effect of numeracy on the likelihood to take the optimal action ($p = 0.241$). Contrary to our initial expectation, numeracy skills matter more to forecast compliance with recommendations than with probabilities.

In the following sections we will analyze the decisions in the Probability condition and the Recommendation condition in more detail.

4.3.3 Strengths and Weaknesses of Probability Forecasts

In Figure 4.5(b), we can see that both high and low numerate subjects comply nearly 100% of the time when the probability is either extremely high or extremely low. But for moderate levels of P (between about 30% and 75%) there is substantial deviation from compliance for both high and low numerate subjects. Both high and low numerate subjects tend to take the cost well below the optimal threshold of 75%. We might hypothesize that the explanation for this pattern is driven by risk aversion. The number of boxes collected in the BRET is a good proxy for subjects' risk appetite. Assuming the constant relative risk aversion utility function $u(x) = x^r$ and given the implied levels of r based on the number of boxes

Figure 4.6 Proportion of Subjects Who Take the Risk in the Probability Condition



collected (see Crosetto and Filippin 2013, Appendix A), we can calculate the expected utility maximizing decision for each subject and each value of P. The result is depicted in Figure 4.6 (averaged over subjects and 5-percentage-point blocks) and indicates that risk aversion alone is not sufficient to explain the observed pattern. Instead, it appears that many subjects, independent of numeracy, fail to follow the optimal strategy when P gets above 50%.

Interestingly, this also happens in the condition with both, recommendation and probability. In fact, it appears that probability information effectively chases out the benefits of recommendation for the high numerate subjects. We present informal analysis of this in Figure 4.7, in which we display word clouds for the four conditions based on the answers to the questions "briefly describe how you have decided when to take the risk and when to take the cost". Figure 4.7(b and d) and Figure 4.7(c and d) indicate that the condition with both probability and recommendation is much more similar to the Probability condition than to the Recommendation condition. Observe that, for both Probability and the Both condition, subjects explain their actions in terms of probability or percent and seemingly ignore the advice when it is given.

Looking more closely at the answers in the Probability condition, we find that 47% of subjects indicate that the rule they followed was to choose to take the cost when the loss probability was above 50%, effectively ignoring the financial implications of their decisions

Figure 4.7 Word Clouds Based on the Request to Explain How the Decisions Were Made

(a) Neither Recommendation Nor Probability

(b) Probability Only

(c) Recommendation Only

(d) Both Recommendation and Probability

Note. For details on the data processing for the word clouds see Supplementary Material 4.G.

which would, if correctly folded into the decision making procedure, suggest the optimal 75% cutoff. This behavior is in line with the finding by Slovic and Lichtenstein (1968), who showed that ratings of a gamble's attractiveness were determined much more strongly by the probabilities of winning and losing than by the expected payoff. This decision heuristic has come to be known as *proportion dominance* Finucane et al. (2003) and has been replicated a number of times (for example, Goldstein and Einhorn 1987, Ordóñez and Benson 1997).

Provided with a probability forecast, subjects seem to follow a simple decision heuristic and compliance is only weakly correlated with numeracy.

4.3.4 Strengths and Weaknesses of Recommendation Forecasts

In Figure 4.5(c), we can see that recommendations work quite well, when the optimal decision is to take the risk. The compliance rates of both high and low numerate subjects in the Recommendation condition are higher than the compliance rates in the Probability condition, when the optimal decision is to take the risk (69% versus 94% and 63% versus 71% for high and low numerate subjects, respectively). However, the difference is only significant for high

but not for low numerate subjects ($p < 0.001$ and $p = 0.103$, respectively). When the forecast implies to take the cost, compliance in the Recommendation condition is significantly lower than in the Probability condition for both high and low numerate subjects ($p < 0.001$).

To further explore the behavior in the Recommendation condition we will estimate panel logistic regression models for the data in the Recommendation condition. All models use random effects to control for the panel structure of our data. The dependent variable in all models is 1 when the siren action (take the cost) is taken and 0 otherwise. The explanatory variable *Opt* is 1 when the optimal action given the forecast corresponds to the siren action and 0 otherwise. Therefore, $(1 - Opt)$ is 1 when the optimal action is to take the status quo action (take the risk). We cross the *Period* variable $(1, \ldots, 100)$ with the indicator variables *Opt* and $(1 - Opt)$ to capture learning effects for taking the siren action and the status quo action, respectively.

Following Bolton and Katok (2017), we define the variable *Errors* to be the total number of previous recommendations that turned out to be incorrect ex post – either the recommendation was to take the risk but the loss did occur or the recommendation was to take the cost but the loss did not occur. So the *Errors* variable is simply the cumulated number of forecast errors observed during the game. We cross the *Errors* variable with the variable *Opt* and with $(1 - Opt)$ to track how subjects react to false alarms. Table 5 summarizes the estimates of the models that we fit for each condition separately.

Model (1) takes a first look at the dynamics of the decision in the Recommendation condition and tracks how subjects are influenced by forecast errors. Here, as in all models, the *Opt* coefficient is positive, indicating that the recommendation improves decision making. However, in Models (3) and (4), the coefficients for Opt become non-significant. We will discuss this below. The coefficient for $Opt \times Period$ is positive and significant, and for $(1 - Opt) \times Period$, it is negative and significant in all models. So on aggregate and with experience subjects learn to take the cost more often when it is optimal to do so and less often when it is not. The coefficient for $Opt \times Errors$ is negative and significant and the coefficient for $(1 - Opt) \times Errors$ is positive and significant, indicating that subjects are less likely to trust the forecast the more errors they observe. The more errors they observe, they take the cost less often when it

Table 4.5 Effect of Numeracy and Forecast Errors in the Recommendation Condition

Explanatory variable	(1)	(2)	(3)	(4)
Opt	3.246***	3.233***	0.128	0.487
	(0.205)	(0.213)	(0.314)	(0.466)
Opt × *Period*	0.036***	0.037***	0.034***	0.034***
	(0.010)	(0.011)	(0.011)	(0.011)
(1 − *Opt*) × *Period*	−0.024***	−0.021**	−0.023***	−0.023***
	(0.008)	(0.008)	(0.008)	(0.008)
Opt × *Errors*	−0.105***	−0.222***	−0.095***	−0.134***
	(0.031)	(0.037)	(0.033)	(0.039)
(1 − *Opt*) × *Errors*	0.072***	0.091***	0.068***	0.054*
	(0.025)	(0.027)	(0.026)	(0.028)
Opt × *Errors* × *Numeracy*		0.030***		0.010**
		(0.004)		(0.005)
(1 − *Opt*) × *Errors* × *Numeracy*		−0.009***		0.005
		(0.003)		(0.004)
Opt × *Numeracy*			0.166	0.017
			(0.183)	(0.199)
(1 − *Opt*) × *Numeracy*			−0.654***	−0.724***
			(0.181)	(0.192)
Constant	−2.611***	−2.684***	−0.178	0.042
	(0.347)	(0.367)	(0.808)	(0.837)
Log-likelihood	-1573	-1511	-1490	-1487
Observations		4700		

Notes. Random-effects logistic regression. Standard errors in parentheses.
* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01, two-tailed.

would be optimal to do so and more often when it is not optimal. So overall, while subjects make better decisions with increasing experience, their compliance rate decreases with the number of false alarms, a behavioral regularity known as the cry wolf effect (Bliss et al. 1995, Meyer and Bitan 2002, Bolton and Katok 2017).

Next, we will analyze how numeracy influences decision making in the Recommendation condition. In Model (2) we add the variables $Opt \times Errors \times Numeracy$ and $(1 - Opt) \times Errors \times Numeracy$. The coefficient to $Opt \times Errors \times Numeracy$ is positive and significant and the coefficient to $(1 - Opt) \times Errors \times Numeracy$ is negative and significant, indicating that subjects with high numeracy skills suffer less from a decrease in compliance with future forecasts as response to previous forecast errors than subjects with low numeracy skills.

