Management Science

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

Designing Buyback Contracts for Irrational But Predictable Newsvendors

Michael Becker-Peth, Elena Katok, Ulrich W. Thonemann,

To cite this article:
Michael Becker-Peth, Elena Katok, Ulrich W. Thonemann, (2013) Designing Buyback Contracts for Irrational But Predictable Newsvendors. Management Science 59(8):1800-1816. http://dx.doi.org/10.1287/mnsc.1120.1662

MANAGEMENT SCIENCE
Vol. 59, No. 8, August 2013, pp. 1800–1816
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1662

Designing Buyback Contracts for Irrational But Predictable Newsvendors

Michael Becker-Peth
Department of Supply Chain Management and Management Science, University of Cologne, D-50923 Cologne, Germany, michael.becker-peth@uni-koeln.de

Elena Katok
Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080, ekatok@utdallas.edu

Ulrich W. Thonemann
Department of Supply Chain Management and Management Science, University of Cologne, D-50923 Cologne, Germany, ulrich.thonemann@uni-koeln.de

One of the main assumptions in research on designing supply contracts is that decision makers act in a way that maximizes their expected profit. A number of laboratory experiments demonstrate that this assumption does not hold. Specifically, faced with uncertain demand, decision makers place orders that systematically deviate from the expected profit maximizing levels. We have added to this body of knowledge by demonstrating that ordering decisions also systematically depend on individual contract parameters and by developing a behavioral model that captures this systematic behavior. We proceed to test our behavioral model using laboratory experiments and use the data to derive empirical model parameters. We then test our approach in out-of-sample validation experiments that confirm that, indeed, contracts designed using the behavioral model perform better than contracts designed using the standard model.

Key words: newsvendor; behavioral operations; experimental; order behavior; contract optimization
History: Received April 12, 2011; accepted September 2, 2012, by Christian Terwiesch, operations management. Published online in *Articles in Advance* February 15, 2013.

1. Introduction

Supply contracts are used to align the incentives of channel members, and there exists a rich body of literature that analyzes how different types of supply contracts can be used to do so (see Cachon 2003 for an overview). Buyers in a channel are typically modeled in one of two ways: either as a monopolist facing a downward-sloping market demand (see Tsay et al. 1998) or as a newsvendor facing an exogenous retail price and random market demand (see Lariviere 1998). We focus on the newsvendor buyers.

A common assumption in the analytical literature is that decision makers choose order quantities that maximize their expected profit.[1] However, a number of authors have shown that this assumption does not necessarily hold with actual decision makers. In laboratory experiments, human decision makers place orders that are inconsistent with an expected profit maximizing behavior (see, e.g., Schweitzer and Cachon 2000, Bostian et al. 2008, Bolton and Katok 2008). The experimental results are robust. In all experiments that we are aware of, demand chasing and anchoring effects have been observed. The demand chasing effect refers to the observation that people adjust order quantities based on previous demand realization. The anchoring effect refers to the observation that people anchor on the mean demand and place orders that are between the mean demand and the expected profit maximizing quantity. Most experiments use a setting that corresponds to a wholesale price contract, but similar results have also been observed with buyback and revenue sharing contracts (Katok and Wu 2009).

There exists a large stream of literature on supply contracting that uses the newsvendor model as a building block and a growing stream of literature from behavioral operations that shows that human subjects do not order according to this model. In this paper, we bridge the gap between these two streams of literature. We develop a behavioral order model that predicts orders of human decision makers more accurately than the newsvendor model. We then use the behavioral model to replace the newsvendor model in a supply contracting model and show that contracts that are based on the behavioral order

[1] Or expected utility. See Eeckhoudt et al. (1995), Keren and Pliskin (2006), and Wang and Webster (2009).

model perform better than contracts that are based on the newsvendor model.

This paper makes three contributions. First, we are the first to document that buyback contracts with the same critical ratio (i.e., the same expected profit maximizing order quantity) but different contract parameters result in different order quantities. The implication is that it does not suffice to analyze the critical ratio to predict order quantities (as suggested by the supply contracting and behavioral operations literature), but that the individual contract parameter values must be taken into account. Second, we provide a behavioral model that explains actual orders more accurately than the standard model. We base our behavioral model on prospect theory and mental accounting (Kahneman and Tversky 1979, Thaler 1985) and conduct laboratory experiments to motivate the model. The insights gained from the analyses help us to understand the decision-making process of human decision makers, and we build on these insights to develop a behavioral model to predict ordering behavior. Third, we show how the behavioral model can be used to design supply contracts that incentivize subjects to place first-best orders. We test our model using out-of-sample validation experiments, in which we successfully incentivize decision makers to place orders that are closer to first-best than under the newsvendor model.

2. Analytical Background

One of the standard settings that the supply contracting literature analyzes is a channel with a single seller and a single buyer (see Cachon 2003 for an overview). The seller has linear production costs of c per unit and no capacity constraint. The buyer faces a random demand with cumulative distribution function $F(\cdot)$ and exogenous retail price r. The seller is offering a supply contract to the buyer, who decides on the order quantity S and places the order with the seller. The seller produces the order and delivers it to the buyer, who sells the product to customers. The order quantity depends on the type and parameters of the supply contract. One commonly analyzed supply contract that can, in principle, be used to maximize the expected channel profit is the buyback contract (Pasternack 1985), and we focus on this contract in this paper.

Under a buyback contract, the seller offers the product to the buyer at a unit wholesale price w that is at least as high as the unit cost c. The buyer can return unsold units to the seller at a unit buyback price b. The seller can salvage the excess inventory at unit salvage value v (which, without loss of generality, we assume to be zero in the remainder of this paper). The buyer faces the classic newsvendor problem (Arrow et al. 1951), and the order quantity that maximizes the expected profit of the buyer is

$$S^*_{NV} = F^{-1}\left(\frac{r-w}{r-b}\right). \tag{1}$$

The term in parentheses is referred to as *critical ratio*. The subscript NV indicates that this is the order quantity under a newsvendor model. The order quantity that maximizes the expected channel profit is

$$S^*_{FB} = F^{-1}\left(\frac{r-c}{r}\right) \tag{2}$$

and is referred to as *first-best order*.

If the optimal order quantity of the buyer is equal to the first-best order, then the channel is said to be *coordinated* because the expected channel profit is maximized. This is the case if the critical ratios of Equations (1) and (2) are the same. Setting the critical ratios equal and solving for b yields

$$b_{FB} = -\frac{c}{r-c}r + \frac{r}{r-c}w, \quad c \le w \le r. \tag{3}$$

The supply contracting literature assumes that buyers order S^*_{NV} according to Equation (1). Then the buyback price b_{FB} maximizes the expected channel profit. However, if buyers deviate from Equation (1), then the buyback price that induces the buyer to place first-best orders will generally be different. From an expected channel profit perspective, a buyback price $b \neq b_{FB}$ might then deliver higher expected channel profits than would $b = b_{FB}$.

3. Development of Behavioral Model

Various authors have shown that human subjects do not order according to the newsvendor model when they are offered supply contracts (e.g., Bostian et al. 2008, Katok and Wu 2009, Bolton et al. 2012, and others). Most experiments have focused on the newsvendor model with a wholesale price contract, which can be thought of as a buyback contract with a buyback price of zero. In laboratory experiments, the average order quantities of decision makers tend to lie between the newsvendor quantity and the mean demand. The literature has identified two main explanations (see, e.g., Bolton and Katok 2008, Bostian et al. 2008, Schweitzer and Cachon 2000). The first explanation is the chasing effect, i.e., the adjustment of current order quantities toward previous demand realizations. The second explanation is the anchoring effect, i.e., the anchoring of orders on the mean demand.

Previous research has observed demand chasing and anchoring effects for a given set of contract parameters but has not analyzed how the values of the contract parameters affect ordering decisions.

However, understanding this relationship is important because a contract designer must determine individual contract parameter values, which can be done effectively only if the relationship between contract parameter values and orders is well understood. We address this issue here.

We start our analyses with a simple experiment in which we offer decision makers two contracts with different parameter values but identical critical ratios (§3.1). It turns out that these contracts result in significantly different order quantities, although their newsvendor quantities are identical. This effect has not been previously documented because previous laboratory experiments investigating newsvendor behavior have focused on a small number of contract parameter value combinations with different critical ratios. We show that the effect can be explained by mental accounting (§3.2) and use the insight gained from the analyses to develop a behavioral order model (§3.3).

3.1. Effect of Contract Parameters on Orders

Consider a buyer who is offered a supply contract with wholesale price w and buyback price b. The supply contracting literature suggests that the decision maker computes the critical ratio $CR = (r - w)/(r - b)$ and determines the newsvendor order quantity according to Equation (1). We take as baseline (null) hypothesis the optimal policy of the newsvendor model:

Hypothesis 1. *The order quantity of buyers is* S^*_{NV}.

The behavioral literature provides evidence that actual mean orders differ from the newsvendor quantity, and we do not expect Hypothesis 1 to be fully borne out. Specifically, the behavioral literature suggests that actual mean orders are between the newsvendor quantity and the mean demand. However, for settings where the newsvendor quantity is equal to the mean demand, the literature suggests that mean orders are equal to the newsvendor quantity.

Another hypothesis concerns the individual contract parameters for a given critical ratio. Various contract parameter combinations (w, b) result in the same critical ratio $CR = (r - w)/(r - b)$ and therefore in the same newsvendor order quantity. In other words, the orders of expected profit maximizing buyers are not affected by the individual contract parameters (w, b) if they result in the same critical ratio. This is formally stated in Hypothesis 2.

Hypothesis 2. *The order quantity of buyers depends on the critical ratio only, not on the individual contract parameters.*

We test the hypotheses by offering subjects two contracts with different contract parameter values but the same critical ratio: Contract I has parameter values $w = 50$ and $b = 0$ and Contract II has parameter values $w = 80$ and $b = 60$. Unit revenue is $r = 100$ and demand is uniformly distributed between 0 and 100. The critical ratio is $CR = 0.50$ under both contracts, resulting in newsvendor orders of $S^*_{NV} = F^{-1}(0.50) = \mu = 50$ under both contracts. Because newsvendor orders are equal to the mean demand, potential deviations of actual orders from the newsvendor quantity cannot be explained as the result of anchoring on the mean demand.

