http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–18
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

From Noise to Bias: Overconfidence in New Product Forecasting

Daniel Feiler,[a] Jordan Tong[b]

[a] Tuck School of Business, Dartmouth College, Hanover, New Hampshire 03755; [b] Wisconsin School of Business, University of Wisconsin-Madison, Madison, Wisconsin 53706

Contact: daniel.feiler@tuck.dartmouth.edu, https://orcid.org/0000-0001-6248-5654 (DF); jordan.tong@wisc.edu, https://orcid.org/0000-0002-1962-058X (JT)

Received: April 7, 2020
Revised: January 25, 2021; March 25, 2021
Accepted: March 30, 2021
Published Online in Articles in Advance: November 5, 2021

https://doi.org/10.1287/mnsc.2021.4102

Copyright: © 2021 INFORMS

Abstract. We study decision behavior in the selection, forecasting, and production for a new product. In a stylized behavioral model and five experiments, we generate new insight into when and why this combination of tasks can lead to overconfidence (specifically, overestimating the demand). We theorize that cognitive limitations lead to noisy interpretations of signal information, which itself is noisy. Because people are statistically naive, they directly use their noisy interpretation of the signal information as their forecast, thereby underaccounting for the uncertainty that underlies it. This process leads to unbiased forecast errors when considering products in isolation, but leads to positively biased forecasts for the products people choose to launch due to a selection effect. We show that this selection-driven overconfidence can be sufficiently problematic that, under certain conditions, choosing the product randomly can actually yield higher profits than when individuals themselves choose the product to launch. We provide mechanism evidence by manipulating the interpretation noise through information complexity—showing that even when the information is equivalent from a Bayesian perspective, more complicated information leads to more noise, which, in turn, leads to more overconfidence in the chosen products. Finally, we leverage this insight to show that getting a second independent forecast for a chosen product can significantly mitigate the overconfidence problem, even when both individuals have the same information.

History: Accepted by Charles Corbett, operations management.
Funding: This work was supported by the Tuck School of Business at Dartmouth College and the BRITE Laboratory at the University of Wisconsin-Madison.
Supplemental Material: Data and supplemental material are available at https://doi.org/10.1287/mnsc.2021.4102

Keywords: judgment & decision-making • new product development • laboratory experiments • random error • decision bias • overconfidence • choice • forecast

1. Introduction

To make decisions under uncertainty, nearly all managers rely on a forecast—a prediction or estimate of a future unknown value. Although computation and automation play an increasingly important role in a firm's forecasting, human judgment remains a critical aspect of many forecasts, particularly when managers face novel situations with less precedence. Given its significance, a vast and important literature has sought to improve our understanding of human forecasting behavior and potential biases therein. Although most researchers have studied forecasting behavior in isolation—as is the natural starting point for any topic—we also know that forecasting rarely exists as a standalone judgment or decision in organizations. Rather, it typically occurs as part of an operational process composed of multiple interconnected decisions.

Taking a process perspective, we study the role that human forecasting plays within a common combination of decisions for new products. We focus on three main features of new product forecasting. First, forecasting new products typically requires significant human interpretation of imperfect predictive information. Historical demand data tend to be limited or nonexistent for new products, and other predictive information may be highly unstructured or inconsistent. Second, forecasting informs the selection of new products, as managers evaluate multiple alternatives, options, and designs and then choose which product will advance to the next stage. Third, managers make downstream operational decisions, but only for the new products selected for launch. Therefore, only the forecast for selected products is ultimately consequential for subsequent operational decision making, and forecasts for nonselected products have no further implications. For example, firms only produce inventory for the products they choose to launch (and not for those they don't). As Girotra et al.

(2010, p. 592) wrote about new product development and innovation, "the metric for effectiveness is the quality of the ideas *selected* as the best" (emphasis added).[1]

By studying forecasting within this process, we contribute new insight into when and why *overconfidence* in new products emerges. Overconfidence has been said to be "the most significant of the cognitive biases" (Kahneman 2011, p. 255) and a consistent problem for new product launches (Simon and Shrader 2012). Moore and Healy (2008) outlined three types of overconfidence; in this paper, we examine overconfidence in the form of overestimation (as opposed to overprecision or overplacement), thereby revisiting the question of why managers who launch new products tend to have inflated expectations of them.

Through a stylized model and a series of experiments, we predict and find evidence for a behavioral challenge that emerges as a consequence of how two psychological primitives interact with selection among alternatives. The first psychological primitive is that, due to cognitive-processing limitations, humans interpret information with noise. They exhibit random error when making inferences from predictive signal information. The second psychological primitive is that humans are statistically naive in that they tend to overly rely on directly available information (more specifically their interpretation of it) and, in doing so, underaccount for the noise that underlies it (from the environment and from themselves). Because people overweight their noisy interpretations of imperfect information signals, but at the same time attempt to select new products for which they have the most favorable forecast, the products they choose to launch tend to feature a positively biased behavioral forecast. In other words, due to a failure in nonregressive thinking, people's more extreme expectations in either direction are too extreme, and product selection chooses from the high end of those forecasts. A self-aware Bayesian appropriately accounts for noise from the environment and from themselves by making a mean-regressing correction that is systematically downward for products that tend to get selected. Consequently, such an agent has unbiased expectations, even for the product for which she has made the most favorable forecast (Smith and Winkler 2006).

We also know that selecting the highest forecast is beneficial relative to selecting a product for which one has lower expectations, which raises the question: Is there a way to retain the benefits of human product selection while mitigating this selection-driven overconfidence? We examine whether and how getting a second individual to independently assess the information for the chosen product and make the final forecast can mitigate the overconfidence effect. We use the model to show that this strategy works particularly well if human interpretation error drives forecast noise, such that two people looking at the same information make different, idiosyncratic interpretations. Where the first person's expectation is too extreme, the second person's expectation would be mean-reverting, even when the two individuals have identical information. However, to the extent that the positive forecast error exists in the information signal itself, rather than in the interpretation of the information, then this process will be less effective in reducing the bias.

We present the results from five controlled laboratory experiments (with an additional study in the appendix), across three populations, providing evidence for the theorized type of overconfidence. Our primary experimental task is inspired by a popular operations management classroom simulation exercise (Hammond 2016), in which students select amongst designs, forecast demands, and then launch and make production decisions for a new cell phone. We ask subjects to first choose which product design to launch based on predictive demand data in the form of a team of analysts' judgments for several product designs. Then, for the product they choose, they must decide how much inventory to produce, where it is costly to overproduce or underproduce. Subjects are ultimately compensated based on the financial performance of the product they choose to launch.

We now preview our main experimental results. We first establish that for a single randomly selected product, individuals tend to be well calibrated, but when the full process is considered, individuals overproduce for the product they selected to launch (Studies 1(a) and 1(b)). We demonstrate that this overconfidence bias is sufficiently problematic that, under certain conditions, choosing products randomly can actually yield higher profits than individuals choosing products themselves (Study 2). We then turn to the question of how an organization might reduce the overconfidence bias while retaining the benefits of nonrandom product selection. To do so, we first isolate mechanism evidence by showing that person-specific interpretation error moderates this overconfidence. Specifically, we show that even when information remains constant from a Bayesian perspective, more complicated information leads to more noise, which, in turn, leads to more overconfidence in the chosen products (Study 3). Finally, we leverage this insight to test a mitigation strategy that should only work in the presence of interpretation error: For a product chosen by one person, getting a second independent forecast from a different person can significantly mitigate the overconfidence problem, even when both individuals have the same information (Study 4).

2. Related Literature and Contributions

Overconfidence has been found to be one of the most pervasive and pernicious biases plaguing managerial decision making (Bazerman and Moore 2012).

Psychologists have found that individuals tend to overestimate their actual performance, overplace their performance relative to others', and falsely perceive excessive precision in their beliefs (overestimation, overplacement, and overprecision; Moore and Healy 2008). Overconfidence has been an important topic of study within behavioral operations management as well, mostly focusing on overprecision. Ren and Croson (2013) argued that overprecision in one's demand forecast in the newsvendor setting can help explain why individuals order too little inventory under high-profit conditions and order too much under low-profit conditions (Schweitzer and Cachon 2000). Li et al. (2017) modeled the possible benefits of overprecision when newsvendors compete with one another. Tong and Feiler (2017) model forecasters as forming demand beliefs based on naive statistics on small random samples and show that, among other behaviors, it leads to overprecision in demand forecasts.

The particular overconfidence explored in this paper is the product of a relationship between uncertainty and choice that has also been noted in several theoretical papers spread across a variety of literatures outside of operations management (Brown 1974, Harrison and Kreps 1978, Harrison and March 1984, Harrison and Bazerman 1995, Van den Steen 2004, Smith and Winkler 2006, Hogarth and Karelaia 2012). In the recent psychology literature, Tong et al. (2018) find evidence of this choice-driven overoptimism as an alternative explanation to wishful thinking. In this paper, we provide new experimental evidence for this type of selection-driven overestimation by applying it to the context of new products, exploring the further implications of interpretation noise, and testing novel predictions for this context.

In our paper, unsystematic random error ultimately leads to systematic overconfidence. In this sense, our paper relates to others where similar phenomena occur. For example, in Erev et al. (1994) and Soll (1996), unsystematic random judgment error explains systematic overconfidence. However, the type of overconfidence considered in these papers is different (overestimating the probability that a judgment is correct), and the general problem structure and organizational process considered in this paper is different.

New product development (NPD) inherently involves a variety of human judgments, from the creativity needed to generate new ideas to the cognitive self-awareness needed to intelligently plan for, and execute, the selected initiatives (Gartner and Thomas 1993, Grushka-Cockayne et al. 2018). In addition to emphasizing the importance of human behavior, researchers studying NPD have also focused heavily on its multistage nature, often explicating the sequential aspects of the process through idea generation, idea assessment, product testing, and product launch to market (Urban et al. 1996, Petersen et al. 2005). Experimental research in NPD has shared this focus on how process affects behavior. For example, Kagan et al. (2018) experimentally examined time-allocation decisions between the phases of ideation and execution, finding that teams who endogenously determine when to transition from ideas to implementation are at greater risk of overall failure. We complement this line of experimental work on new products by examining a different combination of tasks—assessment, selection, and production—which highlights new behavioral phenomena that are important for NPD.