In Model (3) we add the variables $Opt \times Numeracy$ and $(1 - Opt) \times Numeracy$ to the set of explanatory variables in Model (1). The coefficient to $Opt \times Numeracy$ is positive but not significant and the coefficient to $(1 - Opt) \times Numeracy$ is negative and significant. This indicates that when the optimal action is to take the cost, numeracy skills have a positive but not significant effect on the likelihood that subjects chose the siren action. However, when the optimal action is to take the risk numeracy significantly improves performance by lowering the likelihood to take the non-optimal siren action. In Model (3) we also see that the coefficient to Opt becomes non-significant, indicating that recommendations do not improve decision making per se, but require a certain level of numeracy skills. Thus, innumerate subjects almost completely ignore the recommendation.

So far we found that numeracy skills can improve decisions in two ways: numeracy skills increase subject's tolerance toward false alarms (Model (2)) and they improve subject's compliance with status quo action (Model (3)). In the Model (4) we test both ways in one model. The coefficients to $Opt \times Errors \times Numeracy$ and to $(1 - Opt) \times Numeracy$ remain significant, indicating that high numerate subjects suffer less from a decrease in compliance with future siren forecasts as response to previous forecast errors and are in general more likely to take the status quo action when it is optimal to do so than subjects with low numeracy skills.

4.3.5 A Hybrid Forecasting Scheme

Our results show that both probabilities and recommendations have their own behavioral strengths and weaknesses. So the question is, if we can combine probabilities and recommendations in a way to offset their major behavioral shortcomings but keep their benefits. We designed a hybrid forecasting system that aimed to include the benefits from both types of information. The forecast guidance always provides a recommendation, but only additionally provides the probability of the loss event when the recommendation is to take the cost. In other words, the recommendation for default action included recommendation only, while the recommendation for the siren action, also added probability, by the way of an explanation. This idea was based on our previous observations that recommendations are more successful at inducing the default action (take the risk) and providing probabilities is more successful at inducing the siren action (take the cost), we expect the hybrid forecast to capture most of the benefits.

We tested the hybrid forecasting scheme in a follow-up experiment with 50 subjects recruited on MTurk. The experimental procedure was identical to our main experiment. The instructions for the follow-up experiment are identical to the Recommendation condition except for one additional sentence (in italics): "Each round you will be given advice of whether to take the cost or take the risk. *If in a given round the advice is to take the cost, you will be also given the loss probability P that pertains to that round.* The advice has been determined in a way that, on average, if you follow the advice you will earn the most money possible. You are not required to follow the advice." Thus, the information provided to our subjects is identical to the Recommendation condition, when P is lower than 75% and identical to the Both condition, when P is greater than or equal to 75%.

To analyze the success of the hybrid forecast scheme we will estimate two panel logistic regression models one with the pooled data from the Recommendation condition and the follow-up experiment and one with the pooled data from the Both condition and the follow-up experiment. Both models use random effects to control for the panel structure of the data. The estimates of the models are summarized in Table 4.6. The dependent variable is 1 when the siren action (take the cost) is taken and 0 otherwise. The explanatory variables *Opt* and

Table 4.6 Comparison of the Hybrid Forecast With Recommendation and Both

Explanatory variable	(1)	(2)
Opt	3.552***	4.655***
	(0.114)	(0.168)
Opt × *Hybrid*	1.534***	−0.905**
	(0.452)	(0.392)
(1 − *Opt*) × *Hybrid*	−0.528	−1.777***
	(0.448)	(0.364)
Constant	−2.676***	−1.369***
	(0.320)	(0.259)
Log-likelihood	-2872	-3157
Observations	9700	9400

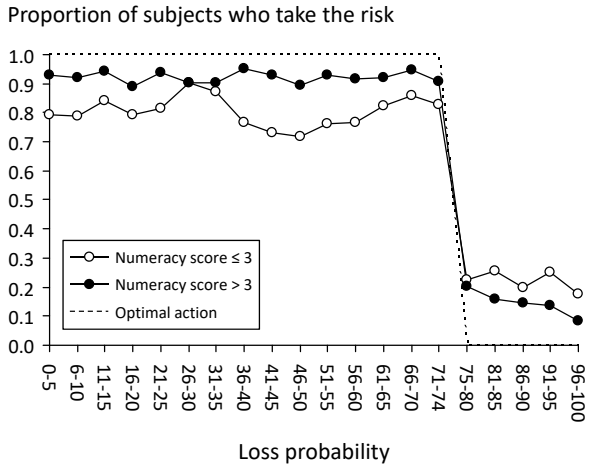
Notes. Random-effects logistic regression. Standard errors in parentheses.
* p-value < 0.10, ** p-value < 0.05, *** p-value < 0.01, two-tailed.

(1 − *Opt*) are defined as above. We add an indicator variable *Hybrid* that is 1 for the data from the follow-up survey and 0 otherwise.

In Model (1) we can see that the hybrid forecast is at least as good as providing recommendations in inducing compliance with the status quo action, when it is optimal. However, compared to recommendations the hybrid forecast is significantly better in inducing the siren action, when it is optimal. In Model (2) we can see that, compared to providing both recommendations and probabilities the hybrid forecast is significantly better in inducing compliance with the status quo action, when it is optimal. However, the hybrid forecast could not unlock the full potential of providing both, when it is optimal to take the siren action. Overall forecast compliance significantly improves (Wilcoxon rank-sum test, $p < 0.01$).

Figure 4.8 displays the proportion of decisions to take the risk separated by the range of numeracy scores. From the figure, we observe that the hybrid forecast guidance moves high numerate subjects largely in the right direction but is less effective with low numerate subjects. This indicates that improving trust in forecasting for low numerates will require a new approach.

Figure 4.8 Proportion of Subjects Who Take the Risk in the Hybrid Condition Separated by the Range of Numeracy Scores



4.4 Discussion and Managerial Implications

We analyzed how forecast guidance interacts with the quality of decisions made by forecast users. We found that numeracy influences trust in the forecast as measured by compliance levels, although this happens in different ways than initially anticipated. While low numerate subjects exhibit substantially less compliance with recommendation forecasts than do high numerates, there is only a modest positive relationship between numeracy and trust with probability forecasts. Both high and low numerate subjects comply nearly 100% of the time when the probability is either extremely high or extremely low. But for moderate levels of P there is substantial deviations from compliance for both high and low numerate subjects. The failure in compliance with probability forecasts can be attributed to effectively ignoring the financial implications and paying too much attention to probabilities. Most subjects, independent of numeracy, seem to follow a 50% rule of thumb, leading them to take the cost when the loss probability exceeds 50%. We also observed that probability forecasts overshadow recommendations, leading to substantial deviation from compliance for moderate levels of P in the Both condition.

The results of our study must be considered within its limitations. First, the subjects were recruited on MTurk and may not be representative for the population. However, the socio-

demographic characteristics of our sample approximately resemble those of the population reported in the 2010 United States census. Second, the experiment used a context-free cost-loss game that differs from reality in several important ways. While the design of study allows us to control for context-dependent biases and therefore increases internal validity, it clearly diminishes external validity. Analyzing how our findings translate to specific contexts and the real world we leave to future research.

The limitations notwithstanding, our study provides a clearer understanding of how forecast guidance interacts with the quality of decisions made by forecast users. We also shed light on the role of numeracy in this interaction. Our findings on behavior may lead to new ideas on how to design forecast guidance and stimulate future research. One promising approach to improve forecast compliance could be the use of graphical displays of numerical information. Graphs summarize and present numerical information in an alternative, but not less precise way. Especially, so-called pictographs seem to be a promising tool for communicating risk to persons with higher and lower numeracy (Galesic et al. 2009, Hess et al. 2011). Pictographs represented by icons showing the frequency of a loss event can be used to illustrate magnitude and convey the notion of randomness (Nelson et al. 2008). Analyzing the effectiveness of such pictographs or other visual displays in a cost-loss game offers interesting opportunities that we leave to future research.