We conduct an in-class experiment with $N = 53$ undergraduate students. Students receive a one-page briefing on the problem and are asked to select order quantities under both contracts (Experiment 1). We do not provide information on demand realizations to avoid demand chasing effects.

Experiment 1. Uniform demand $(0, 100)$, $CR = 0.50$, $S^*_{NV} = 50$, $N = 53$.
Contract I: $w = 50$, $b = 0$.
Contract II: $w = 80$, $b = 60$.

Under Contract I, average actual orders are $\bar{S} = 41.2$ ($\sigma = 23.0$) and are significantly below newsvendor orders of 50 (two-tailed Wilcoxon signed-rank test, $p = 0.009$). Under Contract II, average actual orders are $\bar{S} = 59.4$ ($\sigma = 26.8$) and are significantly above newsvendor orders of 50 (two-tailed Wilcoxon signed-rank test, $p = 0.019$). Therefore, we reject Hypothesis 1. The difference in average actual orders is $\Delta\bar{S} = 59.4 - 41.2 = 18.2$, which is significantly different from zero (two-tailed Wilcoxon signed-rank test, $p < 0.001$). Therefore, we also reject Hypothesis 2.

The results of the experiment provide interesting insights: First, they show that under a buyback contract subjects do not necessarily order between the newsvendor quantity and the mean demand. Second, they show that individual contract parameter values, and not only the critical ratio, affect orders.

Because neither demand chasing nor anchoring can explain these two regularities, some additional effects must be considered to explain actual behavior. We argue that people perform certain cognitive operations when making a decision.

3.2. Mental Accounting

A concept that encapsulates these cognitive processes is mental accounting (Kahneman and Tversky 1979, Thaler 1985), which has been used to explain many effects that are inconsistent with expected profit maximizing behavior. Mental accounting requires the specification of the way elementary outcomes are evaluated, how they are combined, and how the combined outcomes are valued (examples of mental accounting are provided next).

3.2.1. Evaluation of Outcomes. If, consistent with the standard contracting models and standard economic theory, decision makers maximize their expected profit, then people's valuation of income should not be affected by the source of the income but only by the total income. Under a buyback contract, we have two sources of income, income from sales to customers and income from returns to the supplier. If the standard assumption of the contracting literature holds, then people should be indifferent to the source of the income, which is stated in Hypothesis 3.

Hypothesis 3. *People are indifferent to the source of income. They give equal value to income from sales to customers and income from returns to the supplier.*

To test the hypothesis, we use a procedure similar to the one employed in Thaler (1985) and ask subjects about the happiness of Mr. A and Mr. B, where happiness is a proxy for Mr. A's and Mr. B's valuation of an outcome. In Experiment 2, we ask 46 students at the University of Cologne to compare the happiness of Mr. A and Mr. B, where Mr. A receives income from sales only and Mr. B receives income from sales and from returns. The numbers in parentheses show the number of responses.

Experiment 2. Mr. A bought 200 newspapers at 1 euro each, sold 100 of them for 4 euros each, and returned the 100 unsold newspapers to the publisher, receiving no additional compensation and netting 200 euros profit. Mr. B bought 200 newspapers at 1 euro each, sold 100 of them for 3 euros each, and returned the remaining 100 unsold newspapers to the publisher, receiving 1 euro for each and netting 200 euros profit.

Who is happier? Mr. A (6), Mr. B (33), no difference (7).

The results of the experiment show that the source of the income matters to many people, and we reject the standard economic hypothesis that people are generally indifferent to the source of the income (Hypothesis 3).

Our results are in line with the mental accounting arguments that state that different sources of income can be associated with different values (Thaler 1999). For instance, O'Curry (1997; cited in Thaler 1999) shows that people value and use the winnings of an office football pool differently than they do income tax refunds, and Kooreman (2000) shows that changes in child allowance income affect spending differently than do changes in other income categories.

We follow Thaler (1999) and argue that the source of income matters. We model the different values associated with income from sales and returns by multiplying the income from returns by a parameter γ, where $\gamma > 1$ corresponds to a higher valuation of income from returns than from sales and vice versa.[2]

3.2.2. Combination of Elementary Outcomes. There are various ways a decision maker can frame a problem (Tversky and Kahneman 1981). A natural framing of our problem is the standard textbook framing (e.g., Nahmias 2008) of the newsvendor model, where the decision maker computes underage cost (in our setting, the unit profit margin) and overage cost (in our setting, the unit cost of unsold products), and we use this framing in our model, resulting in two separate accounts, one for sales, and one for leftovers.[3]

3.2.3. Value Function. Mental accounting uses the value function of prospect theory (Kahneman and Tversky 1979) to evaluate outcomes (Thaler 1999). Because the elements of the expected profit function of our model are relatively flat and we are looking for a parsimonious model to explain the behavior, we use a linear value function. This makes our model analytically tractable and we have to estimate only a single parameter of the value function (loss parameter β) as opposed to more parameters for the nonlinear value function.

The value that is assigned to the monetary stream x is $v(x) = -\beta[x]^- + [x]^+$, where $[x]^-$ is equal to the absolute value of x if x is negative and 0 otherwise, and $[x]^+$ is equal to x if x is positive and 0 otherwise. If people apply the value function to the total monetary outcome, then the relevant monetary stream x is the total monetary outcome. However, if people use multiple mental accounts, then the value function is applied to each account (Thaler 1999).

[2] Separate concave value functions can also explain the results of Experiment 2 (see Thaler 1985). We analyze this alternative explanation by conducting an additional experiment. We asked 66 students of the University of Cologne about the happiness of Mr. A (Mr. B), where Mr. A (Mr. B) receives 150 (50) euros from sales and 50 (150) euros from returns. Who is happier? Mr. A (38), Mr. B (16), no difference (12). If people apply the same concave value function to both income streams, then Mr. A and Mr. B are equally happy. Because only 12 out of 66 subjects evaluate the happiness of Mr. A and Mr. B equally, people seem not to apply the same value function to both income streams. We note that the results do not suggest that the value function is not concave but suggest that the value functions are different for sales and returns.

[3] We use our framing as opposed to other possible framings that apply, for instance, the value function to the final outcome, because we find indications that our framing is used in the debriefing of Experiment 1. We ask the subjects of Experiment 1 about the rationales behind their decisions. Out of the 43 subjects who provide answers, 28 (65%) subjects state that they base their decisions on the profit margin, 24 (56%) state that they base their decision on the unit cost of unsold products, and 19 (44%) state both. These factors are the most often mentioned factors, followed by the buyback price (11 subjects, 26%), the wholesale price (3 subjects, 7%), and the unit revenue (1 subject, 2%).

3.3. Behavioral Model

Under our framing, people consider the upside and downside potentials of their order decisions separately. The upside potential is the income associated with selling products. For a demand realization d, it is $(r - w)\min(S, d)$, where $\min(S, d)$ is the number of units sold. The downside potential is the loss associated with leftovers—the cost of overage. It can be computed as $-(w - \gamma b)(S - \min(S, d))$, where $(S - \min(S, d))$ is the number of units returned.

We apply the value function to each monetary stream to obtain the value associated with decision S and demand realization d,

$$v(S, d) = v((r - w)\min(S, d)) + v(-(w - \gamma b)(S - \min(S, d))). \quad (4)$$

Because revenues from sales are nonnegative and losses from leftovers are nonpositive, we obtain

$$v(S, d) = (r - w)\min(S, d) - \beta(w - \gamma b)(S - \min(S, d)). \quad (5)$$

In the experiments above, we have focused on situations with a critical ratio of 0.50 and symmetric demand. In such situations, an anchoring effect is not relevant because the optimal order quantity is equal to the mean demand. However, there is ample evidence that people generally anchor on the mean demand. Therefore, we include an anchoring effect in our model. We use the approach of Benzion et al. (2008) and specify an anchoring parameter α that assigns a weight of α to the mean demand and a weight of $(1 - \alpha)$ to the solution without anchoring.[4] Restricting the parameters to $0 \le \alpha \le 1$, $\beta > 0$, and $r > w > \gamma b > 0$ to avoid unreasonable results, it is straightforward to obtain the optimal order quantity of the behavioral model:

$$\hat{S} = (1 - \alpha)F^{-1}\left(\frac{r - w}{r - w + \beta(w - \gamma b)}\right) + \alpha\mu. \quad (6)$$

Next, we conduct an experiment with a variety of critical ratios and contract parameter combinations and then validate the model.

[4] An alternative to introducing the anchoring parameter α directly would be to include a rationale for the ordering bias in the model. For instance, one could follow the Ho et al. (2010) approach and add a cost parameter for the units over-ordered and under-ordered. If the ratio of those parameters is smaller than the true overage and underage costs in high profit conditions (higher in low profit conditions), then orders become biased toward mean demand. The two modeling approaches are similar and result in a similar fit. In our setting, an advantage of the direct approach is that we do not have to hypothesize a rationale for the order bias and that the behavioral parameters α and β and their magnitude can be interpreted as the degree of mean demand anchoring and the degree of loss aversion, interpretations that are frequently used in the behavioral literature.

4. Design of the Main Experiment

We use a laboratory experiment to analyze how the wholesale price and buyback price affect order quantities. We spend the first 15 minutes of the experiment briefing the subjects. The briefing consists of four sections (the online appendix, available at http://www.scmms.uni-koeln.de/25675.html, contains the charts used during the briefing and screenshots of the software):

1. *Problem description* (*six minutes*): We start the briefing by explaining the newsvendor problem. In synopsis, the purchase price to be paid for buying a unit of the product is w talers per unit, it can be sold to customers for $r = 100$ talers per unit, and unsold units can be returned at b talers per unit. Demand is uniformly distributed between 1 and 100 units and independent between periods. To illustrate how profit is computed, we provide an example with $w = 60$, $b = 30$, an order quantity of 10, and a demand of 80, resulting in a profit of 400. We provide a second example with the same wholesale price and buyback price, but an order quantity of 70 and a demand of 20, resulting in a profit of −700.

