PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 27, No. 4, April 2018, pp. 624–643
ISSN 1059-1478 | EISSN 1937-5956 | 18 | 274 | 000624

DOI 10.1111/poms.12823

A Behavioral Remedy for the Censorship Bias

Jordan Tong*
University of Wisconsin-Madison, Wisconsin School of Business, 975 University Ave., Room 4293, Madison, Wisconsin 53706, USA,
jordan.tong@wisc.edu

Daniel Feiler
Dartmouth College, Tuck School of Business, 100 Tuck Hall, Hanover, New Hampshire 03755, USA, df@dartmouth.edu

Richard Larrick
Duke University, Fuqua School of Business, 100 Fuqua Drive, Durham, North Carolina 27708, USA, larrick@duke.edu

Existing evidence suggests that managers exhibit a censorship bias: demand beliefs tend to be biased low when lost sales from stockouts are unobservable (censored demand) compared to when they are observable (uncensored demand). We develop a non-constraining, easily implementable behavioral debias technique to help mitigate this tendency in demand forecasting and inventory decision-making settings. The debiasing technique has individuals record estimates of demand outcomes (REDO): participants explicitly record a self-generated estimate of every demand realization, allowing them to record a different value than the number of sales in periods with stockouts. In doing so, they construct a more representative sample of demand realizations (that differs from the sales sample). In three laboratory experiments with MBA and undergraduate students, this remedy significantly reduces downward bias in demand beliefs under censorship and leads to higher inventory order decisions.

Key words: debias; censored demand; forecasting; newsvendor; behavioral operations
History: Received: July 2017; Accepted: October 2017 by Elena Katok, after 2 revisions.

1. Introduction

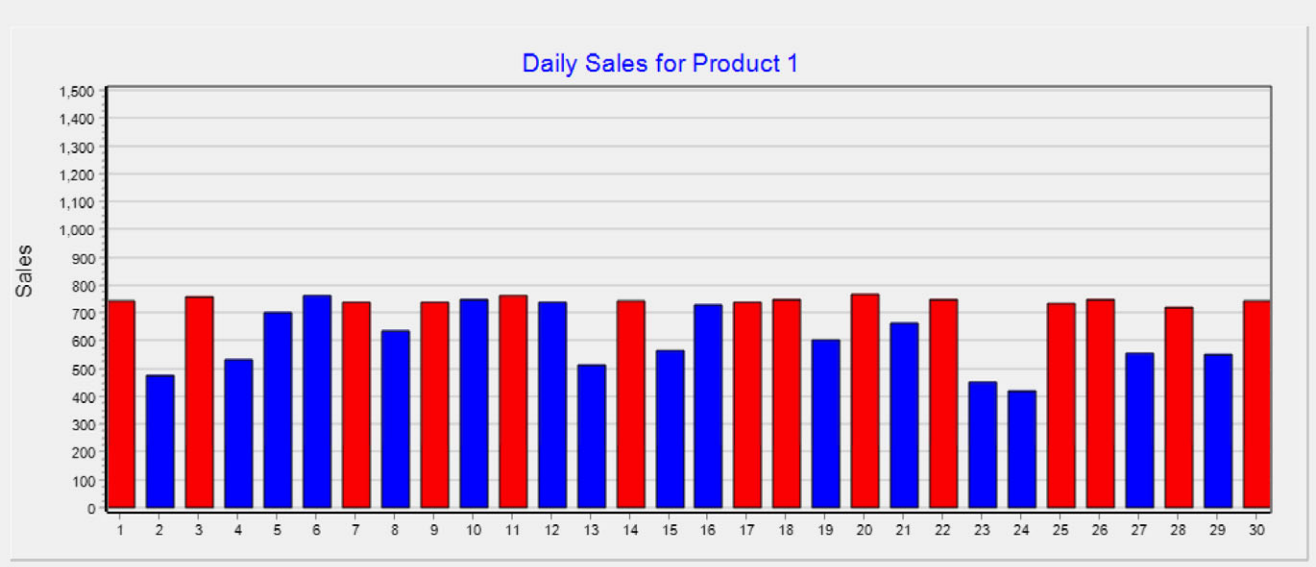
1.1. Censored Demand and the Censorship Bias

When demand for a product exceeds the available inventory, a stockout occurs. In many cases firms cannot perfectly observe lost sales after stockouts, causing the inventory level to effectively censor observations of demand. In such *censored demand* environments, exact demand can be observed when there is sufficient inventory; however, whenever there is a stockout, the number of sales is less than demand and the exact number of lost sales is unknown.

Managers often must rely on censored demand information. For example, brick-and-mortar retailers typically rely on censored point-of-sales (POS) data to make demand forecasting and other decisions (see Chen and Mersereau 2015). Even in information-rich online retail settings, retailers often cannot observe lost sales if they choose to inform customers of stockouts. Not only retailers, but also suppliers frequently rely on censored demand data, whether it be sales from their own customers (usually retailers) or POS data at their retailer partners. As one demand planner at an electronics manufacturer told the authors, even though they place excess demand on backorder, "customers will know that availability isn't good, and they sometimes won't even place their order or may substitute a different product—obviously in this case, we do not end up with visibility and need to make assumptions [about lost sales]." We certainly acknowledge that in some environments stockouts are rare and in other environments stockouts may occur with lost sales being perfectly observable. However, many managers are regularly in a situation in which they are left uncertain about what sales *could have been* after a stockout.

This leads to an important question: Are humans efficient at coping with censored demand? Recent experimental evidence provides a preliminary answer to this question: "no." Demand beliefs have been shown to be systematically lower under censored demand than under uncensored demand. One driver of this *censorship bias* is that censorship creates a misrepresentative sample of historical data which cause underestimation of demand (Feiler et al. 2013). Consider the graph of historical sales data for a given product in Figure 1 (in fact, this is a screenshot from Study 3 in this study). Here, the blue bars indicate no

Figure 1 Historical Sales Data Graph Illustrating Censored Demand, Taken from Study 3 [Color figure can be viewed at wileyonlinelibrary.com]



Note: Red bars indicate days with a stockout, blue bars indicate days with no stockout.

stockout days, while the red bars indicate stockout days. What will the manager think is average daily demand over the month?

In the figure, the average of the *sales* outcomes (the average of all the bars) is lower than the average of the actual *demand* outcomes because each red bar is lower than the demand outcome that day. In fact, for this graph, the average sales is only 667 even though the average demand outcome is 737. If the manager fails to fully account for the misrepresentativeness of sales data, her beliefs about mean demand will be lower than 737, biased toward the average sales 667. Consistent with this idea, Feiler et al. (2013) demonstrated in several experiments that, indeed, when faced with an unknown demand distribution, censorship leads to downward-biased beliefs about the true demand mean. Moreover, because demand beliefs serve as a key input to inventory decisions, downward-biased demand beliefs generated lower inventory order decisions.

Accurate demand beliefs are important because they are used as inputs of inventory decision-making as well as other important decisions, such as investment, budgeting, and staffing. A central objective in operations management is to design systems to improve firm performance and here we consider how one might implement intelligent system design to address the censorship bias.

1.2. Remedying the Censorship Bias

One approach to improving performance is to try to bypass managers altogether through *automation* of demand forecast updating and inventory ordering systems. Researchers in operations management have made great progress on mathematically deriving characteristics of the optimal forecasting and inventory order policies under censored demand (e.g., Ding et al. 2002, Lariviere and Porteus 1999, Lu et al. 2008, Nahmias 1994). Chen and Mersereau (2015) provide a review and history of this literature, including both Bayesian and nonparametric approaches. Nevertheless, in practice, the forecasting task frequently requires managerial judgment (see Fildes and Goodwin 2007, Kremer et al. 2011, Schweitzer and Cachon 2000). While not uniformly better, managers generally provide positive value, for example, they have information which the system does not (Fildes et al. 2009). Indeed, multiple retailers and manufacturers have told the authors that their managers frequently make subjective judgments about lost sales when making judgmental demand forecasts from sales data. For example, the demand forecasting software at the electronics manufacturer mentioned earlier simply assumes sales equals demand, so the company's demand planning team manually adjusts forecasts using their own judgment when they know that customers recently had not placed orders due to stockouts. Automation may be feasible in some cases, but in practice, managerial judgment still plays a large role in demand planning. Therefore, it is important to develop strategies for improving demand estimation when the manager is not bypassed via automation.

A common suggestion for improving performance without bypassing managers is to provide *financial incentives* for good performance. From a firm perspective, there are at least two drawbacks to this approach. First, financial incentives are often expensive and challenging to implement fairly and effectively—in some cases they may not even be viable because they require establishing a fair standard with which to judge decisions. Second, although incentives are a critical tool for overcoming shirking and carelessness (Rydval and Ortmann 2004), they have been shown to be surprisingly ineffective for remedying many cognitive biases (Camerer and Hogarth 1999, Fischoff 1982). These reasons suggest that an alternative to financial incentives may be valuable to firms.

Another possible mechanism to help managers make better decisions is to try to induce them to *think more deliberately*. Sloman (1996) and Stanovich and West (1998) argue that there are two general approaches to making decisions: rapid, automatic, intuitive processes ("System 1") and slower, deliberate, analytic processes ("System 2"). The presumption in this work is that slower thinking allows decision-makers to recognize and correct the systematic biases produced by rapid intuition. And, if individuals take enough time to reflect and calculate, they may be less susceptible to judgment biases. This idea has empirical support in psychology (e.g., Epley and Gilovich 2006) as well as in research on inventory decision-making specifically. For example, researchers have found that individuals who score higher on the cognitive reflection test (Frederick 2005), which is a measure of one's tendency to use "System 2," tend to make better inventory decisions than those who score lower (Moritz et al. 2013, Narayanan and Moritz 2015). This perspective suggests that improving performance may simply be a matter of increasing cognitive effort. However, Camerer and Hogarth (1999) have stressed that greater effort is not always sufficient for improving performance. The benefit of effort is contingent on possessing the appropriate cognitive strategy. In other words, thinking can be like paddling: at some point, more of it is only helpful if it is in the right direction.

How might one improve individuals' thinking in the face of censored demand? Our proposal is to design a *decision infrastructure* that facilitates a helpful thinking process for the manager. Given that the biased demand judgments are theorized to be driven by the misrepresentative sales data, our proposal is to design the infrastructure so that individuals first self-generate a new, more representative demand sample before making demand judgments. Specifically, the debiasing technique (motivated in more detail in the next section) has individuals record estimates of demand outcomes (REDO)—they explicitly record a self-generated estimate of every demand realization, allowing them to record a different demand value than the number of sales in stockout periods.