Supplementary Materials

4.A On-Screen Instructions

These were the on-screen instructions that were shown to the subjects upon accessing the experiment.[2] In this condition, the subjects received both probability and recommendation. When subjects were only given probability, the paragraphs detailing the recommendation were removed (paragraph with dotted border). When the subjects were only given recommendation, the details of the probability were removed (paragraph with solid border). When the subjects were given neither probability nor recommendation, they were only provided with the loss probability across all rounds. The payment information were identical across all conditions.

Payment Information

We want you to give us your best and honest answers to the questions that follow. We value your participation, and offer an incentive on top of the amount you will be paid for this HIT (if you answer the comprehension questions correctly). We will pay it out as a bonus in Mechanical Turk.

During this survey you will play **100 rounds** of a game from which you can earn money. Your profits in this game are expressed in tokens. At the end of the survey the sum of your profits will be converted into U.S. dollars at a rate of $1 per 3,000 tokens; the more tokens you earn, the more money you will make.

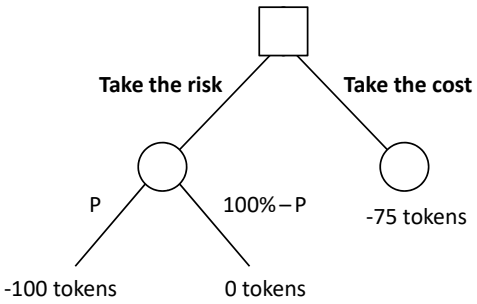
After they game you will answer some questions from which you can earn additional money.

[2] Our instructions are similar to the instructions in Bolton and Katok (2017).

We are academics at a business school, and we always pay as promised. We believe that compensating you is important and also fair, and we hope that you will participate in our future studies.

Instructions

At the beginning of each round of the game, you are given a credit of 150 tokens. You must then decide whether to **take the risk** or **take the cost**. If your decision is to take the risk, there is some probability P of incurring a **loss** of 100 tokens. If your decision is to take the **cost**, you incur a cost of 75 tokens for certain.



Your profit depends on your decision and on whether the loss occurs.

If you take the risk, then either:

Your profit = 150 – 100 = 50 tokens if the loss occurs

or

Your profit = 150 tokens if the loss does not occur.

If you take the cost, then

Your profit = 150 – 75 = 75 tokens regardless of whether the loss occurs or not.

You will play 100 rounds of the game. The probability of loss (P) varies from round-to-round. To determine whether the loss actually occurs in a round, the computer will generate a random number between 0% and 100%, with each number in this range equally likely. If the random number is below or equal to the loss probability for that round, the loss occurs; if it is above the loss probability, the loss does not occur.

For example, suppose the loss probability P for the round is 60%. If the random number comes out to be 65%, the loss does not occur. If the random number comes out to be 40%, the loss occurs.

Information to help you decide

While the loss probability P varies from round-to-round, the average loss probability across all rounds is 50%.

Each round you will be given the loss probability P that pertains to that round.

Each round you will be given advice of whether to take the cost or take the risk. The advice has been determined in a way that, on average, if you follow the advice you will earn the most money possible. You are not required to follow the advice.

Note that the advice does not guarantee that you will make the most money possible in any given round. It is possible that when the advice is take the risk, the loss does occur. It is also possible that the advice is to take the cost, and the loss does not occur.

4.B Comprehension Questions

In this section, we provide the questions and answers from the quiz that our subjects had to pass before they could continue with the main part of the experiment. Out of the 337 subjects who started the quiz, 110 subjects (32.6%) could not answer all questions correctly on the first attempt and were therefore screened out. Out of the 110 subjects who failed the quiz, four subjects had only one correct answer, eleven subjects had two correct answers, 22 subjects had three correct answers, and 73 subjects had four correct answers. We do not find any significant differences in the failure rates across treatments ($\chi^2(3) = 1.2872$, $p = 0.732$).

To make sure that the instructions are clear, please answer the following comprehension questions.

(1) The loss probability P varies from round-to-round.

- ⊙ True
- ○ False

(2) What is the average loss probability across all rounds?

- ○ 40%
- ⊙ 50%
- ○ 60%
- ○ 65%

(3) If you take the risk, then your profit...

- ⊙ ... depends of whether the loss occurs.
- ○ ... is independent of whether the loss occurs.

(4) If you take the cost, then your profit...

- ○ ... depends of whether the loss occurs.
- ⊙ ... is independent of whether the loss occurs.

(5) How many rounds will you play?

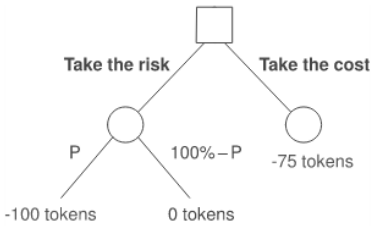
- ○ 25
- ○ 50
- ○ 75
- ⊙ 100

4.C Screen Shots

Decision Screen

Round 1 of 100

You receive a credit of 150 tokens.



The loss probability P for this round is **43%**.

The advice for this round is to **take the risk**.

Please make your choice:

Take the risk | Take the cost

Continue

Result Screen

Round 1 of 100

The loss probability was 43%.

The advice was to take the risk.

You chose to **Take the risk**.

The loss occured.

Your profit is **50 tokens**.

Continue

4.D The Bomb Risk Elicitation Task

In this section, we provide the instructions and results of the BRET (Crosetto and Filippin 2013), that subjects completed after the main part of the experiment.

Instructions

Below you see a field composed of 100 numbered boxes. Exactly one of these 100 boxes contains a bomb. You do not know the bomb's location. You only know that it is equally likely to be in any of the 100 boxes.
Your task is to choose how many boxes to collect. Boxes will be collected in numerical order. So you will be asked to choose a number between 1 and 100.

After you have confirmed your decision the computer will randomly determine the number of the box containing the bomb.

- If you happen to have collected the box where the bomb is located – i.e. if your chosen number is greater than or equal to the drawn number – you will earn zero.
- If the bomb is located in a box that you did not collect – i.e. if your chosen number is smaller than the drawn number – you will earn 10 cents for each collected box.

On the next screen you will be asked to indicate how many boxes you would like to collect. You confirm your choice by clicking 'Continue'.

To make sure that the instructions are clear, please answer the following short quiz.

Suppose that the bomb is located in box 25. How much will you earn, if you collect ...

(a) ... 21 boxes _____

(b) ... 38 boxes _____

(c) ... 62 boxes _____

(d) ... 79 boxes _____

Suppose that the bomb is located in box 75. How much will you earn, if you collect ...

(a) ... 21 boxes _____

(b) ... 38 boxes _____

(c) ... 62 boxes _____

(d) ... 79 boxes _____

The location of the bomb depends on how many boxes you decide to collect.

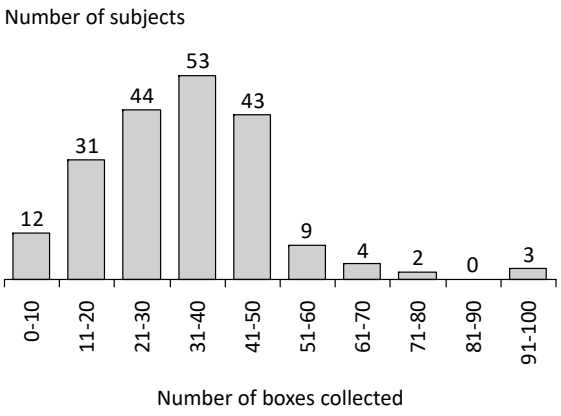
○ Yes

○ No

Results

Figure 4.9 shows the distribution of the number of collected boxes in the BRET. The mean number of boxes collected is 35.0 (SD = 16.1).

Figure 4.9 Distribution of Risk Appetite



4.E Post-Experimental Questionnaire

(1) What is your age?

