Three Essays in Decision Making

by

Ray Weaver

Submitted to the Alfred P. Sloan School of Management
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
at the
Massachusetts Institute of Technology

June 2008

Signature of Author ______________________________
Alfred P. Sloan School of Management
May 29, 2008

Certified by ______________________________
Shane Frederick
Sarofim Family Career Development Professor
Associate Professor of Management Science
Thesis Supervisor

Accepted by ______________________________
Birger Wernerfelt
J.C. Penney Professor of Management Science
Chair, Doctoral Program

Three Essays in Decision Making

by

Ray Weaver

Submitted to the Alfred P. Sloan School of Management
on May 29, 2008 in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Management Science

ABSTRACT

This dissertation is composed of three essays about consumer judgment and decision making.

In Essay 1, I develop a novel explanation for the well-known endowment effect — the tendency for owners to value goods more than non-owners do. According to a prominent explanation for this effect, the prospect of losing possessions creates psychological pain, inducing sellers to demand more than buyers will pay. My alternative account is based on transaction disutility: consumers are reluctant to trade on terms that are disadvantageous with respect to perceived market prices. The endowment effect appears to be caused by inflated selling prices because market prices usually exceed the value of ownership to consumers. But I show that reducing reference prices relieves sellers' transaction disutility, shrinking or eliminating the effect. Moreover, very low reference prices create disutility among buyers, resulting in disparities driven by a reluctance to buy, not to sell.

Essay 2 explores the implications of transaction disutility for consumer preferences. Maximum buying and minimum selling prices are commonly believed to reveal preferences: a consumer who prefers one good over another presumably has a higher reservation price for it. But transaction disutility can distort reservation prices away from underlying values. If alternative measures of preference — such as binary choices between goods — are not regarded by consumers as transactions, they are not subject to such distortions. This difference can create preference reversals, that is, incoherence between explicit choices and the preferences implied by stated reservation prices. I find strong experimental evidence for this proposition.

The "Bayesian Truth Serum" (BTS) is a survey scoring method designed to provide truthtelling incentives for respondents answering multiple choice questions about intrinsically private matters: opinions, tastes, past behavior. My final essay discusses several tests of BTS. In one questionnaire, respondents indicated their familiarity with various items (e.g. electronics brands), one-third of which were nonexistent foils. BTS did in fact reward truthtelling: the scoring method severely penalized "recognition" of foils. Also, survey takers viewed the BTS method as credible: people who were paid for achieving higher BTS scores claimed to recognize fewer foils, even when facing competing incentives to deceive.

Thesis Supervisor: Shane Frederick
Title: Sarofim Family Career Development Professor, Associate Professor of
Management Science

Table of Contents

Biographical Note

Ray Weaver was a National Merit Scholar and was awarded the Woodward Merit Fellowship at Washington University in St. Louis, where he studied Electrical Engineering and Computer Science. At Wash. U., he received the Electrical Engineering department's Distinguished Service Award, and was named to Eta Kappa Nu, the EE honor society. He graduated with dual Bachelor of Science degrees in 1992.

After completing his undergraduate work, Ray joined the information technology practice of Deloitte Consulting. He worked there three years, serving clients in the transportation and health care industries.

Ray matriculated to the MBA program at the Wharton School of the University of Pennsylvania in 1995. At Wharton, he was named a Palmer Scholar, graduating in the top five percent of his class. He received his degree in 1997.

Following his MBA, Ray moved to Silicon Valley to work for Intel, where he helped launch Bluetooth, the wireless communication technology. In early 1999 he moved to Cambridge to join a start-up called Akamai Technologies, where he was the first member of the marketing group. Akamai went public in October 1999 and is now a member of the S&P 500.

In the fall of 2002, Ray left Akamai and matriculated into the doctoral program in the marketing group of MIT's Sloan School of Management. In addition to his research responsibilities, Ray has frequently served as a teaching assistant for undergraduate and master's courses, and was nominated for the Goodwin Medal in 2005.

After completing his PhD, Ray will join the faculty of the marketing unit at Harvard Business School. He lives in Cambridge with his two wonderful daughters, Hayden and Casey.

Acknowledgements

I have been extremely fortunate to learn from such a remarkable group as the marketing faculty at MIT. Many thanks to Duncan Simester, Birger Wernerfelt, John Little, Glen Urban, Michael Braun, Catherine Tucker, Juanjuan Zhang, Barak Libai, Sharmila Chatterjee, and Jin-Gyo Kim. Special thanks to Drazen Prelec, Dan Ariely, John Hauser, and in particular Shane Frederick for investing so much time and energy in me.

Thanks also to my fellow students, who have taught me much and are great friends: Daniel, Nathan, Boris, Nina, Kristina, Cristina, Daria, Han, Zoe, Leonard, On, Robert, Olivier, Jiwoong, Anat, and especially Matt, the best office mate ever. Thanks to my study partners Tanu, Alice, and Ramana, who helped me survive my first year. I am very grateful to my undergraduate research assistants for their help and enthusiasm: Stephanie, Laura, Jonathan, Danielle, and Annie. Thank you to Rosa, Sandra, Drew, Sharon and Hillary for helping me navigate this place.

Many thanks to the people at Akamai Technologies who took a chance on me: Paul Sagan, Tom Leighton, Jonathan Siegel, and Danny Lewin. Without the opportunity to work at Akamai, I would not have had the wherewithal, and maybe not the courage, to go back to school at my age with two young children. We miss you, Danny.

Six years is a long time, and I sometimes felt I was losing my way. I had a lot of help staying on track from friends and family who believed in me, most especially Rachel Brasier, to whom I am very grateful. I also thank Kelly Weaver for her consistent flexibility and teamwork in raising our children.

I thank my wonderful daughters, Hayden and Casey, for both teaching and giving me patience. Girls: I often tell you that most worthwhile things require lots of practice and hard work. I hope you agree that this dissertation is evidence of that.

Finally, I want to thank Arthur Conan Doyle, Malcolm Gladwell, and Richard Dawkins, whose books have inspired me and given me something very, very high to aim for.

- Essay 1 -

Transaction Disutility and the Endowment Effect

I would like to thank seminar participants at Boston University, Emory, Harvard Business School, New York University, San Diego State University, UCLA, and Yale; Drazen Prelec, Matt Selove, Birger Wernerfelt, and especially Shane Frederick for helpful comments.

