Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Overconfidence in Newsvendor Orders: An Experimental Study

Yufei Ren, Rachel Croson

MANAGEMENT SCIENCE
Vol. 59, No. 11, November 2013, pp. 2502–2517
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1715

Overconfidence in Newsvendor Orders: An Experimental Study

Yufei Ren
Department of Economics, Union College, Schenectady, New York 12308, reny@union.edu

Rachel Croson
College of Business, University of Texas at Arlington, Arlington, Texas 76019, croson@uta.edu

Previous studies have shown that individuals make suboptimal decisions in a variety of supply chain and inventory settings. We hypothesize that one cause is that individuals are overconfident (in particular, overprecise) in their estimation of order variation. Previous work has shown theoretically that underestimating the variance of demand causes orders to deviate from optimal in predictable ways. We provide two experiments supporting this theoretical link. In the first, we elicit the precision of each individual's beliefs and demonstrate that overprecision significantly correlates with order bias. We find that overprecision explains almost one-third of the observed ordering mistakes and that the effect of overprecision is robust to learning and other dynamic considerations. In the second, we introduce a new technique to exogenously reduce overprecision. We find that participants randomly assigned to this treatment demonstrate less overprecision and less biased orders than do those in a control group.

Key words: overconfidence; overprecision; newsvendor; experiment; behavioral operations management
History: Received October 6, 2010; accepted January 11, 2013, by Peter Wakker, decision analysis. Published online in *Articles in Advance* May 7, 2013.

1. Introduction/Motivation

One fundamental question in operations management is how individuals make inventory decisions. Theories of optimal ordering typically assume that decisions are made by unboundedly rational, profit-maximizing agents. In contrast, a long literature in psychology and behavioral economics demonstrates that individuals are likely to suffer from a number of biases, which may impact their decision-making process. The emerging field of experimental and behavioral operations management has collected evidence of the existence of these biases in operations management contexts. For example, experimental work has demonstrated that individuals deviate from optimal ordering patterns in multiechelon supply chain settings (Sterman 1989, Croson and Donohue 2006, Steckel et al. 2004, Loch and Wu 2008, Katok and Wu 2009).

This paper examines behavior in a simple newsvendor setting, which was first introduced by Edgeworth (1888) and which has served as the basis of many models of inventory management (for reviews, see Porteus 2002). Optimal ordering behavior is well known and easy to derive in this setting (Arrow et al. 1951) and involves balancing the expected cost of under- and overstocking.

Although this solution is well known and simple to calculate, previous studies have shown that individuals often deviate from it. Schweitzer and Cachon (2000) were the first to demonstrate systematic deviation from optimal orders. They identified consistent but asymmetric deviations; individuals order too little under high-profit conditions, in which the optimal order is higher than the mean of the demand distributions, and order too much under low-profit conditions, in which it is lower. This result has been replicated by numerous studies, including Bostian et al. (2008), Bolton et al. (2012), Benzion et al. (2008), Bolton and Katok (2008), Moritz et al. (2013), and Katok and Wu (2009).

However, we still lack an understanding of what *causes* these biased orders. Understanding the cause of the bias is critical to developing a solution (a debiasing technique). Previous authors have identified (but not fully tested) a number of alternative possible causes for the order bias. Schweitzer and Cachon (2000) show that the asymmetric bias cannot be caused by risk preferences (see also Keren and Pliskin 2006, Kanthen and Huy 2009), reflection effects from prospect theory, waste aversion, stockout aversion, or an underweighting of opportunity costs, but they leave anchoring and minimizing ex post inventory error as remaining possible causes. Su (2008) shows theoretically that the bias could be caused by "trembles" or noisy decision making, and Croson et al. (2011) show theoretically that the bias could be caused by overconfidence (in particular, overprecision). To our knowledge, ours is the first study that has attempted to experimentally test this explanation.

This paper presents two experiments to test the impact of overprecision on order bias. In the first experiment, we collect individual-level measures of overprecision. We present experimental subjects with a basic newsvendor problem and examine the relationship between their level of overprecision and their biased ordering behavior at the individual level. We find that overprecision has a robust and stable correlation with the level of biased orders even after controlling for other typically observed individual biases, such as anchoring and overplacement.

Although some evidence suggests that behavior improves with experience, many studies demonstrate that the order bias is persistent, even after learning (experience) and explicit training in the problem and its solution (e.g., Bolton and Katok 2008). Additional analysis in our first study demonstrates that the effect of overprecision on biased orders is robust to the inclusion of learning.

A second experiment makes two additional contributions. First, we use a new, in-task measure of overprecision in which participants estimate the demand distribution that they are facing. We find that in-task overprecision similarly correlates with order bias. Second, we implement a new technique, the subjective probability interval estimate (SPIES; Haran et al. 2010), which has been shown to exogenously reduce overprecision. We find that participants who are randomly assigned to the SPIES treatment demonstrate less overprecision and less biased orders than do those in a control group. Results from this experiment provide further evidence that overprecision is at least one contributing cause of biased orders in the newsvendor setting and demonstrate a potential debiasing technique that could be used in managerial settings.

2. Biased Newsvendor Orders

In the newsvendor problem, an inventory manager sells a product to customers at a price p per unit. The number of units demanded each day is drawn from a stationary and known distribution $D(\cdot)$. The manager orders his daily inventory, paying marginal cost c per unit, before knowing the realization of demand. Excess units can be salvaged at price s per unit but cannot be saved for the next day. The optimal order Q^* is generated from the critical fractile solution and is $Q^* = F_D^{-1}(\beta)$, where $\beta = (p-c)/(p-s)$ and $F_D^{-1}(\cdot)$ denotes the inverse cumulative distribution function (CDF) of the demand distribution.

The original paper demonstrating biased orders in this setting is Schweitzer and Cachon (2000). They recruited 34 MBA students who had been taught the newsvendor problem in class in the previous semester. Subjects faced 30 rounds of a single-period newsvendor problem with demand drawn from a uniform distribution (1, 300). Schweitzer and Cachon (2000) compared behavior in two different markets; high profit and low profit. The market price was always 12. In the high-profit market, marginal cost was 3, such that the optimal order was 225, or larger than the mean of the demand distribution at 150.5. In the low-profit market, marginal cost was 9, such that the optimal order was 75, or smaller than the mean of the demand distribution at 150.5. This paper was the first to demonstrate the asymmetric ordering bias. Participants ordered significantly less than optimal in the high-profit market and more than optimal in the low-profit market.

This result has since been replicated by numerous scholars. Bolton and Katok (2008) showed that feedback, learning, and experience can improve performance, yet the bias persists. Bolton et al. (2012) showed that although performance by professional inventory managers was slightly better than that of undergraduate students and slightly worse than that of graduate students, the bias persists in multiple subject pools. In Bolton et al. (2012), experimenters told subjects what the optimal order should be, and subjects still placed biased orders.

Evidence consistent with this bias has also been demonstrated in the field. Fisher and Raman (1996) examine the inventory decisions of a firm selling fashion skiwear. They find that inventory managers consistently ordered too little inventory and that profits would have increased by 60% had they ordered optimally. In contrast, Katok et al. (2001) examine the inventory decision of a firm selling maps. They find that inventory managers consistently ordered too much inventory, and an improved system saved a total of more than $800,000 per year. This evidence from the field supports the existence of an asymmetric order bias.

What might (or might not) cause this pattern of biased orders? Schweitzer and Cachon (2000) demonstrate that the pattern of order bias is not consistent with risk aversion, reflection effects from prospect theory, waste aversion, stockout aversion, or the underweighting of opportunity costs. However, many possible causes remain. Schweitzer and Cachon (2000) suggest that either anchoring (individuals anchor on the mean of the demand distribution and insufficiently adjust toward the optimal order) or the desire to minimize ex post inventory error can serve as an explanation for their results.

Croson et al. (2011) suggest a different explanation: overconfidence (and in particular, overprecision). Generally, individuals who are overprecise believe

Table 1 **Estimated γ from Several Newsvendor Experiments**

Authors	Demand distribution	High profit: $Q^* >$ mean of demand γ	Low profit: $Q^* <$ mean of demand γ
Schweitzer and Cachon (2000)	U(1, 300)	0.36	0.20
Benzion et al. (2008)[a]	U(1, 300)	0.37	0.20
Bolton and Katok (2008)	U(0, 100) in *H* condition U(50,150) in *L* condition	0.44	0.48
Bostian et al. (2008)	U(1, 100)	0.64	0.44
Benzion et al. (2008)[a]	N(150, 2,500)	0.22	0.08
Our study[b]	N(100, 900)	0.17	0.34

Source. Croson et al. (2011).
[a]Because of limitations on data availability, this uses data from the first and last 20 rounds.
[b]To ensure comparability with other studies that use a practice session before the experiment begins, we exclude the first five rounds from our data.

that their estimates are more accurate than they truly are (Moore and Healy 2008). In the newsvendor setting, this could lead to an underestimation of the variance of the demand distribution. In their model, true market demand is denoted as $D(\mu, \sigma^2)$. However, overprecise newsvendors believe demand to be $D_O(\cdot)$, a mean-preserving but variance-reducing transformation of the true consumer demand. They assume that $D_O(\cdot)$ is constructed by mixing the true demand distribution with a zero-variance distribution at the mean of demand $F(\mu, 0)$:

$$D_O = \gamma D + (1-\gamma)F(\mu, 0), \quad (0 \le \gamma \le 1).$$

Therefore, D_O shares the same mean as D but has a smaller variance ($\gamma^2\sigma^2$ rather than σ^2).