We also build on a robust literature on behavioral forecasting in operations management. Many decision biases can be traced back to the erroneous beliefs or forecasts that are used as inputs to those decisions. For instance, Kremer et al. (2011) argue that system-neglect biases in demand forecasting will drive biases in inventory decisions, and Feiler et al. (2013) showed that underestimating mean demand due to censorship can drive an underordering bias (see also Becker-Peth and Thonemann 2018 and Siemsen et al. 2018). As in the present paper, some researchers studying behavioral forecasting have emphasized the importance of the process in which forecasting occurs. Kremer et al. (2016) examined hierarchical forecasting and the difference between using forecasts at the superordinate level directly versus forecasting those values indirectly through summing up subordinate-level forecasts (bottom-up). Lee and Siemsen (2017) examined the consequences of decomposing newsvendor decision making across multiple individuals. Flicker (2019) explored how noisy demand forecasts by managers can be translated into near-optimal inventory decisions via algorithms. Tyebjee (1987) also noted the underlying statistical dynamic explored here, pointing out that forecasting methods can only be audited against products that were actually introduced to market (see also Smith and Winkler 2006). We extend this work by focusing on the relationship between behavioral forecasting and product selection and exploring it experimentally. We also carry on a longstanding tradition in management science of assessing the role of human judgment in an organization's forecasting processes (Hogarth and Makridakis 1981, Lawrence et al. 2006).

Building from the existing literature, we believe this paper may help the reader gain several important insights. First, we identify a previously underappreciated source of overconfidence in new product development. We experimentally document this pattern of behavior across five experiments (plus one in the appendix) and three different populations (managers, undergraduate students, and Amazon Mechanical Turk workers). Second, we illustrate that the costs of the effect can be large enough to potentially outweigh the benefits of being able to choose between products in the first place. Third, our behavioral theory and experimental evidence can help a manager anticipate the situations in which this pitfall is especially likely to be problematic. Fourth, our work sheds new light on how organizational task structures and second opinions can overcome overconfidence—and the conditions under which they cannot.

3. Theory Development

In this section, we motivate our model setting and our two key behavioral assumptions. Then, we derive how these two assumptions lead to our hypotheses about overconfidence and mitigation via organizational structure. We conclude with an overview of our experimental studies and how they test our theory.

3.1. Demand Model

We begin by defining our assumptions about demand, which we also follow in our experimental designs. For a given product i, define demand as a random variable $D_i \sim N(\mu_i, \sigma_i)$. The unknown mean demand for each product μ_i is a random variable $\mu_i \sim N(\mu^{means}, \sigma^{means})$, where μ^{means} and σ^{means} are known. For ease of exposition, we assume that the standard deviation for any product is the same, $\sigma_i = \sigma^{demand}$ for all i. In fact, in some experiments, we set $\sigma_i = \sigma^{demand} = 0$ for all i, so that all of the uncertainty is due to the unknown values of μ_i.

3.2. Noise and Statistical Naivety

We assume that individuals cannot directly observe the true demand mean μ_i, but only an indirect signal $S_i \sim N(\mu_i, \sigma^{info})$, $\sigma^{info} > 0$. For example, such an information signal could be of the form of sample observations from D_i. Thus, σ^{info} captures *information noise.* We assume the signals across products are independent.

We now describe our two behavioral forecasting assumptions. First, individuals add *interpretation noise*. Because of cognitive limitations, a given individual j does not perfectly extract the information from the signal, but instead adds mean-zero random error ϵ_{ij}, where we assume $\epsilon_{ij} \sim N(0, \sigma^{interpret})$, $\sigma^{interpret} > 0$, and ϵ_{ij} s independent. We have:

Assumption 1 (Interpretation Noise). *Individual j interprets the information signal as $S_j^b(i) = S_i + \epsilon_{ij}$, rather than S_i.*

Second, we assume that decision makers are *statistically naive*—they directly use their interpretation of the information signal $S_j^b(i)$ as their forecast for product i. We have:

Assumption 2 (Statistically Naive). *Individual j's forecast for product i is $F_j^b(i) = S_j^b(i)$. They do not make a statistical correction to sufficiently account for (a) the information noise σ^{info} or (b) their own interpretation noise $\sigma^{interpret}$.*

One may think of this kind of statistical naivety as a type of base-rate neglect (see Barbey and Sloman 2007 for a review) in a continuous domain, as opposed to the traditional binary domain. In the classic cancer-testing example (Eddy 1982), physicians tend to overly rely on an imperfectly diagnostic positive or negative test result, failing to properly account for the base rate of having cancer in the first place. Similarly, in our scenario, we assume that individuals overly rely on their imperfect interpretation of the noisy demand signal, failing to properly account for the prior distribution of true demand means.

Assumptions 2(a) and 2(b) decompose this naivety assumption into two components. By directly using one's interpretation of the signal, the individual fails to make a statistical adjustment to account for two distinct sources of noise: σ^{info} is due to the environment, whereas $\sigma^{interpret}$ is due to the individual. We will see later that, in order to explain our results, Assumption 2(b) must hold above and beyond 2(a).

It is important to observe that Assumptions 1 and 2 do *not* lead to a biased forecast for any *randomly selected* single product i. Because we have assumed that S_i has mean μ_i, and ϵ_{ij} has mean zero, $F_j^b(i)$ is, in fact, an unbiased estimator for any randomly selected product i. Nevertheless, this behavior is still statistically naive in the sense that it fails to make a correction for the facts that the signal information is noisy, and one's interpretation of that information is noisier still. A Bayesian recognizes the need to account for these sources of noise by relying more on base rate. In other words, a Bayesian would not directly use $S_j^b(i)$ as their forecast, but, rather, would forecast something between μ^{means} and $S_j^b(i)$ (see Appendix B). Such Bayesian corrections for randomly selected products also result in unbiased beliefs, but they yield downward or upward adjustments, depending on where the signal falls relative to the base rate.

3.3. Overconfidence in Selected New Products

Although Assumptions 1 and 2 result in unbiased forecasts for randomly selected products, we now demonstrate how these assumptions lead to an overconfidence bias by accounting for the fact that new product forecasting generally involves a nonrandom selection process. Therefore, we must consider the number of products under consideration and how forecasts affect the selection. Ultimately, only the forecast for the product that gets selected for launch remains an important input for downstream operations decisions.

Let n denote the number of products under consideration. Assume that expected profit is increasing in mean demand μ_i, such that the decision maker chooses the product associated with their highest forecast. Denote i_j^b the index of the product associated with person j's largest forecast, $i_j^b = \arg\max_i \{F_j^b(i)\}_{i=1,\ldots,n}$. Thus, $F_j^b(i_j^b)$ denotes person j's forecast associated with person j's chosen product. The following result states the overconfidence result:

Proposition 1. *Let $n > 1$. Individual j's forecast for individual j's chosen product is biased high—that is, $E[F_j^b(i_j^b)] > \mu_{i_j^b}$.*

For an intuition behind this result, see the left panel of the illustrative diagram in Figure 1. Note that product C has the highest true demand mean. However, because of information and interpretation noise, the individual may erroneously choose product A or B. In this example, the individual errs high for product B, causing them to pick product B instead of product C. In general, the process systematically makes it *more* likely for the manager to pick a product if they err *high* with its forecast (e.g., product B), but *less* likely if they err *low* with its forecast (e.g., product A). As a consequence, chosen products, in expectation, feature positive errors.[2] In contrast, the right panel of Figure 1 illustrates the Bayesian-adjusted forecasts (Appendix B). Here, the Bayesian regresses their forecasts toward the underlying mean of products in general. These adjustments are larger in absolute magnitude for signal interpretations that are more extreme. Note that a mean-regressing Bayesian adjustment does not change the relative ranking of the products and, therefore, does not alter product selection (e.g., product B gets selected in Figure 1 in either case). That is, Bayesian adjustments do not address the overconfidence problem by eliminating suboptimal product choices. Instead, they negate the selection-driven overconfidence by being more mean-reverting (downward) for the products that do get selected.

3.4. Mitigating Overconfidence via Organizational Structure

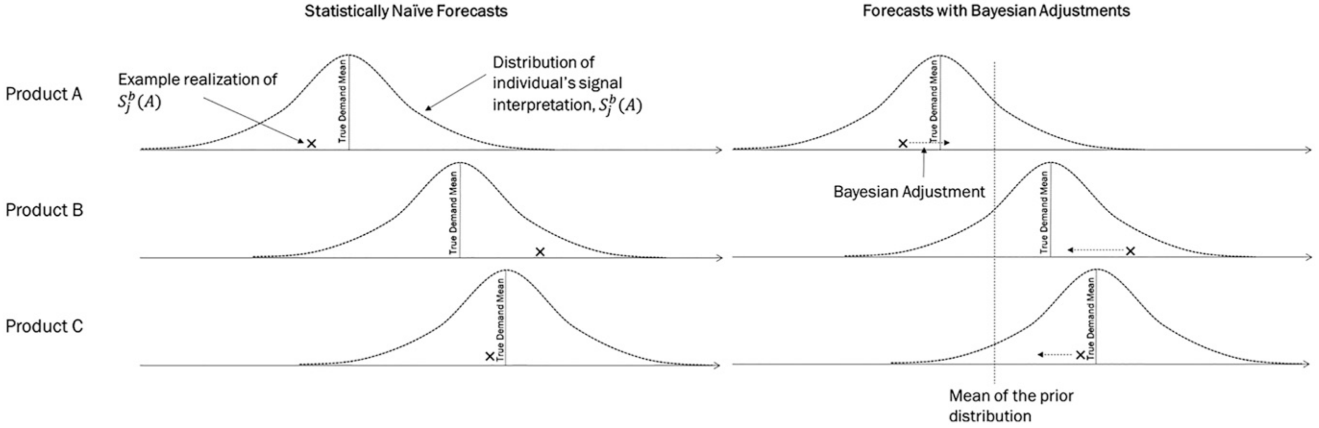
The magnitude of the overconfidence prediction in Proposition 1 does not depend on how much of the noise is due to imperfect information relative to imperfect interpretation of that information. Therefore, distinctly isolating interpretation noise from information noise may seem immaterial. However, we now argue that distinguishing between the two sources has important implications for when and whether organizational structure can be used as an effective mitigation strategy.