ABSTRACT

Acquiring a good seems to increase the value of the good to its owner, as owners generally demand more to sell than non-owners are willing to pay. This *endowment effect* is typically explained in terms of loss aversion: the prospect of losing possessions causes psychological pain for which the premium demanded by sellers is compensation. An alternative explanation is that values are distorted by *transaction disutility*: people are reluctant to trade on terms they consider disadvantageous or unfair. Consumers judge potential trades against perceived market prices, and because market prices usually exceed the value of ownership to most consumers, sellers experience transaction disutility and adjust their reservation prices upward. I show experimentally that reducing reference prices relieves sellers' transaction disutility, shrinking or eliminating the endowment effect. Moreover, further reducing reference prices to levels below consumer values induces transaction disutility in *buyers*, creating buyer-seller disparities driven by a reluctance to buy, not to sell. Thus, the size of the endowment effect is U-shaped as a function of reference price: large when reference prices are either very high or very low, but small for intermediate reference prices. I also find that the effect is smallest among people who most value the good. These results support the contention that an aversion to bad deals, not an aversion to losing possessions per se, causes buying and selling prices to diverge. They also suggest that marketing efforts designed to make favorable reference prices salient are more effective than those that try to instill a sense of ownership in potential customers.

INTRODUCTION

Buying and selling are two sides of the same marketplace coin. Stigler (1966) asserts: *"There is a deep symmetry between buying and selling, and the economic theory of the two acts is identical. (With barter of one commodity for another, an ask and an offer are indivisible: I offer one sheep for one bag of salt)."* Both acts are expressions of the value one places on a good, with money as the medium for this expression. Accordingly, measures of the most a non-owner is willing to pay for some good, and the least an otherwise similar owner demands to give it up, should coincide (Willig, 1976). In fact, however, this presumption is often violated. In a review of 59 studies involving market goods, Horowitz and McConnell (2002) found that minimum selling prices exceeded maximum buying prices by a factor of nearly three.[1]

Thaler (1980) termed this buyer-seller price disparity the *endowment effect*. The effect is commonly attributed to loss aversion (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991): people are assumed to compare potential trades to the status quo, and to feel losses relative to their current holdings more keenly than gains. If the transaction under consideration is the purchase or sale of a coffee mug, owners regard its potential loss as more significant than non-owners regard its potential acquisition. According to this model, selling prices exceed buying prices because sellers charge a premium to offset the psychological pain they anticipate experiencing when they give up their mug.[2]

[1] There are normative reasons why buying and selling prices might diverge, including transaction costs and income effects. Some studies have found persistent disparities even when these explanations are ruled out (e.g. Kahneman, Knetsch & Thaler, 1990); my aim is to better understand gaps that are too large to be fully explained by such factors. Separately, Hanemann (1991) shows that large disparities can normative for goods that lack perfect substitutes, such as personal health. But this result does not generally apply to market goods, which are substitutes for money.

[2] Thaler (1980) uses "endowment effect" to refer both to the finding that selling prices tend to exceed buying prices, and to the theoretical explanation of loss aversion for that finding. To avoid confusion, I use the term only to describe the phenomenon.

I propose an alternative explanation: the endowment effect is caused not by the anticipated pain of losing a possession, but by a reluctance to trade on terms that are perceived to be disadvantageous or unfair — an experience I call *transaction disutility*, following terminology proposed by Thaler (1985). I show that the magnitude of the endowment effect, operationalized as the ratio of selling prices to buying prices, depends critically on the difference between a good's reference price and the value consumers expect to derive from ownership. When this difference is large, so is the gap between buying and selling prices, but when reference prices are similar to valuations, the disparity is smaller or disappears entirely. Furthermore, although the endowment effect appears to be driven by sellers, I demonstrate that this perception is an artifact of typically high reference prices. When reference prices are low relative to valuations, buyers experience transaction disutility, resulting in disparities driven by a reluctance to buy, not to sell. Lastly, I show that the endowment effect is smallest among people who most value a good. All of these results support the transaction disutility model, and none are readily explained by the "pain-of-losing" account of the endowment effect.

THE IMPACT OF TRANSACTION DISUTILITY ON BUYING AND SELLING PRICES

Thaler (1985) proposed that consumers contemplating transactions consider not only the benefits of using the good they might buy or sell, but also the perceived merits of the deal: whether the actual price is higher or lower than they expect. He asked respondents to imagine sitting on a beach with a friend who offered to bring them a bottle of their favorite beer, but needed to know the maximum he should spend on their behalf. Those informed that the beer would be purchased at a fancy hotel would pay an average of $2.65, but others told that it would be bought at a run-down grocery store authorized just $1.50. In other words, the *expectation* to pay less became a *willingness* to pay less. Thaler also showed that transaction utility affects sellers. People asked to imagine selling tickets to a hockey game they could no longer attend demanded more money if the original purchase price was higher. Thaler's examples suggest that buying and selling prices can be swayed by the price that consumers think is fair or reasonable, in a way unrelated to the utility of the good itself.

I propose that transaction utility — in particular, as will be explained later, transaction *dis*utility — is the principal cause of the endowment effect. Reservation prices (or "bids": the maximum a buyer is willing to pay, the minimum an owner demands to sell) are not only expressions of the utility a good confers, but also reflect consumers' beliefs about what transaction price is reasonable. This reasonableness is judged according to some reference price for the good, which may be suggested to consumers (retail price, for example), or which they may generate internally. I propose that people are particularly averse to bad deals with respect to this reference price. Therefore, I assume that buyers are reluctant to pay more than the reference price, and sellers are reluctance to accept less. This drives buying prices downward and selling prices upward, even when buyers and sellers have similar underlying preferences for the good.

Transaction disutility depends on the relationship between a consumer's valuation and the reference price, and therefore may affect buyers and sellers asymmetrically. Suppose, for example, that valuations for a coffee mug vary uniformly between zero and ten dollars. If the mug's reference price is eight dollars, transaction disutility will have a large influence on potential sellers because for most people, selling at their valuations would be a "bad" deal, i.e. below the reference price. Conversely, transaction disutility will have little effect on potential buyers, most of whom could pay their valuations without making a bad deal, which from their perspective is a price *above* the reference price.[3]

Reference prices are usually based on market prices,[4] which typically exceed the value most consumers place on products. Consumers turn down the opportunity to buy most goods in the marketplace. Supermarkets, for example, carry tens of thousands of items, but each shopper buys at most a few dozen on a given trip and repeatedly declines to purchase almost everything else in the store. Like the pain-of-losing account, my account of evaluation disparities implicates sellers rather than buyers. That is, buying prices are assumed to closely mirror underlying valuations, whereas selling prices significantly exceed them. In contrast to the pain-of-losing explanation, however, I attribute this to the incidental fact that reference prices are generally high, rather than to a fundamentally different consideration on the part of sellers, namely an aversion to parting with possessions.[5]

Despite Thaler's proposal that transaction utility can influence buying and selling prices, the concept has not, surprisingly, been explicitly invoked to explain the endowment effect. Some recent research, however, is suggestive of this explanation. Simonson and Drolet (2004) studied the impact of arbitrary anchors, such as digits from one's Social Security Number, on bids. They found that anchors on perceived value influence buyers, whereas anchors on market price influenced sellers. Their conclusion was that personal value and market price both inform stated reservation prices, but are weighted differently: value is more important when buying, and market price more important when selling. This pattern is consistent with my model when

[3] In incentive-compatible experiments, reservation prices are "worst-case scenarios": buyers may well pay less, and sellers receive more, than their stated bids. In this context, transaction disutility can be thought of as the maximum possible suffering, if the worst case is realized.