_____

(2) What is your gender?

- ○ Male
- ○ Female

(3) What is your primary language (i.e., the one you speak most of the time)?

_____

(4) What is the highest level of education you have completed?

- ○ Less than high school degree
- ○ High school graduate (high school diploma or equivalent including GED)
- ○ Vocational/technical school
- ○ Some college but no degree
- ○ Bachelor's degree
- ○ Master's degree
- ○ Doctoral degree (PhD)
- ○ Advanced professional degree (JD, MD, etc.)

(5) Which category best describes your major?[3]

- ○ Arts and Humanities (Arts, Language, Literature, History, Philosophy, etc.)
- ○ Business (Accounting, Finance, Marketing, etc.)
- ○ Engineering and Computer Science (Civil, Electrical, Mechanical, etc.)
- ○ Health and Medicine (Medicine, Nursing, Public Health, etc.)
- ○ Natural Sciences and Mathematics (Biology, Chemistry, Maths, Physics, etc.)
- ○ Social Sciences (Communication, Economics, Politics, Psychology, Sociology, etc.)
- ○ Other (please specify): _____

(6) How would you best describe your current employment status?

- ○ Working (paid employee)
- ○ Working (self-employed)
- ○ Not working (temporary layoff from a job)
- ○ Not working (looking for work)
- ○ Not working (retired)
- ○ Not working (disabled)
- ○ Not working (other): _____

(7) Please indicate the category that best describes your own annual income from all sources before taxes.

- ○ $10,000 and under
- ○ $10,001 to $20,000
- ○ $20,001 to $30,000
- ○ $30,001 to $40,000
- ○ $40,001 to $50,000
- ○ $50,001 to $60,000
- ○ $60,001 to $70,000
- ○ $70,001 to $80,000
- ○ $80,001 to $90,000
- ○ $90,001 to $100,000
- ○ $100,001 to $150,000
- ○ over $150,000

[3] Question 5 was displayed only if the answer to Question 4 was Bachelor's degree or higher.

4.F Correlations Between Variables

We collected data on subjects' numeracy skills and risk preferences; and asked for additional demographic information, including highest level of education completed and own annual income from all sources before taxes. Responses to these questions were categorical. For *Education*, 0 equals "less than high school degree," 1 equals "high school graduate (high school diploma or equivalent including GED)," 2 equals "vocational/technical school," 3 equals "some college but no degree," 4 equals "bachelor's degree," 5 equals "master's degree," 6 equals "doctoral degree (PhD)," and 7 equals "advanced professional degree (JD, MD, etc.)." For *Income*, 0 equals "$10,000 and under," 1 equals "$10,001 to $20,000," 2 equals "$20,001 to $30,000," ... 9 equals "$90,001 to $100,000," 10 equals "$100,001 to $150,000," and 11 equals "over $150,000."

Correlations between variables are provided in Table 4.7. Numeracy scores were higher for men as well as for more educated people in our sample. Incomes were higher for more educated people.

Table 4.7 Correlations Between Variables

Variable	1	2	3	4	5	6	Mean (SD)
1. Numeracy	—						3.74 (1.84)
2. Risk appetite[a]	−.02	—					34.95 (16.07)
3. Age	.01	−.08	—				34.98 (10.90)
4. Education	**.31**	−.02	.12	—			3.44 (1.24)
5. Gender[b]	**−.22**	.03	.06	.02	—		0.55 (0.50)
6. Income	.07	.04	.09	**.24**	−.13	—	3.09 (2.53)

Notes. Coefficients printed in bold are significant ($p < 0.01$), all others are not significant at any level.
[a]number of boxes collected in Crosetto and Filippin's (2013) risk elicitation task.
[b]0 = male, 1 = female.

4.G Data Processing for Word Clouds

The first step of our data processing was to remove punctuation, extra white spaces, and special characters; and convert the text to lower case. After this initial data cleaning, we substituted words and phrases to correct for spelling mistakes and to group words with similar meaning. The substitutions are shown in Table 4.8.

Table 4.8 Substitutions

Substitute	Substituendum	Substitute	Substituendum	Substitute	Substituendum
30	30s	gain	gained, gaining	realize	realized
40	40s	gamble	gambled	reason	reasoned, reasoning
60	60ish	general	generally	regret	regretted
85	high85	generate	generated, generating	reliable	proved, turned
90	90s	guess	guessing, predict	report	reports
a lot	alot	happen	happened, happening	result	results
action	actions	help	helped	risk	riash, risks
actual	actually	hope	hopes, hoping	risk would	riskwould
advice	advices, advised, adviser, advisor, reccomendations, recomendation, recommendation, recommendations, recommended, suggested, suggestion, suggestions	intuition	gut, gut feeling, instinct, intuitive	round	rounds
agree	agreed	keep	keeping	run	ran
alternate	alternated	know	knew, knowing	safe	safer, save
always	every round, every time, everytime	lead	led	scare	scared
always with	every timewith	learn	learned	search	searching
answer	answers	listen	listened	seem	seemed, seems
attempt	attempts	load	loads	select	selected
average	averages	look	looked	show	showed, shown
balance	balanced	lose	loosing, loses, losing, lost	significant	significantly
based	bases	loss	losses	situation	situations
become	becoming	loss and	lossand	sometimes	occasionally, rarely
begin	began	luck	lucky	start	started
benefit	beneficial, benefiting	make	made, making	statistical	statistically
cancel	cancelled	maximize	maximizing	stay	stayed
chance	chances, shot	mean	means	stick	stuck
change	changed, changing	mostly	mainly, most of the time, normally, ordinarily, typically, usually	straight forward though	straightforward-though
choice	choices	non loss	nonlosses	streak	streaks
choose	choosing, chose	notice	noticed, noticing	successive	successively
click	clicked	number	numbers	suspicious	suspiciously
collect	collecting	occasion	occasions	switch	switched

Finally, we removed common stop words like “a”, “and”, “the”, etc.. The list of stop words is shown in Table 4.9. Table 4.10 shows the most common words that subjects used

Table 4.9 Stopwords

a	between	gave	know	on	than	unless	whereupon
able	bro	get	known	only	that	until	wherever
about	but	getting	last	or	thats	up	whether
across	by	give	latter	other	the	upon	which
after	came	given	least	out	their	us	while
again	certain	go	less	over	them	v	whither
against	come	going	let	own	then	very	who
all	could	got	like	per	there	via	whoever
almost	did	had	look	probably	thereafter	vs	whole
along	didnt	have	me	provided	therefor	want	whom
also	do	he	mean	provides	theres	wanted	whose
although	does	help	might	rather	they	wants	why
am	doesnt	here	more	regardless	think	was	will
an	doing	how	most	right	third	wasnt	with
and	done	i	much	said	this	way	within
another	dont	id	must	same	those	well	without
any	down	if	my	saw	though	went	would
anyways	each	ill	myself	second	three	were	woulda
around	either	im	near	see	through	what	wouldnt
as	even	in	never	seem	throughout	whatever	wouldve
at	except	instead	next	should	to	when	x
be	far	into	no	so	too	whence	yet
because	few	is	non	some	took	whenever	you
been	first	isnt	not	somehow	toward	where	your
being	for	it	now	still	try	whereafter	yours
below	forth	its	of	sure	twice	whereas	yourself
best	four	itself	off	take	two	whereby	yourselves
better	from	just	often	taken	under	wherein	

to describe how they have decided when to take the risk and when to take the cost. The table also shows the proportion of subjects per treatment who used the corresponding word to describe their strategy.