Our approach is in the same spirit as the influential book, Nudge (Thaler and Sunstein 2008), which argues that policy makers should be "choice architects" or "information architects" who carefully craft an infrastructure to encourage better decision-making. It argues that an effective perspective for practical and implementable improvement is "libertarian paternalism," in which choice architecture guides improved decision-making without limiting freedom or significantly changing incentives. This nonrestrictive approach is also consistent with the recommendation of forecasting system design experts, who assert that "absolute restrictiveness, where the system is deliberately designed to prohibit the use of particular processes, is dangerous, since the designer is unlikely to be certain that the included processes are the most appropriate to use, especially in a dynamic environment where the underlying conditions of use may change" (Fildes et al. 2009, p. 358).

1.3. Contributions and Related Literature

Our study makes the following contributions. First, we present a theory-driven remedy for the censorship bias in demand beliefs (REDO). Second, we provide experimental evidence with MBA and undergraduate students supporting the effectiveness of this remedy in the well-studied repeated newsvendor setting. Third, we provide additional experimental evidence of REDO's effectiveness in a more general setting of demand estimation from a graph of historical sales data. Our results deliver insight into how censorship affects forecasting and inventory decisions and how to reduce the censorship bias.

This study primarily builds on recent work that has examined human behavior in censored demand settings. Most closely related is Feiler et al. (2013), who document downwardly biased demand beliefs in censored demand settings, and show that this demand belief bias leads to lower order newsvendor order decisions with unknown censored demand and equal overage and underage costs. Zhao et al. (2016) also examine newsvendor decisions with unknown and censored demand and find similar results when overage and underage costs are asymmetric: orders under censored demand are lower than orders under uncensored demand. They also analyze the learning and updating process and provide evidence that people anchor on the last-period sales when making order decisions with censored demand. Finally, Rudi and Drake (2014) study newsvendor decisions with a known demand distribution but with censored demand feedback. They experimentally demonstrated that order decisions were lower when demand feedback was censored (than when it was uncensored) even when participants were explicitly told the demand distribution in advance. None of the above papers study behavioral interventions to reduce the effect of demand censorship, which is the focus of this study.

More generally, this study contributes to the growing body of research that advances our understanding of how people make demand forecasts and inventory decisions and how to improve them by better understanding drivers of behavioral biases. For example, there have been significant efforts to make progress with regard to developing behavioral interventions to reduce the pull-to-center effect in the newsvendor problem. Ren and Croson (2013) found that structuring the demand forecasting task to reduce overprecision (Haran et al. 2010) is partially effective at reducing the pull-to-center effect (Bolton and Katok 2008, Schweitzer and Cachon 2000). Removing the demand-framing of the newsvendor problem (Kremer et al. 2010), forcing individuals to commit to standing orders (Bolton and Katok 2008), manipulating the salience of the psychological costs of leftovers and stockouts (Ho et al. 2010), reducing the frequency of feedback (Lurie and Swaminathan 2009), and decomposing the newsvendor decision into subtasks (Lee and Siemsen 2017) have also proven to be at least partially effective in reducing the pull-to-center effect.

In time-series demand forecasting contexts, Kremer et al. (2011) find that forecasters generally tend to overreact to variations in demand in relatively stable environments, but underreact to them in relatively unstable environments. Later, Moritz et al. (2014) provided evidence that forecaster performance in this setting can be improved by manipulating decision speed to avoid overly fast or slow decisions. In supply chain contexts, researchers have found that sharing point-of-sale demand data and inventory data can mitigate behavioral causes of the bullwhip effect and improve performance even when such information sharing does not affect the optimal policy (Croson and Donohue 2003, 2006). Similarly, educating individuals on the structure of the optimal policy and providing system-wide training have proven at least partially effective at reducing the behavioral bullwhip effect (Croson et al. 2014, Wu and Katok 2006). Researchers have also found that forecast sharing under asymmetric demand information can improve performance due to trust and trustworthiness even if it would not if individuals were purely rational (Özer et al. 2011, 2014), and that changing the structure of contracts can affect supply chain performance even if the contracts would be considered equivalent for purely rational decision-makers (e.g., see Chen et al. 2013, Davis and Leider 2015, Katok and Wu 2009, Zhang et al. 2016).

We complement this larger body of work by examining how to reduce behavioral bias caused by demand censorship through a low-cost intervention that does not constrain decision-making or affect formal incentives or information.

2. A Remedy for the Censorship Bias: Record Estimates of Demand Outcomes

2.1. Theoretical Motivation for REDO

How do censored environments lead to biased judgment and what can we do to nudge managers to improve their thinking? Censored environments have been classified by psychologists as a type of “wicked environment” (Hogarth et al. 2015), in which judgments must be made based on systematically misrepresentative data—in this case, the average observed sales data are systematically biased below the average demand. Because individuals often do not fully adjust their beliefs to account for misrepresentative samples (e.g., Feiler et al. 2013, Juslin et al. 2007, Kareev et al. 2002), we theorize that they will also have downwardly biased demand beliefs when faced with censored demand.

Specifically, building on Feiler et al. (2013), we theorize that in order to estimate mean demand (EMD), people anchor on the mean of the observed sales sample and then may try to adjust upward to take into account the stockout information. While people are likely to be heterogeneous in the magnitude of these adjustments, on average, their adjustments are likely to be insufficient (Feiler et al. 2013). Research has found that when adjusting from an anchor, individuals tend to adjust insufficiently because they have already psychologically dwelled on the anchor as relevant information (Mussweiler and Strack 1999) and then cease the effortful adjustment (Epley and Gilovich 2006) as soon as they reach the edge of some region of plausibility (Quattrone 1982). In the case of censorship, there is also an asymmetry in the concreteness of demand observations, where the low values of demand are known with certainty but the high values of demand are only imagined beyond a lower sales stockout (Feiler et al. 2013). Lastly, in extreme cases, no adjustment from mean sales may occur if an individual does not recognize that lost sales exist.

Building on this hypothesized psychological driver of the censorship bias, we propose a new behavioral remedy for improving demand beliefs under censored demand. The remedy, REDO, requires individuals to explicitly report what they think was each period’s demand. Of course, when there is not a stockout the correct answer to this question is straightforward—it is simply the number sold that period. However, when there is a stockout, there is no way to know exactly how many additional units could have been sold, so one must record a guess (at least as large as the number sold) for demand in that period.

The core idea behind REDO is to change the decision-maker's judgment process so that they do not anchor directly on the observed mean sales when forming beliefs about the true demand mean. Instead, REDO asks the manager to try to adjust every misrepresentative data point before trying to estimate true mean demand. In other words, it nudges people to adjust the sample to make it more representative and then assess its mean, rather than assess the mean of a biased sample and then try to make an adjustment. REDO should facilitate adjustments in three ways: (i) REDO repeatedly reminds people that stockouts imply lost sales; (ii) REDO encourages decision-makers to create a new, more accurate sample by imagining the extent of lost sales for every stockout period; (iii) REDO turns vague stockout information into concrete numbers that help balance the vividness of information between in-stock and stockout periods.

Record estimates of demand outcomes is consistent with several established debiasing perspectives. Hogarth et al. (2015) concludes their discussion of "wicked environments" by stating that in order to improve decision-making, "one should provide experiences that lead to appropriate responses—that is, in kind environments" (p. 383). In line with this perspective, REDO seeks to improve decision-making by helping managers alter the data experienced in the environment to make it more kind before making decisions.

Arkes (1991) argues that the general cognitive process of "consider the opposite" has robust value for many decision biases because it helps break people out of a narrow frame (see also Heath and Heath 2013, Larrick 2009). If a manager's thinking is leading him to conclude that the demand is equal to the number of sales on a day with a stockout, a good corrective strategy is to nudge him to explicitly consider why such a conclusion might be wrong: how many more could he have sold? REDO helps individuals make such considerations which help them to think correctly about censored data.

Finally, REDO can also be viewed as a type of task-decomposition approach (e.g., see Armstrong 1975, Lee and Siemsen 2017, MacGregor et al. 1988, Raiffa 1968). Managers may be better at deciding how far to adjust each sales observation they know is misrepresentative as required in REDO than they are at trying to incorporate all of information about stockouts to decide how far to adjust upward from the observed average sales.

2.2. REDO and Inventory Decisions

While REDO is primarily designed to help improve demand beliefs, it also is likely to have an effect on inventory order decisions. We hypothesize that REDO will reduce the difference in order behavior between censored and uncensored environments. Demand beliefs serve as an important input to order decisions such that higher beliefs generally correspond to higher order decisions. For example, if individuals place orders by anchoring and adjusting from their belief about the demand mean (Schweitzer and Cachon 2000), REDO ought to increase orders by increasing individual's mean demand beliefs. Similarly, if individuals anchor and adjust from the previous sales outcome (Zhao et al. 2016), REDO can increase orders by providing a different and higher anchor—the self-generated guess of the previous demand outcome.

While we expect REDO to reduce the gap between censored and uncensored inventory decision-making, it is important to note that this reduction does not necessarily imply improvement of the profitability of order decisions in all situations because of the pull-to-center effect (Bolton and Katok 2008, Schweitzer and Cachon 2000). Specifically, the pull-to-center effect predicts that even under uncensored demand, orders will be too small for high-profit products but too large for low-profit products. The censorship bias predicts lower orders for all profit levels. REDO focuses on improving demand beliefs and reducing the effects of censorship, but it does not address the pull-to-center effect. Thus, without also addressing the pull-to-center effect, we expect REDO to increase the profitability of orders for high-profit products but not necessarily for low-profit products. We further examine the joint implications of the censorship bias and the pull-to-center effect in section 4 and discuss how one might address them both to improve performance in section 6.

2.3. Overview of Studies testing REDO

Studies 1 and 2 test REDO's effect on demand beliefs and inventory order decisions in the well-studied newsvendor problem. Study 1 tests REDO with MBA students under simple cost conditions in which the overage costs is equal to the underage cost. Study 2 tests REDO with undergraduate students with asymmetric overage and underage costs. It also compares performance against a different task—reestimating mean demand in every period—to help rule out attention to demand as a full explanation for REDO's improvement. Finally, Study 3 isolates REDO's improvement in demand beliefs from the inventory decision-making process by making inventory levels exogenous. It shows how to implement REDO when managers EMD from bar graphs of historical sales data, tests REDO's performance, and benchmarks it against another manipulation requiring similar attention.

3. Study 1

This experiment tests whether REDO improves the accuracy of demand beliefs and examines how it impacts inventory order decisions in a repeated newsvendor problem with unknown and censored demand. We also isolate the censorship bias from the well-known pull-to-center effect by choosing a simple cost setting in which the overage and underage costs are equal.

3.1. Methods

One hundred forty-seven daytime MBA students from a highly ranked American business school participated in the study. The sample was 27.2% female and the average age was 28.8 years. All participants had already taken graduate courses on Probability & Statistics and on Operations Management. In their statistics course, students became familiar with probability distributions and normal curves; in their operations course, they had been taught the newsvendor problem and its optimal ordering solution for known demand distributions.