[4] Prior research has shown that consumers may base their reference prices on the most recently observed price (Winer 1986), the price last paid (Monroe 1973), or the average price of similar products (Emery 1970).

[5] For example, Kahneman, Knetsch and Thaler (1990) assert that "the endowment effect is primarily a problem for sellers; we observe little reluctance to buy but much reluctance to sell." Tversky and Kahneman (1991) add: "the buyers in these transactions do not appear to value the money they give up in a transaction as a loss."

market prices exceed most consumer valuations, which is the norm and which was true in Simonson and Drolet's studies.[6] Using verbal protocols, Brown (2005) examined how consumers justify disparities their own maximum buying vs. minimum selling prices for common goods: chocolate bars, coffee mugs, and notebooks. Explanations related to the pain of losing account were rare, but explanations consistent with transaction disutility were very common. His participants spontaneously generated reference prices by estimating market prices or speculating about how much others might pay, and made bids that gave them goods deals or avoided losses with respect to these reference prices.

This recent research supports my hypothesis, but is not conclusive. These and nearly all studies of evaluation disparities have used stimulus goods with high reference prices, confounding the effect of endowment (sellers are endowed whereas buyers are not) with the effect of transaction disutility (which applies to most sellers, but not most buyers). The approach here is to eliminate this confound by manipulating reference prices experimentally. This leads to two important contributions to our understanding of the endowment effect. First, I show that transaction disutility mediates the endowment effect. In fact, under some circumstances and for some reference prices, I find that the endowment effect disappears altogether. Second, I explain why the effect appears to be asymmetrically driven by sellers, and specify the conditions under which disparities are caused by a reluctance to buy.

Predictions of the model

Assume that a consumer's desire for any given good, given the available substitutes and a budget constraint, can be expressed in monetary terms as his valuation *v*. This valuation, which cannot

[6] In their Study 3, for example, the reported list price of a toaster was $39, but the median stated reservation price was just $23, suggesting that the reference price exceeded most valuations.

be observed directly, corresponds to Thaler's (1985) "acquisition utility": the expected benefits from using the good.[7] Consumers' stated maximum buying prices (b) or minimum selling prices (s) are based on their underlying valuations, but can be distorted by transaction disutility. Specifically, I assume that consumers adopt some reference price r, and experience transaction disutility whenever a trade at their personal valuation would be unfavorable with respect to this reference price: that is, whenever $r < v$ for buyers, and $r > v$ for sellers. Transaction disutility distorts buying prices downward and selling prices upward. In summary:

Buyers: If $r < v$, then $b < v$
If $r > v$, then $b = v$

Sellers: If $r < v$. then $s = v$
If $r > v$, then $s > v$

Figure 1 illustrates this model for a consumer with valuation v. The x-axis depicts a range of possible reference prices the consumer might face. The y-axis plots the consumer's stated maximum buying price or minimum selling price as a function of his valuation and the reference price. If declared reservation prices were independent of reference prices — that is, if transaction disutility played no role — the graphs for both buyer and seller would be horizontal lines at v. Instead, transaction disutility causes selling price to exceed buying price for any $r \neq v$. As the figure shows, the size of the evaluation disparity increases as the difference between the valuation and reference price, $|v - r|$, grows. I use the term *distance* to refer to this measure. (I will later discuss the important special case of r equal to v, where the model predicts no endowment effect.)

[7] To be precise, Thaler's acquisition utility is value minus acquisition price, whereas here valuation is not net of price.

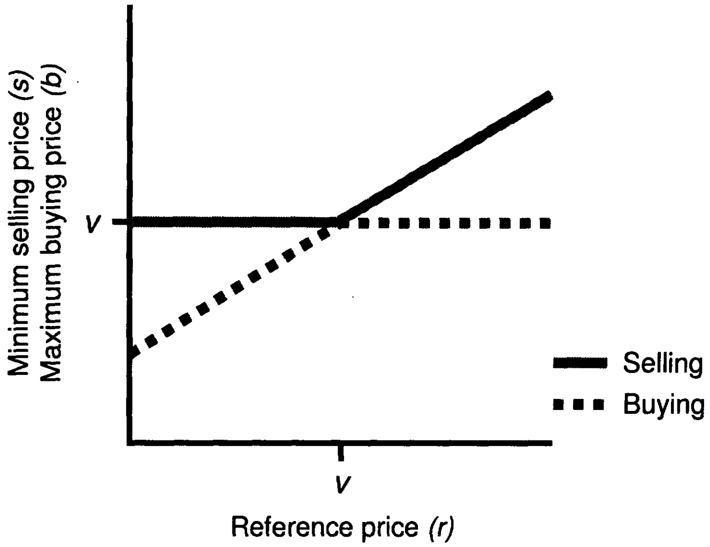


Figure 1. Stated reservation price as a function of valuation and the reference price. Buying prices are deflated when $r < v$, and selling prices are inflated when $r > v$.

This stylized example simplifies reality in several ways that should be noted. First, whereas Figure 1 shows a single individual with some particular valuation, in practice preferences vary, so valuations will follow some distribution.[8] Similarly, people may differ on which reference price they adopt and on how strongly they react to unfavorable reference prices (see Lichtenstein, Netemeyer and Burton, 1990).[9] We can generalize the distance measure to accommodate this heterogeneity: my model predicts that the size of the endowment effect (as measured by the ratio of the average of the sellers' reservation prices to the average of the buyers' reservation prices, *S/B*) grows as the average of the distances over all observed consumers increases.

Note that there is no *positive* transaction utility in the model — no additional utility obtained from transacting on terms that are *favorable* with respect to the reference price. In other words,

[8] So long as ownership status is random, however, this distribution is the same for buyers and sellers.
[9] If the widespread perception that older consumer are "tight" is correct, it may be due to their memory of times when nominal prices were much lower.

regardless of the reference price, I assume that a buyer is never willing to pay more than, nor is a selling willing to accept less than, his underlying valuation. This complication is omitted because I believe that if present, its effects are weak.[10] However, the predictions hold in a generalized model that includes positive transaction utility, provided that negative transaction utility outweighs positive. The model also does not explicitly capture the effect of reference prices on underlying valuations. Reference prices may function as arbitrary anchors as well as signals of a good's quality (Simonson and Drolet, 2004; Nunes and Boatwright, 2004). Such influences are reflected in stated reservation prices, but they do not explain the endowment effect, as they should influence both buying and selling prices.

The transaction disutility model predicts that evaluation disparities increase with distance, i.e. with an increasing difference between valuations and the reference price. Thus, as we manipulate *r* experimentally from very low to very high levels, the buyer-seller price gap should first shrink, then increase. The pain-of-losing account does not make this prediction, because it holds that the endowment effect is caused by a difference in ownership status between buyers and sellers, which difference is consistent as *r* changes.