2. *Exercises* (*three minutes*): To analyze whether the subjects understand the basic relationships of the problem, we ask them six test questions: We ask them to compute the number of units sold, the number of units left over, and the profits for two examples. In Example 1, we use $w = 70$, $b = 10$, order quantity = 70, and demand = 80; in Example 2, we use $w = 70$, $b = 10$, order quantity = 70, and demand = 20. After three minutes, we collect the answers and evaluate them after the experiment: 68% of the subjects answer all questions correctly. Thirty-two percent of the subjects make at least one mistake. Those who make mistakes answer on average 1.6 of the six questions incorrectly.

3. *Presentation of solution* (*three minutes*): After collecting the exercises, we present the correct solutions. The objective of this part of the briefing is to improve the understanding of the underlying problem even further.

4. *Explanation of game* (*three minutes*): During the last three minutes of the briefing, we provide a road map of the game, including information on the earnings that subjects receive after the game and screenshots of the software.

The actual game consists of two phases. In the *warm-up phase*, we present five contracts with different parameters and ask the subjects to place orders. The contracts are offered sequentially; that is, the following contract is displayed only after all subjects have chosen an order quantity. After the order quantities for the first five contracts have been chosen, we provide a summary screen that shows the parameters of the five contracts, the order quantities that are chosen

for each contract, and a random draw of the demand that is identical for all subjects (i.i.d. within subject).

In the *data collection phase*, we collect the data. We present 28 contracts, using the same approach as in the warm-up phase. The sequence in which we show the contracts is randomized. After all subjects have placed 28 orders, we provide a summary screen that shows the parameter values of the 28 contracts, the order quantities, a random draw of the demand (identical across subjects, i.i.d. within subject), and the final earnings.

Note that unlike most previous research, ours does not provide feedback after each decision because the focus of our study is not learning but the subjects' reactions to contract parameters.

We program the experimental software using the *z*-Tree system (Fischbacher 2007) and conduct the experiment at the University of Cologne. Thirty-one student subjects participate, and each subject plays exactly one session. Cash is the only incentive offered. Participants are recruited from the subject pool of the Cologne Laboratory for Economic Research (CLER) with the help of the recruitment software ORSEE (Greiner 2004). At the end of the session, subjects are paid their average individual earnings from the game at a rate of 1 euro per 100 talers. The average earning is 13.54 euros, including a 2.50 euro participation fee for each subject. The session lasts approximately one hour.

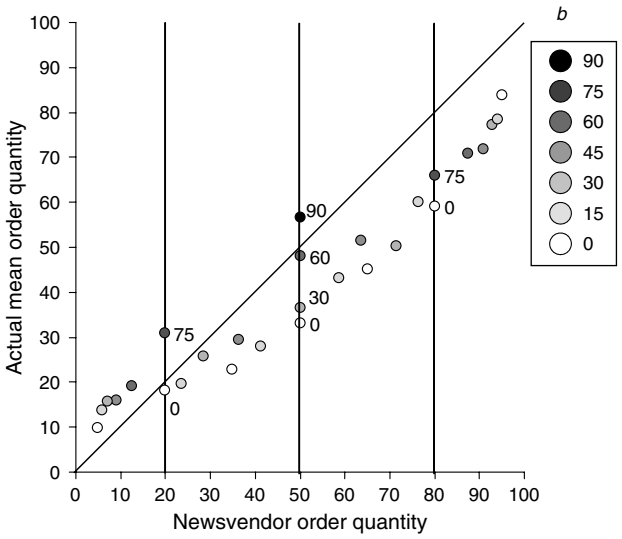
Table 1 shows the w and b combinations we use in the experiment and provides a comparison between newsvendor orders (left panel) and average orders and their standard deviations (right panel). We use values that span an equidistant grid over the feasible region of the contract parameter values.

Table 1 Newsvendor Orders and Actual Mean Orders of the Laboratory Experiment

	Newsvendor orders							Actual mean orders						
	w							*w*						
b	5	20	35	50	65	80	95	5	20	35	50	65	80	95
0	95	80	65	50	35	20	5	84 (18)	59 (22)	45 (16)	33 (15)	23 (11)	18 (11)	10 (9)
15	—	94	76	59	41	24	6	—	78 (18)	60 (19)	43 (17)	28 (11)	20 (11)	14 (11)
30	—	—	93	71	50	29	7	—	—	77 (17)	50 (16)	37 (14)	26 (13)	16 (12)
45	—	—	—	91	64	36	9	—	—	—	72 (21)	51 (18)	29 (14)	16 (12)
60	—	—	—	—	88	50	13	—	—	—	—	71 (19)	48 (22)	19 (13)
75	—	—	—	—	—	80	20	—	—	—	—	—	66 (23)	31 (18)
90	—	—	—	—	—	—	50	—	—	—	—	—	—	57 (23)

Note. Standard deviations in parentheses.

Figure 1 Actual Mean Orders (Averaged Over All Subjects) vs. Newsvendor Orders for Each of the 28 Contracts Used in the Laboratory Experiment



Before we provide statistical analyses in the next section, we take an aggregate view of the data. Figure 1 provides a graphical representation of the results. The 45-degree line corresponds to the actual mean orders being equal to the newsvendor orders. The graph shows that the subjects exhibit the behavior predicted by the behavioral model: First, we observe that subjects tend to order on average more than newsvendor quantities for small critical ratios and less than newsvendor quantities for large critical ratios. This behavior is consistent with an anchoring parameter of $\alpha > 0$. Second, there exists a tendency to order on average less than newsvendor quantities, which is consistent with a loss parameter of $\beta > 1$. Third, average order quantities are different for given critical ratios and increase in the buyback price, which is consistent with a value parameter $\gamma > 1$.

5. Analysis of Behavioral Models

We use the data of the laboratory experiment to estimate the parameters α, β, and γ of our behavioral model. In §5.1, we analyze an aggregate behavioral model that uses one set of behavioral parameters for all subjects. In §5.2, we analyze an individual behavioral model that uses a separate set of behavioral parameters for each subject.

5.1. Aggregate Behavioral Model

We estimate the behavioral parameters α^A, β^A, and γ^A of the aggregate behavioral model

$$S_{nj} = (1-\alpha^A)F^{-1}\left(\frac{r-w_j}{r-w_j+\beta^A(w_j-\gamma^A b_j)}\right) + \alpha^A\mu + u_n + \epsilon_{nj}, \quad (7)$$

where subscript n represents individual participants, subscript j represents contracts, and $u_n \sim N(0, \theta^2)$ and $\epsilon_{nj} \sim N(0, \sigma^2)$ are error terms. The superscript A indicates that this is an aggregate model, i.e., a model where we use a single set of behavioral parameters for all subjects.

We conduct a maximum likelihood estimation of the nonlinear random effects model

$$L(S \mid \alpha, \beta, \gamma, \theta, \sigma) = \prod_{n=1}^{N} \prod_{j=1}^{J} f(S_{nj}; \alpha, \beta, \gamma, \theta, \sigma), \quad (8)$$

where $f(\cdot)$ denotes the probability density function for the order quantity S_{nj}, given the parameters $\alpha, \beta, \gamma, \theta$, and σ according to Equation (7).

The resulting estimates are $\alpha^A = 0.279$, $\beta^A = 1.988$, and $\gamma^A = 1.036$. All three parameters are significantly different from the parameters of the newsvendor model; i.e., α^A is significantly different from 0 and β^A and γ^A are significantly different from 1 ($p < 0.001$ for all parameters).[5] For the maximum likelihood estimation, we assume that the residuals are normally distributed, which is the case with our data (Kolmogorov–Smirnov test, for σ: $p = 0.6502$, θ: $p = 0.5235$; see the online appendix for details).

Next, we analyze the importance of using a full model with three behavioral parameters α^A, β^A, and γ^A as opposed to a reduced model with fewer parameters by comparing the performances of the full and reduced models. Table 2 shows the results and reports the log-likelihoods and *AIC*s (Akaike information criterion) of the models (Akaike 1981). The log-likelihood of the full model (model (7)) is significantly higher than the log-likelihoods of all reduced models (models (1)–(6), χ^2-test, $p < 0.001$). The *AIC*, which controls for the number of estimated parameters, is lower for the full model than for all reduced models, which indicates that the full model provides a better fit than the reduced models and that all three parameters α^A, β^A, and γ^A are statistically justified.[6]

Figure 2 shows that the orders estimated by the aggregate model are closer to average actual orders than the estimate of the newsvendor model (compare Figures 1 and 2). However, individual orders can differ considerably from average orders (see Figures A.1–A.3 in Appendix A, which show the actual orders and the orders predicted by the aggregate model for all subjects). We next show how the heterogeneity of the behavioral parameters can be addressed by estimating separate behavioral parameters for each individual.

[5] If we eliminate the 11 subjects who do not answer all six test questions correctly, we get similar results: $\alpha^A = 0.271$, $\beta^A = 1.720$, and $\gamma^A = 1.024$. Again all three parameters are significantly different from the parameters of the newsvendor model at $p < 0.001$.

[6] Note that γ is also conceptually important to explain the observed order quantities because for $\gamma = 1$ the order quantities are the same for different contracts with the same critical ratio.