The parameter $(1-\gamma)$ thus measures an individual's level of overprecision. Newsvendors with γ equal to 1 ($(1-\gamma = 0)$ are perfectly calibrated (unbiased). Newsvendors with γ smaller than 1 ($(1-\gamma) > 0$) are overprecise; they believe that the distribution of demand is less variable than it truly is. Newsvendors with γ equals 0 ($(1-\gamma) = 1$) are infinitely overprecise, believing that demand is constant at its true mean.

Croson et al. (2011) assume that newsvendors optimize given their (possibly biased) beliefs about the demand distribution. They then compare the orders placed by overprecise newsvendors: Q_O^* with the optimal order Q^*. Their Proposition 1, reprinted below, shows that overprecise newsvendors underorder (relative to optimal) in high-profit conditions and overorder (relative to optimal) in low-profit conditions. This is exactly the pattern that had been observed in previous experiments.

Following their notation, let $F_S(X)$ be the CDF of $X(\cdot)$ such that $X(\cdot) = (D(\cdot) - \mu)/\sigma$. Let m be $F_D(\mu)$, the probability that the actual demand draw is less than expected demand, and β be $(p-c)/(p-s)$. They then show the following:

Proposition 1 (Croson et al. 2011). (a) *If* $\beta > m$, *then* $Q_O^* = \mu + \gamma\sigma F_S^{-1}(\beta) < Q^*$.
(b) *If* $\beta < m$, *then* $Q_O^* = \mu + \gamma\sigma F_S^{-1}(\beta) > Q^*$.
(c) *If* $\beta = m$, *then* $Q_O^* = Q^* = \mu$.

Thus in high-profit conditions, $\beta > m$, $F_S^{-1}(\beta) > 0$, and $Q_O^* < Q^*$. This demonstrates that overprecise newsvendors underorder in high-profit conditions. In low-profit conditions, $\beta < m$, $F_S^{-1}(\beta) < 0$, and $Q_O^* > Q^*$. This demonstrates that overprecise newsvendors overorder in low-profit conditions. This pattern is exactly consistent with that observed by the previous experimental literature.

Based on their model, Croson et al. (2011) use data from previous experiments to estimate the average level of overprecision exhibited by experimental participants. The first five rows of Table 1 are replicated from Croson et al. (2011) and describe the estimated levels of overprecision from four previous experiments. The last row is new to this paper and provides similar estimates using the data generated from our study.

Although this previous work is consistent with the idea that overprecision causes biased orders, this paper provides direct experimental evidence of this relationship. The two experiments we report were both conducted at the Center for Behavioral and Experimental Economic Science lab at the University of Texas at Dallas.[1] In our first experiment, we elicit each individual's levels of overprecision, along with other cognitive biases (e.g., anchoring, overplacement), and compare those levels with observed order bias. We find a significant and expected relationship between overprecision and biased orders that

[1] Of 272 original participants, 18 (6.6%) were excluded from these two studies based on their behavior during the experiment as recorded in the laboratory log, including inattention or violation of the laboratory rules during the experiment (e.g., talking on their cell phone, talking with other participants, etc.).

is robust to controlling for the other biases elicited. In our second experiment, we replicate this result with a different (in-task) measure of overprecision and randomly assign participants to a control or a treatment condition. The treatment condition uses a new technique (SPIES; Haran et al. 2010) that has been shown to exogenously reduce overprecision. We find that participants who are randomly assigned to this treatment demonstrate less overprecision and less biased orders than do those randomly assigned to the control group.

3. Study 1: Experimental Design and Descriptive Results

In the first study, before the experiment began, participants were asked to complete a computer based pre-survey in which we measured their level of overprecision, overplacement, experimenter-generated anchoring, and self-generated anchoring. At the end of the experiment, they were asked to complete a post-survey with socioeconomic questions. These measures will be described in more detail in §4. Appendix A provides a copy of the instructions for the interested reader.

In the newsvendor task, each participant faced 50 independent rounds of a single-period newsvendor game with a known demand distribution, normal with mean 100 and standard deviation 30. At the beginning of the experiment, participants were told the details of the demand distribution and shown several sample histograms and sample draws from this distribution. In each round, participants placed an order for a number of units of product with unit cost c, which varied between treatments. Then the computer randomly drew a demand realization from the demand distribution. Subjects sold their inventory up to that demanded amount and received a price of 10 for each unit sold. Any excess inventory could be returned to the manufacturer for a salvage value of 2 per unit. No inventory could be transferred to the next round. At the end of each round, the computer calculated the profit (loss) of that round and started a new round. Participants were provided with detailed feedback after each round at the bottom of their screen. All the parameters were known to the participants at all times. At the end of the session, participants were paid based on the total profits resulting from their decisions. Average earnings were \$15.40, and sessions lasted on average 40 minutes.[2]

The experiment involved two treatments, using a between-subject design, summarized in the first two columns of Table 2. All participants were in either a high-profit or a low-profit setting. In high-profit settings, the marginal cost of each unit of inventory was 4. The resulting optimal order is 120 units, higher than the mean of demand (100). In low-profit settings, the marginal cost of each unit of inventory was 8 while the other parameters remained the same. The resulting optimal order is 80 units, lower than the mean of demand (100). We will use the data to replicate the previous literature on asymmetric order bias and to examine the role of overprecision on biased orders.

Table 2 Parameters and Number of Participants

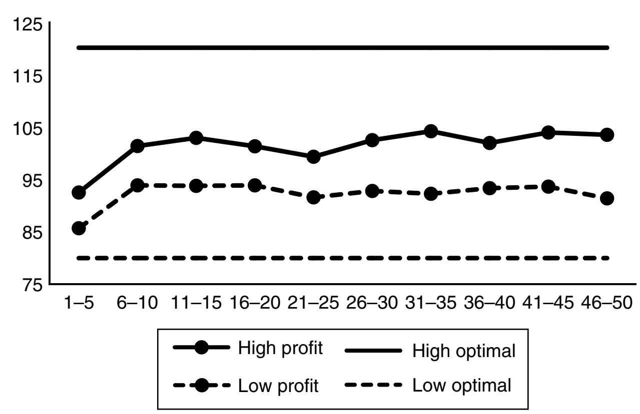
Treatment	Parameters	No. of participants	Average orders
High profit	Price = 10, cost = 4, salvage = 2 Q^* (optimal order) = 120	87	101.32** [22.47]
Low profit	Price = 10, cost = 8, salvage = 2 Q^* (optimal order) = 80	86	92.21** [19.39]

Notes. Standard deviation in brackets. t-test versus optimal.
$^{**}p < 0.01$.

The last two columns of Table 2 describe the number of participants in each treatment as well as the average [standard deviation] of newsvendor orders over the 50 rounds. Our results in aggregate replicate those of previous studies. Orders are too low in the high-profit treatment (101 versus 120) and too high in the low-profit treatment (92 versus 80), and these differences are statistically significant.

Figure 1 depicts average orders for the two treatments in five-round increments. The asymmetric bias of orders can clearly be seen. Although behavior adjusts somewhat over the 50 rounds, improvements are primarily seen between the first five rounds and the remaining rounds, and orders neither converge to nor diverge from optimum as the game progresses. A more detailed analysis of learning in this setting appears in §4.3.

Figure 1 Average Orders in High- and Low-Profit Conditions



[2] After this experiment, subjects were also asked to participate in another experiment. The time and earnings reported here represent the total from the two experiments.

These errors were costly. In the high-profit treatment, order bias reduced profit by 8.4% for 50 rounds (24,214 was actually earned versus the 26,424 that would have been earned had participants ordered optimally), and in the low-profit treatment, order bias reduced profit by 20.6% for 50 rounds (4,917 versus 6,192). These differences are statistically significant; a t-test comparing the distribution of profits against the optimal level yields $t = -10.20$, $p < 0.001$, $n = 87$ for the high-profit condition and $t = -14.50$, $p < 0.001$, $n = 86$ for the low-profit condition.