Studies were conducted via computer simulations. For example screenshots of the studies, see the Appendix. Participants were instructed that they would be running a fictional newspaper-vending business. Each day they needed to buy newspapers for \$1 per unit to stock in their stand and sell for \$2 per unit. At the end of each day any excess newspapers would be discarded for \$0 per unit. Participants were also told that due to the cost structure, the per-unit opportunity cost of underordering (the "underage cost") was \$1, while the per-unit cost of overordering (the "overage cost") was also \$1. These parameters enabled us to study the effects of REDO in the absence of the pull-to-center effect (i.e., failing to properly account for the asymmetry in overage and underage costs). Note that if the demand mean m were known, the expected profit-maximizing order quantity would be to simply order m:

$$\begin{aligned} q^m &= m + \sigma\Phi^{-1}\left(\frac{c_u}{(c_u + c_o)}\right) \\ &= m + 100\Phi^{-1}\left(\frac{1}{2}\right) = m, \end{aligned}$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function. However, because m is unknown, participants must form beliefs about it based on feedback, which differed by condition.

Participants were given the following demand information. Demand for their newspapers each day was normally distributed with a known standard deviation of 100 but an unknown mean. They were told the mean m was between 400 and 800, although it would not change over the course of the game. We randomly assigned participants to a demand mean of either 500, 600, or 700.

The task entailed the following. In each period, participants entered their stocking decision for the day. A demand outcome was then randomly drawn from the true demand distribution. Sales, leftover inventory, and profits for that period were automatically calculated and presented. This process was repeated for 30 periods with all past decisions and outcomes remaining visible.

After the final period, participants were asked, "Now that you have completed all 30 days, please make a guess of what your m was. In other words, what was the true underlying mean demand for your newspapers?" Participants received a \$1 bonus if their estimate was within 10 units of the correct number. While they were playing the game, participants were unaware that they would later face this incentivized final estimation question.

In exchange for participation, a donation of \$5 was made to a school club or charity of their choice. In addition to the final demand estimation bonus, participants could earn additional money based on the amount of profit they earned in the simulation, which they could keep for themselves or also donate. For every \$2000 earned in the game, participants earned \$1 in bonus money for themselves (with partial dollars possible). The total money generated by most individuals was between \$8 and \$13.

3.2. Experimental Conditions

Participants were randomly assigned to one of three conditions, which are outlined in Table 1. In the *Censored* condition, demand each period was censored by the inventory level: the number of sales missed after stocking out was unobservable.

In the *REDO* condition, demand was also censored. The REDO condition was identical to the censored condition except that participants answered one question at the end of each period. After a stockout, they answered: "What is your best guess of what the exact demand was today?" After not stocking out, they answered, "What was the exact demand today?" These answers were recorded in a separate column and remained visible throughout the simulation.

Table 1 Names (in bold) and Descriptions of the Three Experimental Conditions in Study 1

Experimental conditions	*N*	Description	Participant inputs collected
Censored and no intervention	49	Unobservable lost sales after stockouts	30 order decisions, 1 final mean demand estimate
Censored and Record Estimates of Demand Outcomes (**REDO**)	49	Unobservable lost sales after stockouts Estimate (if stockout) or report (if no stockout) that period's demand	30 order decisions, 30 demand outcome estimates from REDO, 1 final mean demand estimate
Uncensored	49	Observable lost sales after stockouts	30 order decisions, 1 final mean demand estimate

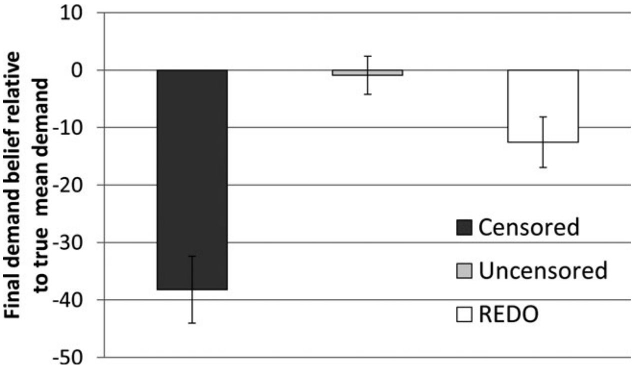
In the *Uncensored* condition, participants could see exactly what demand had been at the end of each period. In this manner, their missed sales from stocking out were observable. The actual demand for each period was presented in a column and remained visible throughout the simulation.

3.3. Results

Mean demand beliefs. See Figure 2 for a graph of final demand beliefs by condition. A series of *t*-tests were conducted to test the effect of experimental condition on final estimates of the underlying demand mean (relative to the true demand mean). These analyses involve one observation per subject. All statistical tests reported here are conducted on average responses by condition. Without any behavioral interventions, orders tended to be lower under censored demand than under uncensored demand. Estimates of the demand mean were significantly lower in the censored condition than in the uncensored condition, $t(96) = 5.56$, $p < 0.001$. This finding replicates the existing work on the censorship bias (Feiler et al. 2013) which has documented that demand beliefs are lower when lost sales are unobservable as opposed to observable.

Individuals in the REDO condition formed more accurate demand beliefs than those in the censored condition. Estimates in the REDO condition were higher than those in the censored condition, $t(96) = 3.51$, $p < 0.001$, but were significantly lower than those in the uncensored condition, $t(96) = 2.10$, $p = 0.04$.

Figure 2 Mean Final Demand Beliefs Across Experimental Conditions in Study 1, with Standard Error Bars



Were the final demand beliefs in each condition different from the true demand mean? To answer this question, one-sample t-tests were conducted for each condition. On average, individuals in the censored, $t(48) = 6.57$, $p < 0.001$, and REDO, $t(48) = 2.84$, $p < 0.01$, conditions significantly underestimated the true mean of demand. The estimates of individuals in the uncensored condition were not significantly different than the true mean demand, $t(48) = 0.28$, $p = 0.78$.

REDO responses. The recorded demand outcome estimates in the REDO condition can shed light on how the remedy improved performance. On average, the mean of participant-generated REDO sample was 23.49 units higher than the mean sales observed by the same individual, $SD = 32.33$, $t(48) = 5.09$, $p < 0.001$. However, on average the mean of the REDO sample was still 22.69 units lower than the true mean demand for that individual, $SD = 16.34$, $t(48) = 9.72$, $p < 0.001$. Among individuals in the REDO condition, the mean of the REDO samples correlated positively with the final estimates of the true demand mean ($r = 0.82$, $p < 0.001$). Thus, individuals who generated higher REDO samples indeed indicated larger demand beliefs. The average of their whole REDO sample was 10.94 units lower than their final estimate of the true mean demand ($SD = 18.91$), which was a significant difference, $t(48) = 4.05$, $p < 0.01$. On the other hand, the average of their REDO sample from their last 10 periods was only 5.77 lower than their final estimate of the true mean demand ($SD = 29.66$), which was not a significant difference, $t(48) = 1.36$, $p = 0.18$. Final estimates of mean demand were significantly closer to the mean of their REDO sample than to the mean of their observed sales, $t(48) = 8.71$, $p < 0.01$. In other words, individuals in the REDO condition appear to use their recent REDO sample to inform their final estimate of the true mean demand.

Statistical benchmark. It is also useful to calculate a statistical benchmark to verify that our study design was executed correctly and to provide a point of comparison for participants' performance when faced with censored demand. Therefore, we also calculated the maximum likelihood estimate for the mean demand, given the sales and stockout data observed by each participant at the end of the game, using the R package EnvStats (Millard 2013). In conditions with censored demand (Censored and REDO), the statistical benchmark was not significantly different from the true mean demand (M = 3.04, SD = 28.17, $t(96)$ = 1.06, p = 0.29.)

Given the same observations of sales and stockouts, participants in the censored condition had final estimates of mean demand that were significantly lower than the statistical benchmark, $t(47)$ = 6.71, $p < 0.001$. Final estimates in the REDO condition were previously shown to be significantly higher than in the censored condition; however, they were still significantly lower than the statistical benchmark, $t(48)$ = 3.60, $p < 0.001$.

Order decisions. We also investigated order behavior in each condition. A plot of average orders by condition can be seen in Figure 3. A linear regression was conducted in SAS with standard errors clustered by individual to account for the non-independence of the multiple observations for each participant. These analyses involve 30 observations per subject. The dependent variable was the number of units ordered for a given period. One individual did not place an order in period 1. The independent variable of primary interest was experimental condition (Censored, REDO, and Uncensored). For the regression, the censored condition was treated as the baseline condition. Dummy variables were included for the REDO and Uncensored conditions, each equal to 1 for observations in the respective conditions and otherwise equal to zero.

To account for the different true demand means faced by participants, and the specific demand draws they received from the true distribution, we controlled for the running average of past demand draws at the time of a given ordering decision. For example, in period 6, this variable was equal to the average of the random demand draws from the first five periods for a given participant. The only exception was that, in period 1, this variable was set equal to 600.

In model 2, we introduce the possibility of time trends that may differ across conditions. We added period (mean-centered), and interaction terms between period and condition. Within the framework of this model, we can then test whether the differences across condition are significant in a given period, accounting for time trends that may be different across conditions.

The regression model results for Study 1 can be seen in Table 2. The results from Model 1 show that orders in the censored condition were significantly lower than those in the uncensored condition, $t(146)$ = 2.81, p = 0.005. Average orders in the REDO condition were significantly higher than those in the censored condition, $t(146)$ = 3.65, $p < 0.001$. Orders in the REDO condition were not significantly different than those in the uncensored condition, $t(146)$ = 1.51, p = 0.13.

Model 2 introduces the effect of period (mean-centered) and interactions between condition and period.