Consider a case in which all forecast noise is *information noise*. Then, two separate and otherwise independent people will be similarly biased because the information itself carries the random error. In this extreme case, involving more people in the process will have no effect on the outcome. In contrast, we now show that if there is significant *interpretation noise*, then involving more people in the process can potentially impact the outcome, even if all people have access to the same information. Specifically, one can mitigate the overconfidence bias of Proposition 1 by having a second person make an independent forecast for the chosen product using the same information. When individuals examine the same information (i.e., S_i is the same for any person j) and interpretation noise is independent across individuals (i.e., ϵ_{ij} and $\epsilon_{i\hat{j}}$ are independent for any persons j and $\hat{j}$), we have:

Proposition 2. *Let $n > 1$ and $\hat{j} \neq j$. Consider individual j's chosen product. Then, individual $\hat{j}$'s forecast is less positively biased than individual j's forecast—that is, $E[F_{\hat{j}}^b(i_j^b)] < E[F_j^b(i_j^b)]$.*

The intuition behind Proposition 2 is similar to a regression to the mean effect. People are idiosyncratic and may randomly interpret the information for a given product in an overly positive or negative light.

Figure 1. Illustrative Example Comparing Statistically Naive Forecasts with Bayesian-Adjusted Forecasts



However, a product that someone chooses is not random—it is systematically more likely to suffer from positive interpretation error from the chooser. A new person's forecast for that same product, however, does not suffer from the selection effect and receives a random interpretation noise draw. It will therefore tend to regress downward, toward the true mean.

The predicted effectiveness of the above result as a mitigation strategy depends on the relative magnitudes of information noise and interpretation noise. When the interpretation noise dominates (i.e., $\sigma^{interpret}/\sigma^{interpret}+\sigma^{info} \to 1$), then this mitigation strategy completely eliminates the overconfidence bias. However, when the information noise dominates (i.e., $\sigma^{info}/\sigma^{interpret}+\sigma^{info} \to 1$), then this mitigation strategy does not decrease bias at all. This result illustrates the critical importance of distinguishing between information noise and interpretation noise: As their relative importance changes, this mitigation strategy goes from being a complete solution to having absolutely no remedial effect on the overconfidence bias.

3.5. Experiments Overview

We now present a series of studies exploring the above theory. First, we test whether there exists overconfidence behavior consistent with Proposition 1. We find that people overproduce in the product they chose to launch from a set of seven options, but do not overproduce when presented one randomly selected product. Such overproduction occurs whether participants recommend production levels after they decide which product to launch (Study 1(a)) or before they decide which product to launch (Study 1(b)).

In Study 2, we show that the magnitude and consequence of the overconfidence effect can be surprisingly large. Specifically, we show that it is possible for this overconfidence effect to lead to such large overproduction in downstream operational decisions that it can outweigh the benefit of allowing the human to try to pick a good product in the first place.

In Study 3, we focus on the mechanisms driving the overconfidence effect. We manipulate only the interpretation noise component (Assumption 1) by varying the task complexity, while holding the formal information signal constant. We show that increasing task complexity, which increases interpretation noise, is sufficient to generate a significantly larger overconfidence bias for chosen products. Moreover, by drawing comparisons to the information signals, we find supporting evidence that subjects do not make sufficient statistical corrections for interpretation noise (Assumption 2(b)) above and beyond corrections for information noise (Assumption 2(b)).

Finally, having identified the interpretation noise as a separate and meaningful forecast noise component that can drive this form of overconfidence, Study 4 leverages this driver to test the prediction from Proposition 2. We find that a second independent opinion indeed reduces the overconfidence bias. Namely, people invest less in a given product when they were not the ones who chose to launch it, even if they have the exact same information.

4. Study 1(a): Overestimation of Chosen Product

We first test for the existence of the main overconfidence effect in a simple experiment in which participants decide how many units to produce for a new product launch based on demand information.

4.1. Methods

4.1.1. Design and Procedures. There was an attention check before entering the survey. First, participants were asked if they were willing to read instructions. If they indicated that they were not, then their participation was terminated, and no further data were collected from them. Second, there was a reading-comprehension attention check, in which participants needed to read a paragraph about a person's life and then answer three questions about it. Participants were allowed two attempts to answer these questions correctly. If they were unable to get these questions right on two attempts, then they exited the study, and no further data were collected from them.

Participants were randomly assigned to one of two conditions. In the *Choice condition*, participants were presented with seven possible products, from which one would be chosen to launch. In the *No Choice condition*, participants were shown only one product.

Participants completed the task via a browser-based user interface and were informed of the following information. In a hypothetical business, they would need to decide how many units of a product to produce based on how many units of it they think people will want to buy (i.e., expected demand). Their firm was considering seven possible new products to launch, and the firm wanted the participant's opinion about each product. Through written descriptions and histograms, they were informed that:

1. The potential future demand for a given product under consideration is uncertain.

2. The distribution of true mean demand of all products is $\mu_i \sim N(200, 10)$ (shown in a histogram).

3. The business hired four "professional forecasters" to provide independent datapoints (i.e., forecasts) for each product. These datapoints are independent across products and across forecasters.

4. Each professional forecaster provides one datapoint per product with accuracy of $N(\mu_i, 45)$ (shown in a histogram).

For each product, one can interpret the sample mean of these four datapoints as the "signal" S_i because it is a sufficient statistic for a rational Bayesian decision maker to properly update their belief (see Appendix B and Casella and Berger 2002). Thus, each $S_i \sim N(\mu_i, 45/\sqrt{4})$. To summarize in the notation of Section 3, in this experiment we set $\mu^{means} = 200$; $\sigma^{means} = 10$; $\sigma^{info} = 45/\sqrt{4} = 22.5$; and $\sigma^{demand} = 0$.

Depending on the experimental condition, the software randomly selected either seven products (Choice condition) or one product (No Choice condition) from the aforementioned true distribution of products. For each randomly selected product, the participants saw the simulated forecasters' datapoints.

In the *Choice condition*, participants looked at the datapoints (see example in Figure 2) and decided which of the seven products they wished to select for launch. They earned $0.01for each unit of true future demand of the product they chose (generally between $1 and $3) and, therefore, would earn more bonus if they picked a product with higher true demand. In the *No Choice condition*, there was no product selection, and participants simply proceeded to the next phase, although at the end of the study, they would also earn $0.01 for each unit of true future demand for their product.

Presented simultaneously was the prompt to make the production decision. Participants were informed that their goal was to produce a number of units as close to the true demand of the (selected) product as possible. They would lose $0.01 from their bonus for every unit that the number they produced was off from the true future demand, high or low. They looked at the data in the table (Figure 2) and answered how many units they wanted to produce of the (selected) product. To be clear, in both conditions, they earned a bonus equal to the true demand of the product multiplied by $0.01, but lost $0.01 from their bonus for every unit their "number produced" was off from the true demand (high or low).

Participants were asked to report their gender (open-entry), age, ethnicity, and education level. They were also given feedback on their performance, including seeing the true demands for the relevant product.

4.1.2. Participants and Preregistration. The sample consisted of 404 individuals recruited through Mechanical Turk. We obtained a sample in which 85% of participants had at least an associate's degree. The sample was 44% female and 18% non-White, with mean age of 36.7 (SD = 8.2). We targeted a sample size of 200 per cell, and no analyses were conducted until final data collection was completed. This study was preregistered at AsPredicted.org.[3]

4.2. Results

The dependent variable is production error, which we define as the production decision minus the true demand for the product. Positive values imply overproduction, and negative values imply underproduction.

In the Choice condition, the number of units produced was on average +21.69 units relative to the true demand (SE = 1.77) versus only +6.13 units in the No Choice condition (SE = 1.81), a significant difference between conditions, $t(402) = 6.14$, $p < 0.001$.[4]

In both conditions, participants produced more units than perfectly rational decision makers would have produced, given the same forecast datapoints ($p < 0.001$ in both Choice and in No Choice); however, this difference was 15.63 units larger in the Choice condition than in the No Choice condition, $t(402) = 6.64$, $p < 0.001$. Perfectly rational decisions were not significantly different from the true demand in the Choice condition (M = 0.38, SE = 0.65, $t(201) = 0.59$, $p = 0.56$) or in the No Choice condition (M = 0.46, SE = 0.64, $t(201) = 0.71$, $p = 0.48$).

4.3. Study 1(a) Summary

This study provides initial evidence of greater overestimation for products chosen from a set than for a single product. This effect manifested itself in overproduction for the new product *chosen* for launch. We also replicate this study with a sample of managers (see appendix).

5. Study 1(b): Overestimation Before Choice

According to our theory, product selection need not precede the production decision for an individual to

Figure 2. (Color online) Screenshot of Example Forecaster Datapoints in Study 1

Forecaster's Name	Product #1	Product #2	Product #3	Product #4	Product #5	Product #6	Product #7
Angela	178	221	208	177	226	229	238
Brian	268	187	277	172	186	271	195
Connie	185	182	225	165	195	312	142
Derek	232	195	246	219	230	244	230

exhibit overconfidence in the selected product. The mechanism applies also when hypothetical plans are made for all products before the selection occurs. To verify this prediction, we ran an iteration of the previous experiment in which participants recommended a production decision for each possible product, if it were to be launched. Only later did they learn that their firm would like them to select which of the products to launch, which they then did.