Table 2 **Likelihoods and AICs for Different Aggregate Behavioral Models**

	Model						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
α^A	0.282 (0.02)			0.257 (0.02)	0.298 (0.02)		0.279 (0.02)
β^A		3.698 (0.24)		1.802 (0.14)		4.186 (0.30)	1.988 (0.16)
γ^A			1.014 (0.005)		1.034 (0.004)	1.021 (0.004)	1.036 (0.003)
θ^A	11.70	15.08	12.07	8.75	12.63	15.44	8.89
σ^A	14.80	15.82	17.01	14.52	14.45	15.71	14.06
Loglike^A	−3,679	−3,679	−3,733	−3,591	−3,598	−3,674	−3,565
AIC^A	7,364	7,364	7,473	7,191	7,204	7,356	7,140

Note. Standard errors in parentheses.

5.2. Individual Behavioral Model

To take the heterogeneity of individuals into account, we estimate the behavioral parameters α_n, β_n, and γ_n for each subject n individually. We define the individual behavioral model

$$S_{nj} = (1 - \alpha_n) F^{-1}\left(\frac{r - w_j}{r - w_j + \beta_n(w_j - \gamma_n b_j)}\right) + \alpha_n \mu + u_n + \epsilon_{nj}. \quad (9)$$

We refer to this model as the individual behavioral model because it uses one set of parameters for each individual. The model also uses subject specific variances $\epsilon_{nj} \sim N(0, \sigma_n^2)$.

Figure 2 **Actual and Predicted Mean Orders of the Aggregate Behavioral Model**

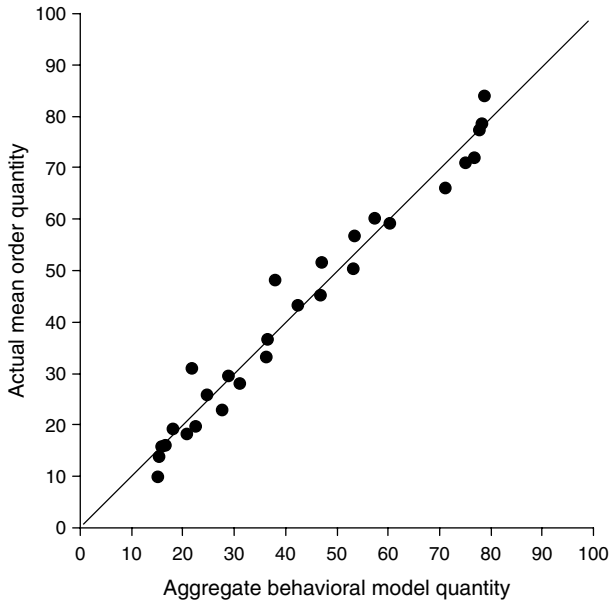


Table 3 Estimated Individual Preferences

	Parameter					Parameter			
Subject	α	β	γ	σ	Subject	α	β	γ	σ
1	0.00	1.57***	1.019*	1.000	17	0.15***	35.09***	1.054***	0.826
2	0.14***	4.66***	1.050***	0.762	18	0.00	2.83***	1.035***	1.366
3	0.07	3.65***	0.848*	1.191	19	0.79***	2.91	1.065***	1.313
4	0.40***	2.48*	0.931	1.630	20	0.22*	2.89*	1.053***	2.616
5	0.03	1.39	0.984	2.203	21	0.02	1.20**	1.011	0.818
6	0.16*	2.14***	1.050***	1.676	22	0.11**	2.41***	0.993	0.973
7	0.19***	1.30**	1.003	0.902	23	0.00	1.00	1.000	0.031
8	0.67***	3.33	0.932	1.714	24	0.58***	1.57**	0.983	0.623
9	0.29***	4.74***	1.057***	1.477	25	0.21**	2.37**	1.053***	1.699
10	0.04	0.93	1.052***	1.529	26	0.29***	13.58***	0.702***	0.440
11	0.11	0.98	1.002	2.229	27	0.26***	1.41**	0.921*	0.865
12	0.47***	0.83	1.074***	1.653	28	0.18***	1.68***	0.985	0.676
13	0.30***	2.03***	0.999	1.098	29	0.00	1.90***	1.037***	1.339
14	0.34***	0.60***	1.067***	1.165	30	0.00	1.02**	1.001	0.054
15	0.17*	2.70***	1.049***	1.768	31	0.35**	4.75**	1.054***	1.076
16	0.54***	9.37**	1.054***	1.541					

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

We note that the residuals are only approximately normally distributed (see Figures B.1 and B.2 in Appendix B). The use of maximum likelihood estimation with normally distributed errors is justified because the distribution of the residuals is fairly symmetric and so the estimates are consistent. The resulting individual parameters α_n, β_n, and γ_n and the standard deviations of the error terms σ_n of each individual are shown in Table 3.

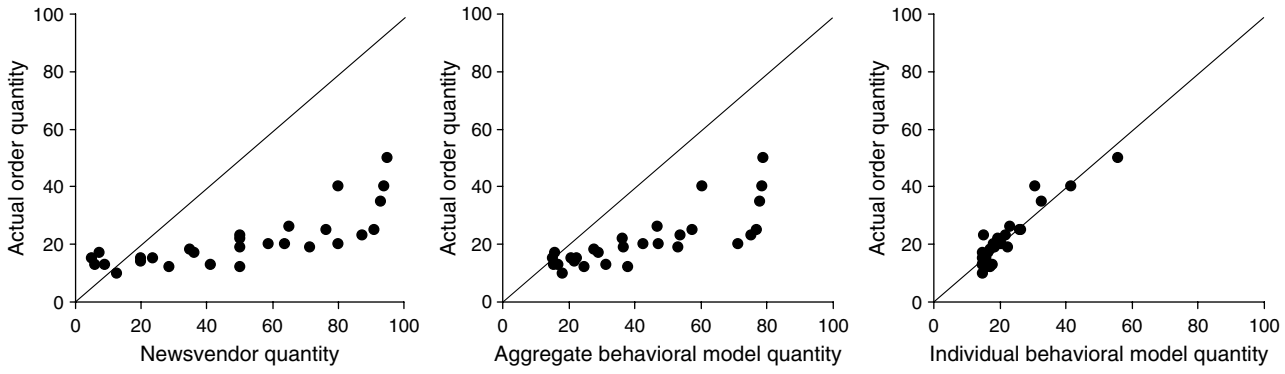
The aggregated model is nested in the individual model, and we can compare the fits of both models. The log-likelihood of the individual behavioral model is $L^I_{\log} = -2{,}999$, which is significantly higher than the log-likelihood of the aggregate behavioral model of $L^A_{\log} = -3{,}565$ (χ^2 test, $p < 0.001$). The *AIC* of the individual behavioral model of $AIC^I = 6{,}248$ is lower than the *AIC* of the aggregate behavioral model ($AIC^A = 7{,}140$). We conclude that the individual behavioral model provides a better fit than the aggregate behavioral model.

Finally, we test if subject-specific variances of the error terms are necessary. We repeat the analysis above for a model with a common variance of the error term for all subjects. The *AIC* of this model is 6,795, which is greater than the *AIC* of the model with subject-specific variances ($AIC^I = 6{,}248$). The log-likelihood is −3,302 and significantly smaller than the log-likelihood of the model with subject-specific variances ($L^I_{\log} = -2{,}999$, likelihood ratio test: $p < 0.001$). The model with subject-specific variances has a better fit than the model with a common variance, and we will rely on it in our subsequent analyses.

We illustrate how the different models work using subject 26 (Figure 3). The figure shows the predicted and actual orders under three models: the newsvendor model, the aggregate behavioral model, and the individual behavioral model. The closer the dots are to the 45-degree line, the better the fit of the model.

The left graph of the figure shows the fit of newsvendor model. We see that actual orders of subject 26 are greater than newsvendor orders for small critical ratios and smaller than newsvendor orders for large critical ratios, which is modeled in the individual behavioral model by an anchoring parameter

Figure 3 Predicted vs. Actual Orders of Subject 26 for the Newsvendor Model, the Aggregate Behavioral Model, and the Individual Behavioral Model



$\alpha_{26} = 0.287$. In general, actual orders tend to be below newsvendor orders, which is consistent with a loss parameter of $\beta > 1$ and the individual behavioral model uses $\beta_{26} = 13.583$. Finally, order quantities are different for given critical ratios (e.g., $CR = 0.50$), which is modeled in the individual behavioral model by $\gamma_{26} = 0.70$.

The newsvendor model assumes $\alpha = 0, \beta = 1$, and $\gamma = 1$, values that are very different from the values of the individual model. As the left graph of Figure 3 shows, the newsvendor model performs poorly.

The center graph of Figure 3 shows the fit of the aggregate behavioral model. It indicates that the aggregate behavioral model performs better than the newsvendor model. The aggregate behavioral model uses $\alpha^A = 0.279$, $\beta^A = 1.988$, and $\gamma^A = 1.036$. The values for α and β of the aggregate behavioral model are an improvement verses the values of the newsvendor model, but the value of γ is even worse than in the newsvendor model: subject 26 values income from sales higher than income from returns, which is rare and the opposite of what is used in the aggregate model.

With an aggregate behavioral model, such individual preferences are not appropriately modeled, whereas they can be appropriately modeled with an individual behavioral model. The right graph of Figure 3 shows how the individual behavioral model performs and illustrates that is provides a much better fit than the aggregate behavioral model and the newsvendor model.

We note that subject 26 is an extreme case. The parameters of this subject differ more from the parameters of the newsvendor model and the aggregate behavioral model than the parameters of most other subjects do. We use subject 26 to illustrate how individual behavioral models can handle such extreme cases. In Appendix A, we provide data on how the three models perform for each of the 31 subjects.

6. Validation Experiments

Above, we showed that a behavioral model can be fit, such that it models actual ordering behavior reasonably well. However, up to this point, our analyses were in-sample, meaning that we fit a model and measure its fit on the same set of contracts and subjects. We next proceed to analyze the accuracy of out-of-sample predictions of the behavioral models and the benefit of using the behavioral models instead of the newsvendor model.

As a way to motivate the new set of experiments, we consider a channel with a single seller and multiple buyers. A contract designer is interested in designing contracts that maximize the expected channel profit (achieving a first-best solution). To this end, the contract designer analyzes historical order quantities and estimates the behavioral parameters α, β, and γ of the buyers. Then she uses the behavioral model to determine contract parameters that result in first-best orders.