Figure 3 **Average Orders (divided into intervals of five periods) in Study 1 by Condition at the Three Unknown True Demands Means, *m* [Color figure can be viewed at wileyonlinelibrary.com]**

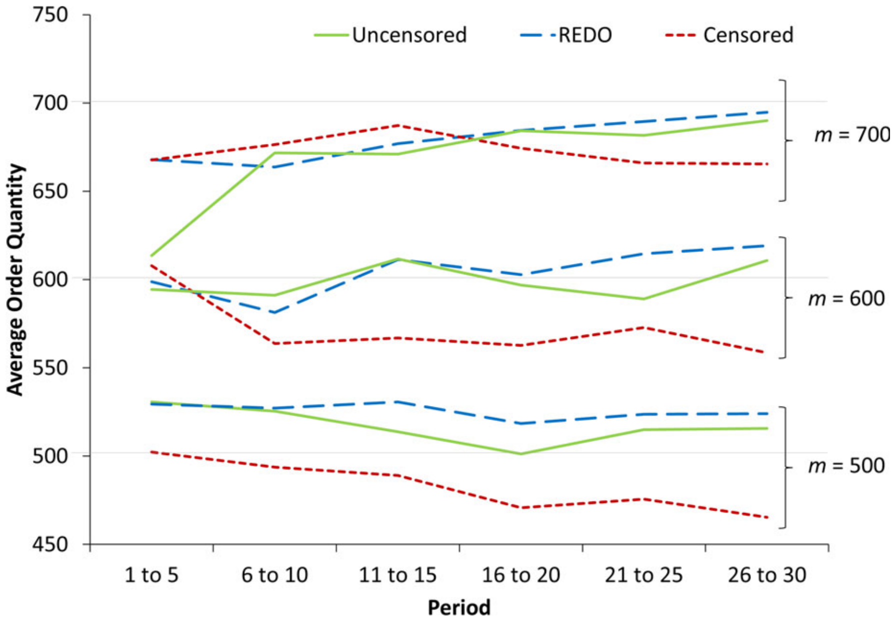


Table 2 The Regression Models of Inventory Ordering in Study 1

	DV: Inventory order	
	(1)	(2)
Intercept	100.78**	100.64**
	(18.94)	(18.97)
Running Avg. Demand	0.79**	0.79**
	(0.03)	(0.03)
Censored condition	(Baseline condition for the regression)	
Uncensored condition	19.32**	19.31**
	(6.87)	(6.87)
REDO condition	27.64**	27.64**
	(7.57)	(7.58)
Period		−0.97*
		(0.38)
Period × Uncensored		1.77**
		(0.50)
Period × REDO		1.76**
		(0.56)

Notes: The numbers in parentheses are standard errors clustered by subject. The variable Running Avg. Demand is the cumulative average of random demand draws at the time of the decision, except in period 1, in which it is equal to 600. The variables Uncensored and REDO are equal to 1 if a participant is in the respective condition and are otherwise equal to 0. The variable Period is mean-centered (30 periods). The number of observations is 4409 and the degrees of freedom for *t*-tests are 146.
*Denotes $p < 0.05$ and **denotes $p < 0.01$.

The main effects of experimental condition remain significant in Model 2, which can be interpreted directly from the table because, due to mean-centering, the case of period = 0 is the average period. Notably, Model 2 also predicts that, in the last period (30), orders will be 35.5 units higher in the uncensored condition than the censored condition ($SE = 7.68$), $t = 4.62$, $p < 0.001$, and 51.29 units higher in the REDO condition than in the censored condition ($SE = 9.93$), $t = 5.16$, $p < 0.001$.

Although our theory makes no specific predictions about the effect of REDO over time, it may be useful to note *ex-post* whether the benefits of REDO over Censored depended on period. Model 2 (see Table 2) shows that the difference between REDO and Censored was increasing as period increased. Plots of the data show that orders in the Uncensored and REDO conditions remained relatively flat over time (relative to the Running Average Demand), while orders in the censored condition gradually decreased over time. Therefore, in this experiment, the censorship bias was exacerbated over time and REDO insulated individuals from this tendency to make worse orders over time with censored demand.

Since overage and underage costs were equal in this experiment, q^m was equal to the true mean demand for each participant. Where did orders fall relative to the benchmark of q^m? Orders in the censored condition were significantly lower than q^m (the true demand mean), $t(146) = 3.84$, $p < 0.001$. However, orders in the uncensored condition, $t(146) = 1.08$, $p = 0.28$, and REDO condition, $t(146) = 0.70$, $p = 0.48$, were not significantly different from q^m (the true demand mean).

We also note that the observed downward bias in order decisions in the censored condition relative to the uncensored condition *cannot* be explained by the optimal joint order and demand learning policy, although formulating such an optimal policy is not the focus of the present study. To see why, observe that under the uncensored condition, the myopic policy is optimal: the order decision in the current period does not affect future periods, so one should place an order quantity to optimize expected profits in the current period. However, under censored demand, the current period order affects the potential demand information one receives: larger orders can potentially provide more information which may help inform future orders. For this reason, the optimal order quantity is actually *larger* than the myopic order under censored demand, especially in early periods (e.g., see Lariviere and Porteus 1999), although some recent evidence suggests that such "information stalking" usually does not yield large value (Besbes et al. 2015).

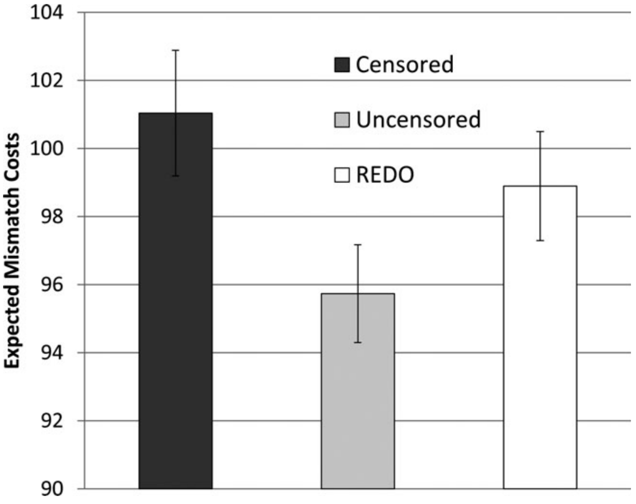
Expected mismatch costs. For inventory decisions, maximizing expected profit is equivalent to minimizing the expected costs of having inventory that is mismatched with demand. Each order that an individual makes has an associated expected mismatch cost given the true demand distribution that they face, which we can estimate using standard techniques (e.g., see Cachon and Terwiesch 2013). We tested whether expected mismatch costs of orders differed by experimental condition. We conducted a linear regression of the effect of condition on expected mismatch costs with standard errors clustered by individual (see Figure 4).

Expected mismatch costs were significantly lower in the uncensored condition than in the censored condition, $t(146) = 2.26$, $p = 0.03$. In the REDO condition, expected mismatch costs were not significantly higher than in the uncensored condition, $t(146) = 1.46$, $p = 0.15$, but were also not significantly lower than in the censored condition, $t(146) = 0.87$, $p = 0.38$.

3.4. Study 1 Discussion

Individuals underestimated demand when facing censored data. However, when they reported an estimate of demand outcomes in every period (REDO), this bias decreased. This pattern was similarly reflected in inventory decisions which were biased downward when demand was censored but were improved with REDO. In line with the argument that REDO works by helping participants imagine a more representative sample, the REDO sample's mean was closer to the true demand mean than the sales sample.

Figure 4 **Mean Expected Mismatch Costs by Condition in Study 1 with Bars for Robust Standard Errors Clustered by Individual**



Moreover, those participants whose REDO samples were more representative of true demand also tended to report better demand beliefs and made less-downwardly biased order decisions.

In Study 1, participants in the REDO condition answered 30 questions about demand before their final mean demand estimate, while participants in the Censored condition answered zero questions about demand before their final mean demand estimate. Therefore, one might question how much of REDO's effectiveness is driven by increased and more frequent attention to demand. The following study seeks to rule out the possibility that REDO improves demand estimation performance by merely forcing participants to answer more questions about demand.

Study 1 also tested REDO with a fairly sophisticated subject pool: MBA students with completed coursework in Operations Management and Probability & Statistics at a highly ranked business school. Furthermore, the challenge was kept relatively simple in that the cost parameters were set to make the overage and underage cost equal. Under these conditions, there is clear evidence that the censorship bias exists separately from the pull-to-center effect and that REDO can help mitigate the censorship bias. It also suggests that REDO yields improvement even if decision-makers are already well educated with relevant knowledge in operations and statistics.

4. Study 2

Study 2 replicates the results of Study 1 and also benchmarks the effectiveness of REDO relative to another censored condition that requires similar amounts of attention to demand. We introduce a new "effort control" condition, under which participants face censored demand and are required to update their mean demand belief after every period. In this way, we seek to provide supporting evidence that REDO's effectiveness is not only due to getting participants to think more often about demand, but by helping them to think better about demand. Specifically, REDO facilitates people to correct the sample before estimating the mean, while EMD does not.

Study 2 also extends the results of Study 1 to consider the effectiveness of REDO on demand beliefs under asymmetric overage and underage costs. The asymmetric cost setting also allows us to explore the consequences of REDO on inventory decisions under conditions in which the pull-to-center effect is present. We focus on the common case in which the underage cost is greater than the overage cost, although we also discuss how the pull-to-center effect ought to effect inventory decisions in the opposite case.

4.1. Methods

One hundred seventy-four undergraduate students participated in the study; the average standard aptitude test scores among this population were at the 98th percentile nationally. According to self-reports, the sample was 51.6% female and 47.9% Caucasian; the average age was 20.3 (SD = 1.71), with a small number of individuals opting to not disclose their gender, ethnicity, or age. The most common academic majors represented in the sample were economics, government, and engineering.

Participants played a simulation very similar to that in Study 1 (see screenshot in the Appendix). However, in this experiment participants bought newspapers for $1 per unit to stock in their stand and sold them for $3 per unit. Excess newspapers were discarded for $0 per unit. Therefore, the per-unit opportunity cost of underordering (the "underage cost") was $2, while the per-unit cost of overordering (the "overage cost") was $1.

Participants were told the mean would stay the same for the duration of the game, and were told that it was somewhere between 400 and 800. In this experiment, we randomly assigned each participant an integer between 500 and 700 which would be their unknown stationary true demand mean. Identical to experiment 1, individuals made inventory decisions for 30 periods, receiving feedback that remained visible for the remainder of the game. After 30 periods, they provided an estimate of their true mean demand, m.

If the true demand mean m were known, the expected profit-maximizing order quantity would be to order more than m due to the larger costs of underage versus overage. Specifically, given a known m the expected profit-maximizing order quantity every period is:

$$q^m = m + 100\Phi^{-1}\left(\frac{2}{(2+1)}\right) \cong m + 43.1,$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function. Participants earned bonus compensation between \$5 and \$20, distributed via electronic Amazon gift cards, proportional to their profit performance adjusted for their respective demand mean.

4.2. Experimental Conditions

Participants were randomly assigned to one of four conditions (see Table 3). The first three conditions were identical to Study 1: *Censored*, *REDO*, and *Uncensored*. We also included a new condition, EMD as another point of comparison.

In the EMD condition, compared to in REDO, participants also faced censored demand but answered a different question at the end of each period, "What is your best estimate of m (the underlying mean demand) now?" These answers were recorded in a column and remained visible throughout the simulation. The EMD provides a point of comparison for REDO to test whether simply asking any question about demand each period is sufficient to improve performance or if REDO is inducing people to think in a helpful way. While both EMD and REDO require the participant to answer 30 questions about demand, REDO nudges people to correct the observed sales to capture lost sales before estimating the mean, while EMD does not.