5.1. Methods

5.1.1. Design and Procedures. Participation was conditional on passing the same attention check as in Study 1(a). All participants experienced the *Pre-Choice Production Decision condition*, which was a variation on Study 1(a) as follows. Participants were shown the four signals (professional forecasters' datapoints) for each of seven potential new products. For one product at a time (in a random order), the firm asked the participant, "If this product is selected for launch, how many units should we produce?" Next, participants were informed that their firm was now determining which one of those seven products to launch. The firm now asked the participant to select the product that will be most profitable. Participants were then shown the same seven products again (with their respective datapoints from the forecasters) and selected which product to launch.

5.1.2. Participants. The sample consisted of 101 participants recruited from Mechanical Turk. Through targeted recruiting, we obtained a sample in which 85% of participants had at least an associate's degree. The sample was 44% female and 70% Caucasian, with mean age of 36.3 (SD = 11.2). We targeted a sample size of 100, and no analyses were conducted until final data collection was completed.

5.2. Results

The dependent variable again is production error: The production decision minus the true demand for the given product. A regression with standard errors clustered by participant shows that production errors were significantly more positively biased for chosen products than for nonchosen products (M = 21.96, SE = 2.26), $t(100) = 9.73$, $p < 0.001$. The same result holds with fixed effects for participants included in the model. Descriptively, on average, the product that was subsequently chosen to be launched had a production decision that was 28.15 units higher than true demand for that product. Production decisions for products that were ultimately not chosen were only 6.12 higher than the true demands for those products, on average (see endnote 3).

5.3. Study 1(b) Summary

Consistent with our theory, we find evidence that overconfidence in the selected product can occur even when the judgment of, and production decision for, that product is formed before the actual product selection occurs. This result suggests that the observed effect cannot be accounted for by cognitive dissonance (i.e., wanting to be consistent with past choices) or imagined costs of abandonment because the inflated production decision occurred before knowing which product would be chosen.

6. Study 2: Profit Consequences of the Overconfidence

In this experiment, we examine the potential profit consequences of this overconfidence effect by considering the impact of product selection and overconfidence on each component of a typical inventory management profit function. Given a chosen product i, production quantity q_i, and demand D_i, we assume a standard profit function:

$$\Pi(q_i, D_i) = mD_i - \left\{c_o[q_i - D_i]^+ + c_u[D_i - q_i]^+\right\}, \qquad (1)$$

where $m > 0$ denotes the profit margin per unit sold, c_o the unit overage cost, and c_u the unit underage cost.[5] The first term mD_i is commonly referred to as the *maximum profits*: the profits for product i if there were no mismatch between q_i and D_i. The second term, which is in brackets, is the *mismatch cost*: the consequences of the production decision being unequal to demand, either overshooting or undershooting it.

Could there be cases where overconfidence is so costly that we would predict higher profit from *random* product selection than from *human* product selection? On the one hand, random product choice is clearly detrimental because humans can generally choose products with higher demand, increasing the maximum profit portion of the profit function. On the other hand, Studies 1(a) and 1(b) illustrated that letting humans choose can lead to overproduction in chosen products, which yields costly overage and a higher mismatch cost portion of the profit function. As a consequence, there is a trade-off between these two dynamics. Specifically, when the cost of having quantity-demand mismatch is large relative to the benefits of choosing a higher-demand product from the set (i.e., when c_u and c_o are large relative to m), it may be possible that letting humans choose the product leads to worse profit performance.

By this logic, we predicted: (a) higher mismatch costs when individuals chose the product themselves versus random choice; (b) higher maximum profits when individuals chose the product themselves versus random choice; and (c) when c_o is large relative to

m, lower profits when individuals chose the product themselves versus random choice.

6.1. Methods

6.1.1. Design and Procedure. The experiment was similar to that in Study 1(a), with a couple of important changes. First, we consider a profit function with a large c_u and c_o relative to m (c_u and c_o three times larger than m). Second, we implement a new condition in which the product is randomly selected from the participant's set, as opposed to letting the participant choose.

Participants were randomly assigned to one of two conditions: Human chooses product or Chance chooses product, as shown in Table 1. As in Study 1(a), after reading the instructions, seven products were presented, each with four independent datapoints, which served as noisy signals of their true demand.

In the *Human Choice condition*, participants chose which product to launch, knowing that they would earn $m = \$0.01$ for each unit of true future demand of the product they chose.

In the *Chance Choice condition*, participants were shown the set of seven products and were informed that their company had decided to let Chance the Dog choose which product to sell. It was stated that "Chance the Dog is not smart. He simply chooses a product at random with equal chance." Participants were informed that they would earn \$0.01 for each unit of true future demand of the product chosen by Chance, as in the other condition (i.e., $m = \$0.01$). Participants then clicked to advance to the next screen, which stated "One moment please … Chance is randomly picking one of the products … " for five seconds before automatically advancing. Participants then were shown which product Chance had randomly selected.

Both conditions then moved on to the production Decision Phase. For the chosen product, participants determined a production decision. As previously, their goal was to produce a number of units as close to the true demand of the chosen product as possible. In this experiment, they would lose \$0.03 from their bonus for every unit that the number they produce was off from the true future demand, either high or low (i.e., $c_u = c_o = \$0.03$).

6.1.2. Participants. The sample consisted of 386 participants recruited from Mechanical Turk. Through targeted recruiting, we obtained a sample in which 94% of participants had at least a bachelor's degree. The sample was 51% female and 77% Caucasian, with mean age of 37 ($SD = 11.4$). We targeted a sample size of 200 per condition, and no analyses were conducted until final data collection was completed.

6.2. Results

6.2.1. Product Selection. The true demand for the selected product was 204.2 in the Human Choice condition ($SE = 0.69$) and 200.0 in the Chance Choice condition ($SE = 0.74$), a significant difference, $t(384) = 4.14$, $p < 0.001$. Therefore, people were statistically significantly better than random at selecting products with higher demand.

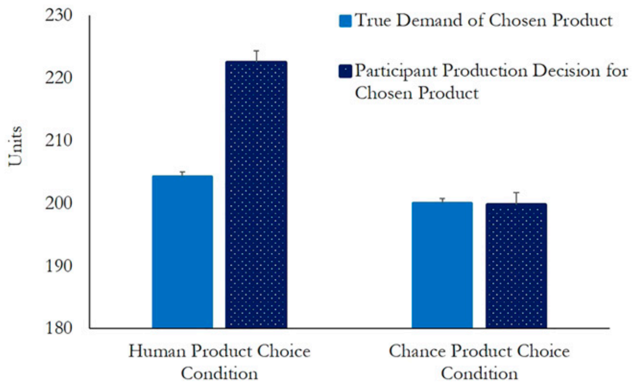
6.2.2. Production Decisions. The dependent variable is production error: production decision minus the true demand for the chosen product. There was a significant difference between conditions, $t(384) = 6.89$, $p < 0.001$ (see Figure 3). In the Human Choice condition, the production decisions were 18.21 units higher than the true demand on average ($SE = 1.86$), a significant bias toward overproduction, $t(198) = 9.77$, $p < 0.001$. In the Chance Choice condition, the production decisions were not significantly different from the true demand of the selected product on average ($M = -0.22, SE = 1.86$),$t(186) = 0.12$, $p = 0.91$. However, consistent with our theory, within the chance condition, if chance happened to select the product with the highest average signal, then the participants' production decisions were 19.52 units higher than the true demand on average ($SE = 4.03$), a significant bias toward overproduction, $t(29) = 4.84$, $p < 0.001$.

6.2.3. Profit. Figure 4 shows how product selection and product decisions combined to impact net profits in each condition. Human choice selects products with higher true demand than random choice, thereby achieving significantly higher maximum profits, $t(384) = 4.14$, $p < 0.001$. However, when the participants chose which product to launch, they subsequently overproduced units, whereas when Chance selected which product to launch, the participants' production decisions were unbiased. Consequently, consistent with our theory, the mismatch costs were higher when products were selected by humans than by Chance, $t(384) = 2.44$, $p = 0.015$. We found that these two forces net out such that participants with

Table 1. Study 2 Experimental Conditions

Human choice	Chance choice
Participants choose which product they want to produce and then decide a production quantity	A random process chooses which product to produce, and then the participant decides a production quantity

Figure 3. (Color online) Study 2 Production Decisions by Condition



random product selection earned more profit than participants who chose for themselves which product to launch, $t(384) = 1.62$, $p = 0.1$.

6.3. Study 2 Summary

This study illustrates how, under certain conditions, it is possible for the overconfidence effect to lead to such large overproduction in downstream operational decisions that it can outweigh the benefit of allowing the human to try to pick a good product in the first place.

7. Study 3: Information Complexity and Interpretation Error

In Section 3, we theorized that there are two distinct sources of noise that contribute to overconfidence in chosen products: information noise and interpretation noise (Assumption 1). In this experiment, we manipulate only the interpretation noise—holding the information noise constant. We test whether doing so is sufficient to exacerbate the bias toward overconfidence in chosen products, as our theory predicts. In this way, we test whether individuals sufficiently account for their own interpretation noise (Assumption 2(b)) above and beyond any adjustments they make for information noise (Assumption 2(a)).

7.1. Methods

7.1.1. Design and Procedure. This experiment featured a 2 × 2 mixed design. The first experimental manipulation, No Product Choice versus Product Choice, was a within-subject factor (each participant completed both conditions in a random order). The second experimental manipulation, low versus high information complexity, was a between-subjects factor (participants were randomly assigned to one information complexity condition). See Table 2 for a summary of these conditions.