Prediction 1: Evaluation disparities are a U-shaped function of the reference price, whose minimum is found where the average distance between consumer valuations and the reference price is minimized.

The model also predicts that reference price manipulations will affect buyers and sellers differently depending on the range over which *r* is varied. For $r < v$, transaction disutility affects buyers but not sellers; for $r > v$, it affects sellers but not buyers.

Prediction 2: For low reference prices, changing the reference price affects buying prices more than selling prices. For high reference prices,

[10] People do occasionally buy things they don't need because the deal is "too good to pass up." But positive transaction utility among *sellers* seems exceptionally rare: I have not heard of someone selling a good for less than he values it because of a very low reference price.

changing the reference price affects selling prices more than buying prices.

A question of particular interest is whether the model predicts that the endowment effect will ever disappear entirely for certain reference prices. Put another way, does *S/B* = 1 at the bottom of the U-shaped function? Assuming we measure the effect using the *means* of the buying and selling prices, the answer is in general no. As long as there is some heterogeneity in valuations, average distance is non-zero for all values of *r*, so *S* is always larger than *B*.[11] Using *medians*, however, the result is different. Suppose *r* is equal to the median of the (unobservable) distribution of underlying valuations. Then the median valuation is not distorted for either buyers or sellers: some buying prices are deflated, but remain above the median, and some selling prices are inflated, but remain below the median. An example of this special case is shown in Figure 2. As the figure shows, mean *S* and *B* are clearly distorted, but their medians are not.[12] This analysis leads to my final prediction:

Prediction 3: **Measured over a sufficiently large range, there is some intermediate reference price at which there is no endowment effect on a median basis, i.e. at which median buying and selling price are equal.**

[11] If the distribution of valuations is sufficiently narrow, however, average distance will be small and *S/B* will approach 1 when *r* is close to this distribution's mean.

[12] Although Figure 2 shows a uniform distribution of valuations, the prediction holds for any underlying distribution.

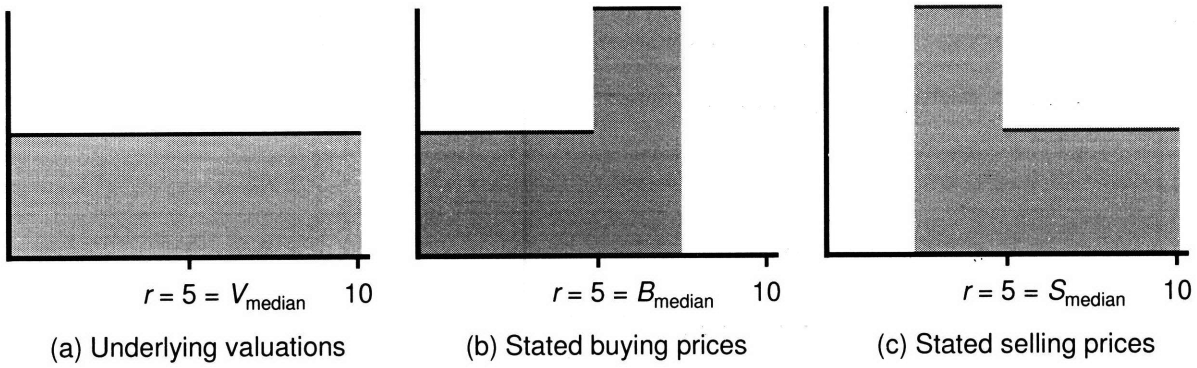


Figure 2. Distributions of buying and selling prices assuming valuations V are distributed uniform[0, 10]. When $r = V_{median}$, means are distorted, but medians are not.

Before turning to the experimental results, I first specify how we will characterize reference prices as "low" or "high." I use these labels to refer to the position of the reference price relative to the average consumer valuation. Though v is a latent construct, the ordinal relation between average buying prices, selling prices, and valuations is known: reference prices distort B downward and S upward. Thus r is "low" when it is less than B and S, high when it exceeds B and S, and moderate when it falls between the two.

EXPERIMENTAL EVIDENCE

Study 1: High vs. moderate reference price

Because natural reference prices tend to be high, the first study examines the effect of varying r from moderate to high. I predict that the endowment effect will be larger for high r, and that the reference price manipulation will influence sellers more than buyers.

Method. I used a 2 (role: buyer or seller) × 2 (reference price: moderate or high) between subjects design. Participants (N=125) were recruited for laboratory sessions at two universities.

The stimulus good was "movie theater" candy — the large boxes sold at theater concession stands. In the moderate *r* condition, subjects were told, "As a point of reference, the Target store in Watertown sells this candy for $1.49 per box." In the high *r* condition, I informed subjects, "As a point of reference, the Harvard Square Theater sells this candy for $4.00 per box." Both statements were true.[13]

Each participant was randomly assigned to be either a potential buyer or potential seller, and to receive either the moderate or high reference price.[14] After viewing their options — Raisinets, Milk Duds, Goobers, and Jelly Belly Sours — buyers were asked the most they would pay for a box of their preferred candy, whereas sellers were informed that they would be given a box of their preferred candy, and were asked the least they would demand to sell it. Additional onscreen instructions then explained the Becker-DeGroot-Marschak (BDM) (1964) incentive-compatible elicitation procedure by which the actual transaction would be determined. To avoid introducing additional reference prices, the range from which the BDM price would be drawn was not revealed.[15] After everyone finished reading these instructions, a researcher passed around boxes of the four types of candy for inspection, reiterated that the outcomes were real and that it was in each participant's interest to reveal his true reservation price, and answered any questions. Subjects then privately submitted their bids.

After recording their bids, a computer program randomly generated a BDM price for each participant. The program reported this price, and the experiment's outcome, onscreen: buyers traded if their bid exceeded the BDM price, and sellers traded if their bid was less than the BDM price. Finally, all trades were executed. Buyers who traded paid by deducting their purchase

[13] The pattern of reservation prices supports the characterization of these reference prices as medium and high, respectively: about half of the bids exceeded $1.49, but nearly all of them (56/60) were below $4.00.
[14] Within each experimental session, everyone was either a buyer or seller, and understood they were transacting with the experimenter rather than with each other. Reference prices were communicated to the subjects privately.
[15] In fact, BDM prices were drawn from a uniform distribution between zero and five dollars.

price from a participation fee; all subjects had been informed of this fee beforehand to eliminate liquidity concerns.

Results. An ANOVA reveals significant main effects of role (buyers = \$1.35, sellers = \$2.24, $p < .0001$) and reference price (moderate = \$1.34, high = \$2.13, $p < .0001$).[16] Notably, the interaction term is also significant ($p < .03$), indicating that increases in reference price affected sellers more than buyers. This difference is, of course, reflected in the magnitude of the endowment effect. For moderate *r*, selling prices exceeded buying prices by a factor of 1.37 (\$1.58 vs. \$1.15; $t_{59} = -1.87$, $p = .07$ by a two-tailed test). For high *r*, this gap increased to 1.87 (\$2.88 vs. \$1.54; $t_{62} = -4.05$, $p < .0001$).