In our validation experiments, the objective is to incentivize first-best orders such that the expected channel profit is maximized. The reason that the first-best solution is an important one to study is that it offers more potential benefits to both parties than any other solution that is not first-best. Our approach, in general, can be used to induce any desired solution (such as a manufacturer-optimal or retailer-optimal solution); but for demonstration purposes, and given the conceptual importance of the first-best solution, we use our approach here to induce the retailer to place first-best orders. The objective of achieving first-best is appropriate, for instance, in situations where the seller and buyer belong to the same organization. For other settings and objectives, such as decentralized settings in which sellers design contracts to maximize their own expected profit subject to a reservation profit constraint of the buyer, the approach would be similar, but different target quantities would be incentivized.

In §6.1, we consider a setting in which the contract designer uses customized contracts for each buyer. She estimates the behavioral parameters for each buyer and uses individual behavioral models to design contracts that are tailored to each buyer. If legal or other restrictions prohibit the use of different contracts for different buyers, the contract designer must offer the same contract to all buyers. In §6.2, we consider this setting and use the aggregate behavioral model to compute contract parameters that incentivize, averaged over all buyers, first-best order quantities.

The approaches of §§6.1 and 6.2 assume that the behavioral parameters α, β, and γ are given (or have been estimated). Given the behavioral parameters, i.e., given the biases of the buyers, the contract designer chooses a contract that incentivizes the buyers to order first-best.

Instead of taking the bias of the buyers into account when designing a contract, a contract designer could attempt to de-bias the buyers by training them in the newsvendor model. If subjects do not order according to the newsvendor model because they cannot translate the problem parameters into expected profit maximizing orders and their objective to maximize expected profits, then training would shift the orders of the buyers closer to the newsvendor solution. Under contracts designed with the newsvendor model, trained buyers would place orders that are closer to first-best than would untrained buyers. In §6.3 we analyze the effectiveness of training.

The monetary incentives in the validation experiments are the same as in the main experiment

(1 euro per 100 taler average profit plus a 2.50 euro participation fee).

6.1. Individual Behavioral Contracts

In our validation experiment, we use a two-phase approach. In Phase 1, we offer subjects several contracts and collect data on their order behavior. We then use the order data to estimate an individual behavioral model. In Phase 2, we use the individual behavioral model to determine contract parameters that incentivize first-best orders. We refer to these contracts as *behavioral contracts* (as opposed to contracts that are designed using the newsvendor model, which we refer to as *newsvendor contracts*). We offer the behavioral contracts to the subjects and analyze their performance.

6.1.1. Phase 1: Estimation of Behavioral Preferences. The objective of Phase 1 is to estimate the parameters of the individual behavioral model. We use the same general experimental setup as in the previous experiments with 30 new subjects. Nineteen of the 30 subjects answer all six test questions correctly and exhibit a good understanding of the underlying problem. In the validation, we use only these 19 subjects because these subjects are more representative of real buyers than are subjects who have trouble understanding the problem.[7]

In Phase 1, we offer all subjects 19 contracts. Table 4 shows the contracts we use, the newsvendor order quantities, the actual mean orders placed by the subjects, and, in parentheses, the standard deviations of the orders. Figure 4 presents the results visually.

We use the orders of the $N = 19$ subjects under the $J = 19$ contracts and estimate an individual behavioral model (Equation (9)). The parameter values are shown in Table 5. So at the end of Phase 1, we have estimates of the behavioral parameters of each individual, which allow us to predict how order quantities are affected by the contract parameters.

6.1.2. Phase 2: Incentivizing Subjects to Order First-Best. The objective of Phase 2 is to incentivize first-best order quantities. To cover a reasonable range of critical ratios, we choose critical ratios $CR = 0.20$, 0.29, 0.41, 0.50, 0.63, 0.71, 0.76, and 0.80. Note that the critical ratios are a subset of those used in Phase 1. We use the same critical ratios in the Phase 2 as in Phase 1 because this allows us to compare the performances of the newsvendor contracts in Phase 1 with that of the behavioral contracts in Phase 2.[8]

[7] We note that including the 11 subjects who answered at least one test question incorrectly does not alter the main results and all conclusions remain the same as reported below.

[8] We assume production costs of $c = 100(1 - CR)$ such that the newsvendor order quantities in Phase 1 correspond to first-best order quantities we want to incentivize.

Table 4 **Newsvendor Orders and Actual Mean Orders in Phase 1 of the Validation Experiment**

	Newsvendor orders					Actual mean orders				
	w					*w*				
b	20	35	50	65	80	20	35	50	65	80
0	80	65	50	35	20	49 (24)	36 (21)	28 (20)	19 (12)	17 (15)
15	94	76	59	41	24	78 (18)	54 (21)	37 (16)	26 (12)	18 (11)
30	—	93	71	50	29	—	71 (20)	47 (19)	30 (13)	24 (10)
45	—	—	91	64	36	—	—	74 (23)	45 (19)	28 (16)
60	—	—	—	88	50	—	—	—	75 (23)	38 (20)

Note. Standard deviations in parentheses.

For each critical ratio and each subject, we determine behavioral contracts that incentivize the subject to order first-best quantities by setting the behavioral model (Equation (9)) equal to the first-best order quantity (Equation (2)). If no solution exists, e.g., because a subject is too loss averse or anchors too much to incentivize the subject to place high orders, then we choose the contract parameters that incentivize orders that are as close to first-best orders as possible. We offer eight customized contracts to each of the 19 subjects. Because there is large heterogeneity in the behavioral parameters, there is also large heterogeneity in the individual contracts. (The online appendix shows the contract parameters we use.)

We do not inform subjects about our contract optimization approach; i.e., we do not explain that we

Figure 4 **Actual Mean Orders vs. Newsvendor Orders in the Validation Experiment**

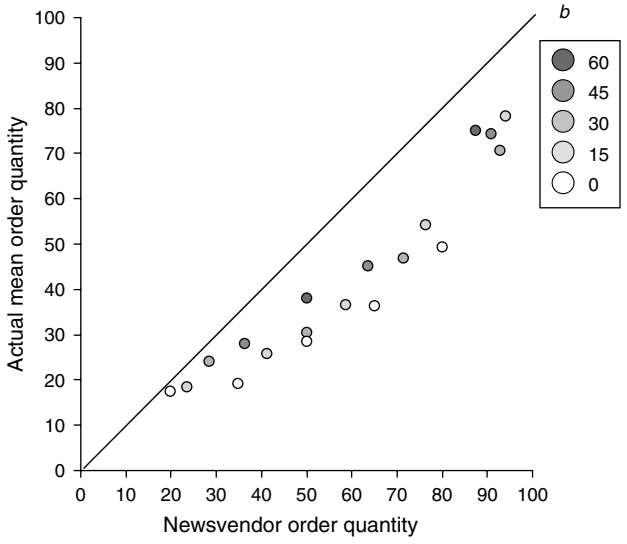


Table 5 Individual Behavioral Parameters of Subjects in the Validation Experiment

	Parameter					Parameter			
Subject	α	β	γ	σ	Subject	α	β	γ	σ
1	0.220***	5.46***	1.020	1.000	11	0.000	4.11***	1.089***	1.162
2	0.453***	4.29	1.046	3.117	12	0.243*	1.10	1.200***	2.106
3	0.365***	2.70**	1.041	1.812	13	0.119**	2.73***	1.071***	0.998
4	0.153*	4.49**	1.054**	1.563	14	0.000	0.97	1.051	1.303
5	0.032	53.85*	0.783	0.679	15	0.214*	2.11**	0.849	1.842
6	0.043	5.08***	1.077***	1.304	16	0.375***	2.88***	1.118***	1.490
7	0.447***	4.98**	1.098***	1.689	17	0.208***	2.90***	0.936	0.979
8	0.163**	0.95	1.010	1.057	18	0.383**	4.83	1.104***	3.608
9	0.009	0.99	0.998	0.091	19	0.329***	3.93***	1.050	1.549
10	0.198***	12.52***	1.009	1.024					

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

use their orders under the first 19 contracts to analyze their order behavior and that we take their order behavior into account when designing the following eight contracts. We automate the order analysis and contract optimization such that there is no recognizable time delay between the last treatment of Phase 1 and the first treatment of Phase 2. Therefore, subjects have no indication that optimization takes place between phases and are not aware that two phases exist.

Were we to inform subjects about our optimization approach, they could, in theory, anticipate this and place first phase orders that do not maximize their utility but instead attempt to maximize their utility from both phases of the experiment. Their ability to do this would critically depend on their beliefs about the workings of our optimization algorithm, which they have had no reasonable basis to form. So our implementation of the validation experiment assumes that Phases 1 and 2 behavior are independent, which is a reasonable approach for a first study that uses behavioral insights to design contracts. We do not analyze gaming effects in this paper but note that this could be an interesting topic for future research.