The instructions and task were very similar to the preceding experiments, with two notable exceptions. First, rather than making production decisions, here, individuals simply estimated the true future demand of products with incentives for accuracy in their estimates. Given that our theorized mechanism is driven by a positive bias in the forecast for the chosen product, we wanted to ensure that the effect would emerge, even in the absence of subsequent production decisions.

More importantly, participants were randomly assigned to one of two information complexity conditions. These two conditions were identical in terms of the mathematical value of the information provided, but varied in how difficult it was to interpret the information. In the *Low information complexity condition*, for each product, participants observed one informative datapoint with distribution $N(\mu_i, 15)$, in the form of a single forecaster's guess of true demand. In the *High information complexity condition*, for each product, participants observed nine informative datapoints with distribution $N(\mu_i, 45)$, in the form of nine independent forecasters' guesses of true demand.

As in the previous experiments, all forecasts were explained to be completely independent across products and across forecasters. For a rational Bayesian decision maker, these two conditions are equivalent in that they contain the same information value. The information value contained in the High information complexity with nine independent forecasters is $\sigma^{info} = 45/\sqrt{9} = 15$, which is equal to that contained in the Low information complexity condition. The average of the nine datapoints in the High information complexity condition is a sufficient statistic for the signal (see Casella and Berger 2002).

Participants were informed that they would observe and use the professional forecasters' best guess of

Figure 4. (Color online) Descriptive Decomposition of Study 2 Profits by Condition

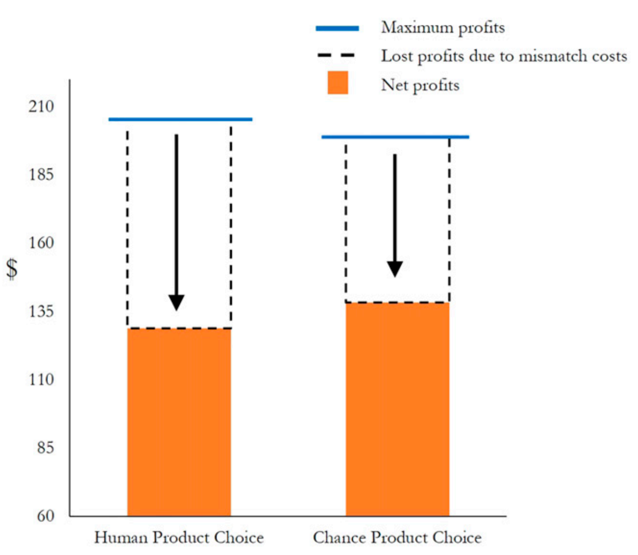


Table 2. Study 3 Experimental Conditions (2×2 Design)

	No product choice (n = 1)	Product choice (n = 7)
Low information complexity ($\sigma^{interpret}$ low)	Participants are shown only one product on its own and do not have any choice. Demand information is provided via a single data point, which requires little cognitive processing and leaves little/no room for interpretation error.	Participants are shown seven products and have to choose one single product. Demand information for each product is provided via a single data point, which requires little cognitive processing and leaves little/no room for interpretation error.
High information complexity ($\sigma^{interpret}$ high)	Participants are shown only one product on its own and do not have any choice. Demand information must be interpreted from nine less diagnostic data points, which, taken together, are equally informative as the single data point in the low information complexity condition. Requires more cognitive processing and leaves room for interpretation error.	Participants are shown seven products and have to choose one single product. Demand information for each product must be interpreted from nine less diagnostic data points, which, taken together, are equally informative as the single data point in the low information complexity condition. Requires more cognitive processing and leaves room for interpretation error.

each product's future true demand, in order to (i) pick the product they think would have the highest demand and/or (ii) forecast a product's future true demand. Participants then completed the tasks for the No Choice condition and the Choice condition in a random order.

In the *No Choice condition*, participants were shown forecasts for a single product and were asked to give their best guess of the true demand for that product. They received a bonus of 50 cents minus one cent for every unit their forecast was off. They completed five rounds of the No Choice condition, making an incentivized forecast in each.

In the *Choice condition*, participants were shown forecasts for seven products. They then needed to (i) choose the product that they thought had the highest true demand, receiving $1 if they chose correctly; and then (ii) give their best guess of the true demand for that product, receiving a bonus of 50 cents minus one cent for every unit they were off (min. of 0). They completed one round of the Choice condition.[6]

After all the rounds of the experiment were completed, participants were shown performance feedback for each round: their estimate and the true demand for each.

7.1.2. Participants. We recruited subjects from a survey panel of managers via ROI Rocket, as in the replication of Study 1(a) (see appendix). The subject pool was individuals currently in management positions that were full-time employed, U.S. citizens, and had at least a bachelor's degree. We received 478 participants after targeting 100 participants per condition; no analyses were conducted until data collection was complete.

7.2. Results

7.2.1. Manipulation Check: Information Complexity Generates Interpretation Noise. If our manipulation was successful, then we should see higher variance in forecast errors under high information complexity compared with under low information complexity. Looking within the No Choice condition (where participants see only one product at a time), for each participant, we computed the standard deviation of their five forecast errors, which is an estimate for $\sqrt{\sigma^{info^2} + \sigma^{interpret^2}}$. As expected, the average standard deviation of participant forecast errors was higher in the high information complexity condition ($M = 20.33, SE = 0.68$) than in the low information complexity condition (M = 17.02, SE = 0.62), $t(476) = 3.61$, $p <$ 0.001. This difference provides evidence supporting a successful manipulation of interpretation noise because the information noise was constant between conditions ($\sigma^{info} = 15$). Assuming independence, these estimates imply that $\sigma^{interpret}$ is about 13.7 in the high information complexity condition and about 8.0 in the low information complexity condition.[7]

7.2.2. Information Complexity Moderates Overconfidence in Chosen Products. Each forecast by a participant was an observation, and we clustered standard errors by participant to account for the repeated measures. The dependent variable was the forecast relative to the true demand for a given product. Positive values represented overestimation, and negative values represented underestimation. The independent variables were a dummy variable for choice condition—Choice versus No Choice—a dummy variable for information complexity condition—High versus Low—and the interaction between the two. See Figure D.1 in Appendix D for histograms by condition.

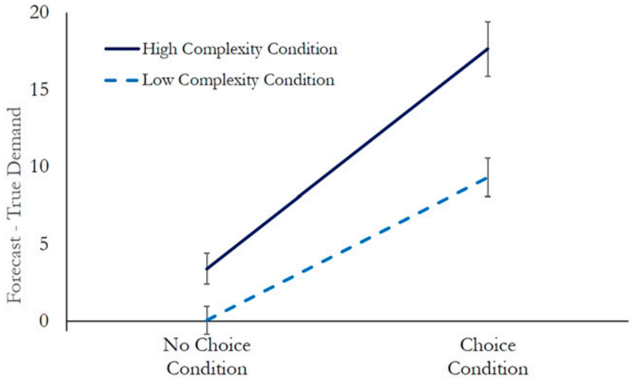
There was a significant interaction between information complexity and choice, $t(477) = 2.17$, p = 0.02. Consistent with our theory, the effect of choice on overconfidence was larger with high information complexity than with low information complexity (see Figure 5). Simple effects tests reveal that with low complexity, forecasts with choice were 9.26 units higher than forecasts with no choice (SE = 1.25), $t(477) = 7.41$, $p < 0.001$. With high complexity, forecasts with choice were 14.25 units higher than forecasts with no choice (SE = 2.17), $t(477) = 8.04$, $p < 0.001$. The same results hold with subject fixed effects included in the model. We note a small positive bias in the No Choice-High Information Complexity cell of the 2 × 2 study design that was not predicted by our theory (see Figure 5).

Finally, we verify that a perfectly rational decision maker, given the exact same information as the participants, would have made unbiased forecasts relative to the true demands, $M = -0.07$, $SE = 0.38$, $t(477) = 0.18$, $p = 0.85$. Its average difference from truth is unaffected by information complexity, $t(476) = 1.43$, $p = 0.15$.

7.2.3. Comparison with Information Signals. Comparisons to the information signal provide another test of whether subjects are able to sufficiently make a statistical correction for their own interpretation noise (Assumption 2(b)) above and beyond any correction, or lack thereof, for the information noise (Assumption 2(a)). Namely, Assumption 2(b) implies that (1) interpretation noise can cause forecasts for chosen products to be even higher than the information signal (which is already biased high due to the selection process); and (2) greater interpretation noise leads to an even larger positive bias relative to the information signal. These predictions require Assumption 2(b); they cannot be derived under Assumption 2(a) only.

Our results support these predictions. Under high information complexity, forecasts of a chosen alternative were significantly higher than its associated information signal, $M = +9.18$, $SE = 1.90$, $t(240) = 4.83$, $p < 0.001$, and this gap was significantly larger with high information complexity than with low information complexity, $t(476) = 2.91$, $p = 0.003$. Under low information complexity, there was a much smaller, only marginally significant difference between forecasts and the information signal, $M = +2.23$, $SE = 1.40$, $t(236) = 1.63$, $p = 0.10$.[8] These findings further suggest that interpretation noise contributes to overconfidence above and beyond information noise.

Figure 5. (Color online) Study 3 Forecasts by Condition



7.3. Study 3 Summary

We found that, whereas forecasts of any single product were relatively unbiased, forecasts of chosen products were significantly positively biased. When the noisy objective signal came with more information complexity (introducing more interpretation error), participants added more interpretation noise in their assessment of each product, leading to forecasts of chosen products that were more positively biased. In fact, individuals' forecasts were biased even above the information signal, which is strong evidence that individuals are unable to make a statistical correction to account for their own interpretation noise.