This study may be viewed as an empirical verification of Thaler's (1985) aforementioned "beer on the beach" thought experiment, with the extension of the study to sellers as well as buyers. As with the beer, the suggested source of the candy (movie theater or big box retailer) is irrelevant to the consumption experience. Nevertheless, transaction disutility causes a distinctive pattern of reservation prices.

Study 2: Manipulating reference prices from low to high

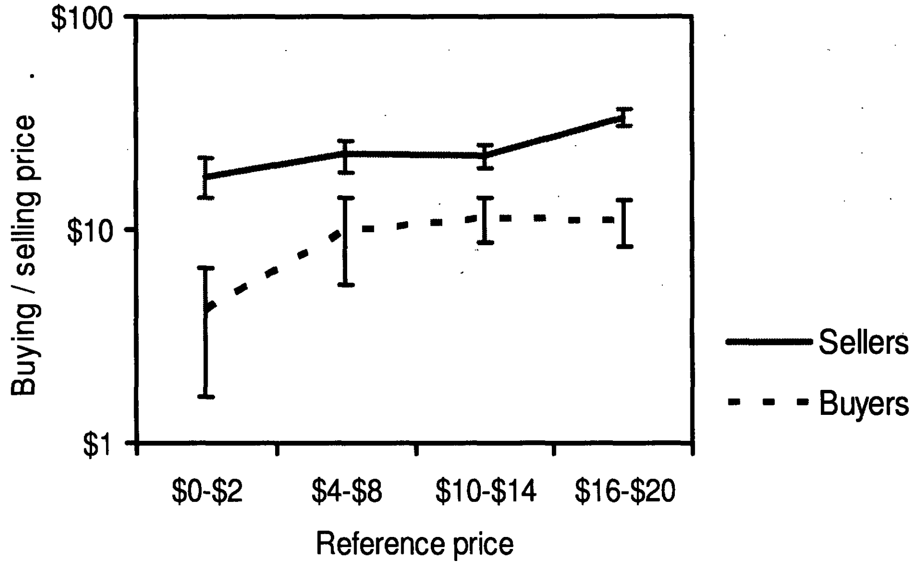
The results of Study 1 are consistent with the transaction disutility model, but to fully test all three predictions developed above, in Study 2 I manipulate the reference price over a broader range, from low to high. Because participants might find claims of a very low market price implausible and therefore reject them, here I manipulate *r* by citing the reservation prices of

[16] For all studies, reported means are based on reservation prices transformed using the Winsor procedure (Barnett and Lewis, 1978). Within each condition, the highest five percent of responses were changed to the 95th percentile response, and the lowest five percent to the 5th percentile. This transformation avoids undue influence by extreme responses without discarding data.

anonymous others. I also use a good whose market price is likely unknown to subjects: a 1925-D buffalo nickel in "very fine" condition.

Method. Participants (N=352) were recruited from an online panel and from spectators at an outdoor concert along the Charles River. They were randomly assigned to the role of either buyer or seller and told, "A randomly chosen person who took this survey before you would [pay at most] [sell for as little as] $*x*." In fact, x was drawn from a uniform distribution that varied from $0 to $20 in two-dollar increments. Sellers were asked to imagine they had won such a coin in a raffle. Participants were encouraged to report their true reservation prices given their current wants, needs, and financial situation.

Results. Figure 3 shows how the manipulations of the reference price affected mean selling and buying prices (top panel) and their resulting ratio (bottom panel). (In this figure, the 11 different reference prices are aggregated into four bands because the number of subjects who received each reference price is small.) These results confirm the model's prediction that the magnitude of the endowment effect is a U-shaped function of r. When $r < v$, buyers drive the price gap, because their reluctance to pay more than the reference price causes them to underbid their valuations. This downward pressure on buying prices subsides as r increases, and buying prices approach valuations for moderate r. When $r > v$, sellers drive the gap, because they are reluctant to sell below the reference price — that is, for high values of r buying prices mirror underlying valuations, but transaction disutility inflates selling prices.



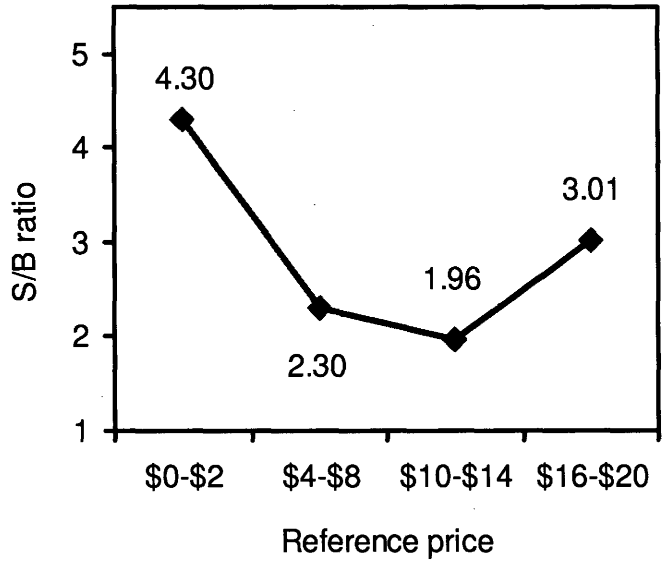


Figure 3. Buying and selling prices (top) and the resulting evaluation disparities (bottom) for the buffalo nickel study. Reservation price's are plotted on a log scale to make it easier to see relative changes as *r* varies.

I conducted two one-way ANOVA tests to confirm the different effects on buying and selling prices as *r* was varied from low-to-moderate vs. moderate-to-high levels. Comparing subjects who received very low (*r* = $0-$2) vs. low/moderate (*r* = $4-$8) reference prices, we find a significant interaction effect ($p < .04$) between role and reference price. This confirms the prediction that r influences buyers more than sellers over this range. For the moderate/high ($r =$ \$10-\$14) vs. very high ($r =$ \$16-\$20) reference price comparison, the interaction is marginally significant ($p < .10$), supporting the prediction that r influences sellers more than buyers when $r > v$.

Note that although Figure 3 measures the size of the endowment effect as the ratio of average selling to average buying prices, ANOVA tests are based on differences, not ratios. In the data reported herein, larger ratios are always accompanied by larger differences. That differences also follow a U-shaped pattern is a strong test of the model, because as r becomes small, the difference between S and B might decrease simply because of a scale effect.