One anonymous referee worried that experimental participants might get bored in this 50-round task and that they might thus place orders that deviate from optimum to relieve this boredom. This is a reasonable and testable hypothesis. If individuals were figuring out the optimum order but then purposefully deviating from optimum to relieve boredom, we would expect to see orders that diverged from optimal over time. Our analysis in §4.3 finds the opposite effect, that order bias *decreases* over the 50 rounds of the game.

A second test for boredom involves comparing orders across individuals over time. If participants are purposefully diverging from optimal because of boredom, we should see orders of different participants being relatively close together in the early rounds of the game, and then diverging from each other in the later rounds as different individuals deviated either up or down to relieve their boredom. This would predict an increase in the *variation* of orders placed between subjects as the game progressed. We tested this hypothesis by calculating the variance of orders between subjects in each round and regressing this variance on round number. We find that the variance of orders between subjects is decreasing over time; in the high-profit treatment that variance is 1,477 in round 1 and 371 in round 50, and in the low-profit treatment, that variance is 1,201 in round 1 and 360 in round 50. This pattern is contrary to what would be predicted by the boredom hypothesis. For a more careful analysis, we regress the variance of orders between subjects in each round on the round number. The coefficients are significant and negative for both treatments ($\beta = -6.09$, $p < 0.001$, $n = 50$ for high-profit and $\beta = -5.76$, $p < 0.001$, $n = 50$ for low profit).

4. Study 1 Continued: The Impact of Overprecision

In the beginning of the experiment, participants were asked to complete tasks to measure their level of overprecision as well as their level of overplacement, experimentally generated anchoring, and self-generated anchoring. Appendix B depicts the instructions and instruments used. In this section we describe these measures briefly and demonstrate the robust relationship between overprecision and order bias, controlling for these other biases. We then examine learning in this setting and demonstrate (and calibrate) the robust relationship between overprecision and order bias, controlling for learning and other dynamics.

4.1. Controlling for Other Biases

4.1.1. Overprecision. Our variable of interest is overprecision. We use the classic overprecision measure, following Russo and Schoemaker (1990). This measure involves 10 general-knowledge questions and captures the extent to which individuals believe that their estimates are more accurate than they truly are. Participants are asked to respond with 90% confidence intervals for each question. Thus they are not asked to guess the answer but instead to offer a low and a high estimate of the answer such that the true answer will lie within their range 90% of the time. If individuals were well calibrated, the true answer should lie within their confidence intervals for 9 of the 10 questions. However, the robust finding from decades of previous research is that individuals exhibit overprecision in this task (Moore and Healy 2008), and thus the true answer lies within their confidence interval, on average, for 4 or 5 of the questions rather than 9.

For each participant, we calculate the number of questions for which the true answer lies in the range provided by the participant and subtract this number from 9. Thus a correctly calibrated participant would have an overprecision score of 0, whereas an extremely overprecise participant would have an overprecision score of 9. No participant was underprecise (which would have resulted in an overprecision score of -1). Thus higher numbers means more overprecision. Levels of overprecision as measured in this task have been shown to correlate with performance in financial markets (Biais et al. 2005). Deaves et al. (2009) similarly show that overprecision increases the likelihood of individuals to trade.

In addition to its predictive properties, previous research has demonstrated that an individual's level of overprecision is stable, both over time and across domains. For example, Jonsson and Allwood (2003) test subjects' overprecision level in two different domains (word knowledge and spatial knowledge) on three different occasions (each two weeks apart). Their results showed significant and stable correlation of a given subject's performance in each of the two domains and over the three elicitation times. Similarly, in Glaser et al. (2012), participants provided interval estimates for general-knowledge questions, economic and financial questions, and stock market predictions. Individual bias across these tasks was similarly stable across domains. Other papers that have demonstrated stability of overprecision over time and domains include Stankov and Crawford (1996), Crawford and Stankov (1996), Soll (1996), Thompson and Mason (1996), Stankov and Crawford (1997), Juslin and Olsson (1997), Stankov (1999), Bornstein and Zickafoose (1999), Klayman et al. (1999), Kleitman and Stankov (2001), and Pallier et al. (2002). Jonsson and Allwood (2003, p. 561), in reviewing this literature, note that "most research has pointed to stability."

4.1.2. Overplacement. In addition, we elicited three other measures of individual's biases. The first is overplacement (or the better than average effect). This captures the extent to which individuals believe that they are better than others, when they are not, and was measured with one simple question (Alicke and Govorun 2005). Participants were asked to "consider all the other participants in this experiment (not only those here today)." They were then asked to estimate "what percentage of them will make more money than you will?" We use the answer for this question as the measure of overplacement. Subjects with a smaller answer exhibit higher levels of overplacement.[3]

4.1.3. Anchoring. The second and third measures involved anchoring. Epley and Gilovich (2001) identify two types of anchoring: experimenter generated and self-generated. Our measure of these two types of anchoring involved four experimenter-anchored questions, two in which the true answer was higher than the anchor, and two in which the true answer was lower than the anchor, taken from Jacowitz and Kahneman (1995), along with four self-generated anchoring questions, two in which the true answer was higher than the anchor, and two in which the true answer was lower than the anchor, taken from Epley and Gilovich (2005). Our measure of experimenter-generated anchoring is determined by calculating the standardized anchoring level for each question: (true value − answer)/(true value − anchor), averaged within an individual across the four experimenter-anchored questions. A similar procedure generated the self-generated anchoring measure. Higher numbers indicate a greater influence of the anchor on an individual's response.

In previous behavioral operations management papers, researchers used regression techniques to examine whether participants were anchoring their orders on the mean (Schweitzer and Cachon 2000) or on the previous period's realization of demand (Katok and Wu 2009). Our measure does not capture either of these phenomena but instead captures an underlying personality trait, how much the individual is affected by anchors generally. In the analysis that follows, we will examine the effect of overprecision on order bias and examine the robustness of the effect in the presence of controls for overplacement and anchoring.

4.1.4. Overprecision Robustly Predicts Ordering Error. Our dependent measure is the difference between each individual's decision (Q_{it}) and the optimal order (Q_i^*) in each of the 50 rounds. For the low-profit condition (where orders are above optimal), we use this measure. For the high-profit condition (where orders are below optimal), we multiply this measure by −1 (switch the sign). This will be our primary dependent variable, referred to as individual error. We then regress this measure of individual error on our measure of overprecision and our other measures as controls (overplacement, self-generated anchoring, and experimenter-generated anchoring). All regressions also include a control for the high-profit condition (H), which equals 1 in the high profit condition and 0 otherwise. Our regression equation is thus

$$I_L(Q_{it} - Q^*) = \alpha + \Sigma\beta_i X_i + \phi H + \eta_i + \varepsilon_{it}, \qquad (1)$$

where I_L equals 1 in the low-profit condition and −1 in the high-profit condition. We have 50 observations for each individual; thus we use a random effects generalized least squares (GLS) regression to estimate Equation (1), where η_i captures unobserved individual heterogeneity.[4] We used STATA 12 to conduct all the analyses in this paper. Results are shown in Table 3.[5]

We find a positive and significant effect of overprecision on the level of individual error (column (1)), suggesting that more overprecise individuals exhibit greater order bias in the experiment. Columns (2)–(5) further test the robustness of our results by including both overprecision and the other biases (individually in columns (2)–(4) and together in column (5)). In all regressions, the coefficients on overprecision remain significant and of approximately the same magnitude. This analysis thus suggests that overprecision is significantly and robustly related to order bias.

[3] Benoit and Dubra (2011) suggest that overplacement may be a result of rational Bayesian updating rather than a cognitive bias per se.

[4] As we mentioned in §3, the variation of Q_{it} decreases over time. One might be concerned that this would bias our results. To check this, we used a mixed effect regression model, which assumes the standard deviation of the error term ε_{it} varies with time. Our results remain the same and are available from the authors. We thank an anonymous referee for suggesting this robustness check.

[5] These results and all those reported later in this paper are robust to including individual demographic characteristics such as gender, age, education, risk aversion, and number of math courses taken as controls in the regressions.

Table 3 Demonstrating the Impact of Overprecision and Its Robustness (Random Effects)

	(1)	(2)	(3)	(4)	(5)
Overprecision	0.722*	0.723*	0.721*	0.730*	0.728*
	[0.361]	[0.362]	[0.361]	[0.360]	[0.362]
Overplacement		−0.003			0.005
		[0.049]			[0.050]
Self-generated anchoring			−0.123		−0.120
			[0.184]		[0.185]
Experimenter-generated anchoring				0.006	0.007
				[0.005]	[0.005]
Constant	8.459**	8.606*	8.196**	8.846**	8.328*
	[2.280]	[3.655]	[2.318]	[2.294]	[3.659]
Control for treatment	Yes	Yes	Yes	Yes	Yes
N	8,650	8,650	8,650	8,650	8,650
R^2	0.031	0.031	0.032	0.034	0.035

Note. Standard errors in brackets.
$^{*}p < 0.05$; $^{**}p < 0.01$.