8. Study 4: Mitigating Overestimation via Organizational Structure

This experiment differs from the first four in two primary ways. First, it is designed to test the mitigation strategy motivated by Proposition 2: Getting a second independent forecast for a chosen product can significantly mitigate the overconfidence problem. Reflecting on the results of Study 2, such a strategy is desirable because it mitigates the mismatch cost under human choice without resorting to random choice (i.e., while retaining the maximum profit benefits of human choice). Recall that the key driver of this prediction is in the interpretation noise (Assumption 1) manipulated in Study 3—two people can look at the same information and make person-specific, idiosyncratic random errors. An independent person adds no value if overconfidence is driven only by information noise and people have access to the same information.

Second, we vary the look and feel of this experimental task. Our first four experiments built on one another using the same task set-up. Although it is helpful to demonstrate how an effect reliably emerges within a given task format, it is also helpful to show the generalizability of the phenomenon to other formats. In this experiment, although the underlying mathematical structure of the set-up remains the same, we provide the information datapoints by showing line graphs over time, as opposed to "expert opinions," as in the previous experiments. In this way, we simulate the common managerial experience of examining plots of preliminary data to make inferences and projections about future outcomes under ambiguity and uncertainty—realistic features that can induce interpretation noise in practice.

8.1. Methods

8.1.1. Design and Procedure. Participants were assigned, in an alternating fashion, to one of two conditions (see Table 3). The first participant was assigned to the *Choice condition*, in which she examined plots of preliminary demand data for six products, selected the one she expected to have the highest demand next period, and then made an inventory decision for that product. The second participant was then assigned to the *Independent condition* and was unknowingly yoked to the first participant. This participant was shown the same six products, but was simply asked to make an inventory decision only for the particular product that the aforementioned participant had selected. This process was repeated for 20 rounds.[9]

In a given round in the *Choice condition*, participants were presented with six products under consideration, and for each of these products, they were shown a line graph of preliminary demand data from the first nine periods (see an example in Figure 6). Participants attempted to select the product that they believed would have the highest demand next period because doing so would enable them to achieve more profit.

In contrast to our previous experiments, participants were not explicitly told how the demands were generated. This ambiguity mimics much of reality; a manager often observes a plot of preliminary performance, but cannot be sure about the exact underlying data-generating process. Of course, this design choice means that the researcher cannot clearly define normative behavior in this study, in contrast to the previous studies. However, it does not detract from our ability to compare behavior between conditions, which is the main focus of this study. The environment was, in fact, quite simple. Each product had a unique mean demand, which was randomly selected $\mu_i \sim N(100, 10)$. Then, for a given product, its demand in each period was an independent draw from $N(\mu_i, 30)$.

After selecting a product to sell, participants were shown the inventory decision screen, on which the other five graphs remained visible, but the selected product's graph was highlighted. The participant was then prompted to make an inventory quantity decision for the chosen product. Inventory decisions were incentivized such that the individual earned \$1 times demand minus an overage penalty \$1 per unit of inventory that exceeded demand; and \$1 for each unit the inventory fell short of demand. These values were an in-game currency that was later translated proportionally into bonuses that ranged from \$1 to \$10. As in previous experiments, these parameters incentivize one to try to choose the product that has the highest forecasted demand for the next period and to make an inventory decision equal to the forecast.

Between every round, participants received outcome feedback: They observed the realized demand for the product they selected and the consequent profit and personal earnings they achieved in that round.

The *Independent condition* differed from the Choice condition only in the following ways. Participants did not make product selections, but instead made independent inventory decisions for the products selected by participants in the Choice condition. This process was achieved by assigning individuals to a condition in an alternating manner. For example, the first participant was assigned to the Choice condition, in which she chose which product to launch and then decided how many units of that product to procure (repeated for 20 rounds). The second participant was then assigned to the Independent condition and simply made inventory decisions for each of the 20 products selected by the first participant (without deciding which product to launch). In this way, subjects were paired across conditions, and both individuals in a given pair made inventory decisions for an identical series of products. The information on the screen at the time of the inventory decision was identical across the two conditions. Participants were not informed of this linkage across participants. In the Independent condition, they were simply informed that they had been tasked with making inventory decisions for particular products.

8.1.2. Participants. Participants were undergraduate and graduate students at a major U.S. university, who signed up through an online recruitment system. Given that there are 20 observations per participant in this study, we targeted a sample size of 50 participants per condition, resulting in a total of 106 participants. No analyses were conducted until data collection was complete. Among these, 93% were full-time students, and 66% worked part-time or full-time jobs (74% female and 39% non-White). They received \$5 for participation and could earn up to \$10 more (the bonus was proportional to the profits they accumulated in the game via their inventory decisions). The average total earnings per participant was \$12.94.

8.2. Results

We examined whether the inventory decisions were systematically higher in the Choice condition than in the Independent condition (see Figure 7). In the following series of regression models, the dependent variable is the inventory decision, with standard errors clustered by participant. As predicted, in a univariate regression, inventory decisions were significantly higher in the Choice condition than in the Independent condition, $\beta = 6.89$, $SE = 2.11$, $t(105) = 3.27$, $p = 0.001$. With round and an interaction between round and experimental condition included in the model, there was no interaction between round and condition, $\beta = 0.008$, $SE = 0.17$, $t(105) = 0.05$, $p = 0.96$,

Table 3. Study 4 Experimental Conditions

Choice	Independent
Participants are shown six products and have to choose one single product. The same participant then makes an inventory decision for that chosen product.	Each participant is shown the *same* six products as a participant in the Choice condition, but take the choice made by their matched participant in the *choice* condition as exogenously given. They then make their own independent inventory decision for that product.

and the effect of experimental condition remained, $\beta = 6.89$, $SE = 2.11$, $t(105) = 3.27$, $p = 0.001$.[10] Therefore, we do not see evidence that individuals can readily learn to reduce this bias from experience with round-by-round feedback.

Then, including product fixed effects in the model, inventory decisions remained significantly higher in the Choice condition than in the Independent condition, $\beta = 6.89$, $SE = 1.58$, $t(105) = 4.37$, $p < 0.001$. In other words, even for a given product (and its associated graph), one should expect the inventory decision to be 6.89 units higher if that product was chosen by the same decision maker than if it was not.

We can alternatively include "pair" fixed effects in the model to test the effect of condition within each pair that was yoked across conditions by design. Recall that the two individuals in each yoked pair ultimately saw the same exact series of products for which to make decisions. Again, we find significantly higher inventory decisions in the Choice condition than in the Independent condition, $\beta = 6.89$, $SE = 1.10$, $t(105) = 6.25$, $p < 0.001$. All of these results hold if we remove round and the interaction between round and condition from the model, reflecting the stability of the effect. Therefore, we do not find evidence that the effect grows or diminishes over time.

8.3. Study 4 Summary

Inventory decisions were systematically higher when the individual both chose which product to sell and made the investment decision for the chosen product than when the investment decision was made by an independent person from the product choice process. Notably, the significant difference between these cases existed, even though participants were paired across conditions such that they made inventory decisions

Figure 6. (Color online) Study 4 Screenshot

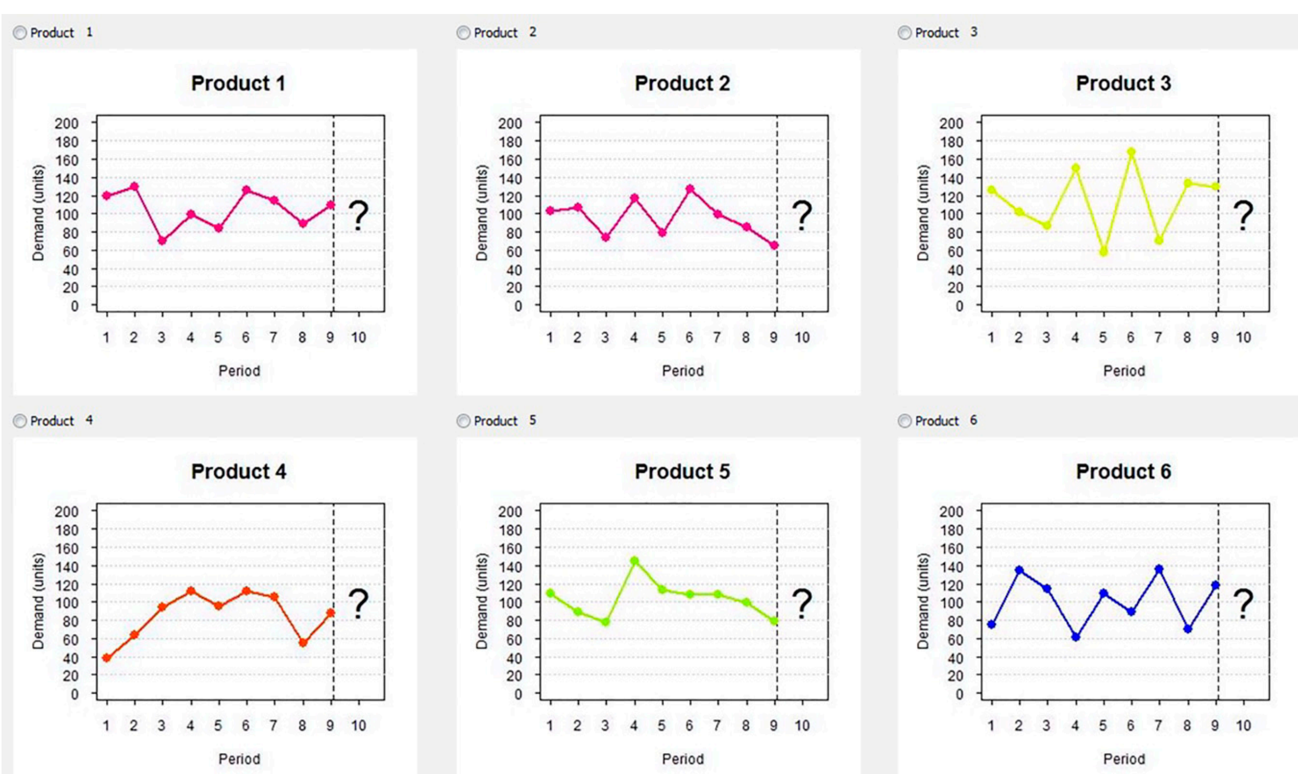
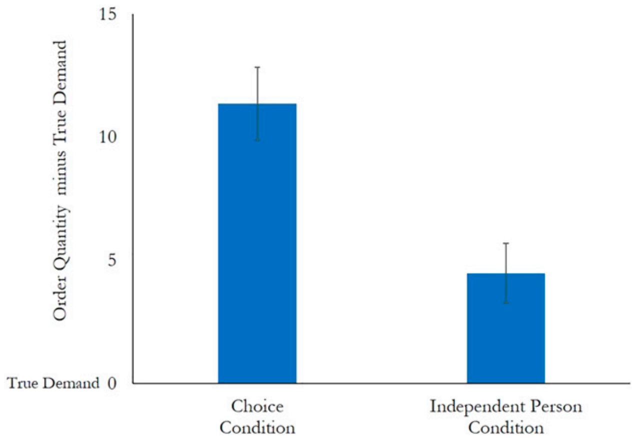