The results of the validation experiment are shown in Figure 5. The left panel shows the first-best quantities that we want to incentivize versus the mean actual orders placed. Each dot represents the mean orders of the 19 subjects for one of the eight critical ratios used in Phase 2. The 45-degree line indicates where mean actual orders are equal to first-best quantities. The graph shows that the mean actual orders are much closer to first-best quantities than in Phase 1 (see Figure 4 for the mean orders in Phase 1). In Phase 2, the average deviation of the mean actual orders from the target quantities is 5.0. For the same set of critical ratios, the average deviation in Phase 1 is 16.4; i.e., behavioral contracts have an average deviation from first-best that is 69.5% below the average deviation of newsvendor contracts. The average

Figure 5 Mean Order Quantities with Behavioral Contracts in the Validation Experiment

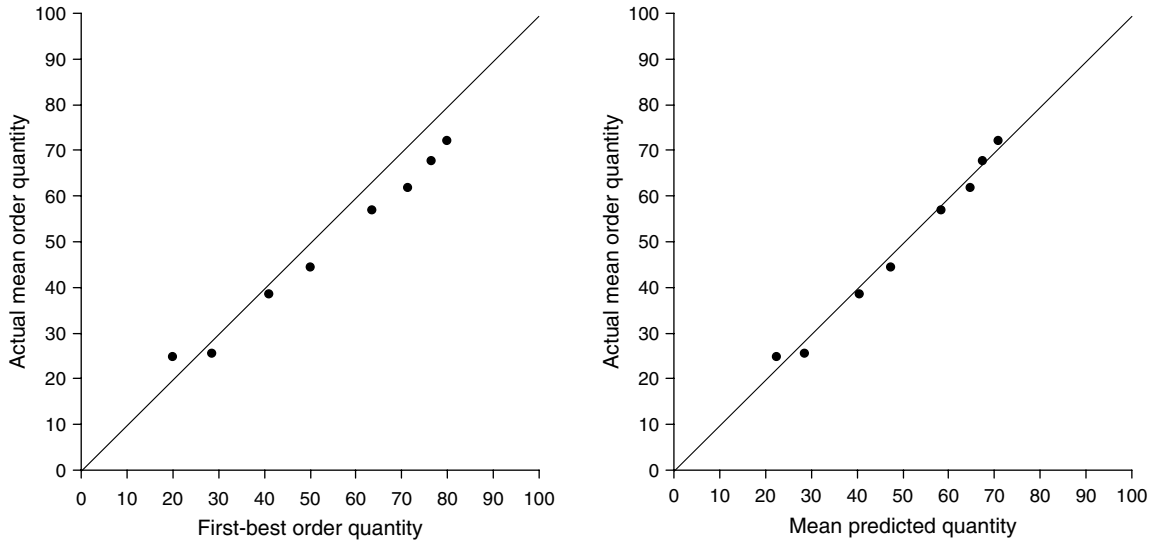


Table 6 Expected Profits in Validation Experiments for the Eight Critical Ratios Used in All Phases

		Phase 2		Phase 3
Validation experiment	Phase 1	Individual	Aggregate	Trained
Individual behavioral contracts	1,365	1,533 (+12.3%)		
Aggregate vs. individual behavioral contracts	1,389	1,546 (+11.3%)	1,519 (+9.4%)	
Training vs. individual behavioral contracts	1,426	1,579 (+10.7%)		1,503 (+5.4%)

absolute deviation of subjects in Phase 2 is significantly lower than in Phase 1 (Wilcoxon sign-rank test of mean deviations, $p = 0.002$).

Although the performance of Phase 2 shows a substantial improvement over Phase 1, the left panel of Figure 5 indicates that there is still some bias remaining: For large first-best quantities, mean actual orders tend to be below first-best quantities. A closer look at the preferences of the individuals (Table 5) provides an explanation: Some subjects cannot be incentivized to order large quantities. Consider, for example, subject 7. Subject 7 has a strong anchoring bias ($\alpha = 0.447$) and an upper bound on the order quantity of this subject is $(1-\alpha)F^{-1}(1)+\alpha\mu = 0.553 \cdot 100 + 0.447 \cdot 50.5 \approx 78$. In other words, we cannot incentivize order quantities above 78 for subject 7.

Although the left panel of Figure 5 shows how well the behavioral model can be used to reach a target quantity, it is not appropriate for analyzing the predictive accuracy of the behavioral model. Consider again subject 7 and the treatment with critical ratio $CR = 0.80$. We would like to incentivize an order quantity of 80, but any contract incentivizes orders of at most 78 for this subject. Along with subject 7, there are other subjects that cannot be incentivized to order 80 units and who are therefore offered contracts that incentivize the maximum quantity possible. Averaged over all subjects, we are incentivizing for $CR = 0.80$ an average order quantity of 71.0; i.e., the behavioral model predicts average orders quantities of 71.0. The mean actual orders for $CR = 0.80$ are 72.1 and close to the prediction.

The right panel of Figure 5 compares the mean predicted orders and mean actual orders of Phase 2 and can be used to analyze the accuracy of the prediction. From the graph, we can see that actual mean orders are close to the predicted orders. The average difference between predicted and actual mean orders is 1.2 units and 93% below the average deviation of the newsvendor model in Phase 1.

6.1.3. Comparisons of Profits. Besides mean order quantities, an important performance indicator is expected channel profit. Expected channel profit is affected not only by the mean order quantity but also by order variability. For instance, constantly ordering 50 units results in a higher expected channel profit than alternating between 25 and 75 units. In general, expected channel profit is decreasing in order variability.

To quantify the monetary benefit of using the behavioral contracts, we compare the average expected channel profits in Phases 1 and 2 for the treatments that are used in both phases (compare Table 6, row labeled *Individual behavioral contracts*). In Phase 1, with newsvendor contracts, subjects place orders that result in an average expected channel profit of 1,365. In Phase 2, with individual behavioral contracts, subjects place orders that result in an average expected channel profit of 1,533 (+12.3%), which is significantly higher than in Phase 1 (Wilcoxon signed-rank test, $p < 0.001$). Therefore, using the behavioral model instead of the newsvendor model results in significant improvements of expected channel profits.

6.2. Aggregate vs. Individual Behavioral Contracts

In situations in which it is not possible to offer each buyer an individual contract, the contract designer can use the aggregate behavioral model for contract design. Under aggregate behavioral contracts, all buyers are offered the same contract to incentivize a given target order quantity. We analyze aggregate behavioral contracts in a second validation experiment.[9]

In Phase 1 of the experiment, we expose $N = 24$ new subjects to the same $J = 19$ contracts as in the first validation experiment. Then we estimate an aggregate behavioral model (Equation (7)) as well as an individual behavioral model (Equation (9)). We use both the aggregate and individual behavioral models in a single validation experiment because this allows us to analyze the performance differences between both models using a within subject comparison.

In Phase 2, for the same critical ratios as in Phase 2 of the first validation experiment, we design eight

[9] We made a minor modification in the experimental design compared to the first validation experiment. We ask all subjects the same test questions as in the first validation experiment. However, in this and the next validation experiment, subjects cannot start on Phase 1 before they have answered all questions correctly. The optimization of the parameters again was unknown and not recognizable for the subjects.

contracts using the aggregate behavioral model and, for each subject, we also design eight contracts using the individual behavioral models. Then each subject is exposed to the eight aggregate behavioral contracts (the same for all subjects) and the eight individual behavioral contracts (generally different for each subject). Before we actually offer the 16 contracts to a subject, we randomize the sequence in which they are offered to avoid order effects.

The results are summarized in Table 6 in the row labeled *Aggregate vs. individual behavioral contracts*. The expected profits under the aggregate behavioral contracts are 9.4% higher in Phase 2 than in Phase 1 (Wilcoxon signed-rank test, $p < 0.001$). The expected profits under individual behavioral contracts are 11.3% higher in Phase 2 than in Phase 1 (Wilcoxon signed-rank test, $p < 0.001$). The difference in expected profits of the individual and aggregate behavioral model is $11.3\% - 9.4\% = 1.9\%$ and only weakly significant (Wilcoxon signed-rank test, $p = 0.096$). So the aggregate model (somewhat surprisingly) performs nearly as well as the individual model, which speaks to the practicality of our approach: In real contracting arrangements, implementing an aggregate model is likely to be both more practical and less vulnerable to gaming than implementing an individual model.

6.3. Training vs. Individual Behavioral Contracts

We argue that subjects generally have behavioral preferences that differ from those of the newsvendor model. According to our model, these behavioral parameters belong to the subjects' personality traits—subjects place biased orders intentionally and not by mistake. However, an alternative explanation of the order pattern that we observe, and that as modelers we should always consider, is a lack of good understanding of the underlying newsvendor model. We address this issue in an additional validation experiment where we provide training on the newsvendor model.

The validation experiment has three phases. Phases 1 and 2 are the same as those of the first validation experiment; i.e., in Phase 1, we expose the subjects to 19 different contracts to estimate the behavioral parameters of each subject and in Phase 2 we offer each subject eight individualized contracts. At the beginning of Phase 3, before any order is placed in Phase 3, we provide subjects with additional training on the newsvendor model. After training, we offer subjects again eight out of the 19 contracts of Phase 1. We use the same eight contracts that are individualized in Phase 2 to analyze how training affects the subjects' orders.

As part of the training, we first explain the newsvendor model in detail by essentially repeating the initial briefing. Next, using a numerical example with $w = 67$, $b = 40$, and $r = 100$, we illustrate the effect of the order quantity on expected profits and show how the order quantity affects expected sales, expected returns, minimum and maximum profits, and loss probabilities. Then we explain how the expected profit maximizing order quantity can be computed (standard textbook derivation), illustrate the approach using an example, and explain that compensation in the actual experiment will be based on average profit. The training slides are contained in the online appendix.

Before subjects can place orders in Phase 3, we ask them to solve an exercise ($w = 80$, $b = 60$, $r = 100$, demand uniform between 1 and 100). We ask them to determine the expected profit maximizing order quantity and to determine additional performance indicators for order quantities of 40 and 50: expected sales and returns, minimum and maximum profits, and loss probabilities. Subjects can continue only after they have answered all questions correctly. We use this approach to ensure that the subjects are able to compute the newsvendor quantity from the data and that they are aware of the consequences of their order decision on potential losses and on expected revenue streams.

In the validation experiment, we use 19 new subjects. The results are summarized in Table 6 in the row labeled *Training vs. individual behavioral contracts*. The results for Phases 1 and 2 are similar to those of the first validation experiment. The increase in expected channel profits from Phase 1 to Phase 2 is 10.7% (Wilcoxon signed-rank test, $p < 0.001$). The results for Phase 3 show that training weakly increases expected channel profits over Phase 1 by 5.4% (Wilcoxon signed-rank test, $p = 0.059$).