Under uncensored demand, although we expected demand beliefs to be unbiased, we expected order decisions to be downwardly biased toward the demand mean due to the pull-to-center effect. However, we expected both order decisions and demand beliefs to be further downwardly biased under censored demand. Finally, we expected REDO to yield significantly higher demand beliefs and order decisions than under both the censored demand and EMD treatments.

4.3. Results

Mean demand beliefs. Given that participants had been randomly assigned to a demand distribution with mean between 500 and 700, each participant's true demand mean was subtracted from their estimate of their underlying demand mean. Figure 5 shows the average final demand beliefs relative to the true demand mean, for each of the four condition.

We began analyses by examining our primary interest: whether REDO can improve the accuracy of demand beliefs. A series of t-tests were conducted to test whether the final beliefs about the underlying demand mean was different across experimental conditions. Two individuals did not submit a final estimate of the underlying demand mean and were therefore omitted from this analysis (one in REDO and one in EMD). Estimates of the demand mean in the censored condition were significantly lower than estimates in the uncensored condition, $t(91) = 3.75$, $p < 0.001$.

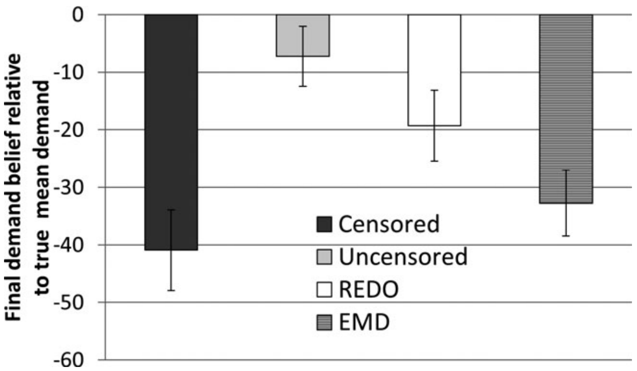
Record estimates of demand outcomes significantly improved demand beliefs. Estimates in the REDO condition were significantly higher than those in the censored condition, $t(86) = 2.24$, $p = 0.03$, and not significantly different than those in the uncensored condition, $t(79) = 1.50$, $p = 0.14$. However, final estimates of the demand mean in the EMD condition were not significantly different than those in the censored condition, $t(89) = 0.88$, $p = 0.38$, and were significantly lower than those in the uncensored condition, $t(82) = 3.29$, $p = 0.002$.

The average estimates of the mean demand were also compared to the true demand mean. On average, individuals in the censored, $t(49) = 5.84$, REDO, $t(37) = 3.13$, and EMD, $t(40) = 5.69$, conditions underestimated the true mean demand, $p < 0.01$ for each. The estimates of participants in the uncensored condition were not significantly different than the true mean demand, $t(42) = 1.38$, $p = 0.17$.

Table 3 Names (in bold) and Descriptions of the Three Experimental Conditions in Study 2

Experimental conditions	*N*	Description	Participant inputs collected
Censored and no intervention	50	Unobservable lost sales after stockouts	30 order decisions, 1 final mean demand estimate
Censored and Record Estimates of Demand Outcomes (**REDO**)	39	Unobservable lost sales after stockouts Estimate (if stockout) or report (if no stockout) that period's demand	30 order decisions, 30 demand outcome estimates from REDO, 1 final mean demand estimate
Censored and Estimate Mean Demand (**EMD)**	42	Unobservable lost sales after stockouts Estimate the underlying mean demand after every period	30 order decisions, 30 mean demand belief updates, 1 final mean demand estimate
Uncensored	43	Observable lost sales after stockouts	30 order decisions, 1 final mean demand estimate

Figure 5 Mean Final Demand Beliefs Across Experimental Conditions in Study 2, Shown with Standard Error Bars



REDO responses. Next, we focused our attention on the responses of the individuals in the REDO condition. REDO improved beliefs about mean demand, but were the recorded and estimated demand outcomes generated by individuals in the REDO condition more representative of the true underlying demand than sales outcomes? Indeed, on average, the mean of the REDO sample was 18.3 units higher than the mean sales observed, $SD = 12.03$, $t(38) = 9.98$, $p < 0.001$. However, the average of the REDO sample was still 25.8 units lower than the true mean demand, $SD = 23.57$, $t(38) = 6.91$, $p < 0.001$.

The mean of their self-generated REDO sample closely corresponded to their final estimate of mean demand. The average of their REDO sample was only 6.06 units lower than their final estimate of the true mean demand ($SD = 30.31$), which was not a significant difference, $t(38) = 0.52$, $p < 0.60$. Similarly, when looking only at the last 10 periods of their REDO sample, the average was only 6.52 units lower than their final estimate of the true mean demand ($SD = 29.12$), which was not a significant difference, $t(38) = 1.38$, $p = 0.18$. Furthermore, across individuals, the mean of the REDO sample correlated positively with their final estimates of the true demand mean ($r = 0.60$, $p < 0.001$). Final estimates of mean demand were also closer to the mean of their REDO sample than to the mean of their observed sales, $t(38) = 2.00$, $p = 0.05$. Therefore, the set of numbers individuals enter when engaging with the behavioral intervention seem to shape the ultimate perceptions individuals develop of the underlying demand.

Statistical benchmark. We again calculated the maximum likelihood estimate for the mean demand by the end of the game for each player who faced censored demand (the Censored, REDO, and EMD conditions), using the same technique as in Study 1. The statistical benchmark was not significantly different from the true mean demand ($M = -3.34$, $SD = 24.93$), $t(129) = 1.53$, $p = 0.13$.)

In the censored condition, final estimates of mean demand were significantly lower than the statistical benchmark given the exact same observations of sales and stockouts, $t(48) = 6.01$, $p < 0.001$. In the REDO condition, despite being significantly higher than in the censored condition, as previously shown, final estimates of mean demand were still lower than the statistical benchmark, $t(48) = 6.01$, $p < 0.001$. Final estimates were also lower than the statistical benchmark in the EMD condition, $t(40) = 5.70$, $p < 0.001$.

Order decisions. To investigate order decisions, a linear regression was conducted in SAS with standard errors clustered by individual.[1] Refer to Table 4. The dependent variable was the individual's order decision for a given period. The key independent variable was experimental condition (Censored, REDO, EMD, Uncensored). As explanatory variables, three dummy variables were included in the model, one for each of the REDO, EMD, and Uncensored conditions, thereby treating the Censored condition as the baseline case.

As in the previous study, to account for the variation in true demand means, and specific demand draws, we controlled for the running average of past demand draws at the time of a given ordering decision. Once again, in period 1, this variable was set equal to 600.

Table 4 The Regression Model of Inventory Ordering in Study 2

	DV: Inventory order	
	(1)	(2)
Intercept	126.08**	126.11**
	(25.01)	(25.09)
Running Avg. Demand	0.76**	0.76**
	(0.04)	(0.04)
Censored condition	(Baseline condition for the regression)	
Uncensored condition	25.96**	25.96**
	(8.89)	(8.89)
REDO condition	18.41*	18.41*
	(9.11)	(9.12)
EMD condition	−1.89	−1.88
	(8.73)	(8.73)
Period		1.27**
		(0.47)
Period × Uncensored		−0.52
		(0.56)
Period × REDO		−1.25*
		(0.63)
Period × EMD		−0.72
		(0.60)

Notes: The numbers in parentheses are standard errors clustered by subject. The variable Running Avg. Demand is the cumulative average of random demand draws at the time of the decision, except in period 1, in which it is equal to 600. The variables Uncensored, REDO, and EMD are equal to 1 if a participant is in the respective condition and are otherwise equal to 0. The variable Period is mean-centered (30 periods). The number of observations is 5160 and the degrees of freedom for *t*-tests are 171.
*Denotes $p < 0.05$ and **denotes $p < 0.01$.

In Model 2, we introduce the possibility of time trends that may differ across conditions. We added period (mean-centered) and interaction terms between period and condition. Within the framework of this model, we can then test whether the differences across condition are significant in a given period, accounting for time trends that may be different across conditions.

How did experimental conditions differ from one another with respect to orders? Orders in the censored condition were significantly lower than those in the uncensored condition, $t(171) = 2.92$, $p < 0.01$. Orders in the REDO condition were significantly higher than those in the censored condition, $t(171) = 2.02$, $p = 0.04$, and not different from those in the uncensored condition, $t(171) = 0.99$, $p = 0.32$.

On the other hand, estimating the mean demand each period did *not* improve decision-making. Orders in the EMD condition were not different than those in the censored condition, $t(171) = 0.22$, $p = 0.83$, and were lower than those in the uncensored condition, $t(171) = 3.87$, $p < 0.01$.

Model 2 shows that the effect holds while accounting for period (mean-centered). The model predicts significantly higher orders in the REDO condition than in the censored condition in the average period. As for the previous experiment, here we make atheoretic observations on the effect of REDO over time. Model 2 (see Table 4) shows that the difference between REDO and Censored was decreasing as period increased. Plots of the data show that the difference between those two conditions was extremely large in the first five periods, moderately large in periods 6–10, and then stabilized at a smaller gap for the remaining periods, with the average improvement of REDO over Censored in the last five periods being 8.21 units. When forcing a linear trend on the difference between REDO and Censored, the model predicts the difference to reduce to 0.33 in period 30 ($SE = 12.01$), $t(171) = 0.03$, $p = 0.98$.

Overall, orders tended to be much lower than q^m. In all four feedback conditions, orders were significantly lower than q^m (demand mean + 43.1), $p < 0.001$ for each. This behavior is consistent with the well-documented "pull-to-center" effect—individuals do not sufficiently account for asymmetry between overage and underage costs and therefore order too close to the demand mean—which has been studied with a known demand distribution (Bolton and Katok 2008, Schweitzer and Cachon 2000). While orders in all conditions were biased low, orders in the uncensored and REDO conditions were significantly higher than in the censored and EMD conditions, suggesting better performance in the former two conditions.

In summary, individuals who observed censored demand and received the REDO intervention behaved similarly to individuals with uncensored demand. However, individuals who observed censored demand and performed the EMD task behaved similarly to individuals with censored demand and no additional task.