Figure 7. (Color online) Study 4 Inventory Decisions by Condition



for the exact same products, based on the exact same preliminary demand plots. These results suggest that one can reduce the overconfidence problem driven by the interpretation noise documented in Study 3 without resorting to pure random choice as in Study 2. The fact that the order quantity was higher than the true demand, on average, in the Independent Person condition is broadly consistent with the notion that there also existed some information noise in the graph itself that affected both the product chooser and the independent judge.

9. General Discussion

In this paper, we studied decision behavior in a common combination of tasks for new products: selection, forecasting, and production. Across five experiments, we found that people overweight their noisy interpretations of imperfect information signals while selecting the new products for which they have the most favorable forecast. As a consequence, the products they choose to launch tend to feature a positively biased behavioral forecast, which can have significant performance consequences for downstream operational decisions. In certain cases, this form of overconfidence can be so severe that it offsets the benefit of human product selection (versus random selection) in the first place. We provided mechanism evidence by manipulating the interpretation noise through information complexity. Even when the information is equivalent from a Bayesian perspective, more complicated information leads to more forecast noise, which, in turn, leads to more overconfidence in the chosen products. Finally, we leveraged this insight to show that getting a second independent forecast for a chosen product can significantly mitigate the overconfidence problem, even when both individuals have the same information.

By deepening our understanding of this type of overconfidence for new products, our paper identifies when it is most likely to be an important managerial issue to address. Our results suggest that this type of overconfidence is likely to be most problematic when (1) there are a large number of product ideas to choose from (large n); (2) information for products is limited and particularly noisy (σ^{info} is large); (3) information richness and complexity is high, leaving considerable room for human interpretation ($\sigma^{interpret}$ is large); (4) the downstream operational cost of a biased-high forecast is large; and (5) the organizational process is one under which a person's forecast both determines which product gets launched and is also used to calibrate downstream operational decisions for that product.

9.1. Mitigation Strategies

9.1.1. Altering the Organization's Task Structure. Our results highlight the possibility of mitigation through modifying organizational structure. Specifically, because idiosyncratic interpretation noise is a driver of the overconfidence bias, getting an independent forecast from another person for the chosen product can help mitigate the effect. A mitigation strategy that works by modifying the process rather than the people is appealing because it may be more reliable and cost-effective than trying to move each individual person's thinking toward perfect rationality. In Study 4, we reduced bias by getting one more person's judgment for one product, for a total of two people and $n + 1$ product judgments. A natural question, then, is: Given organizational constraints of people, resources, and time for making judgments, what is the optimal way to organize people across the forecast-selection-production task combination to maximize value? We view this as a promising direction for future work.

9.1.2. Altering the Individual's Decision Process. Our results suggest that some established mitigation strategies aimed at improving the cognitive judgments of individuals may also help reduce the overconfidence problem we identify. Namely, we predict that those strategies that help reduce noise or address statistical naivety should also help to mitigate the overconfidence problem in chosen products. There are several behavioral strategies for reducing noise, such as averaging the judgments of multiple people to capture the "wisdom of the crowd" (e.g., Mannes et al. 2014) or even averaging the judgments of "the crowd within" one person (Herzog and Hertwig 2009) have proven effective. Similarly, strategies that aim to improve probabilistic judgment through "kind" information representation (Soll et al. 2015) may help reduce the statistical naivety problem. For example, Barbey and Sloman (2007) find that they can reduce base-rate neglect when they present probabilities in terms of nested sets of individuals. The development of similar visual representation strategies in a continuous domain is a potentially fruitful area of future research. Statistical modeling approaches to enforce a Bayesian adjustment may also hold promise. However, repeatedly eliciting subjective priors of sufficient quality—which is required for the Bayesian solution—is generally difficult for practical implementation (Yelland et al. 2010).

Standardizing a decision maker's process across products may also help mitigate the mechanism driving overconfidence in this paper. For example, a manager's forecast may differ depending on whether they get the relevant information shown in a spreadsheet versus presented to them in a PowerPoint. If instituting a standardized process can reduce the chance that the product choice was affected by the *way* the manager happened to analyze the product's information, then it can reduce the overconfidence identified here. Interestingly, such a standardized process should help, even if it only increases the correlation in interpretation random errors for that person, but still does not improve the average accuracy of her judgments. The intuition is that if a person's interpretation random errors are perfectly correlated across products, then the errors will not affect the person's selection between products, thereby circumventing the selection-driven overconfidence.

9.2. Limitations and Concluding Remarks

In this paper, we derived a behavioral theory through formal modeling and tested it via experiments. Doing so enables precision and internal validity, but it certainly does not fully capture the richness and scope of the situational factors that a manager may face in practice. Future work could explore how idea generation for new products informs the forecasting and selection steps, an element of new product development that we notably do not study here. Likewise, we examine a process with a single selection step and a single selected product, as opposed to multistep screening and multiproduct launch. Future work could examine optimal stage-gate processes given the behavioral biases we document, as well as how forecasting behavior changes when multiple products are selected for launch.

All told, it is clear that costly overconfidence can emerge as a product of how uncertainty, selection, and production interplay. This suggests that by studying how tasks interact within operational processes to generate behavioral biases, we can better understand the seemingly immutable impulse toward overconfidence that managers seem to display in their expectations for the courses of action that they have chosen to take.

Appendix A. Proposition Proofs

Proof of Proposition 1. Define $(\sigma^{forecast})^2 = (\sigma^{info})^2 + (\sigma^{interpret})^2$. For any product i and person j, we can think of $F_j^b(i)$ as a signal with which one can make a Bayesian update to the prior distribution of μ_i. Specifically, we have

$$E[\mu_{i_j^b} \mid F_j^b(i)] = \frac{(\sigma^{means})^2(\sigma^{forecast})^2}{(\sigma^{means})^2 + (\sigma^{forecast})^2}\left(\frac{\mu^{means}}{(\sigma^{means})^2} + \frac{F_j^b(i)}{(\sigma^{forecast})^2}\right),$$

(see Casella and Berger 2002). Because the expected value operator is linear, with some algebra, we can show that

$$\begin{aligned} E[F_j^b(i_j^b) - \mu_{i_j^b}] &= E[F_j^b(i_j^b)] \\ &\quad - \frac{(\sigma^{means})^2(\sigma^{forecast})^2}{(\sigma^{means})^2 + (\sigma^{forecast})^2}\left(\frac{\mu^{means}}{(\sigma^{means})^2} + \frac{E[F_j^b(i_j^b)]}{(\sigma^{forecast})^2}\right) \\ &= \left(E[F_j^b(i_j^b)] - \mu^{means}\right)\frac{(\sigma^{forecast})^2}{(\sigma^{means})^2 + (\sigma^{forecast})^2}. \end{aligned}$$

Now, $E[F_j^b(i)] > \mu^{means}$ so long as $n > 1$, so $E[F_j^b(i_j^b) - \mu_{i^b}] > 0$. □

Proof of Proposition 2. By definition, $\hat{j}$'s forecast for j's product choice is no greater than $\hat{j}$'s forecast for the product $\hat{j}$ would choose—that is, $F_{\hat{j}}^b(i_j^b) \le F_{\hat{j}}^b(i_{\hat{j}^b})$. In expectation, this inequality is strict if $\sigma^{interpret} > 0$—that is, $E[F_{\hat{j}}^b(i_j^b)] \le E[F_{\hat{j}}^b(i_{\hat{j}^b})]$. Finally, because individuals are homogeneous, $E[F_{\hat{j}}^b(i_{\hat{j}^b})] = E[F_j^b(i_j^b)]$. Thus, $E[F_{\hat{j}}^b(i_j^b)] < E[F_j^b(i_j^b)]$. □

Appendix B. Bayesian Benchmark

A Bayesian recognizes that their interpretation of the signal, $S_j^b(i)$ is a noisy signal of the true demand mean μ_i. Therefore, instead of directly using it as their forecast, they use it to update their belief about the prior. Specifically, the Bayesian benchmark is defined as follows:

$$\begin{aligned} F_j^*(i) &= E[\mu_i \mid S_j^b(i)] \\ &= \frac{(\sigma^{means})^2(\sigma^{forecast})^2}{(\sigma^{means})^2 + (\sigma^{forecast})^2}\left(\frac{\mu^{means}}{(\sigma^{means})^2} + \frac{S_j^b(i)}{(\sigma^{forecast})^2}\right), \end{aligned}$$

where we define $(\sigma^{forecast})^2 = (\sigma^{info})^2 + (\sigma^{interpret})^2$. (We again refer the reader to Casella and Berger 2002.)

Note that the Bayesian is mean-regressing, and more so when $\sigma^{interpret}$ is larger. When we refer to the "perfectly rational" Bayesian benchmark in our studies, we assume $\sigma^{interpret} = 0$.