To gain a better understanding of the effect of training on order behavior, we use maximum likelihood estimation to estimate aggregate behavioral models for Phases 1 and 3. For the estimations, we use only the eight critical ratios used in both phases. For Phase 1, we obtain $\alpha = 0.335$, $\beta = 3.46$, and $\gamma = 1.16$; for Phase 3, we obtain $\alpha = 0.257$, $\beta = 2.06$, and $\gamma = 1.10$. The results show that training has an effect on the behavioral parameters, but only β is significantly different in Phase 3 from Phase 1 ($p < 0.01$ for β, $p > 0.2$ for α and γ), and we conclude that training might move the behavioral parameters toward the values of the newsvendor model. However, all three parameters of the aggregate behavioral model in Phase 3 remain significantly different from those of the newsvendor model ($p < 0.001$), which indicates that training removes some of the decision bias but does not eliminate it. Bolton et al. (2012) found similar results for training in their study.

The results of the experiment also show that the individual contracts applied to untrained subjects

(Phase 2) perform better than newsvendor contracts applied to trained subjects (Phase 3). In the experiment, expected profit in Phase 2 is 5.3% higher than in Phase 3 (Wilcoxon signed-rank test, $p = 0.046$).

7. Conclusion

There exists a large body of literature on supply contracting that assumes that people place orders according to the newsvendor problem. We have seen that generally this assumption does not hold and that decision makers anchor on the mean demand, are averse to losses, and value different income streams differently. The behavioral model we propose takes these three effects into account and provides a more realistic building block for contract design than the newsvendor model does.

There are managerial implications from our research. It shows that people respond irrationally to supply contracts but that their responses can be reasonably well predicted. Contract designers who are aware of this might consider choosing contracts with high buyback prices and high wholesale prices rather than contracts with low buyback prices and low wholesale prices. Our research indicates that such contracts would be preferred by many buyers. However, there are also buyers who prefer the opposite, and the task of the contract designer is to classify the buyer. Because peoples' behavioral preferences differ, we cannot provide recommendations that hold universally. However, we can provide the general recommendation to realize that people often value different income streams differently, that they frame a contract, and that they place a different value on gains than on losses, information that can be valuable in contract design.

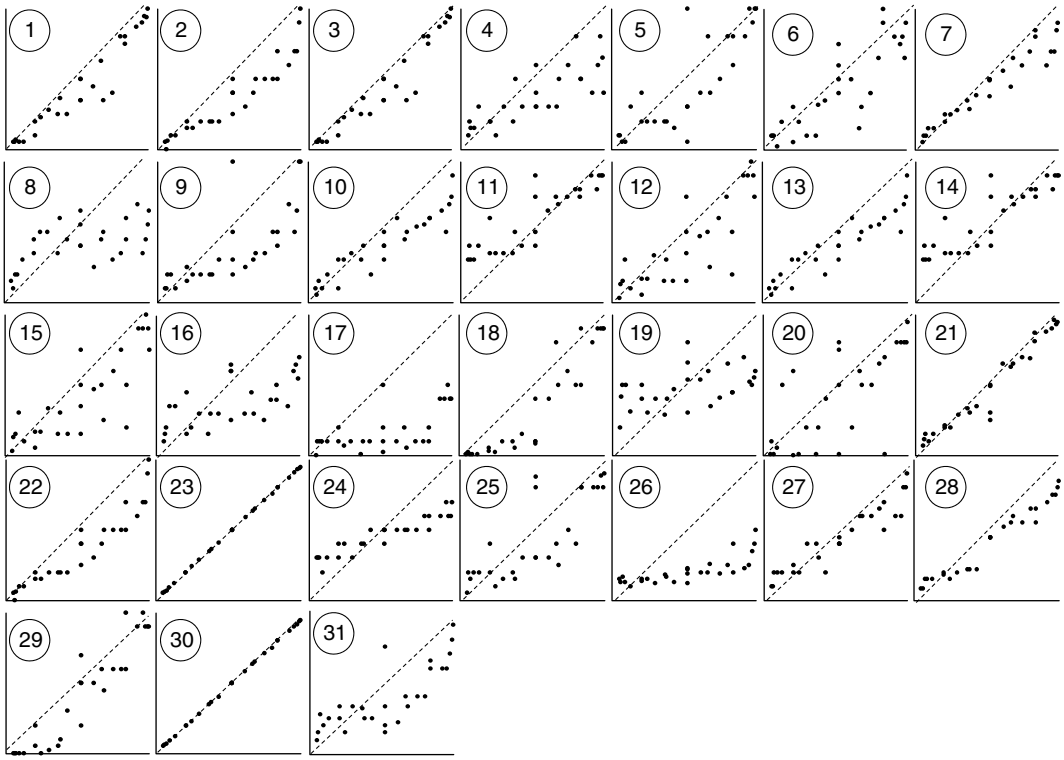
We analyze order behavior in a laboratory environment, which allows us to isolate the effects we are interested in. In reality, additional effects are likely to be relevant, such as social preferences (e.g., fairness and trust; see Cui et al. 2007), competition, and the frequency of the interaction. Although our research provides some insights into the effect of some of the relevant factors on order behavior, there is ample room for future research to analyze additional factors and develop decision support models that map reality more closely than our model does.

Acknowledgments

The authors gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft through the DFG-Research Group "Design and Behavior" and its members for useful comments.

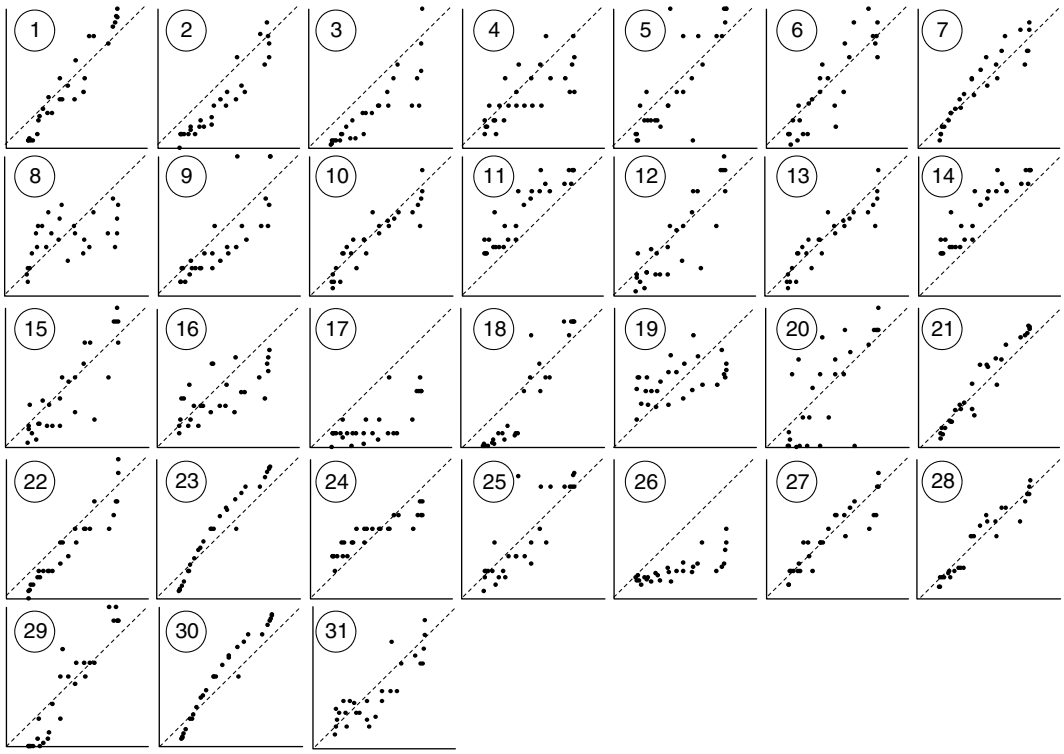
Appendix A. Predicted vs. Actual Orders by Subject

Figure A.1 **Newsvendor Orders vs. Actual Orders**



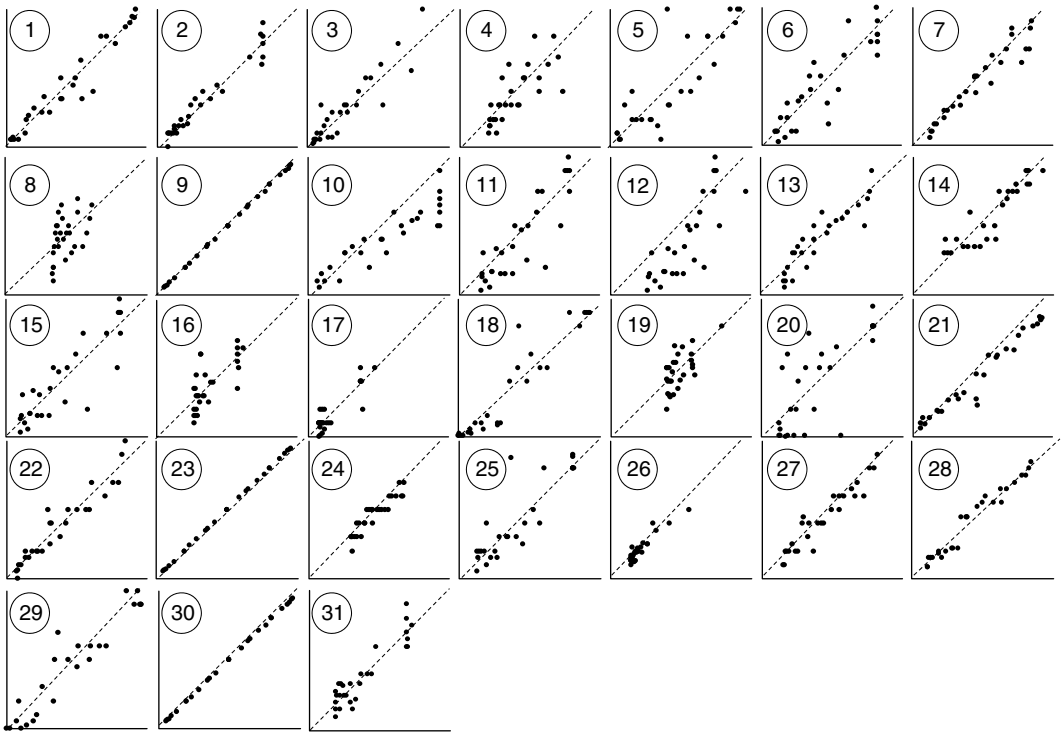
Notes. The scale of both axes is 0–100. *x*-Axis: newsvendor quantity. *y*-Axis: actual order quantity. The numbers in the circles denote the subject. Each dot corresponds to an order.