4.2. Calibrating the Size of the Effect

Although the analysis above demonstrates a robust relationship between overprecision and order bias, a more nuanced picture can be obtained with a more complicated analysis that allows us to calibrate the size of the effect of overprecision on order bias. We follow Benzion et al. (2008) and Bostian et al. (2008) by analyzing the following regression:

$$Q_{it} - \mu = (\alpha + \beta OP_i + \phi H)(Q_i^* - \mu) + \eta_i + \varepsilon_{it}. \quad (2)$$

The notation is similar to that of Equation (1): Q_{it} is individual i's order in round t; μ is the mean of the demand distribution; Q_i^* is the optimal order, given the conditions facing individual i; $Q_i^* - \mu$ is the difference between that optimal order and the mean of the demand distribution; and OP_i is individual i's level of overprecision.

This regression allows us to directly assess the effect of overprecision on accuracy. The parameter α captures the accuracy of the orders. If $\alpha = 1$, then participants are ordering the optimal amount. If $\alpha = 0$, then participants are ordering the mean of the demand distribution. This equation (setting $\beta = 0$) has been used by previous authors to demonstrate biased orders (Benzion et al. 2008, Bostian et al. 2008). We estimate this formalization (setting $\beta = 0$) using maximum likelihood estimation (MLE) with random effects to control for the panel structure of our data (assuming the existence of unobserved individual heterogeneity η_i) and report the results in Table 4, column (1).

We next relax our assumptions, estimating β to capture the effect of overprecision on order accuracy. If β is negative, then there is a negative relationship between the individual's level of overprecision and his order accuracy; more overprecision yields less accuracy. We report the result of the relaxed regression of Equation (2) in Table 4, column (2), estimating both α and β simultaneously. As before, we use MLE with random effects and control for the treatment (high versus low).

Table 4 Measuring the Size of the Effect (Random Effects)

	(1)	(2)
Accuracy (α)	0.390**	0.577**
	[0.065]	[0.113]
Overprecision (β)		−0.036*
		[0.018]
Control for treatment	Yes	Yes
N	8,650	8,650
Log likelihood	−37,214	−37,212

Notes. Standard errors in brackets. Results are robust to including overplacement, self-generated anchoring, and experimenter-generated anchoring as control variables.
$^{*}p < 0.05$; $^{**}p < 0.01$.

The results from this analysis confirm our previous findings. First, there exists order bias. In column (1), the coefficient α is significantly different from 1 ($t = -9.33$, $p < 0.001$). This effect is of a similar magnitude as in previous experimental work. Our coefficient on α is 0.39. Benzion et al. (2008) used a uniform demand distribution (1, 300) and estimated α in the range 0.20 (low profit) to 0.37 (high profit). In different data, using a normal demand distribution (150, 2,500), they estimated α in the range 0.08 (low profit) to 0.22 (high profit).[6] Bostian et al. (2008) used a uniform demand distribution (1, 100) and estimated α in the range 0.43 (low profit) to 0.65 (high profit).

Second, we again find a significant and negative relationship between the level of order bias and the individual's level of overprecision. In column (2), the coefficient β is significantly different from zero. Thus, as an individual's level of overprecision increases, her level of accuracy decreases. The results from column (2) in Table 4 allow us to calibrate the size of

[6] These results are based on the first and last 20 rounds.

the effect of overprecision on order bias. The coefficient β is statistically significantly negative, suggesting that the more overprecise the individual is, the less accurate ordering decisions he makes. In theory, our overprecision measure varies from -1 to 9. A perfectly calibrated respondent (overprecision of 0) would thus have an order accuracy of almost 0.6. Our average order accuracy is 0.4, suggesting that around one-third of the mistakes in ordering can be attributed to overprecision (as accuracy increases from 0.4 to 0.6, mistakes decrease from 0.6 to 0.4).

4.3. Learning

Previous work on the newsvendor has documented improved performance as the game progresses (Schweitzer and Cachon 2000, Bolton and Katok 2008), but even with extensive experience, biased orders remain (Bolton et al. 2012). We ask not whether individuals learn to order more accurately but whether they learn to overcome the effects of overprecision. In other words, if we allow for learning, does the effect of overprecision on biased orders remain?

To answer this question, we conduct an analysis controlling for time trends (time trend effect) and the difference between orders and the previous round's realized demand (chasing demand effect). As previous studies have suggested (Bostian et al. 2008, Katok and Wu 2009), this model captures the effects of insufficient adjustment and minimizes ex post inventory errors.[7] As in §4.2, we use MLE with random effects and control for the treatment. Our results from these analyses are consistent with previous studies on learning. They show that our participants slowly improve their order accuracy over time and chase previous demand. However, even in the presence of these intertemporal effects, overprecison remains a statistically significant influence on order accuracy.

As one anonymous referee suggested, we also test whether the effect of overprecision diminishes over time. We conduct the same analysis as in §4.1, using only observation from the last 10 rounds. The results are very similar to those previously observed; the coefficient on overprecision was nonsignificantly larger in the last 10 rounds than in the game overall. In summary, we find that individuals do learn in this task; their order bias reduces and order accuracy improves over time. However, overprecision continues to cause order errors, even in the presence of learning, and these effects do not decrease over time.

5. Study 2: Reducing Overprecision

The results from Study 1 demonstrate a robust relationship between overprecision and order bias, controlling for other biases and for learning. However, in that study, overprecision is measured rather than manipulated. In this second study, we use a newly developed technique to exogenously reduce overprecision in a random sample of experimental participants (SPIES; Haran et al. 2010). We then compare ordering behavior between the two treatments.[8] In addition to providing causal evidence of our hypothesized link between overprecision and order bias, the SPIES mechanism, if effective, can serve as a debiasing technique for overprecise managers.

In addition to the introduction of SPIES, in this experiment, we use a new, in-task measure of overprecision.[9] Instead of the typical 10-task elicitation of Russo and Schoemaker (1990), we ask participants in this experiment to estimate the median, 5th percentile, and 95th percentile of the demand distribution. This will provide our in-task estimate of overprecision. Below, we discuss each of these two facets of this second study.

5.1. Exogenously Reducing Overprecision: SPIES

Previous work has found overprecision to be an extremely robust bias; many researchers have attempted to eliminate it, with minimal success (e.g., Fischhoff 1982, McGraw et al. 2004, Alpert and Raiffa 1982, Block and Harper 1991). A few previous papers have demonstrated tools for reducing overprecision, including Lichtenstein and Fischhoff (1980) (training and feedback), Arkes et al. (1987) (discussion of responses with peers), and Koriat et al. (1980) (considering contrary evidence). Plous (1995) found that group judgments yielded less overprecision than individual judgments.[10]

More recently, Haran et al. (2010) have developed a promising method that significantly reduces overprecision within the individual. Their tool, SPIES, elicits likelihood estimates over the entire range of possibilities rather than simply asking for the 90% confidence internal. In the SPIES elicitation, participants are asked a question and are provided with a range of possible answers (e.g., less than 10, 11–40, 41–50, more than 51). They are then asked to provide probabilities for each of the possible answers. The intuition behind SPIES is that participants are forced to consider the

[7] See also Lau and Bearden (2013) and Kremer et al. (2010) for a current debate on the existence of these intertemporal effects.

[8] We thank our anonymous referees and associate editor for suggesting that we use an exogenous mechanism to further demonstrate the impact of overprecision on order accuracy.

[9] We thank our anonymous reviewers and associate editor for suggesting that we use an in-task measure of overprecision to further demonstrate the impact of overprecision on order accuracy.

[10] Other papers have identified interventions that can reduce the two other facets of overconfidence, overplacement and overoptimism, including Kahneman and Lovallo (1993) and Cooper et al. (1988) on taking an outsider's perspective and Larrick (2004) and Mussweiler et al. (2000) on considering the opposite.

tails of the probability distribution of their estimates, which decreases overprecision.

Haran et al. (2010) demonstrate that SPIES significantly reduces overprecision. Participants who use SPIES generate significantly more accurate 90% confidence intervals (88.35% accuracy rate) than do participants who directly estimate a 90% confidence interval (or directly provide 5% and 95% fractals; 73.79% accuracy rate). In our experiment, we will use SPIES as our treatment variable to exogenously reduce the level of overprecision, and we will investigate its impact on individual error.