Table 4.10 Common Words Used by Subjects to Describe Their Strategy

Word	Neither	Probability	Recommendation	Both
risk	74.0	80.0	42.6	54.5
percent	18.0	80.0	6.4	68.2
cost	42.0	43.3	29.8	40.9
50	20.0	48.3	10.6	38.6
probability	8.0	53.3	6.4	36.4
loss	30.0	28.3	14.9	34.1
advice	0.0	0.0	63.8	29.5
mostly	18.0	15.0	23.4	22.7
always	30.0	10.0	21.3	4.5
taking	28.0	6.7	19.1	9.1
time	28.0	6.7	21.3	6.8
decide	18.0	18.3	14.9	6.8
choose	18.0	10.0	8.5	15.9
feel	4.0	11.7	14.9	13.6
occur	18.0	10.0	4.3	9.1
follow	0.0	0.0	31.9	11.4
sometimes	10.0	6.7	4.3	13.6
chance	8.0	6.7	4.3	13.6
lose	12.0	6.7	6.4	4.5
intuition	4.0	8.3	6.4	11.4
random	16.0	5.0	6.4	0.0
anything	2.0	15.0	2.1	6.8
profit	18.0	1.7	4.3	2.3
75	14.0	5.0	2.1	4.5
average	14.0	1.7	10.6	0.0
above	0.0	11.7	0.0	13.6
round	10.0	3.3	8.5	4.5
win	18.0	0.0	2.1	2.3
token	10.0	1.7	8.5	2.3
lower	0.0	11.7	0.0	6.8
60	0.0	11.7	0.0	6.8
based	4.0	1.7	4.3	11.4
guaranteed	16.0	0.0	2.1	0.0
happen	12.0	0.0	0.0	0.0

Chapter 5

Conclusion

In this dissertation, we applied behavioral economic engineering approaches to improve decision making. In the three main chapters of the dissertation, we analyzed how different performance metrics, supply contracts, and forms of forecast guidance affect human decision making and how we can improve performance. This chapter summarizes the key results of the three main chapters and provides directions for future research.

In **Chapter 2**, we analyzed how performance metrics used in inventory management affect human decision making. We considered two equivalent inventory metrics, days of supply and inventory turn rate. While the relationship between days of supply and inventory value is linear, inventory turn rate and inventory value have a reciprocal (non-linear) relationship. Human decision makers use differences in these metrics to estimate inventory reduction. As a result inventory reductions that are evaluated based on inventory turn rate are over-valued. This misperception can be avoided by using days of supply instead of inventory turn rate. In our first study, we showed that using the days of supply metric improves performance in investment decisions compared to using the inventory turn rate metric and that this effect persists with experience. In our second study, we showed that this misperception can cause subjects to work harder at reducing inventory. In a real effort task, subjects invested 28% more effort under the inventory turn rate metric than under the days of supply metric. In our third study, individuals chose higher order cost under the inventory turn rate metric than under the days of supply metric. We also found that subjects with high cognitive reflection more frequently decide optimally than those with low cognitive reflection. Thus, if the behavioral

superior metric cannot be used, decision making can still be improved by activating System 2 thinking of the decision makers. This can be supported, for instance, by reducing the emotional and cognitive load, by avoiding time pressure, and by avoiding multi-tasking during decision making. Overall, however, our findings suggest that debiasing the decision maker (activating System 2 thinking) is less beneficial than debiasing the environment (choosing the right metric).

In **Chapter 3**, we analyzed human decision making under service level and under wholesale price contracts. We showed that service level contracts can be parameterized, such that they have steep expected profit functions, relative to other commonly studied contracts such as the wholesale price contract. We argued that this property increases the salience of the actual costs and induces a debiasing effect. As a result, under the steep service level contract, the average order quantity was 66% closer to optimality and standard deviation of order quantities was 42% lower than under a mathematically comparable wholesale contract. In our experiment, the efficiency that human subjects achieved under a service level contract was almost 10% higher than the achieved efficiency under a wholesale price contract. Efficiency under the wholesale price contract was even lower than that of the mean demand heuristic (ordering the expected demand in every period), albeit not significantly. Thus, ignoring underage and overage costs and ordering mean demand in each period would result in a similar efficiency to what subjects achieved in the lab under a wholesale price contract. These results highlight that it is important to consider aspects of human behavior when designing supply contracts.

In **Chapter 4**, we analyzed compliance rates (trust) for forecast guidance provided as probabilities as well as recommendations in a simple take-the-risk or take-the-cost decision game. We found that high numerate subjects are more likely to comply with the forecast than low numerate subjects. Nevertheless the observed behavior in our study was contrary to our initial expectations in important ways. First, low numerate subjects exhibit substantially less compliance with recommendation forecasts than do high numerates. Second, there is only a modest positive relationship between subjects' numeracy skills and trust in probability forecasts. Both high and low numerate subjects complied nearly 100% of the time when the

probability was either extremely high or extremely low. But for moderate levels of the loss probability (between about 25% and 75%) there was substantial deviation from compliance for both high and low numerate subjects. We found that, when given probability information, subjects followed a 50% rule of thumb (taking the cost when the loss probability was above 50%), leading to systematic biases. Subjects seem to ignore financial implications (compare Chapter 3) and pay too much attention to probabilities, which can be seen as a *proxy attribute* (Chapter 2). Our results show that both probabilities and recommendations have their own behavioral strengths and weaknesses. We designed a hybrid forecasting system that blends probabilities and recommendations in a way to offset the major behavioral shortcomings. We observed that the new design moves high numerates largely in the right direction but is less effective with low numerates.

The field of behavioral operations explores how individuals make decisions in operations contexts and how those decisions compare to normative predictions of analytical models. We have shown that human decision makers often do not give financial implications full consideration or even ignore them, especially in situation in which the economic consequences of deviating from normative predictions are less severe. We have also shown how behavioral economic engineering can help to improve decision making, given a better understanding of behavioral regularities. Investigating how existing behavioral theory translates to the operations domain and identifying behavioral regularities is important part of future research in behavioral operations. However, research in behavioral operations should go one step further and also aim to design mechanisms that take behavioral aspects into account in order to improve human decision making in operations contexts. Controlled laboratory experiments are a great starting point to design and test mechanisms. The next step would be to test the proposed designs in the real world, because successful mechanisms in the lab can fail in the field. Therefore, conducting field experiments to further test the robustness of mechanisms is an important step for future research in the field of behavioral operations.

References

Alan Y, Gao GP, Gaur V (2014) Does inventory productivity predict future stock returns? A retailing industry perspective. *Management Science* 60(10):2416–2434.

Arellano C, DiLeonardo A, Felix I (2017) Using people analytics to drive business performance. A case study. *McKinsey Quarterly* July 2017. Retrieved from http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/using-people-analytics-to-drive-business-performance-a-case-study?cid=eml-web.

Ariely D (2010) *Predictably irrational. The hidden forces that shape our decisions* (New York, NY: Harper), 1 edition.

Arkes HR (1991) Costs and benefits of judgment errors. Implications for debiasing. *Psychological Bulletin* 110(3):486–498.

Arrow KJ, Harris T, Marschak J (1951) Optimal inventory policy. *Econometrica* 19(3):250–272.

Bagnoli M, Bergstrom T (2005) Log-concave probability and its applications. *Economic Theory* 26(2):445–469.

Batt RJ, Terwiesch C (2015) Waiting patiently. An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

Bearden JN, Murphy RO, Rapoport A (2008) Decision biases in revenue management. Some behavioral evidence. *Manufacturing & Service Operations Management* 10(4):625–636.

Becker-Peth M, Katok E, Thonemann UW (2013) Designing buyback contracts for irrational but predictable newsvendors. *Management Science* 59(8):1800–1816.

Becker-Peth M, Thonemann UW (2016) Reference points in revenue sharing contracts. How to design optimal supply chain contracts. *European Journal of Operational Research* 249(3):1033–1049.

Bell DE (1985) Disappointment in decision making under uncertainty. *Operations Research* 33(1):1–27.

Bendoly E (2013) Real-time feedback and booking behavior in the hospitality industry. Moderating the balance between imperfect judgment and imperfect prescription. *Journal of Operations Management* 31(1-2):62–71.

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

Bendoly E, Swink M (2007) Moderating effects of information access on project management behavior, performance and perceptions. *Journal of Operations Management* 25(3):604–622.