Figure A.2 Aggregate Behavioral Model vs. Actual Orders



Notes. The scale of both axes is 0–100. *x*-Axis: aggregate behavioral model quantity. *y*-Axis: actual order quantity. The numbers in the circles denote the subject. Each dot corresponds to an order.

Figure A.3 Individual Behavioral Model vs. Actual Orders



Notes. The scale of both axes is 0–100. *x*-Axis: individual behavioral model quantity. *y*-Axis: actual order quantity. The numbers in the circles denote the subject. Each dot corresponds to an order.

Appendix B. Distribution of Residuals

Figure B.1 Residual Plots for the Aggregate Behavioral Model

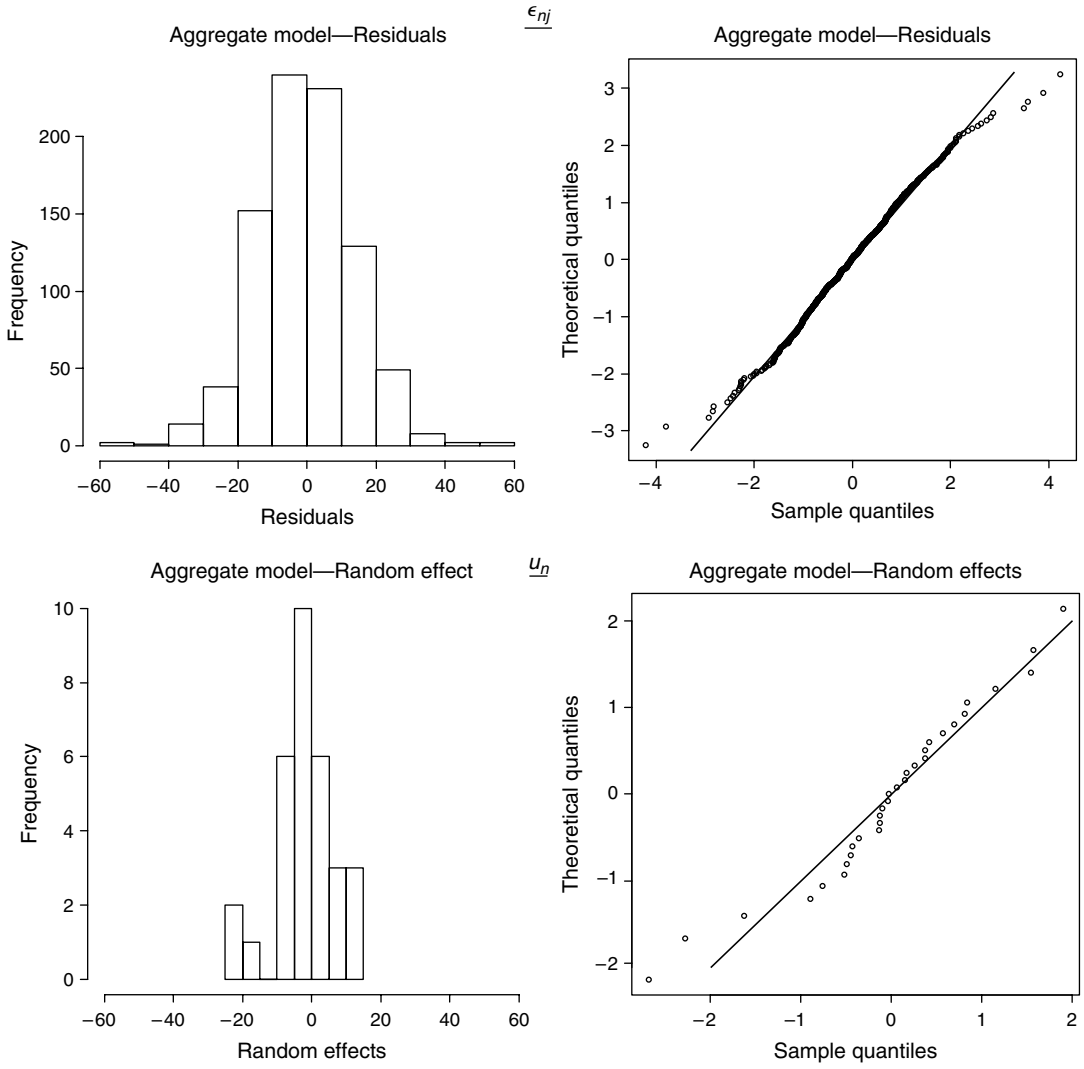
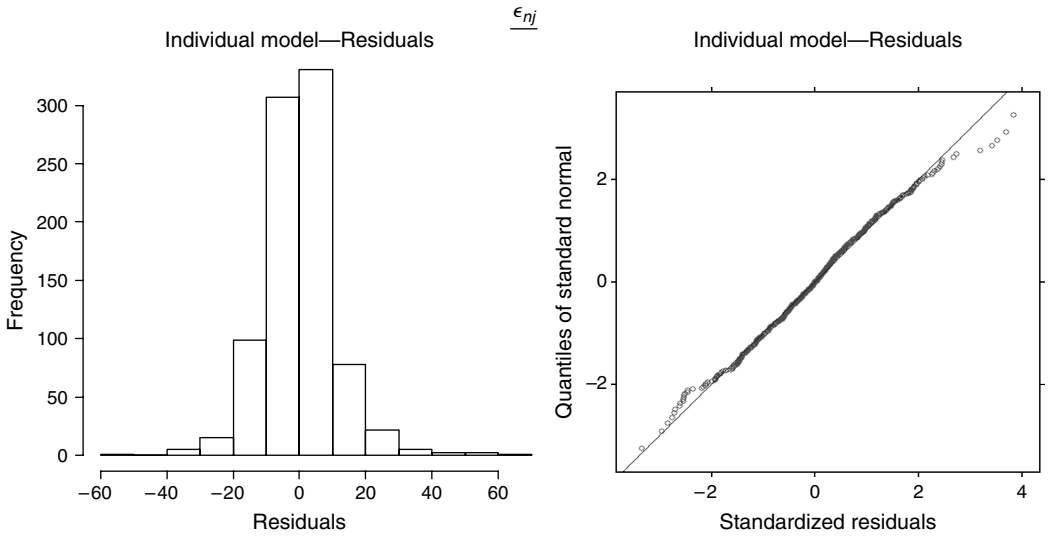


Figure B.2 Residual Plots for the Individual Behavior Model



The Kolmogorov–Smirnov test indicates no significant deviation from normal distribution for the residuals of the aggregate behavioral model ($p = 0.6502$). The random intercept is also not significantly different from normal distribution ($p = 0.5235$). Figures B.1 and B.2 show histograms and q–q plots of the residuals and random errors.

For the individual behavioral model, the test indicates a deviation from normal distribution (due to strong kurtosis = 4.22) ($p < 0.01$). The use of a maximum likelihood estimation assuming normally distributed errors is justified because the distribution of the residuals is fairly symmetric (skewness = 0.59) and so the estimates are consistent. Note that the random intercepts are zero by definition for the individual model so we leave them out here.

References

Akaike H (1981) Likelihood of a model and information criteria. *J. Econometrics* 16(1):3–14.

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

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

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

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

Bostian A, Holt C, Smith A (2008) The newsvendor "pull-to-center effect": Adaptive learning in a laboratory experiment. *Manufacturing Service Oper. Management* 10(4):590–608.

Cachon G (2003) Supply chain coordination with contracts. Graves S, de Kok T, eds. *Handbooks in Operations Research and Management Science: Supply Chain Management* (North-Holland, Amsterdam), 229–339.

Cui TH, Raju JS, Zhang ZJ (2007) Fairness and channel coordination. *Management Sci.* 53(8):1303–1314.

Eeckhoudt L, Gollier C, Schlesinger H (1995) The risk-averse (and prudent) newsboy. *Management Sci.* 41(5):786–794.

Fischbacher U (2007) *Z*-tree: Zurich toolbox for ready-made economic experiments. *Experiment. Econom.* 10(2):171–178.

Greiner B (2004) The online recruitment system ORSEE 2.0—A guide for the organization of experiments in economics. Working Paper Series in Economics 10, University of Cologne, Cologne, Germany.

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

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

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

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

Kooreman P (2000) The labeling effect of a child benefit system. *Amer. Econom. Rev.* 90(3):571–583.

Lariviere MA (1998) Supply chain contracting and coordination with stochastic demand. Tayur S, Ganeshan R, Magazine M, eds. *Quantitative Models for Supply Chain Management* (Kluwer Academic Publishers, Norwell, MA), 233–268.

Nahmias S (2008) *Production and Operations Analysis* (McGraw-Hill/Irwin, New York).

O'Curry S (1997) Income source effects. Working paper, DePaul University, Chicago.

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

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

Thaler R (1985) Mental accounting and consumer choice. *Marketing Sci.* 4(3):199–214.

Thaler R (1999) Mental accounting matters. *J. Behav. Decision Making* 12(3):183–206.

Tsay A, Nahmias S, Agrawal N (1998) Modeling supply chain contracts: A review. Tayur S, Ganeshan R, Magazine M, eds. *Quantitative Models for Supply Chain Management* (Kluwer Academic Publishers, Norwell, MA), 299–336.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.

Wang CX, Webster S (2009) The loss-averse newsvendor problem. *Omega* 37(1):93–105.