5.2. An In-Task Measure of Overprecision

In Study 1, our measure of overprecision was based on the task of Russo and Schoemaker (1990). Our goal in that study was to find a task that was exogenous and unrelated to the newsvendor problem, in order to identify a stable individual trait of overprecision. However, the question remains whether overprecision as measured in that task is truly related to overprecision in the newsvendor task, particularly to participants' underestimation of the variation of the demand distribution.

In this study, therefore, we substitute a new, in task measure of overprecision directly related to estimates of the demand distribution that participants face. Responses to this measure of overprecision are again used to predict order bias. This second study thus provides a robustness check on Study 1.

We base our in-task measure on previous work in behavioral finance, in which experimental participants are asked to estimate median, lower, and upper bounds of stock market forecasts. In Ben-David et al. (2010), the authors show that CFO estimates of the median, lower, and upper bounds for next year's S&P 500 returns predict firms' investment decisions. Firms with more overprecise CFOs have more aggressive investment behaviors.

Our in-task measure of overprecision elicits the participants' "best guess" of the next period's demand realization as well as their beliefs about the 5th percentile and the 95th percentile of the demand distribution (denoted here as $x(0.05)$ and $x(0.95)$, respectively). We follow Keefer and Bodily (1983) and calculate $IN_i = (x(0.95) - x(0.05))/3.2$. This describes an individual's estimation of the standard deviation of the demand distribution. We note that this measure is the reverse of overprecision: less overprecise individuals have higher imputed volatilities; they believe that the standard deviation of the demand distribution is higher than what more overprecise individuals believe. For analysis purposes, we will use IN_i as our in-task measure of overprecision, with higher IN_i denoting less overprecision. We thus predict a negative impact of IN_i on order bias; higher IN_i means less overprecision.

5.3. Experimental Design and Implementation

As in Study 1, each participant was asked to play 50 independent rounds of single-period newsvendor game with a known demand distribution (normal distribution with mean 100 and standard deviation 30). In each round, participants placed an order. Then the computer randomly generated a realized demand, and profits were calculated and earned. The experiment lasted around 30 minutes. The average payment was \$11.59.[11]

All participants responded to the in-task measure of overprecision, based on Ben-David et al. (2010). At the beginning of every five rounds, we asked participants the following:

Tell us what you think demand will be in the next round:

1. *My best guess is that demand will be*:
2. *There is a 1-in-20 chance that demand will be less than*:
3. *There is a 1-in-20 chance that demand will be more than*:

Participants recorded their beliefs about the demand distribution and then proceeded to the newsvendor task. These beliefs were not incentivized, consistent with the previous literature and to keep them parallel to the overprecision measure of Russo and Schoemaker (1990) from Study 1.

This second study involves two treatments. In the control treatment, participants simply responded to this in-task measure of overprecision and made ordering decisions. In the SPIES treatment, they were also asked to respond to the SPIES question before each in-task measure (every five rounds). Appendix C contains the instructions used in these two treatments.

In the SPIES treatment, participants were asked the probability (between 0% and 100%) that demand in the next round would be lower than 10, 11–40, 41–70, 71–100, 101–130, 131–160, 161–190, and greater than 191. Because the true demand distribution is normal with mean 100 and standard deviation 30, these categories were chosen to help participants to estimate the probability for intervals that are one, two, or three standard deviations away from the mean. Participants estimate, for each interval, the probability that actual demand in the next round would fall into the interval. Participants' probabilities had to sum to 100% before proceeding, as in Haran et al. (2010).

5.4. Experimental Results

5.4.1. Orders. As in Table 2, Table 5 describes the number of participants in each treatment as well as the average [standard deviation] of newsvendor

[11] Here, a second experiment was not run at the end: thus these sessions were shorter and paid somewhat less than those in the first study.

Table 5 Orders in Control and SPIES Treatments

	High profit		Low profit	
Treatment	Average orders	No. of participants	Average orders	No. of participants
Control	102.76** [21.26]	17	95.19** [17.30]	18
SPIES	108.05** [16.36]	16	90.35** [21.24]	19

Notes. Standard deviation in brackets. t-test versus optimal.
$^{**}p < 0.01$.

orders over the 50 rounds. Similar to our results in Study 1, orders are lower than optimal in the high-profit treatment (103, 108 versus 120) and higher than optimal in the low-profit treatment (95, 90 versus 80). All of these differences are statistically significant.

These errors were also costly. In the high-profit control treatment, profit is reduced by 7.4% (24,470 was actually earned versus the 26,424 that would have been earned had participants ordered optimally). This difference is statistically significant using a t-test ($t = -3.75$, $p < 0.002$, $n = 17$) and is similar to the lost profits in the high-profit treatment in Study 1 of 8.4%. In the high-profit SPIES treatment, profit is reduced by 4.3% (25,287 versus 26,424); the difference also is statistically significant ($t = -5.07$, $p < 0.001$, $n = 16$) but is less than the 7.4% lost profit in the control condition.

In the low-profit control treatment, profit is reduced by 22.3% (4,810 was actually earned versus the 6,192 that could have been earned had participants ordered optimally). This difference is statistically significant using a t-test ($t = -10.59$, $p < 0.001$, $n = 18$) and is similar to the lost profits in the low-profit treatment in Study 1 of 20.6%. In the low-profit SPIES treatment, profit is reduced by 20.8% (4,905 versus 6,192); the difference also is statistically significant ($t = -8.56$, $p < 0.001$, $n = 19$) but is again lower than the 22.3% lost profit in the control condition.

To compare the data across Study 1 and Study 2, we follow Schweitzer and Cachon (2000) and calculate *adjustment scores* in each treatment to capture order error. In the high-profit treatment, this is the order minus the mean of the demand divided by the optimal order minus the mean of demand. In the low-profit treatment, this is the mean of demand minus the order divided by the mean of demand minus the optimal order. We find that the adjustment score (level of order error) in Study 1 is the same as the level of order error in the control treatment of Study 2 (t-test, $t = 1.32$, $p = 0.19$, $n_1 = 8{,}650$, $n_2 = 1{,}750$). This suggests that the inclusion of the in-task overprecision measure did not significantly improve individuals' ordering behavior. In contrast, the level of order error in Study 1 is significantly higher than the level of order error in the SPIES treatment of Study 2 (t-test, $t = -8.00$, $p < 0.001$, $n_1 = 8{,}650$, $n_2 = 1{,}750$). This suggests that the introduction of SPIES did indeed improve participants' ordering behavior.

However, the focus of this second study is the comparison of average orders between the control and SPIES treatments. This comparison will be done more formally below, but a simple t-test comparing adjustment scores demonstrates that individuals randomly assigned to the SPIES condition ordered significantly closer to optimal than individuals randomly assigned to the control condition (t-test, $t = -7.84$, $p < 0.001$, $n_1 = 1{,}750$, $n_2 = 1{,}750$). This preliminary evidence suggests that the SPIES procedure, which has been previously shown to exogenously reduce overprecision, also exogenously improves ordering behavior.

5.4.2. In-Task Measure of Overprecision. Each participant in this study completed the in-task measure of overprecision 10 times. For the high-profit condition, the in-task measure of overprecision for the control group is 24.93 and for the SPIES group is 33.90. The difference between these two groups is significant (t-test, $t = -6.57$, $p < 0.001$, $n_1 = 170$, $n_2 = 160$). For the low-profit condition, the in-task measure of overprecision for the control group is 24.03 and for the SPIES group is 29.06. The difference between these two groups is also significant (t-test, $t = -3.78$, $p < 0.001$, $n_1 = 180$, $n_2 = 190$). This result thus reinforces the finding of Haran et al. (2010) that the SPIES procedure significantly decreases overprecision.[12]

Furthermore, the variation of the in-task measure of overprecision reduces over time as experimental participants get more comfortable with the elicitation mechanism. In round 1, the variance of the in-task measure of overprecision is 213, which is the highest across all 10 elicitations. Round 46 (the last in-task measure of overprecision) has a variance of 142, which is the smallest across all 10 elicitations. To most accurately measure a participant's level of overprecision, we thus use the last elicitation (from round 46).[13]

[12] Haran et al. (2010) do not provide an external measure of overprecision for participants in the SPIES treatment; instead, they use the responses from the SPIES elicitation to generate an individual's 90% confidence interval, which is then compared to 90% confidence intervals that are more traditionally elicited. When we follow their procedure, we again find that SPIES-generated interval estimates are significantly less overprecise than are the interval estimates generated by the control group; for the high-profit condition, $t = -8.56$, $p < 0.001$, $n_1 = 170$, $n_2 = 160$ and for the low-profit condition, $t = -9.20$, $p < 0.001$, $n_1 = 180$, $n_2 = 190$.