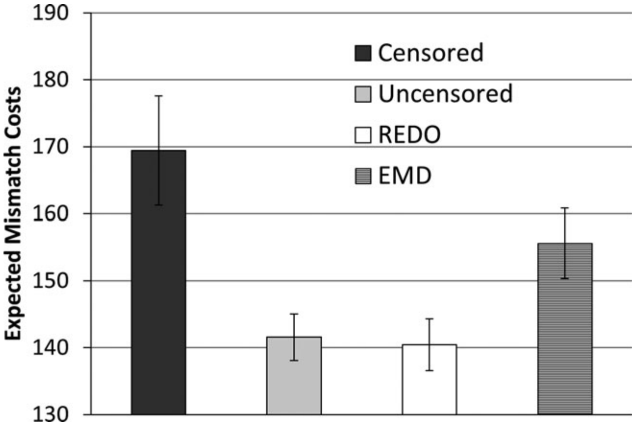
Expected mismatch costs. We tested whether expected mismatch costs of orders differed by experimental condition. A linear regression of the effect of condition on expected mismatch costs with standard errors clustered by individual was conducted. The expected mismatch costs by condition can be found in Figure 6. Expected mismatch costs were significantly higher in the censored condition than in the uncensored condition, $t(171) = 3.12$, $p = 0.002$. REDO significantly decreased expected mismatch costs relative to the censored condition, $t(171) = 3.19$, $p = 0.002$. There was no difference in mismatch costs between the REDO and uncensored conditions, $t(171) = 0.21$, $p = 0.83$.

On the other hand, mismatch costs in the EMD condition were not significantly different than those in the censored condition, $t(171) = 1.42$, $p = 0.16$, and were significantly worse than those in the REDO condition, $t(171) = 2.31$, $p = 0.02$, and uncensored condition, $t(171) = 2.21$, $p = 0.03$.

4.4. Study 2 Discussion

Does REDO reduce the censorship bias even in settings in which overage and underage costs are asymmetric? Similar to Study 1, Study 2 provides evidence that REDO effectively improves demand beliefs under asymmetric overage and underage costs. It also increases order decisions such that it lowers the difference between orders in the censored and uncensored conditions.

Figure 6 Mean Expected Mismatch Costs by Condition in Study 2 with Bars for Robust Standard Errors Clustered by Individual



The evidence suggests that the benefit of REDO is not simply from simply asking participants an additional question about demand each period: no such improvement was found in the EMD condition, in which individuals recorded an estimate of the underlying demand mean each period. This evidence supports the idea that REDO helps people to not just think harder, but in the right direction.

It is noteworthy that in this study, in which the underage cost was greater than the overage cost, censored demand led to inventory orders that were not only biased below q^m but also below the true demand mean (i.e., biased downward beyond the "center" in the pull-to-center effect). This suggests that censored demand affects order decisions above and beyond the pull-to-center effect mechanism, consistent with the downwardly biased demand beliefs. In the opposite setting in which the overage cost is greater than the underage cost, it is natural to conclude that the censorship bias will at least partially counteract the pull-to-center effect: the downward-biased demand beliefs may cancel out the upward bias from the pull-to-center effect. Of course, an effective inventory decision in such a case would be the result of a happy coincidence rather than intelligent decision-making, and the demand beliefs will likely lead to other undesirable consequences.

Finally, although the magnitude of the pull-to-center effect with uncensored demand is not the focus of the present study, it is worth observing that the pull-to-center effect is very strong in this experiment: in the uncensored condition orders were not significantly larger than mean demand. We suspect that the reason we find a stronger pull-to-center effect than traditionally found in the literature (e.g., Bolton and Katok 2008, Schweitzer and Cachon 2000) is that demand was unknown. Subjects' attention was largely on determining the unknown demand mean, which reduced their attention toward the asymmetric cost-balancing task. Also, the overall upward trend in ordering over time is consistent with existing evidence that suggests that subjects' pull-to-center effect reduces over time with a known demand distribution and uncensored feedback (Bolton and Katok 2008).

5. Study 3

Study 3 tests whether REDO improves performance in a different setting: inferring mean demand based on graphical sales data in review. Managers commonly use historical graphs of sales data to inform not only inventory decisions, but also budgeting, pricing, promotions, and availability decisions. In Study 3, we consider how to implement REDO in this kind of common graphical format. We then test whether REDO improves demand beliefs in this managerial setting, benchmarking it against a Censored condition and an Effort control condition.

Study 3 also tests whether REDO improves demand beliefs independently from the inventory decision-making process. We isolate REDO's ability to improve demand beliefs from the inventory decision-making process by making the inventory levels *exogenous*. If the way REDO improved demand beliefs in the repeated newsvendor game of Studies 1 and 2 was only through the inventory decision-making process, then it will no longer be effective when the inventory levels are exogenous. If, on the other hand, the increase in the inventory levels in Studies 1 and 2 were driven (at least in part) by REDO's ability to improve beliefs about demand as we theorize, then REDO should also improve demand beliefs even when no inventory decisions are required. In this manner, Study 3 sheds light on the behavioral mechanism by which REDO improved performance in Studies 1 and 2.

5.1. Methods

We targeted a sample size of 200 subjects by recruiting for a pre-determined 13 laboratory sessions (capacity 20 per session). A total of 210 subjects participated in the study, all undergraduate (85%) or graduate (15%) students. The majority (92%) of subjects were full-time students; 60% also had a part-time or full-time job. Two-thirds of the sample were female. Sixty-four percent self-identified as White, 28% as Asian, and 3% as Black. Seventy-seven percent selected English as their first language; 93% had lived in the United States for at least 1 year, 77% for at least 5 years.

Participants played the role of a business analyst whose job is to read sales graphs and answer questions about the data. The simulated task was programmed using Delphi (see screenshots in the Appendix).

Each participant completed seven rounds within the same condition. Across all conditions, in each round, the participant was shown a bar graph that contained historical sales data for 30 days for a simulated product. Each bar was also colored either blue (indicating the product did not sell out that day) or red (indicating the product did sell out that day).

Participants were informed that (i) each sales graph reflects a randomly selected product (ii) within each graph, daily customer demand is independent and normally distributed with standard deviation 200, but unknown mean, and (iii) the true mean demands for products are normally distributed with mean 750, standard deviation 50.

In contrast to Studies 1 and 2, there were no inventory decisions in this experiment. For all products, the daily inventory levels were exogenously simulated with mean 750, standard deviation 10. Thus, if a randomly generated product had a high (low) demand mean, then the graph showed a high (low) stockout rate.

Across all conditions, the incentivized task in each round was to look at the sales graph and answer the question "What do you think is the average daily customer demand over the 30 days?" In addition to a $5 participation compensation, participants were paid a bonus of $10—(sum of errors × $.01). At the end of the game, participants received feedback on their overall accuracy over the seven rounds. There was no feedback between rounds.

5.2. Experimental Conditions

Participants were randomly assigned to one of three conditions (see Table 5). In all three conditions, participants try to EMD from the graph of censored data. However, we vary the tasks that the system asks the participant to complete before answering this question.

In the base condition (*Censored*), participants are not asked to complete any tasks before guessing the average demand.

In *REDO*, participants are asked to record an estimate of the demand for every day for which there was a stockout (i.e., for all the red bars.) After recording an estimate for a red bar, the graph updated to reflect the demand outcome estimate for that day and the bar turned blue. The participant could start over at any time, but otherwise the sales levels for stockout days did not remain visible once they adjusted the red bars. Finally, once the participant finishes placing an estimate for each red bar and all the bars are blue, they answer the question of what they think is the average demand for the product. We designed REDO in this manner in order to encourage the participant to ignore the original sales data when making their final demand estimate.

Finally, in the effort control condition (Effort), participants were also asked to record estimates for some bars before making a final guess for the mean demand as in the REDO condition. However, in contrast with the REDO condition, in the effort condition, these estimates were simply the number of sales in certain days, and were not always for stockout days. We programmed Effort to first calculate the number of stockouts in the graph. Then the program would ask the participant to simply record the sales for that same number of randomly selected days. Although this task does not require making any adjustments, the participant still must exert effort to try to accurately estimate the size of each bar and to record a number. In this way, relative to REDO, the effort condition holds constant the number of estimations occurring and the level of engagement generated by the intervention, but does not facilitate the construction of a more representative sample. We hypothesized that REDO would improve demand beliefs over both the censored condition and the effort control condition.

5.3. Results

Mean demand beliefs. A linear regression was conducted in SAS with standard errors clustered by individual to account for the non-independence of multiple observations for each participant.[2] The dependent variable was the adjusted mean demand estimate: a graph's true underlying demand mean was subtracted from the participant's estimate of the true demand mean for that graph. The independent variable of interest was the experimental condition: REDO, Censored, or Effort. REDO was designated as the baseline condition with dummy variables included for the other two conditions. Specifically, Censored equal to 1 corresponded to being in the censored condition, Effort equal to 1 corresponded to being in the effort condition, and both condition variables being equal to zero corresponded to being in the REDO condition.

The results from two regression models can be seen in Table 6. From model 1, mean demand beliefs in the

Table 5 Names (in bold) and Descriptions of the Three Experimental Conditions in Study 3

Experimental conditions	*N*	Description	Participant inputs collected
Censored and no intervention	66	Observe the censored sales and stockout data. Then, estimate mean demand	1 final mean demand estimate (for each of seven graphs)
Censored and Record Estimates of Demand Outcomes (**REDO**)	65	Observe the censored sales and stockout data. Redraw each sales bar associated with a stockout by recording an estimate of demand for those days. Then, estimate mean demand	<30 daily demand outcome estimates for stockout days from REDO, 1 final mean demand estimate (for each of seven graphs)
Censored with **Effort** control task	75	Observe the censored sales and stockout data. Record the sales for randomly selected days (programmed to be equal to the number of stockout days) by reading the bar graph. Then, estimate mean demand	<30 random daily sales estimates equal to the number of stockout days, 1 final mean demand estimate (for each of seven graphs)

Table 6 The Regression Models of Mean Demand Estimates in Study 3

	DV: Estimate of mean minus true mean	
	(1)	(2)
Intercept	−57.99**	−57.12**
	(7.92)	(7.75)
REDO condition	(Baseline condition for the regression)	
Censored condition	−23.80*	−24.75*
	(10.99)	(10.89)
Effort condition	−28.63**	−28.68**
	(10.16)	(10.02)
Round		−1.70
		(1.51)
Round × Censored		1.86
		(2.11)
Round × Effort		0.06
		(2.09)

Notes: The numbers in parentheses are standard errors clustered by subject. The variable Round is mean-centered (each participant faced seven distinct graphs, about which they made an estimate). The variables Censored and Effort are equal to one if a participant is in the respective condition and are otherwise equal to 0. The number of observations is 1437 and the degrees of freedom for *t*-tests are 205.
*Denotes $p < 0.05$ and **denotes $p < 0.01$.

REDO condition were significantly higher than those in the censored condition, $t(205) = 2.16$, $p = 0.03$. There was no significant difference between mean demand beliefs in the censored and effort conditions, $t(205) = 0.48$, $p = 0.63$. Mean demand beliefs in censored were also significantly higher than those in the effort condition, $t(205) = 2.81$, $p = 0.005$. Since estimates of the true mean demand were significantly biased low in all three conditions, $p < 0.001$, the higher estimates in the REDO condition represented a significant improvement over the estimates in the other two conditions.