Benzion U, Cohen Y, Peled R, Shavit T (2008) Decision-making and the newsvendor problem. An experimental study. *Journal of the Operational Research Society* 59(9):1281–1287.

Benzion U, Cohen Y, Shavit T (2010) The newsvendor problem with unknown distribution. *Journal of the Operational Research Society* 61(6):1022–1031.

Beyth-Marom R (1982) How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting* 1(3):257–269.

Bilham EG (1922) A problem in economics. *Nature* 109(2733):341–342.

Bliss JP, Gilson RD, Deaton JE (1995) Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics* 38(11):2300–2312.

Bolton G, Stangl T, Thonemann UW (2017) Decision making under service level contracts. The role of cost saliency. Working paper, University of Cologne, Germany.

Bolton GE, Katok E (2008) Learning by doing in the newsvendor problems. A laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management* 10(3):519–538.

Bolton GE, Katok E (2017) Cry wolf or equivocate? Credible forecast guidance in a cost-loss game. *Management Science* Published online in articles in advance: January 5, 2017. https://doi.org/10.1287/mnsc.2016.2645.

Bolton GE, Ockenfels A (2012) Behavioral economic engineering. *Journal of Economic Psychology* 33(3):665–676.

Bolton GE, Ockenfels A, Thonemann UW (2012) Managers and students as newsvendors. *Management Science* 58(12):2225–2233.

Bostian AA, Holt CA, Smith AM (2008) Newsvendor "pull-to-center" effect. Adaptive learning in a laboratory experiment. *Manufacturing & Service Operations Management* 10(4):590–608.

Bryant GD, Norman GR (1980) Expressions of probability. Words and numbers. *New England Journal of Medicine* 302(7):411.

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's Mechanical Turk. A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1):3–5.

Burgman MA (2016) *Trusting judgements. How to get the best out of experts* (Cambridge, UK: Cambridge University Press), 1 edition.

Cachon GP (2003) Supply chain coordination with contracts. de Kok AG, Graves SC, eds., *Supply chain management. Design, coordination and operation*, volume 11 of *Handbooks in operations research and management science*, 227–339 (Amsterdam, The Netherlands: Elsevier).

Cachon GP, Lariviere MA (2005) Supply chain coordination with revenue-sharing contracts. Strengths and limitations. *Management Science* 51(1):30–44.

Cachon GP, Terwiesch C (2013) *Matching supply with demand. An introduction to operations management* (New York, NY: McGraw-Hill), 3 edition.

Camerer CF, Ho TH (1999) Experience-weighted attraction learning in normal form games. *Econometrica* 67(4):827–874.

Caplice C, Sheffi Y (1994) A review and evaluation of logistics metrics. *International Journal of Logistics Management* 5(2):11–28.

Carpenter J, Matthews PH, Schirm J (2010) Tournaments and office politics. Evidence from a real effort experiment. *American Economic Review* 100(1):504–517.

Charness G, Halladay B (2017) BE and EE. Cousins but not twins. *Journal of Behavioral Economics for Policy* 1(2):27–31.

Chen CM, Thomas DJ (2016) Inventory allocation in the presence of service level agreements. Working paper, Bucknell University, Lewisburg, PA.

Chen H, Frank MZ, Wu OQ (2005) What actually happened to the inventories of American companies between 1981 and 2000? *Management Science* 51(7):1015–1031.

Chen H, Frank MZ, Wu OQ (2007) U.S. retail and wholesale inventory performance from 1981 to 2004. *Manufacturing & Service Operations Management* 9(4):430–456.

Chen K, Bendle N, Soman D (2017) Policy by design. The dawn of behaviourally-informed government. Behavioural economics in action, Rotman School of Management, University of Toronto, Canada.

Cohen SA, Kulp S, Randall T (2007) Motivating supply chain behavior. The right incentives can make all the difference. *Supply Chain Management Review* 11(4):18–24.

Cokely ET, Galesic M, Schulz E, Ghazal S (2012) Measuring risk literacy. The Berlin numeracy test. *Judgment and Decision Making* 7(1):25–47.

Copperberg (2013) Spare parts business platform (Elite Hotel Marina Tower, Stockholm, Sweden), February 7–8, 2013.

Cronin MA, Gonzalez C, Sterman JD (2009) Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes* 108(1):116–130.

Crosetto P, Filippin A (2013) The "bomb" risk elicitation task. *Journal of Risk and Uncertainty* 47(1):31–65.

Dane E, Pratt ME (2007) Exploring intuition and its role in managerial decision making. *Academy of Management Journal* 32(1):33–54.

Davis AM (2015) An experimental investigation of pull contracts in supply chains. *Production and Operations Management* 24(2):325–340.

Davis AM, Katok E, Kwasnica AM (2011) Do auctioneers pick optimal reserve prices? *Management Science* 57(1):177–192.

Davis AM, Katok E, Kwasnica AM (2014) Should sellers prefer auctions? A laboratory comparison of auctions and sequential mechanisms. *Management Science* 60(4):990–1008.

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

Denes-Raj V, Epstein S (1994) Conflict between intuitive and rational processing. When people behave against their better judgment. *Journal of Personality and Social Psychology* 66(5):819–829.

Deutscher Bundestag (2015) Ziel der Arbeitsgruppe "wirksames Regieren" sowie Aufgaben der drei im Bundeskanzleramt eingestellten Experten und neutrale Aufklärung der Bürger. Retrieved from http://dipbt.bundestag.de/extrakt/ba/WP18/672/67298.html.

Eckerson WW (2011) *Performance dashboards. Measuring, monitoring, and managing your business* (Hoboken, NJ: John Wiley & Sons), 2 edition.

Elmaghraby WJ, Katok E, Santamaría N (2012) A laboratory investigation of rank feedback in procurement auctions. *Manufacturing & Service Operations Management* 14(1):128–144.

Engelbrecht-Wiggans R, Haruvy E, Katok E (2007) A comparison of buyer-determined and price-based multiattribute mechanisms. *Marketing Science* 26(5):629–641.

Epstein S (1994) Integration of the cognitive and the psychodynamic unconscious. *American Psychologist* 49(8):709–724.

Erlenkotter D (1990) Ford Whitman Harris and the economic order quantity model. *Operations Research* 38(6):937–946.

Evans JSB (2008) Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59(1):255–278.

Fehr E, Goette L (2007) Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review* 97(1):298–317.

Finucane ML, Peters E, Slovic P (2003) Judgment and decision making. The dance of affect and reason. Schneider SL, Shanteau J, eds., *Emerging perspectives on judgment and decision research*, 327–364, Cambridge series on judgment and decision making (Cambridge, UK: Cambridge University Press).

Fischbacher U (2007) z-Tree. Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2):171–178.

Fischer GW, Damodaran N, Laskey KB, Lincoln D (1987) Preferences for proxy attributes. *Management Science* 33(2):198–214.

Fox CR, Tversky A (1998) A belief-based account of decision under uncertainty. *Management Science* 44(7):879–895.

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

Gächter S, Johnson EJ, Herrmann A (2010) Individual-level loss aversion in riskless and risky choices. Working paper, University of Nottingham, UK.

Galesic M, Garcia-Retamero R, Gigerenzer G (2009) Using icon arrays to communicate medical risks. Overcoming low numeracy. *Health Psychology* 28(2):210–216.

Ghazal S, Cokely ET, Garcia-Retamero R (2014) Predicting biases in very highly educated samples. Numeracy and metacognition. *Judgment and Decision Making* 9(1):15–34.

Gill D, Prowse V (2012) A structural analysis of disappointment aversion in a real effort competition. *American Economic Review* 102(1):469–503.

Goldstein WM, Einhorn HJ (1987) Expression theory and the preference reversal phenomena. *Psychological Review* 94(2):236–254.

Green E (2014) Why do Americans stink at math? *New York Times Magazine* July 23, 2014. Retrieved from http://nyti.ms/1nTvjER.