Recall that the transaction utility model implies that if r is varied over a wide enough range, we should observe some value for which the medians of the reservation prices among buyers and sellers will be equal. This is in fact what we find: for $r =$ \$8, the median selling price and median buying price are both \$10 — there is no endowment effect for this moderate reference price. More broadly, replacing means with medians reproduces the U-shaped function of Figure 3: $S_{\text{med}}/B_{\text{med}}$ is 3.00 for $r =$ \$0-\$2, 1.25 for $r =$ \$4-8, 1.50 for $r =$ \$10-\$14, and 2.05 for $r =$ \$16-\$20.

Study 3: Self-generated reference prices

In many studies demonstrating the endowment effect, reference prices are not explicitly provided. In such cases, it is not obvious that transaction disutility underlies the effect. I propose that it does: in the absence of an explicit reference price, consumers spontaneously adopt one — their estimate of the market price, for example — and this reference price influences buying and selling prices according to my model. Some previous research supports this hypothesis. Epley and Gilovich (2001) found that people use self-generated anchors when making estimates. Asked what year Washington was elected President, for example, many people first recalled 1776 (the year the U.S. declared independence) and adjusted upward (see also Brown 2005). In Study 3 I test this prediction by eliciting reservation prices for a pair of vouchers for domestic tickets on American Airlines without providing any reference price.

Method. One hundred twenty-four people waiting along the Boston Esplanade for the city's annual 4th of July fireworks display were given ice cream bars in exchange for completing a short survey. Each participant was randomly assigned to the role of buyer or seller, and was asked his reservation price for a pair of flight vouchers. Participants were later asked to estimate the price the airline would charge for such vouchers. This order ensured that bids would be influenced by estimates of market price only if consumers spontaneously considered them.

Results. The median market price estimate for the vouchers was $300.[17] Using these estimates as proxies for internally generated reference prices, I compared subjects whose estimates are below the median (averaging $214) with those whose estimates exceed the median (averaging $416).[18] Among those whose market price estimates are below the median, the endowment effect is negligible (*S/B* = 1.06), as selling prices scarcely exceed buying prices ($221 vs. $208; $t_{51} = -0.49$, $p = .61$). For the above-the-median group, the endowment effect is substantial (*S/B* = 1.61), as selling prices significantly exceed buying prices ($451 vs. $280; $t_{69} = -3.51$, $p < .001$). Note that in the below-median group, bids are very similar to market price estimates, suggesting that distance is small for these participants, i.e. that reference prices are

[17] Notably, market price estimates were not affected by role (mean of $315 for buyers vs. $344 for sellers, $t_{122} = -1.04$, $p = .30$). This suggests that ownership alone did not change the perceived value of the vouchers.

[18] Estimates of exactly $300 were assigned to the high group, but the results are qualitatively unchanged if the assignment is reversed, or if these people are omitted.

similar to underlying valuations. The transaction disutility model predicts that the endowment effect is minimized under these circumstances, a prediction confirmed by these results. Note that we find the predicted pattern even though this measure of *r* is a noisy proxy for the reference prices participants actually used, which were likely based in part on unobserved factors.

An ANOVA reveals main effects of role (buyers = \$249, sellers = \$353, $p = .003$), reference price group (high = \$365, moderate = \$215, $p < .0001$), and their interaction ($p = .01$): as before, changes in the reference price over this range affected sellers more than buyers. If we had ignored the heterogeneity in reference prices, we would have overlooked an important fact: the endowment effect observed in the group as a whole is driven entirely by the subset of people with above-average reference prices, who (justifiably or not) are reluctant to sell their vouchers for less than they believe American Airlines does.

Study 4: Fans vs. non-fans

The previous studies demonstrate that the endowment effect is increasing in the average distance between consumers' valuations and the reference price. I have demonstrated this relationship by manipulating *r* (Studies 1 and 2) and by exploiting natural heterogeneity in *r* (Study 3). Of course, because distance is a function of both *v* and *r*, the model also predicts changes in the magnitude of the effect if we hold *r* constant and manipulate or exploit natural variations in *v*. I take this approach in Study 4, by comparing the buyer-seller disparity among people with high valuations ("fans") to that for people with low valuations ("non-fans"). If both groups adopt a high reference price, distance will be smaller for the fans, whose valuations are closer to *r*. Therefore, in this case I predict that the endowment effect will be smaller among fans than non-fans. From the perspective of the pain-of-losing account of the endowment effect, this prediction is counterintuitive, as it suggests that people who value a good most highly will experience the most pain from giving it up. However, according to the transaction disutility model, selling prices should be roughly similar for fans and non-fans, because both groups are similarly reluctant to sell below the reference price; but buying prices should reflect underlying values, and thus be higher for fans.

Method. The stimulus good in this study is a Nintendo Wii home video game system—a product for which we can expect considerable heterogeneity in valuations. The participants (134 students who volunteered to complete a paper and pencil questionnaire) were randomly assigned to the role of either buyer or seller. All were truthfully informed, "As a point of reference, the typical street price of a Wii is currently about $350." Subjects gave their maximum buying (or minimum selling) price and answered two questions designed to measure underlying valuations independently of these stated reservation prices: which (if any) video game systems they already own, and the number of hours per week they play video games. Everyone who owns one or more video game systems and reported non-zero playing time was coded as a "fan," by which criterion 49% of respondents are fans of the Wii.[19]

Results. Table 2 shows the mean reservation prices for both groups. The disparity between buying and selling prices is significant among both non-fans ($t_{66} = -6.55$, $p < .0001$) and fans ($t_{64} = -2.39$, $p < .02$). However, as predicted, the endowment effect is larger among non-fans (S/B = 2.28) than among fans (1.28). This difference is confirmed by an ANOVA test, which shows main effects of role (buyers = $159, sellers = $253, $p < .0001$), fan status (non-fans = $187, fans = $225, $p < .01$), and their interaction ($p < .01$). The specific pattern of reservation prices also confirms my predictions: *B* is significantly higher for fans than for non-fans, but *S* is

[19] The results are qualitatively unchanged if we define fans according to either video game system ownership or playing time alone. They are also unchanged if we exclude people who already own a Wii, of whom there were 13 (10% of the sample).

not (t_{65} = -.19, *n.s.*). Contrary to the pain-of-losing account, we find no evidence that fans are more reluctant to part with possessions.

	Non-Fans	**Fans**
Buyers	$110	$201
Sellers	$251	$255
S/B	2.28	1.28

Table 2. Reservation prices for a Nintendo Wii.