[13] Our in-task measure of overprecision allowed participants to name any numbers they liked and thus resulted in significantly more heterogeneity than the overprecision measure used in Study 1

Table 6 **Demonstrating the Impact of Overprecision and SPIES (Random Effects)**

	(1)	(2)	(3)
Imputed volatility (IN_i)	−0.284**		−0.238*
	[0.106]		[0.113]
SPIES		−5.05*	−3.04
		[2.573]	[3.688]
Constant	20.44**	15.30**	20.75**
	[3.372]	[2.208]	[3.376]
Control for treatment	Yes	Yes	Yes
N	3,500	3,500	3,500
R^2	0.033	0.020	0.038

Note. Standard errors in brackets.
$^*p < 0.05$; $^{**}p < 0.01$.

Following the previous analysis, we regress individual error on our new in-task measure of overprecision and the SPIES treatment:

$$I_L(Q_{it} - Q^*) = \alpha + \beta_{In-OP} IN_i + \beta_S SPIES + \phi H + \eta_i + \varepsilon_{it}. \quad (3)$$

As in Study 1, the dependent variable is the difference between the order placed and the optimal order; a measure of order bias and I_L equals −1 for the high-profit condition and 1 for the low-profit condition. Here IN_i is the individual in-task measure of overprecision. Remember that higher levels of IN_i mean lower levels of overprecision and less order bias; thus we expect this coefficient will be negative. *SPIES* indicates whether the participant was randomly assigned to the SPIES treatment (= 1) or the control treatment (= 0). If *SPIES* is effective in decreasing order bias, this coefficient will also be negative. As in §4.2, we use random effects GLS regression to estimate Equation (3). Results are found in Table 6, column (1).

Our in-task measure of overprecision is significantly predictive of order accuracy. Higher imputed volatility (which translates to lower in-task overprecision) leads to significantly less order bias, as seen in column (1). This result enforces our finding in Study 1 that overprecision is a significant factor causing biased decisions in the newsvendor environment. Decision makers who have narrower estimates of the demand distribution demonstrate more biased decisions.[14] Also consistent with our predictions, the *SPIES* coefficient is negative and significant. Experimental participants who were randomly assigned to the SPIES treatment placed significantly less biased orders than those who were not, as seen in column (2).

5.4.3. Mediation Analysis. Consistent with Haran et al. (2010), we argue that the path through which SPIES decreases order bias is through a reduction in levels of overprecision. To test this argument, we use mediation analysis (Baron and Kenny 1986), examining whether IN_i mediates the effect of *SPIES* on order bias.

In the first step, we demonstrate that *SPIES* has a significant total effect on order bias, relying on the regression from Table 6, column (2). In the second step, we regress *SPIES* on imputed volatility (IN_i) at the individual level. We find a significant relationship ($\beta = 8.48$, $p < 0.003$, $n = 70$), which satisfies the second requirement of the mediation analysis. In the third step, we include both *SPIES* and IN_i as predictors of order bias, as shown in Table 6, column (3). We find that IN_i remains significant, while *SPIES* becomes insignificant and the coefficient of *SPIES* decreases (from −5.05 to −3.04), suggesting that the impact of *SPIES* on order bias is mediated by its effect on overprecision.

The final step in mediation analyses is to demonstrate that the (indirect) mediated effect is itself statistically significant. Historically, this has been done via the use of a single-level Sobel test (MacKinnon et al. 2002). That test returns a significant result ($p < 0.003$). However, Krull and MacKinnon (2001) point out that this test is not appropriate for data with a panel structure.

Instead, we follow the procedure of Krull and MacKinnon (2001), which uses bootstrapping to estimate a multilevel version of the Sobel test. This also returns a statistically significant result; the estimated indirect effect of *SPIES* on orders as mediated through IN_i is −2.02, with a 95% confidence interval from −1.61 to −2.41.[15] We can thus reject the null hypothesis that the indirect effect is zero at the $p < 0.05$ level. This satisfies the fourth and final requirement for mediation analysis.

In summary, Study 2 provides two important advances over Study 1. First, we provide a robustness check of our results using an in-task measure of overprecision where participants are asked to estimate the distribution of demand in the newsvendor problem rather than asked to provide confidence intervals for general-knowledge questions. We find that this in-task measure is similarly predictive of individual-level order bias in the newsvendor setting. This result provides further evidence for overprecision as an explanation of order bias. Second, we introduce a new treatment (SPIES), which has been demonstrated to reduce overprecision (Haran et al. 2010). We show that experimental participants who

(the 10-question response from Russo and Schoemaker 1990). We removed an additional 11 individuals who were outliers in their overprecision measure.

[14] These results also hold when we control for learning effects, as in §4.3. Results are available from the authors.

[15] These results are based on 5,000 replications, and a similar result holds for 10,000 replications.

are randomly assigned to this treatment place significantly less biased orders than those who are randomly assigned to a control condition. This treatment exogenously manipulates (rather than simply measures) the level of overprecision that experimental participants exhibit and demonstrates that lower levels of overprecision cause lower order bias. This suggests that SPIES "works" in improving the accuracy of orders. Finally, we show that the effect of SPIES on order bias is mediated by its impact on the individual's level of overprecision. The results suggest that the *reason* that SPIES works is because it alleviates the extent of overprecision exhibited by the newsvendors.

6. Summary and Discussion

This paper seeks to explain an asymmetric bias observed in orders placed in newsvendor settings. Individuals overorder under lowprofit conditions where the optimal order is less than the mean of the demand distribution and underorder under high-profit conditions where the optimal order is greater than the mean of the demand distribution (Schweitzer and Cachon 2000, Bostian et al. 2008, Bolton et al. 2012, Benzion et al. 2008, Bolton and Katok 2008, Katok and Wu 2009). We propose a potential cause for this result (overprecision) and provide two experiments to test our explanation.

In Study 1, we measure participants' level of overprecision as well as levels of overplacement and anchoring. We examine the relationship between overprecision and the biased orders placed in the newsvendor problem, controlling for individual biases in these other measures. We find a significant and robust relationship between overprecision and order bias, where more overprecise individuals place more biased orders. This relationship is also robust to learning. Our rough calibration suggests that overprecision accounts for around one-third of the ordering mistakes individuals make.

In Study 2, we use an in-task measure of overprecision, and furthermore exogenously manipulate overprecision using SPIES (Haran et al. 2010). We demonstrate that experimental participants randomly assigned to the SPIES treatment place significantly more accurate (less biased) orders than do those assigned to the control treatment and that this effect is mediated by the SPIES influence on overprecision. We also demonstrate the continued relationship between overprecision (here, measured in-task) and order bias, providing a robustness check on our results from Study 1.

Every study has limitations, and ours is no exception. Our conclusions come from a laboratory experiment with a student population. Additional data from the field, using inventory managers, would serve as an important external validity check. Furthermore, time constraints in our experiment limited the number of psychological regularities we could elicit. Additional work might also elicit personality variables like need for control, optimism, or extroversion, which may similarly explain biased newsvendor orders, either by directly affecting orders or by influencing an individual's level of overprecision. Conceptually, one might imagine that these biases are themselves related. As one anonymous referee suggested, the in-task measurement of overprecision and the SPIES treatment used in Study 2 may themselves affect the level of anchoring that an individual exhibits because it pushes mean demand into the foreground. We note that these two features were absent from Study 1 and that orders in the control condition of Study 2 were not significantly different than those of Study 1, suggesting that this is not a major concern. However, we agree that future research would be enhanced by an investigation of potential interactions between biases.

Those limitations notwithstanding, we believe that our results make a contribution in moving beyond documenting the existence of biased orders in inventory problems to identifying the underlying causal mechanisms that generate these biases.

7. Conclusion

This research identifies a causal mechanism of order bias (overprecision), which suggests a general class of debiasing techniques that inventory managers might consider to improve their ordering decisions. Understanding the mechanism is critical for developing an effective debiasing strategy, for improving inventory management decisions and the resulting efficiency and profits of firms. Our results suggest that managers should search for mechanisms or interventions that can reduce overprecision and can expect to see corresponding decreases in order bias and increases in profitability. In our second study we identify and test one such intervention, the SPIES technique, and demonstrate its effectiveness. Our results suggest that SPIES (or a SPIES-like procedure) can significantly improve performance in this inventory setting.

Acknowledgments

The authors thank the editorial team for helpful and constructive comments. They also thank participants at the Behavioral Operations Management Conference for useful feedback in earlier versions of this paper. Finally, the authors are grateful to the management and staff of the Center for Behavioral and Experimental Economic Science laboratory at the University of Texas at Dallas. All remaining errors or omissions are the authors' own.