Green LV, Kolesar PJ (2004) Improving emergency responsiveness with management science. *Management Science* 50(8):1001–1014.

Greiner B (2004) An online recruitment system for economic experiments. Kremer K, Macho V, eds., *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht*, volume 63, 79–93 (Göttingen, Germany).

Gruen TW, Corsten DS, Bharadwaj S (2002) Retail out-of-stocks. A worldwide examination of extent, causes and consumer responses. Grocery Manufacturers of America, Washington, DC. Retrieved from http://www.uccs.edu/Documents/tgruen/GMA_2002_%20Worldwide_OOS_Study.pdf.

Gurmankin AD, Baron J, Armstrong K (2004) The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. *Medical Decision Making* 24(3):265–271.

Harris FW (1990) How many parts to make at once. *Operations Research* 38(6):947–950, reprinted from Harris FW (1913) How many parts to make at once. *Factory, The Magazine of Management* 10(2):135–136.

Harrison A, New C (2002) The role of coherent supply chain strategy and performance management in achieving competitive advantage. An international survey. *Journal of the Operational Research Society* 55(3):263–271.

Harrison GW (1989) Theory and misbehavior of first-price auctions. *American Economic Review* 79(4):749–762.

Haruvy E, Katok E (2013) Increasing revenue by decreasing information in procurement auctions. *Production and Operations Management* 22(1):19–35.

Hausman WH (2004) Supply chain performance metrics. Harrison TP, Lee HL, Neale JJ, eds., *The practice of supply chain management. Where theory and application converge*, 61–73, International series in operations research and management science (Boston, MA: Springer).

Hess R, Visschers VHM, Siegrist M (2011) Risk communication with pictographs. The role of numeracy and graph processing. *Judgment and Decision Making* 6(3):263–274.

Ho TH, Lim N, Cui TH (2010) Reference dependence in multilocation newsvendor models. A structural analysis. *Management Science* 56(11):1891–1910.

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *American Economic Review* 92(5):1644–1655.

Hopp WJ (2004) Fifty years of Management Science. *Management Science* 50(1):1–7.

Hoppe EI, Kusterer DJ (2011) Behavioral biases and cognitive reflection. *Economics Letters* 110(2):97–100.

Hsee CK, Yu F, Zhang J, Zhang Y (2003) Medium maximization. *Journal of Consumer Research* 30(1):1–14.

Johnson EJ, Shu SB, Dellaert BGC, Fox CR, Goldstein DG, Häubl G, Larrick RP, Payne JW, Peters E, Schkade D, Wansink B, Weber EU (2012) Beyond nudges. Tools of a choice architecture. *Marketing Letters* 23(2):487–504.

Kagan E, Leider S, Lovejoy WS (2017) Ideation–execution transition in product development. An experimental analysis. *Management Science* Published online in articles in advance: April 20, 2017. https://doi.org/10.1287/mnsc.2016.2709.

Kagel JH, Kim C, Moser D (1996) Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior* 13(1):100–110.

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

Kahneman D, Frederick S (2002) Representativeness revisited. Attribute substitution in intuitive judgment. Gilovich T, Griffin D, Kahneman D, eds., *Heuristics and biases. The psychology of intuitive judgment*, 49–81 (New York, NY: Cambridge University Press).

Kahneman D, Slovic P, Tversky A, eds. (1982) *Judgment under uncertainty: Heuristics and biases* (Cambridge, UK: Cambridge University Press).

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

Karelitz TM, Budescu DV (2004) You say "probable" and I say "likely". Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied* 10(1):25–41.

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.

Katok E, Thomas DJ, Davis AM (2008) Inventory service-level agreements as coordination mechanisms. The effect of review periods. *Manufacturing & Service Operations Management* 10(4):609–624.

Katok E, Wu DY (2009) Contracting in supply chains. A laboratory investigation. *Management Science* 55(12):1953–1968.

Keeney RL, Raiffa H (1976) *Decisions with multiple objectives. Preferences and value tradeoffs.* Wiley series in probability and mathematical statistics (New York, NY: John Wiley & Sons).

Kőszegi B, Rabin M (2006) A model of reference-dependent preferences. *Quarterly Journal of Economics* 121(4):1133–1165.

Klayman J, Brown K (1993) Debias the environment instead of the judge. An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition* 49(1-2):97–122.

Kocabiyikoglu A, Gogus CI, Gonul MS (2015) Revenue management vs. newsvendor decisions. Does behavioral response mirror normative equivalence? *Production and Operations Management* 24(5):750–761.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Science* 62(10):3023–3038.

Kremer M, Minner S, van Wassenhove LN (2014) On the preference to avoid ex post inventory errors. *Production and Operations Management* 23(4):773–787.

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

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

Kutner M, Greenberg E, Baer J (2006) National Assessment of Adult Literacy (NAAL). A first look at the literacy of America's adults in the 21st century. Retrieved from https://nces.ed.gov/NAAL/PDF/2006470.PDF.

Lariviere MA, Porteus EL (2001) Selling to the newsvendor. An analysis of price-only contracts. *Manufacturing & Service Operations Management* 3(4):293–305.

Larrick RP (2004) Debiasing. Koehler DJ, Harvey N, eds., *Blackwell handbook of judgment and decision making*, 316–337, Blackwell handbooks of experimental psychology (Malden, MA: Blackwell Publishing).

Larrick RP, Soll JB (2008) The MPG illusion. *Science* 320(5883):1593–1594.

Lee YS, Siemsen E (2016) Task decomposition and newsvendor decision making. *Management Science* Published online in articles in advance: September 9, 2016. https://doi.org/10.1287/mnsc.2016.2521.

Lipkus IM, Samsa G, Rimer BK (2001) General performance on a numeracy scale among highly educated samples. *Medical Decision Making* 21(1):37–44.

Loch CH, Wu Y (2007) *Behavioral operations management* (Boston, MA: Now Publishers).

Loomes G, Sugden R (1986) Disappointment and dynamic consistency in choice under uncertainty. *Review of Economic Studies* 53(2):271–282.

Lurie NH, Swaminathan JM (2009) Is timely information always better? The effect of feedback frequency on decision making. *Organizational Behavior and Human Decision Processes* 108(2):315–329.

Marcus Evans (2013) Optimizing the international spare parts management in the machinery industry (Steigenberger Hotel Berlin, Germany), January 24–25, 2013.

Meyer J, Bitan Y (2002) Why better operators receive worse warnings. *Human Factors* 44(3):343–353.

Moritz BB, Hill AV, Donohue KL (2013) Individual differences in the newsvendor problem: Behavior and cognitive reflection. *Journal of Operations Management* 31(1-2):72–85.

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

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

Nelson W, Reyna VF, Fagerlin A, Lipkus I, Peters E (2008) Clinical implications of numeracy. Theory and practice. *Annals of Behavioral Medicine* 35(3):261–274.

Obama B (2015) Executive order 13707. Using behavioral science insights to better serve the American people. *Federal Register* 80(181):56365–56367.

Ockenfels A, Selten R (2014) Impulse balance in the newsvendor game. *Games and Economic Behavior* 86(July):237–247.

Ockenfels A, Selten R (2015) Impulse balance and multiple-period feedback in the newsvendor game. *Production and Operations Management* 24(12):1901–1906.

Ockenfels A, Sliwka D, Werner P (2015) Bonus payments and reference point violations. *Management Science* 61(7):1496–1513.

OECD (2016) Skills matter. Further results from the survey of adult skills. OECD Skills Studies, OECD Publishing, Paris, France. Retrieved from http://dx.doi.org/10.1787/23078731.

Oechssler J, Roider A, Schmitz PW (2009) Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization* 72(1):147–152.

Ordóñez L, Benson L III (1997) Decisions under time pressure. How time constraint affects risky decision making. *Organizational Behavior and Human Decision Processes* 71(2):121–140.