Appendix C. Replication of Study 1(a) with Manager Sample

To test robustness of the effect, we replicated Study 1a with a sample of managers.

C.1. Methods

C.1.1. Design and Procedures. Following Study 1(a), participants were randomly assigned to one of two conditions. In the *Choice condition*, participants were presented with seven possible products, from which one would be chosen to launch. In the *No Choice condition*, participants were shown only one product (see Study 1(a) for complete details).

C.1.2. Participants. The sample consisted of 415 managers recruited through an online survey company, ROI Rocket. Participants were U.S. citizens, full-time employed in management positions, and had at least a bachelor's degree. In exchange for participating in the five- to 15-minute survey, individuals received \$5 to \$9 based on performance. We targeted a sample size of 200 per condition, and no analyses were conducted until final data collection was completed. The sample was 49% female and 27% non-White, and the average age was 45.8 ($SD = 11.4$).

C.2. Results

The dependent variable is production error: the production decision minus the true demand for the product. The results fully replicate those of Study 1(a).

In the Choice condition, the number of units produced was on average +18.22 units relative to the true demand ($SE = 2.41$) versus only +5.83 units in the No Choice condition ($SE = 1.88$), a significant difference between conditions, $t(413) = 4.05$, $p < 0.001$. In both conditions, participants produced more units than perfectly rational Bayesian decision makers would have produced, given the same forecast datapoints ($p < 0.001$ in Choice and $p = 0.003$ in No Choice); however, this difference was 12.5 units larger in the Choice condition than in the No Choice condition, $t(413) = 4.29$, $p < 0.001$. Perfectly rational Bayesian decisions were not significantly different from the true demand in the Choice condition ($M = -0.34$, $SE = 0.59$, $t(208) = 0.58$, $p = 0.56$) or in the No Choice condition ($M = -0.22$, $SE = 0.54$, $t(205) = 0.41$, $p = 0.68$).

Appendix D. Supplementary Figure

Figure D.1. (Color online) Study 3: Histograms of Forecasts by Condition

	Low Complexity Condition		High Complexity Condition	
	No Choice Condition	Choice Condition	No Choice Condition	Choice Condition
+40 and over	5%	6%	8%	16%
+30 to +40	3%	7%	5%	12%
+20 to +30	5%	11%	8%	19%
+10 to +20	14%	25%	15%	11%
-10 to +10	42%	37%	35%	29%
-10 to -20	15%	8%	12%	7%
-20 to -30	9%	3%	8%	3%
-30 to -40	3%	1%	4%	2%
-40 and below	4%	2%	4%	2%

Cells contain percentage of responses in that condition that fell in the indicated range. Ranges are based on their estimates of mean demand relative to the actual mean demand.

Endnotes

[1] For long-standing products, initial forecasting and product selection are temporally distant from the ongoing operational decisions and typically are then made by different individuals. We also note that for products with a long demand history, organizations can much more easily employ automated forecasting algorithms.

[2] This result is similar in structure to the Winner's Curse (Kagel and Levin 1986). However, in the Winner's Curse, there is one good with multiple individuals competing for it through bidding in an auction. Here, there is only one individual choosing between multiple alternatives.

[3] https://aspredicted.org/blind.php?x=qq9bi9.

[4] Note that we did not predict this small, but statistically significant, positive bias in the No Choice condition. However, this is exactly why implementing the No Choice case as our control condition was important. It accounts for any baseline of bias and enables us to test for our predicted dynamic over and above that baseline.

[5] Newsvendor models often assume $m = c_u$. However, in general, m may be larger or smaller than c_u (e.g., c_u could capture backorder costs, procurement from a third party, or loss of goodwill). For our purposes, the key comparison is between c_o and m.

[6] As in the previous experiments, each product was a random draw from the product distribution. Then, each forecast shown to the participant was a random draw centered at that true value with standard deviation according to experimental condition, as specified above.

[7] Similarly, the root-mean-square error for participants in the No Choice condition were higher in the high information complexity condition ($M = 22.76$, $SE = 0.80$) than in the low information complexity condition ($M = 19.23$, $SE = 0.74$), $t(476) = 3.22$, $p = 0.001$.

[8] Consistent with our theory, the average signal of a chosen alternative was significantly positively biased, $M = 7.75$, $SE = 0.65$, $t(477) = 11.85$, $p < 0.001$. This signal bias was unaffected by information complexity, $t(476) = 1.11$, $p = 0.27$.

[9] The study was programmed in Delphi and conducted in a behavioral laboratory on computers (see the supplemental material for screenshots of the interface).

[10] The virtual identicality of the effect of condition is a product of the near-zero beta coefficient on the interaction.

References

Barbey AK, Sloman SA (2007) Base-rate respect: From ecological rationality to dual processes. *Behav. Brain Sci.* 30(3):241–254.

Bazerman MH, Moore DA (2012) *Judgment in Managerial Decision Making*, 8th ed. (John Wiley & Sons, Hoboken, NJ).

Becker-Peth M, Thonemann UW (2018) Behavioral inventory decisions: The newsvendor and other inventory settings. Donohue K, Katok E, Leider S, eds. *The Handbook of Behavioral Operations, Wiley Series in Operations and Management Science* (John Wiley & Sons, Hoboken, NJ), 393–432.

Brown KC (1974) A note on the apparent bias of net revenue estimates for capital investment projects. *J. Finance* 29(4):1215–1216.

Casella G, Berger RL (2002) *Statistical Inference*, vol. 2 (Duxbury, Pacific Grove, CA).

Eddy DM (1982) Probabilistic reasoning in clinical medicine: Problems and opportunities. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 249–267.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over-and underconfidence: The role of error in judgment processes. *Psych. Rev.* 101(3):519–527.

Feiler DC, Tong JD, Larrick RP (2013) Biased judgment in censored environments. *Management Sci.* 59(3):573–591.

Flicker B (2019) Managerial insight and "optimal" algorithms. Working paper, University of South Carolina, Columbia, SC.

Gartner WB, Thomas RJ (1993) Factors affecting new product forecasting accuracy in new firms. *J. Product Innovation Management* 10(1):35–52.

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Management Sci.* 56(4):591–605.

Grushka-Cockayne Y, Erat S, Wooten J (2018) New product development and project management decisions. Donohue K, Katok E, Leider S, eds. *The Handbook of Behavioral Operations, Wiley Series in Operations and Management Science* (John Wiley & Sons, Hoboken, NJ), 367–392.

Hammond J (2016) Global supply chain management simulation v2. Harvard Business Publishing Education 8623-HTM-ENG, Boston.

Harrison JR, Bazerman MH (1995) Regression to the mean, expectation inflation, and the winner's curse in organizational contexts. Kramer RM, Messick DM, eds. *Negotiation as a Social Process: New Trends in Theory and Research* (Sage Publications, Inc., Thousand Oaks, CA), 69–94.

Harrison JM, Kreps DM (1978) Speculative investor behavior in a stock market with heterogeneous expectations. *Quart. J. Econom.* 92(2):323–336.

Harrison JR, March JG (1984) Decision making and postdecision surprises. *Admin. Sci. Quart.* 29(1):26–42.

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psych. Sci.* 20(2):231–237.

Hogarth RM, Karelaia N (2012) Entrepreneurial success and failure: Confidence and fallible judgment. *Organ. Sci.* 23(6):1733–1747.

Hogarth RM, Makridakis S (1981) Forecasting and planning: An evaluation. *Management Sci.* 27(2):115–138.

Kagan E, Leider S, Lovejoy WS (2018) Ideation–execution transition in product development: An experimental analysis. *Management Sci.* 64(5):2238–2262.

Kagel JH, Levin D (1986) The winner's curse and public information in common value auctions. *Amer. Econom. Rev.* 76(5):894–920.

Kahneman D (2011) *Thinking, Fast and Slow* (Macmillan, London).

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

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

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

Lee YS, Siemsen E (2017) Task decomposition and newsvendor decision making. *Management Sci.* 63(10):3226–3245.

Li M, Petruzzi NC, Zhang J (2017) Overconfident competing newsvendors. *Management Sci.* 63(8):2637–2646.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107(2):276–299.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psych. Rev.* 115(2):502–517.

Petersen KJ, Handfield RB, Ragatz GL (2005) Supplier integration into new product development: Coordinating product, process and supply chain design. *J. Oper. Management* 23(3-4):371–388.

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

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Managemet Sci.* 46(3):404–420.

Siemsen E, Moritz B, Goodwin P (2018) Forecast decisions. *The Handbook of Behavioral Operations, Wiley Series in Operations and Management Science* (John Wiley & Sons, Hoboken, NJ), 433–458.

Simon M, Shrader RC (2012) Entrepreneurial actions and optimistic overconfidence: The role of motivated reasoning in new product introductions. *J. Bus. Venturing* 27(3):291–309.

Smith JE, Winkler RL (2006) The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science* 52(3):311–322.

Soll JB (1996) Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organ. Behav. Human Decision Processes* 65(2):117–137.

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*, vol. 2 (John Wiley and Sons, Chichester, UK), 924–951.

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

Tong J, Feiler D, Ivantsova A (2018) Good choice, bad judgment: How choice under uncertainty generates overoptimism. *Psych. Sci.* 29(2):254–265.

Tyebjee TT (1987) Behavioral biases in new product forecasting. *Internat. J. Forecasting* 3(3-4):393–404.

Urban GL, Weinberg BD, Hauser JR (1996) Premarket forecasting of really-new products. *J. Marketing* 60(1):47–60.

Van den Steen E (2004) Rational overoptimism (and other biases). *Amer. Econom. Rev.* 94(4):1141–1151.

Yelland PM, Kim S, Stratulate R (2010) A Bayesian model for sales forecasting at Sun Microsystems. *Interfaces* 40(2):118–129.