DISCUSSION

I have argued that the endowment effect is principally caused by transaction disutility — a reluctance to sell for less than (or pay more than) reference prices. An advantage of this model is that it provides a unified account for many stylized facts that formerly have been treated separately. First, it explains why there is no endowment effect for money or proxies for money such as induced value tokens (Kahneman, Knetsch and Thaler, 1990).[20] To explain this finding, Tversky and Kahneman (1991) and Novemsky & Kahneman (2005) propose that people do not experience loss aversion when exchanging goods "as intended." My explanation, by contrast, follow directly from the distance construct: everyone shares a common valuation and reference price for money — namely, its face value. There is no buyer-seller price disparity because no one is trading on unfavorable terms. Similarly, merchants experience no endowment effect for commercial goods because they sell at the market price almost by definition. Market prices fluctuate, of course, but as Kahneman, Knetsch and Thaler (1986) argue, sellers' reference prices are likely to adjust rapidly. There may be a limit to these adjustments, as when market price

[20] Brown (2005) offers a similar interpretation of the induced value tokens study.

declines below the merchant's cost. The resulting transaction disutility may be one reason that third-party liquidators often assume control of inventories during "going out of business" sales.[21]

Differences in distance also explain some variation in the degree of the endowment effect. For example, we expect an especially large effect for goods that are scarce or extravagant, because market prices exceed (most people's) underlying values by wide margins. This is indeed the typical finding. For example, Carmon and Ariely (2000) observed $S/B = 14.5$ for tickets to the men's Final Four college basketball tournament, for which market prices can reach thousands of dollars.

Horowitz and McConnell's (2002) meta-analysis confirms this pattern: across 201 studies, they found that the endowment effect is smallest for goods that are similar to cash, larger for market goods, and largest for public and non-market goods. This relationship is straightforward to interpret in terms of my model: usually, the less a good is like cash, the higher r is above typical values of v, which inflates selling prices and increases evaluation disparities.

The model may also help reconcile disagreement about whether buyers experience loss aversion for the money they give up in transactions, an idea some reject (Tversky and Kahneman, 1991; Novemsky and Kahneman, 2005) and others support (Bateman et al., 1997; Bateman et al., 2005). My account permits a kind of buyers' loss aversion under particular circumstances: when reference prices are low relative to valuations. This is likely to occur only for goods that many people value in excess of the market price, and hence purchase frequently. Notably, the two examples cited by Bateman and colleagues as evidence for loss aversion in buying, Coca-Cola and chocolates, fit this profile. Conversely, we would not expect loss aversion in buying for goods that few people value enough to buy at the market price, such as university coffee mugs. (I conjecture that many college students buy Coke and chocolate, but

[21] I thank Carey Morewedge for this example.

few buy mugs bearing their school's insignia.) In other words, the question of whether there is "loss aversion in buying" may miss the bigger picture that price gaps are driven by transaction disutility, which is usually manifested as inflated selling prices, but in some cases can also result in deflated buying prices.

Fairness or a "surplus maximization heuristic"?

Either of two distinct psychological processes might cause the transaction disutility that consumers experience. One of these, as I have suggested, is that paying more for or selling for less than the reference price violates people's sense of fairness. Considerable evidence for this explanation comes from Kahneman, Knetsch and Thaler (1986), who explored norms of fairness that can constrain firm profits. One of their key findings is a pervading sense of dual entitlement: the firm is entitled to its reference profit, and the consumer is entitled to the terms of the reference transaction, which terms may be based on market prices, posted prices, or personal transaction history.

Another possibility is that consumers implicitly believe that the reference price is available to them in some outside market. If so, then regardless of one's personal valuation, it is sensible to refuse to trade at a price that is much worse than *r*, as doing so would reduce one's surplus from the transaction. This reasoning is rational under two conditions: that the consumer reasonably believes he can access this supposed outside market, and that the transaction costs of doing so are less than the difference between his own valuation and *r*. There are surely cases in which these conditions are met. Suppose, for example, that you have just bought a new car and want to sell your old one. Although the old car may now be of little value to you — you can't drive two cars at once — it is in your interest to largely ignore this fact, and base your reservation price instead on the Kelley Blue Book price or some similar measure of the car's market value.

Rational surplus maximization, however, cannot explain many observed instances of the endowment effect, particularly when the stakes are small and the transaction costs of accessing the outside market would exceed any potential gain. In Study 1, for example, sellers in the high *r* condition demanded $1.34 more than buyers would pay ($2.88 vs. $1.54) for movie theater candy. It is unlikely that this difference reflects a rational belief on the part of people in the high *r* condition that they should hold out for more because they hold the valuable option of selling their candy to some patron at a movie theater. Moreover, if rational surplus maximization inflates sellers' reservation prices above underlying values, it should similarly inflate *buying* prices, because forward-looking buyers will consider their expected profit from reselling at the reference price. That buyers apparently do not consider such potential resales is perhaps explained by Camerer, Ho and Chong's (2004) cognitive hierarchy model of behavior, according to which people act strategically but have limited foresight. The authors find that across various games, people look ahead only about 1.5 steps on average.

It may be, however, that transaction disutility is the result of a surplus maximization *heuristic*: people bid myopically and as if some outside market offering the reference price is available to them, even when it is not. According to the heuristic explanation, consumers don't suffer any moral outrage from the idea of trading on disadvantageous terms. Rather, they have difficulty distinguishing when the reference price provides valid information about what sort of deal they might strike for the good elsewhere. Ultimately, the fairness and surplus maximization heuristic explanations may be closely related: perhaps paying more than or settling for less than the reference price feels unfair because such trades violate a norm that we are entitled to make the deals that are most profitable to us.

Other explanations for the endowment effect

Several other accounts of evaluation disparities have recently been proposed. I will briefly discuss two of these in light of the proposed account.

Focus on the forgone. Carmon and Ariely (2000) propose that the endowment affect is due to differences in attentional focus: sellers focus on the forgone utility of owning the good (e.g. the lost experience of attending a basketball game), whereas buyers focus on the opportunity costs of the purchase — the forgone benefits of whatever they might have purchased instead. Based on this hypothesis, the authors predict that reference price manipulations will affect buyers (who are focused on the money) more than sellers (who are focused on the good). They found empirical support for this prediction: as they increased the reported face value of Final Four basketball tickets from $15 to $45, buying prices increased from about $65 to about $95, but selling prices remained essentially unchanged at $175.

These results are consistent with the transaction disutility model. Low reference prices ($r < v$) influence buyers more than sellers, and even Carmon and Ariely's highest reference price ($45) was in fact well below median buying prices, and thus below valuations. Importantly, however, the focus on the forgone model does not accommodate my findings, because I have shown that in many cases (and in my view, the more typical scenario), manipulations of r affect sellers more than buyers.

Uncertainty. Several investigators have explored the role of uncertainty in the endowment effect. Simonson and Drolet (2004) suggest that the effect is driven by uncertainty about the desire to trade, particularly among sellers. Owners who are unsure about whether they wish to sell demand more than do those who have no such doubts, widening the gap between buyers and sellers. Okada (2007) attributes the effect to uncertainty about a product's value coupled with "conservatism" on the part of both buyers and sellers. By this account, uncertainty about a product's true value depresses willingness to pay because buyers fear paying more than the product is worth to them, and elevates compensation demanded because sellers fear accepting less than it is worth. For example, a potential iPod buyer might worry the device will skip while she jogs, while an iPod owner might worry that he hasn't fully realized the benefits of downloading podcasts.