These patterns are also present in model 2, which includes round (mean-centered) and round by condition variables. In this experiment, the role of time represents experience with different, independent products with no feedback; by contrast, in the previous two experiments, time represented repeated experience with the same product and feedback. As evidenced by model 2 (see Table 6), time (i.e., Round) had no effect on the magnitude of difference between the REDO and censored conditions. Plots of the data show stable patterns over time: estimates of mean demand in the censored condition were consistently biased low, but were consistently higher in the REDO condition. For example, in the last round of the experiment, estimates of mean demand in the REDO condition were on average of 26.31 units higher than in Censored condition.

REDO responses. In the REDO condition, participants reported estimates of daily demand in periods with stockouts. On average, the mean of the REDO sample was 96.16 units higher than the mean sales observed by the same individual, $SD = 66.62$, $t(64) = 11.64$, $p < 0.001$. In fact, the mean of the REDO sample was not significantly different than the true mean demand for that graph; on average, it was only 15.29 higher than the true mean, $SD = 76.84$, $t(64) = 1.60$, $p = 0.11$. The mean of the REDO sample correlated positively with the final estimates of the true demand mean ($r = 0.29$, $p < 0.001$). The average of their REDO sample was 76.31 higher than their estimate of the true mean demand, $SD = 100.93$, which was significant, $t(64) = 6.10$, $p < 0.001$. Final estimates of mean demand were significantly closer to the mean of their REDO sample than to the mean of their initially observed sales, $t(64) = 2.62$, $p < 0.01$.

Statistical benchmark. We verified the study design by evaluating the performance of the same statistical benchmark used in Studies 1 and 2 for each graph that a participant observed in Study 3. Overall, the statistical benchmark had an average error of −1.32 with standard deviation 44.06, which was not different from zero, $t(143) = 1.14$, $p = 0.26$. In all three conditions, final estimates of mean demand were significantly lower than the statistical benchmark given the exact same observations of sales and stockouts, $p < 0.001$ for each.

5.4. Study 3 Discussion

Study 3 provides evidence that REDO helps to improve demand beliefs in graphical sales data in review task, which serves as an input for a variety of managerial decisions. In this way, it helps broaden the conditions in which REDO is applicable beyond the newsvendor setting and beyond a periodic review setting. The study also implements exogenously determined inventory levels (as opposed to endogenously determined inventory levels in the previous two studies) thereby isolating demand estimation as an important driving factor for the success of REDO.

6. Discussion

In many cases, operations managers face censored demand that leads to a potentially costly censorship bias: demand beliefs are biased low under censored demand. By explicating the psychological underpinnings of bias caused by censored demand, we have proposed a behavioral remedy, REDO, for reducing the censorship bias and improving demand beliefs in a nonrestrictive manner. REDO involves having people estimate demand realizations, thereby helping to create a more representative demand sample. Three experimental studies provided evidence of REDO's effectiveness in helping subjects form more accurate demand beliefs in repeated newsvendor and graphical data-in-review settings.

Our proposed remedy not only helps us shed light on why the censorship bias occurs, but it also has practical value. REDO can be implemented directly into system architecture (as they were in our experiments) or indirectly through training and education. In fact, the authors have used our simulations to teach students both the perils of the censorship bias and the effectiveness of REDO as a practical solution. The nonrestrictive nature of our remedy makes it particularly valuable in situations where it is difficult to solve for the optimal forecast analytically, such as non-stationary settings or settings that are difficult to specify mathematically.

This study also has implications for the goal of improving inventory ordering decisions, which has received a significant amount of attention in the behavioral operations literature. However, we stress that REDO is designed to reduce the bias on demand beliefs due to censorship; it is *not* intended to address ordering biases that are present even with known and uncensored demand distributions. Specifically, the intervention is not intended to be effective at eliminating the pull-to-center effect or high-variance ordering behavior (e.g., demand chasing). Future work may investigate the efficacy of implementing our remedy, which is designed to improve demand beliefs with unknown and censored demand, in combination with another intervention aimed at reducing newsvendor biases that occur even with known and uncensored demand distributions. For example, practically, one might use REDO to help elicit a better demand forecast and then automate the cost-balancing task, which is relatively simple once a demand forecast distribution is provided (Schweitzer and Cachon 2000).

It is also worth mentioning that although our first two studies take place in the well-studied repeated newsvendor setting, one could also apply REDO in the same way in other periodic review inventory control settings, such as the base stock model setting. In these settings, we would anticipate that REDO would continue to improve subjects' demand estimations. However, it is difficult to make clear predictions about the net impact on inventory order behavior in all settings because inventory decisions in other non-newsvendor settings are typically more complicated and less well-documented.

There may also be other context-specific improvements that can be made to REDO that may be worth investigating. In our experiments, REDO does not provide the user with any additional information or calculations. Nevertheless, if the system designer has reliable information, adjusting REDO to provide this additional useful information may lead to further improvement. For example, the system could provide as a default REDO value the uncensored observation of a similar product from last year. Such additional information may help individuals self-generate even more representative samples while still providing managers freedom to make modifications (which is important for getting people to use the recommendations, see Dietvorst et al. 2016). However, REDO as presented in our experiments (without providing any additional information) has the advantage of being able to accommodate dynamic environments in which the user may have better information than the system designer. It might also be possible to alter REDO to reduce the effort required, for example, by simply having people estimate the average number of lost sales over all the stockout periods before making a final demand estimate. We conjecture, however, that such an intervention would not be as effective because it does not as thoroughly increase the concreteness of the data in stockout periods nor nudge the participant to mentally simulate the lost sales for every period.

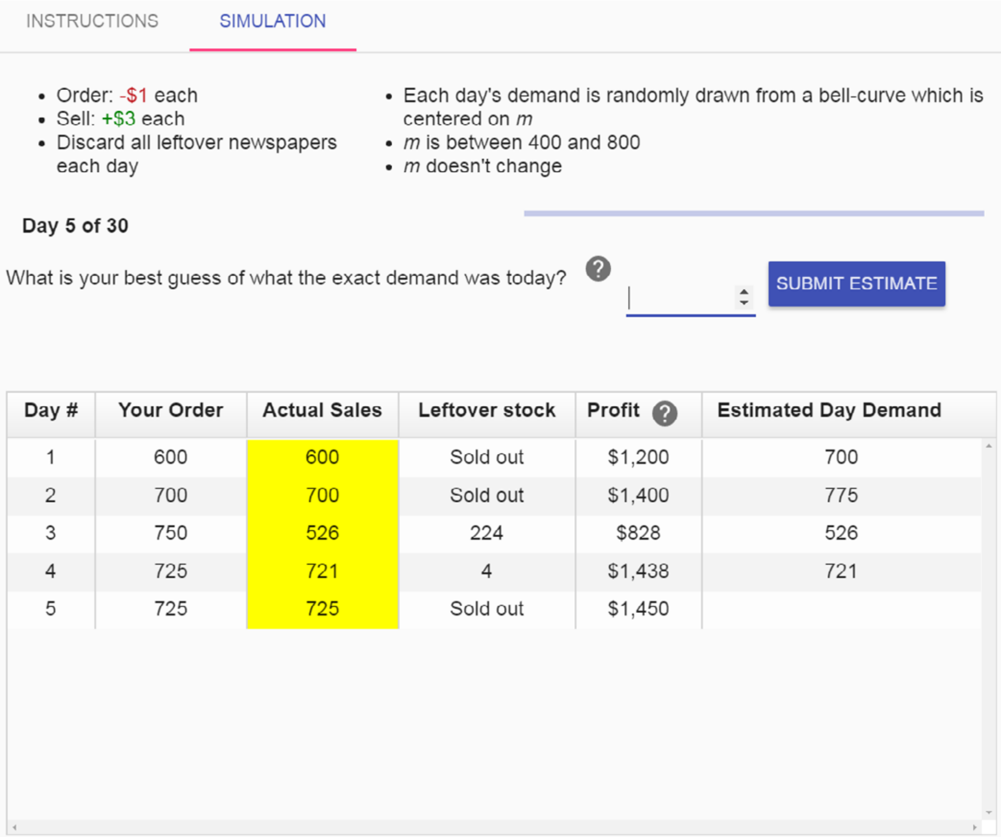
Lastly, the general approach taken in this study to develop REDO may be a good starting point for debiasing in other "wicked environments" both in operations contexts and more generally. For example, the censorship bias may also occur in process analysis: a step's observed production rate is limited by the bottleneck capacity, so managers may underestimate capacities of non-bottleneck stages. Similarly, managers in new product development may be subject to survivorship bias: only successful projects are taken to completion, so managers may overestimate the return of new projects. Having managers explicitly self-generate plausible outcomes that are unobserved in such wicked environments to try to correct for the misrepresentative nature of wicked environments is a potentially fruitful direction for debiasing efforts.

Acknowledgments

The authors gratefully acknowledge helpful feedback and discussions with Jeannette Song, Peter Lukszys, the participants of the Decision Processes Seminar at the Wharton School and at INFORMS 2015. The authors would also like to thank the review team and especially the anonymous Senior Editor for their valuable feedback. Anatasia Ivantsova assisted in programming the experimental interface in Study 3. This research received financial support from the BRITE laboratory at the University of Wisconsin-Madison.

Appendix

Figure A1 Screenshot of the Simulation Interface in the Repeated Newsvendor Setting, in This Case the REDO Condition in Study 2 [Color figure can be viewed at wileyonlinelibrary.com]



- Order: -$1 each
- Sell: +$3 each
- Discard all leftover newspapers each day

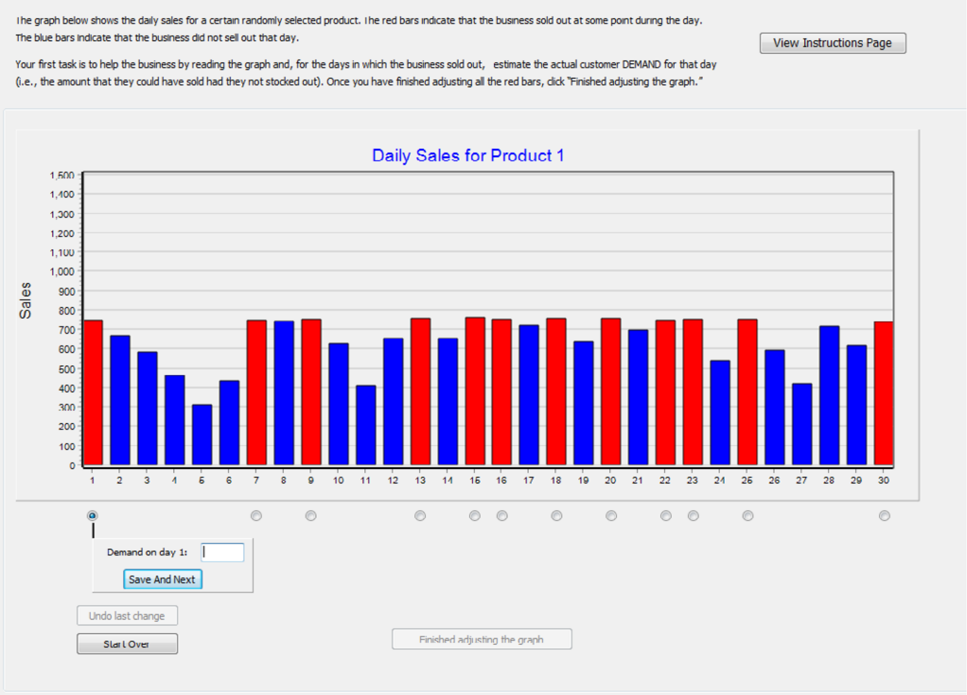
- Each day's demand is randomly drawn from a bell-curve which is centered on *m*
- *m* is between 400 and 800
- *m* doesn't change

Day 5 of 30

What is your best guess of what the exact demand was today?