Appendix A. Instruction, Decision, and Results Screens

Task 3

Instructions

Your earnings from the session today will be determined by your decisions in this task. In this task you will earn profits in experimental dollars. At the end of the session each 10,000 experimental dollars will be worth 3.0 U.S. dollars.

In this task, you will act as a seller of a product. In each round, you need to decide how much of this product to order from a supplier. You will then have the opportunity to sell your product to consumers.

Each unit you order costs 4.0 experimental dollars. Each unit you sell earns 10.0 experimental dollars. The number of units you sell will depend on market demand and how much you ordered. At the end of each round, you will return any unsold units to your supplier and get back 2.0 experimental dollars per unsold unit (salvage value). Units cannot be saved for following rounds.

You will be making this decision for 50 rounds.

Please press CONTINUE when you are finished.

CONTINUE

Task 3 (cont.)

Market Demand

In each round, the computer will randomly select a market demand quantity from a known distribution which will be described later.

The market demand in any round is independent of the market demand in previous or later rounds. Thus, if demand is small (or large) today, this will not affect whether demand is small (or large) tomorrow.

Profit

If market demand is *greater* than the amount you ordered, then you will sell the number of units you ordered. If market demand is *smaller* than the amount you ordered, then you will sell the number of units demanded and return the rest to your supplier.

Your profits for this session will be the sum of your profits in each round. Your profit in each round is the number of units you sell times 10.0 experimental dollars each, minus the number of units you purchased times 4.0 experimental dollars each, plus the number of unsold units you returns to your supplier times 2.0 experimental dollars each.

Your profits equal: Quantity sold X 10.0 - Quantity ordered X 4.0 + Unsold quantity X 2.0

After you place an order, you will be told the market demand and your resulting profit. You will also see a table of your past decisions and resulting profits.

Please press CONTINUE when you are finished. Press REVIEW to begin the instructions again.

REVIEW CONTINUE

Example

This table contains an example of 50 rounds of market demand.

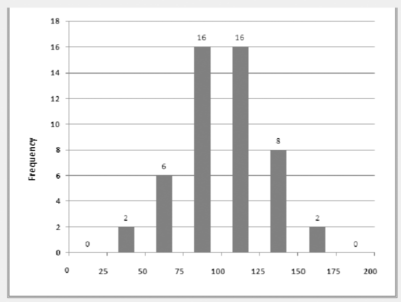
Round	Demand	Round	Demand	Round	Demand	Round	Demand
1	74	14	103	27	57	40	66
2	92	15	58	28	86	41	125
3	88	16	141	29	101	42	106
4	104	17	148	30	83	43	111
5	116	18	129	31	46	44	51
6	165	19	102	32	122	45	92
7	157	20	94	33	143	46	79
8	91	21	91	34	112	47	83
9	76	22	120	35	104	48	117
10	96	23	120	36	94	49	52
11	124	24	76	37	34	50	78
12	125	25	118	38	84		
13	125	26	112	39	129		

To see a histogram of this example, please press CONTINUE.

CONTINUE

Example

This is a histogram of the example.



To see another example of 50 rounds of market demand, please press ANOTHER EXAMPLE. To continue without more examples, please press CONTINUE.

ANOTHER EXAMPLE CONTINUE

Round
1 of 50

Game Information

Selling Price: 10.0
Buying Cost: 4.0
Salvage value for unsold units: 2.0

Enter Your Order:

REVIEW Press REVIEW to review the instructions. PLACE ORDER

Please use only whole numbers.

The market demand will be randomly drawn from a normal distribution with a mean(average) of 100 and a standard deviation of 30.

Reminder: Your profits equal: Quantity sold X 10.0 - Quantity ordered X 4.0 + Unsold quantity X 2.0

Your history

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
Round	Quantity Ordered	Order Cost	Demand	Quantity Sold	Selling revenue	Unsold Quantity	Salvage Revenue	Profit	Total Profit
		= (1) X 4.0		= min [(1), (3)]	= (4) X 10.0	= (1) - (4)	= (6) X 2.0	= (5) - (2) + (7)	

Round
1 of 50

Results

Selling Price: 10.0
Buying Cost: 4.0
Salvage value for unsold units: 2.0

Quantity Ordered:	100	**Order Cost:**	400.0
Randomly drawn demand:	104		
Quantity Sold:	100	**Selling revenue:**	1000.0
Unsold quantity:	0	**Salvage Revenue:**	0.0
		Profit in this round:	600.0

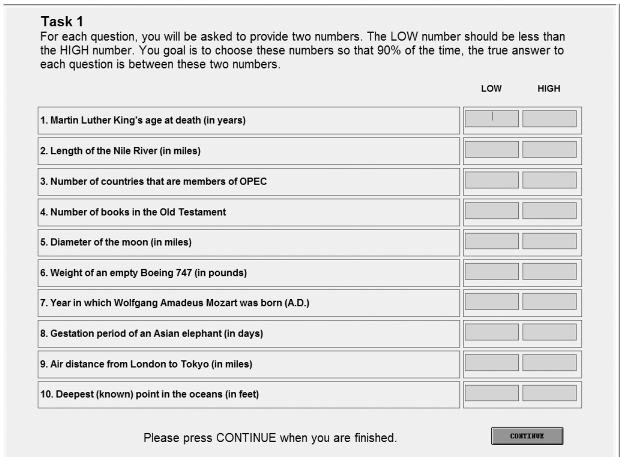
Your history

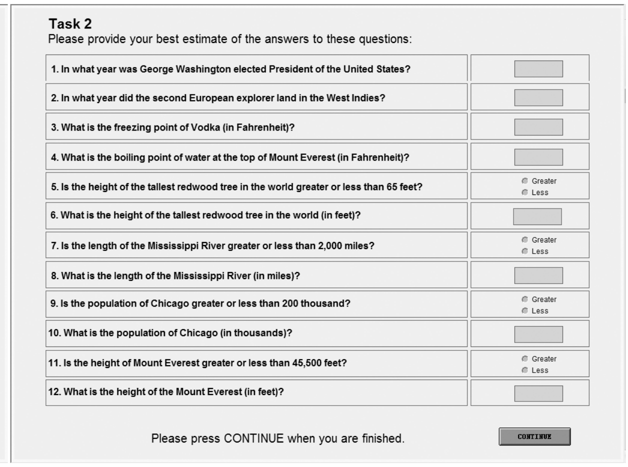
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
Round	Quantity Ordered	Order Cost	Demand	Quantity Sold	Selling revenue	Unsold Quantity	Salvage Revenue	Profit	Total Profit
		= (1) X 4.0		= min [(1), (3)]	= (4) X 10.0	= (1) - (4)	= (6) X 2.0	= (5) - (2) + (7)	
1	100	400	104	100	1000.0	0	0.0	600.0	600.0

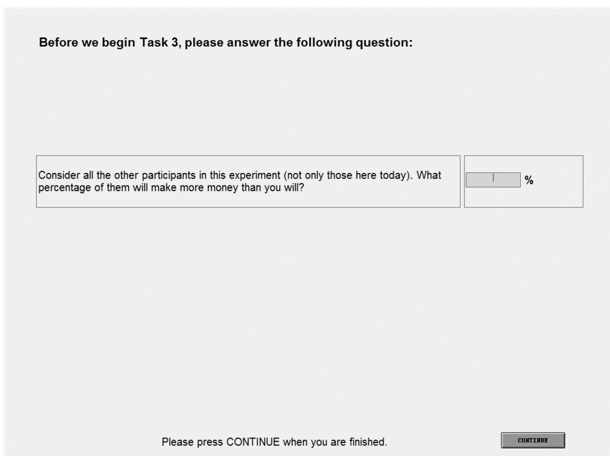
Please press CONTINUE to begin the next round.