Özer Ö, Zheng Y, Chen KY (2011) Trust in forecast information sharing. *Management Science* 57(6):1111–1137.

Paolacci G, Chandler J (2014) Inside the turk. Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* 23(3):184–188.

Paolacci G, Chandler J, Ipeirotis PG (2010) Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5):411–419.

Papke LE, Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11(6):619–632.

Parmenter D (2010) *Key performance indicators. Developing, implementing, and using winning KPIs* (Hoboken, NJ: John Wiley & Sons), 2 edition.

Pasternack BA (1985) Optimal pricing and return policies for perishable commodities. *Marketing Science* 4(2):166–176.

Peters E, Hibbard JH, Slovic P, Dieckmann NF (2007) Numeracy skill and the communication, comprehension, and use of risk-benefit Information. *Health Affairs* 26(3):741–748.

Peters E, Västfjäll D, Slovic P, Mertz CK, Mazzocco K, Dickert S (2006) Numeracy and decision making. *Psychological Science* 17(5):407–413.

Pinker EJ (2007) An analysis of short-term responses to threats of terrorism. *Management Science* 53(6):865–880.

Ren Y, Croson RTA (2013) Overconfidence in newsvendor orders. An experimental study. *Management Science* 59(11):2502–2517.

Reyna VF, Nelson WL, Han PK, Dieckmann NF (2009) How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin* 135(6):943–973.

Rosenboim M, Shavit T, Cohen C (2013) Do bidders require a monetary premium for cognitive effort in an auction? *Journal of Socio-Economics* 42:99–105.

Rosling K (2002) Inventory cost rate functions with nonlinear shortage costs. *Operations Research* 50(6):1007–1017.

Roth AE (1995) Introduction. Kagel JH, Roth AE, eds., *The handbook of experimental economics* (Princeton, NJ: Princeton University Press).

Roulston MS, Bolton GE, Kleit AN, Sears-Collins AL (2006) A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting* 21(1):116–122.

Roulston MS, Smith LA (2004) The boy who cried wolf revisited. The impact of false alarm intolerance on cost-loss scenarios. *Weather and Forecasting* 19(2):391–397.

Rousseau DM, Sitkin SB, Burt RS, Camerer CF (1998) Introduction to special topic forum. Not so different after all. A cross-discipline view of trust. *Academy of Management Journal* 23(3):393–404.

Rudi N, Drake D (2014) Observation bias. The impact of demand censoring on newsvendor level and adjustment behavior. *Management Science* 60(5):1334–1345.

Scheele LM, Thonemann UW, Slikker M (2017) Designing incentive systems for truthful forecast information sharing within a firm. *Management Science* Published online in articles in advance: August 30, 2017. https://doi.org/10.1287/mnsc.2017.2805.

Schwartz LM, Woloshin S, Black WC, Welch HG (1997) The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine* 127(11):966–972.

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

Shunko M, Niederhoff J, Rosokha Y (2017) Humans are not machines. The behavioral impact of queueing design on service time. *Management Science* Published online in articles in advance: February 10, 2017. https://doi.org/10.1287/mnsc.2016.2610.

Sieke MA, Seifert RW, Thonemann UW (2013) Designing service level contracts for supply chain coordination. *Production and Operations Management* 21(4):698–714.

Simon HA (1955) A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1):99–118.

Simon HA (1957) *Models of man. Social and rational* (New York, NY: John Wiley & Sons, Inc).

Simon HA (1959) Theories of decision-making in economics and behavioral science. *American Economic Review* 49(3):253–283.

Sloman SA (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119(1):3–22.

Slovic P, Lichtenstein S (1968) Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology: General* 78(3, Part 2):1–18.

Slovic P, Lichtenstein S (1971) Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Decision Processes* 6(6):649–744.

Slovic P, Monaham J, MacGregor DG (2000) Violence risk assessment and risk communication. The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior* 24(3):271–296.

Smith VL (1962) An experimental study of competitive market behavior. *Journal of Political Economy* 70(2):111–137.

Smith VL (1976) Experimental economics. Induced value theory. *American Economic Review* 66(2):274–279.

Soll JB, Keeney RL, Larrick RP (2013) Consumer misunderstanding of credit card use, payments, and debt. Causes and solutions. *Journal of Public Policy and Marketing* 32(1):66–81.

Soll JB, Milkman KL, Payne JW (2015) A user's guide to debiasing. Keren G, Wu G, eds., *The Wiley Blackwell handbook of judgment and decision making*, volume 2, 924–951 (Maden, MA: John Wiley & Sons, Ltd).

Stangl T, Thonemann UW (2017) Equivalent inventory metrics. A behavioral perspective. *Manufacturing & Service Operations Management* 19(3):472–488.

Stanovich KE, West RF (2000) Individual differences in reasoning. Implications for the rationality debate? *Behavioral and Brain Sciences* 23(5):645–726.

Sterman JD (2002) All models are wrong. Reflections on becoming a systems scientist. *System Dynamics Review* 18(4):501–531.

Sting FJ, Loch CH, Stempfhuber D (2015) Accelerating projects by encouraging help. *Sloan Management Review* 56(3):33–41.

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management* 10(4):566–589.

Supply Chain Academy (2013) The supply chain executive academy (Capability Center, Munich, Germany), October 24–25, 2013.

Svenson O (1970) A functional measurement approach to intuitive estimation as exemplified by estimated time savings. *Journal of Experimental Psychology: General* 86(2):204–210.

Svenson O (2008) Decisions among time saving options. When intuition is strong and wrong. *Acta Psychologica* 127(2):501–509.

Thaler RH (1980) Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization* 1(1):39–60.

Thaler RH (1985) Mental accounting and consumer choice. *Marketing Science* 4(3):199–214.

Thaler RH, Sunstein CR (2008) *Nudge. Improving decisions about health, wealth and happiness* (London, England: Penguin).

The Royal Swedish Academy of Sciences (1978) The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2002. Press release. October 16, 1978. Retrieved from http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/press.html.

The Royal Swedish Academy of Sciences (2002) The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1978. Press release. October 9, 2002. Retrieved from http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/press.html.

The Royal Swedish Academy of Sciences (2017) The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2017. Press release. October 9, 2017. Retrieved from https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2017/press.html.

Thompson JC (1952) On the operational deficiencies in categorical weather forecasts. *Bulletin of the American Meteorological Society* 33(6):223–226.

Thonemann UW, Behrenbeck K, Diederichs R, Großpietsch J, Küpper J, Leopoldseder M (2003) *Supply Chain Champions. Was sie tun und wie Sie einer werden. (in German).* Financial Times Deutschland (Wiesbaden, Germany: Gabler), 1 edition.

Toplak ME, West RF, Stanovich KE (2011) The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition* 39(7):1275–1289.

Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psychological Bulletin* 76(2):105–110.

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

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning. The conjunction fallacy in probability judgment. *Psychological Review* 90(4):293–315.

van Dijk F, Sonnemans J, van Winden F (2001) Incentive systems in a real effort experiment. *European Economic Review* 45(2):187–214.

Wallsten TS (1986) Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General* 115(4):348–365.

Weber EU (1994) From subjective probabilities to decision weights. The effect of asymmetric loss functions on the evaluation of unvertain outcomes and events. *Psychological Bulletin* 115(2):228–242.

Wu DY, Chen KY (2014) Supply chain contract design. Impact of bounded rationality and individual heterogeneity. *Production and Operations Management* 23(2):253–268.

Yamagishi K (1997) When a 12.86% mortality is more dangerous than 24.14%. Implications for risk communication. *Applied Cognitive Psychology* 11(6):495–506.

Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? An empirical study. *Management Science* 63(1):1–20.

Zhang Y, Siemsen E (2016) A meta-analysis of newsvendor experiments. Revisiting the pull-to-center asymmetry. Working paper, University of Cincinnati, Cincinnati, OH.

Zipkin PH (2000) *Foundations of inventory management* (Boston, MA: McGraw-Hill).