Day #	Your Order	Actual Sales	Leftover stock	Profit	Estimated Day Demand
1	600	600	Sold out	$1,200	700
2	700	700	Sold out	$1,400	775
3	750	526	224	$828	526
4	725	721	4	$1,438	721
5	725	725	Sold out	$1,450	

Figure A2 Screenshot of the Simulation Interface for the Graphical Sales Data in Review Setting, in This Case the REDO Condition in Study 3 [Color figure can be viewed at wileyonlinelibrary.com]



The graph below shows the daily sales for a certain randomly selected product. The red bars indicate that the business sold out at some point during the day. The blue bars indicate that the business did not sell out that day.

Your first task is to help the business by reading the graph and, for the days in which the business sold out, estimate the actual customer DEMAND for that day (i.e., the amount that they could have sold had they not stocked out). Once you have finished adjusting all the red bars, click "Finished adjusting the graph."

Notes

[1] Regression diagnostics identified orders of two participants as outliers. The diagnostics were conducted on average orders relative to the respective true demand mean. The Studentized residuals for the two observations were −5.94 (ordering an average of 320.7 units below their true demand mean in the Uncensored condition) and 4.69 (ordering an average of 217.7 units above their true demand mean in the Censored condition), far beyond the traditionally recommended cutoff of ±3.0, suggesting that they were inflating the standard errors in the model. These were the only two observations beyond the cutoff (orders relative to true mean among all other participants: $M = -10.5$, $SD = 45.9$, $SE = 3.4$). Similarly, the CovRatios were 0.41 and 0.64, implying that they were generating significant instability in the model and exaggerating the standard errors of parameter estimates. They were omitted for the analyses reported here; we suspect these participants did not read the instructions. With them included the same patterns hold, although statistical comparisons between conditions are less significant due to the standard errors being inflated by an average of 17% per comparison.

[2] Across two participants, five estimates were missing.

References

Arkes, H. R. 1991. Costs and benefits of judgment errors: Implications for debiasing. *Psychol. Bull.* **110**(3): 486–498.

Armstrong, J. S. 1975. The use of the decomposition principle in making judgments. *Org. Behav. Hum. Decis. Process* **14**(2): 257–263.

Besbes, O., J. Chaneton, C. C. Moallemi. 2015. The exploration-exploitation trade-off in the newsvendor problem. Working paper, Columbia Business School.

Bolton, G. E., E. Katok. 2008. Learning by doing in the newsvendor problem: A laboratory investigation of the role of experience and feedback. *Manuf. Serv. Oper. Manag.* **10**(3): 519–538.

Cachon, G., C. Terwiesch. 2013. *Matching Supply with Demand: An Introduction to Operations Management*, 3rd edn. McGraw-Hill, Singapore.

Camerer, C. F., R. M. Hogarth. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk Uncertainty* **19**(1): 7–42.

Chen, L., A. J. Mersereau. 2015. Analytics for operational visibility in the retail store: The cases of censored demand and inventory record inaccuracy. N. Agrawal, S. A. Smith, eds. *Retail Supply Chain Management*. Springer, New York, 79–112.

Chen, L., A. G. Kök, J. D. Tong. 2013. The effect of payment schemes on inventory decisions: The role of mental accounting. *Management Sci.* **59**(2): 436–451.

Croson, R., K. Donohue. 2003. Impact of POS data sharing on supply chain management: An experimental study. *Prod. Oper. Manag.* **12**(1): 1–11.

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

Croson, R., K. Donohue, E. Katok, J. Sterman. 2014. Order stability in supply chains: Coordination risk and the role of coordination stock. *Prod. Oper. Manag.* **23**(2): 176–196.

Davis, A. M., S. G. Leider. 2015. Contracts and capacity investment in supply chains. Working paper, Cornell University.

Dietvorst, B. J., J. P. Simmons, C. Massey. 2016. Overcoming algorithm aversion: People will use algorithms if they can (even slightly) modify them. *Management Sci.*, Forthcoming, https://doi.org/10.1287/mnsc.2016.2643

Ding, X., M. L. Puterman, A. Bisi. 2002. The censored newsvendor and the optimal acquisition of information. *Oper. Res.* **50**(3): 517–527.

Epley, N., T. Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychol. Sci.* **17**(4): 311–318.

Feiler, D. C., J. D. Tong, R. P. Larrick. 2013. Biased judgment in censored environments. *Management Sci.* **59**(3): 573–591.

Fildes, R., P. Goodwin. 2007. Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces* **37**(6): 570–576.

Fildes, R., P. Goodwin, M. Lawrence, K. Nikolopoulos. 2009. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *Int. J. Forecast.* **25**(1): 3–23.

Fischoff, B. 1982. Debiasing. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, NY, 422–444.

Frederick, S. 2005. Cognitive reflection and decision making. *J. Econ. Perspect.* **19**(4): 25–42.

Haran, U., D. A. Moore, C. K. Morewedge. 2010. A simple remedy for overprecision in judgment. *Judgm. Decis. Making* **5**(7): 467–476.

Heath, C., D. Heath. 2013. *Decisive: How to Make Better Choices in Life and Work*. Random House, New York.

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

Hogarth, R. M. 2001. *Educating Intuition*. University of Chicago Press, Chicago, IL.

Hogarth, R. M., T. Lejarraga, E. Soyer. 2015. The two settings of kind and wicked learning environments. *Curr. Dir. Psychol. Sci.* **24**(5): 379–385.

Juslin, P., A. Winman, P. Hansson. 2007. The naive intuitive statistician: A naive sampling model of intuitive confidence intervals. *Psychol. Rev.* **114**(3): 678–703.

Kareev, Y., S. Arnon, R. Horwitz-Zeliger. 2002. On the misperception of variability. *J. Exp. Psychol. Gen.* **131**(2): 287–297.

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

Kremer, M., S. Minner, L. N. Van Wassenhove. 2010. Do random errors explain newsvendor behavior? *Manuf. Serv. Oper. Manag.* **12**(4): 673–681.

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

Lariviere, M. A., E. L. Porteus. 1999. Stalking information: Bayesian inventory management with unobserved lost sales. *Management Sci.* **45**(3): 346–363.

Larrick, R. P. 2009. Broaden the decision frame to make effective decisions. E. A. Locke, ed. *Handbook of Principles of Organizational Behavior*. Wiley and Sons, Hoboken, NJ, 461–480.

Lee, Y., E. Siemsen. 2017. Task decomposition and newsvendor decision making. *Management Sci.* **63**(10): 3226–3245. https://doi.org/10.1287/mnsc.2016.2521.

Lu, X., J. S. Song, K. Zhu. 2008. Analysis of perishable-inventory systems with censored demand data. *Oper. Res.* **56**(4): 1034–1038.

Lurie, N. H., J. M. Swaminathan. 2009. Is timely information always better? The effect of feedback frequency on decision making. *Organ. Behav. Hum. Decis. Process.* **108**(2): 315–329.

MacGregor, D., S. Lichtenstein, P. Slovic. 1988. Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. *Organ. Behav. Hum. Decis. Process.* **42**(3): 303–323.

Millard, S. P. 2013. *EnvStats: An R Package for Environmental Statistics*. Springer, New York. ISBN 978-1-4614-8455-4, http://www.springer.com.

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

Moritz, B. B., E. Siemsen, M. Kremer. 2014. Judgmental forecasting: Cognitive reflection and decision speed. *Prod. Oper. Manag.* **23**(7): 1146–1160.

Mussweiler, T., F. Strack. 1999. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *J. Exp. Soc. Psychol.* **35**(2): 136–164.

Nahmias, S. 1994. Demand estimation in lost sales inventory systems. *Nav. Res. Log.* **41**(6): 739–758.

Narayanan, A., B. B. Moritz. 2015. Decision making and cognition in multi-echelon supply chains: An experimental study. *Prod. Oper. Manag.* **24**(8): 1216–1234.

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

Özer, Ö., Y. Zheng, Y. Ren. 2014. Trust, trustworthiness, and information sharing in supply chains bridging China and the United States. *Management Sci.* **60**(10): 2435–2460.

Quattrone, G. A. 1982. Overattribution and unit formation: When behavior engulfs the person. *J. Pers. Soc. Psychol.* **42**(4): 593–607.

Raiffa, H. 1968. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Addison-Wesley, Oxford.

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

Rudi, N., D. Drake. 2014. Observation bias: The impact of demand censoring on newsvendor level and adjustment behavior. *Management Sci.* **60**(5): 1334–1345.

Rydval, O., A. Ortmann. 2004. How financial incentives and cognitive abilities affect task performance in laboratory settings: An illustration. *Econ. Lett.* **85**(3): 315–320.

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

Sloman, S. A. 1996. The empirical case for two systems of reasoning. *Psychol. Bull.* **119**(1): 3–22.

Stanovich, K. E., R. F. West. 1998. Individual differences in rational thought. *J. Exp. Psychol. Gen.* **127**(2): 161–188.

Thaler, R. H., C. R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Books, New York.

Wu, D., E. Katok. 2006. Learning, communication and the bullwhip effect. *J. Oper. Manag.* **24**(6): 839–850.

Zhang, Y., K. Donohue, T. H. Cui. 2016. Contract preferences for the loss averse supplier: Buyback versus revenue sharing. *Management Sci.* **62**(6): 1734–1754.

Zhao, Y., X. Zhao, Z. J. M. Shen. 2016. On learning process of a newsvendor with censored demand information. *J. Oper. Res. Soc.* **67**: 1200–1211.