CONTINUE

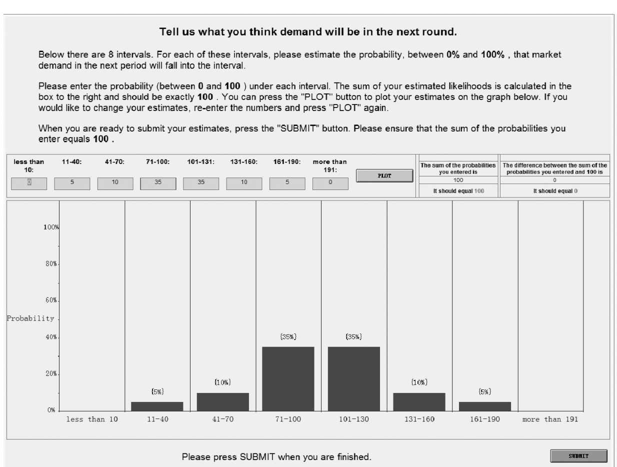
Appendix B. Psychological Tasks

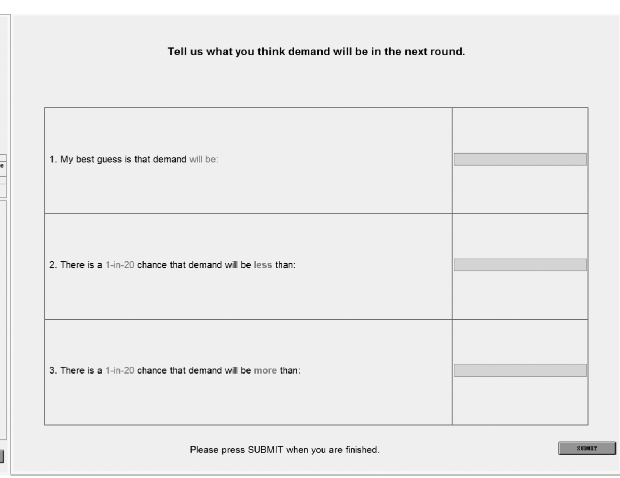






Appendix C. In-Task Overprecision Measurement and SPIES





References

Alicke MD, Govorun O (2005) The better-than-average effect. Alicke MD, Dunning DA, Krueger JI, eds. *The Self in Social Judgment* (Psychology Press, Philadelphia), 85–106.

Alpert M, Raiffa H (1982) A progress report on the training of probability assessors. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, New York), 294–305.

Arkes HR, Christensen C, Lai C, Blumer C (1987) Two methods of reducing overconfidence. *Organ. Behav. Human Decision Processes* 39(1):133–144.

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

Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Personality Soc. Psych.* 51(6):1173–7782.

Ben-David I, Graham JR, Harvey CR (2010) Managerial miscalibration. NBER Working Paper 16215, National Bureau of Economic Research, Cambridge, MA. http://www.nber.org/papers/w16215.pdf.

Benoit JP, Dubra J (2011) Apparent overconfidence. *Econometrica* 79(5):1591–1625.

Benzion U, Cohen Y, Peled R, Shavit T (2008) Decision-making and the newsvendor problem: An experimental study. *J. Oper. Res. Soc.* 59(9):1281–1287.

Biais, B, Hilton D, Mazurier K, Pouget S (2005) Judgmental overconfidence, self-monitoring, and trading performance in an experimental financial market. *Rev. Econom. Stud.* 72(2):287–312.

Block RA, Harper DR (1991) Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organ. Behav. Human Decision Processes* 49(2):188–207.

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

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

Bornstein BH, Zickafoose DJ (1999) I know I know it, I know I saw it: The stability of the confidence-accuracy relationship across domains. *J. Experiment. Psych.: Appl.* 5(1):76–88.

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

Cooper AC, Woo CY, Dunkelberg WC (1988) Entrepreneurs' perceived chances for success. *J. Bus. Venturing* 3(2):97–109.

Crawford JD, Stankov L (1996) Age differences in the realism of confidence judgments: A calibration study using tests of fluid and crystallized intelligence. *Learn. Individual Differences* 8(2):84–103.

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

Croson R, Croson D, Ren Y (2011) The overconfident newsvendor. Working paper, University of Texas at Dallas, Richardson.

Deaves R, Luders E, Luo GY (2009) An experimental test of the impact of overconfidence and gender on trading activity. *Rev. Finance* 13(3):555–575.

Edgeworth FY (1888) The mathematical theory of banking. *J. Royal Statist. Soc.* 51(1):113–127.

Epley N, Gilovich T (2001) Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psych. Sci.* 12(5):391–396.

Epley N, Gilovich T (2005) When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *J. Behav. Decision Making* 18(3):199–212.

Fischhoff B (1982) Debiasing. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, New York), 422–444.

Fisher M, Raman A (1996) Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* 44(1):87–99.

Glaser M, Langer T, Weber M (2012) True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *J. Behav. Decision Making*, ePub ahead of print October 12, http://onlinelibrary.wiley.com/doi/10.1002/bdm.1773/abstract.

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgment Decision Making* 5(7): 467–476.

Jacowitz KE, Kahneman D (1995) Measures of anchoring in estimation tasks. *Personality Soc. Psych. Bull.* 21(11):1161–1166.

Jonsson A-G, Allwood CM (2003) Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality Individual Differences* 34(4):559–574.

Juslin P, Olsson H (1997) Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psych. Rev.* 104(2):344–366.

Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: A cognitive perspective on risk and risk taking. *Management Sci.* 39(1):17–31.

Kanthen HK, Huy C (2009) Supply chain coordination models with retailer's attitudes towards risk. Fleischmann B, Borgwardt K, Klein R, eds. *Operations Research Proceedings 2008* (Springer, New York), 191–196.

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

Katok E, Lathrop A, Tarantino W, Xu S (2001) Using dynamic programming-based DSS to manage inventory at Jeppesen. *Interfaces* 31(6):54–65.

Keefer DL, Bodily SE (1983) Three-point approximations for continuous random variables. *Management Sci.* 29(5):595–609.

Keren B, Pliskin JS (2006) A benchmark solution for the risk-averse newsvendor problem. *Eur. J. Oper. Res.* 174(3):1643–1650.

Klayman J, Soll JB, Gonzales-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what and whom you ask. *Organ. Behav. Human Decision Processes* 79(3):216–247.

Kleitman S, Stankov L (2001) Ecological and person-oriented aspects of metacognitive processes in test-taking. *Appl. Cognitive Psych.* 15(3):321–341.

Koriat A, Lichtenstein S, Fischhoff B (1980) Reasons for overconfidence. *J. Experiment. Psych.: Human Learn. Memory* 6(2):107–118.

Kremer M, Minner S, Luk VW (2010) Do random errors explain newsvendor behavior? *Manufacturing Service Oper. Management* 12(4):673–681.

Krull JL, MacKinnon DP (2001) Multilevel modeling of individual and group level mediated effects. *Multivariate Behav. Res.* 36(2):249–277.

Larrick RP (2004) Debiasing. Koehler DJ, Harvey N, eds. *Blackwell Handbook of Judgment and Decision Making* (Blackwell Publishers, Oxford, UK), 316–339.

Lau N, Bearden JN (2013) Newsvendor demand chasing revisited. *Management Sci.* 59(5):1245–1249.

Lichtenstein S, Fischhoff B (1980) Training for calibration. *Organ. Behav. Human Decision Processes* 26(2):149–171.

Loch CH, Wu Y (2008) Social preferences and supply chain performance: An experimental study. *Management Sci.* 54(11): 1835–1849.

MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V (2002) A comparison of methods to test mediation and other intervening variable effects. *Psych. Methods* 7(1):83–104.

McGraw AP, Mellers BA, Ritov I (2004) The affective costs of overconfidence. *J. Behav. Decision Making* 17(4):281–295.

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

Moritz B, Hill A, Donohue K (2013) Individual differences in the newsvendor problem: Behavior and cognitive reflection. *J. Oper. Management* 31(1–2):72–85.

Mussweiler T, Strack F, Pfeiffer T (2000) Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality Soc. Psych. Bull.* 26(9):1142–1150.

Pallier G, Wilkinson R, Danthiir V, Kleitman S, Knezevic G, Stankov L, Roberts RD (2002) The role of question format and individual differences in the realism of confidence judgments. *J. General Psych.* 129(3):257–299.

Plous S (1995) A comparison of strategies for reducing interval overconfidence in group judgments. *J. Appl. Psych.* 82(4): 443–454.

Porteus EL (2002) *Foundations of Stochastic Inventory Theory* (Stanford Business Books, Stanford, CA).

Russo JE, Schoemaker PJH (1990) *Decision Traps: Ten Barriers to Brilliant Decision-Making and How to Overcome Them* (Simon & Schuster Press, New York).

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

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.

Stankov L (1999) Mining in the "no man's land" between intelligence and personality. Ackerman PL, Kyllonen PC, Roberts RD, eds. *Learning and Individual Differences: Process, Trait, and Content Determinants* (American Psychological Association, Washington, DC), 315–338.

Stankov L, Crawford JD (1996) Confidence judgments in studies of individual differences. *Personality and Individual Differences* 21(6):971–986.

Stankov L, Crawford JD (1997) Self-confidence and performance on cognitive abilities. *Intelligence* 25(2):93–109.

Steckel JH, Gupta S, Banerji A (2004) Supply chain decision making: Will shorter cycle times and shared point-of-sale information necessarily help? *Management Sci.* 50(4):458–464.

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Sci.* 35(3):321–339.

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing Service Oper. Management* 10(4):566–589.

Thompson WB, Mason SE (1996) Instability of individual differences in the association between confidence judgments and memory performance. *Memory Cognition* 24(2):226–234.