In both cases, the authors find that reducing uncertainty shrinks the endowment effect. But they do not explicitly consider the influence of reference prices. It may be, for example, that people who "have decided to sell" a good adopt a much lower *r* (such as zero) than do uncertain sellers. Moreover, in some cases uncertainty in the reference price itself may contribute to the effect. Contrast the results of providing or withholding a reference price in two follow-up studies. For a 20-gallon gasoline card, there was no difference in the endowment effect according to whether or not subjects were given a reference price. For a buffalo nickel, however, giving a reference price made a significant difference, reducing the size of the effect from $S/B = 9.0$ to $S/B = 3.2$. In the first case, explicitly stating *r* made no difference because people already have good knowledge of gasoline prices. But few people know what a buffalo nickel is worth, so informing them of the market price eliminates much uncertainty about *r*.[22]

[22] In a review of endowment effect studies, Sayman and Öncüler (2005) concluded that informing participants about prices has no effect on the size of price gaps. My results qualify that conclusion considerably.

Limits of the transaction disutility model

The results of some studies suggest that transaction disutility alone may not fully account for all observed instances of the endowment effect. One such study was conducted by Knetsch (1989), who randomly endowed half of a group of subjects with coffee mugs and the other half with candy bars, then asked subjects if they would like to trade. Most mug owners preferred to keep their mugs, and most candy owners preferred to keep their candy. It is not obvious that people consider such trades transactions, so it is unclear whether my account explains this reluctance to trade. Perhaps, however, consumers view these trades as transactions whose currency is not money, but rather some other good: a mug owner might (implicitly) view himself as both a potential buyer of candy with "mug money," and the seller of a mug for "candy money." If consumers are uncertain about the relevant reference prices and motivated to avoid regret, they may assign a higher *r* to the good they have been endowed with. Consistent with this explanation, Chapman (1998) found that trade rates were higher for similar goods (two different kinds of candy) than for dissimilar goods. Similar goods may imply similar reference prices, reducing the reluctance to trade.

Strahilevitz and Loewenstein (1998) studied the effect of ownership history on reservation prices. They varied the duration of ownership of stimulus goods, e.g. by endowing some subjects with a coffee mug, then endowing others 20 minutes later. Selling prices increased with ownership duration. Others initially endowed with an object were later made to forfeit it and non-owners with a history of ownership bid more than those with no previous ownership.

Although the transaction disutility model explains the observed evaluation disparities in the data reported here, the studies by Knetsch and Strahilevitz & Loewenstein suggest that in some circumstances, endowment alone does matter. Future research can help determine the degree to which transaction utility vs. the pain of losing possessions creates the endowment effect under various conditions.

Implications for marketing practice

The present research suggests that willingness to pay is driven more by the perceived terms of the transaction than by perceived endowment. Therefore, marketing efforts designed to reassure customers they are getting a good deal may be more effective than those that provide a "sample" of the experience of ownership. Moreover, promotions that instill a sense of ownership, such as free trial periods, can have the perverse effect of suggesting a very low reference price (free!), inducing transaction disutility and making consumers reluctant to buy after the trial period ends. There are of course other benefits (and costs) of these marketing activities, including increasing product awareness. I do not suggest that trial discounts are necessarily bad, only that managers should give the potential negative impact of low perceived reference prices more consideration. The negative side effect of free samples and trials is likely to be strongest when target customers are unfamiliar with the product or the category, since people are especially suggestible when they have little knowledge of the appropriate reference price. Marketers can mitigate this effect by clearly stating the post-trial price, in contrast to the common practice of relegating this information to the "fine print."

For many products, marketers have wide latitude in suggesting relevant reference prices. Firms should discourage the adoption of low reference prices, either by discrediting comparisons to cheaper alternatives, as Porsche attempts with the tagline "there is no substitute"; or by making comparisons difficult, as supermarkets do in creating a "store within a store" for organic foods to physically separate them from their lower-priced conventional counterparts.

Conversely, firms should encourage parallels to other expensive goods. For example, an ad for Celebrity Salons argued that an $80 haircut is a wiser investment than a $1,000 suit, because you wear the suit once a month, but you wear your hair every day.

Firms that operate marketplaces, such as auction websites, make money by matching buyers and sellers. Because transaction disutility reduces trade volume, these firms might increase profits by taking steps to alleviate concerns over potentially bad deals. For example, they might provide data about past transaction prices for similar products. My research also suggests a new product that marketplace operators can profit from: "transaction insurance," which would protect consumers against making particularly egregious deals, similar to the way electronics stores offer low-price matching guarantees.

CONCLUSION

Considerable recent effort has been invested in furthering understanding of loss aversion, of which the endowment effect is one manifestation (Novemsky and Kahneman, 2005; Camerer, 2005; Ariely, Huber, and Wertenbroch, 2005). This research can contribute to this understanding. An aversion to losses requires a reference point against which gains and losses are measured, and although it is well known that reference points are labile, the extent to which the endowment effect depends on external reference prices — not mere ownership status alone — has been understated. Field evidence is instructive: past research has shown reference-dependent preferences in choosing brands (Winer 1986), selling real estate (Genesove and Mayer, 2001), renting apartments (Simonsohn and Loewenstein, 2006), realizing investment losses (Odean 1998), and a general tendency to view transactions in nominal rather than real monetary terms (Shafir, Diamond and Tversky, 1997). All of these examples involve a reluctance to trade *on terms that are unfavorable with respect to reference prices*, not simply a reluctance to part with endowments.

Transaction disutility may also help explain the absence of an endowment effect in some studies (c.f. Plott and Zeiler, 2005). The proposed model implies that the market context consumers adopt when setting their reservation prices affects the size of the gap between buying and selling prices. If circumstances are such that consumers adopt reference prices that are not disadvantageous, or that they disregard reference prices altogether, then there may be no perceived loss from bidding one's valuation. Future research will help determine the conditions under which references prices are given more or less weight, leading to a larger or smaller endowment effect.

REFERENCES

Ariely, D., J. Huber, and K. Wertenbroch (2005). "When do losses loom larger than gains?" *Journal of Marketing Research, 42*, 134-138.

Barnett, V., and T. Lewis (1978). *Outliers in Statistical Data.* New York: Wiley.

Bateman, I.J., D. Kahneman, A. Munro, C. Starmer, and R. Sugden (2005). "Testing competing models of loss aversion: An adversarial collaboration." *Journal of Public Economics, 89*, 1561-1580.

Bateman, I., A. Munro, B. Rhodes, C. Starmer, and R. Sugden (1997). "A test of the theory of reference-dependent preferences." *Quarterly Journal of Economics, 112*, 479-505.

Becker, G.M., M.H. DeGroot, and J. Marschak (1964). "Measuring utility by a single-response sequential method." *Behavioral Science, 9*, 226-232.

Brown, T.C. (2005). "Loss aversion without the endowment effect, and other explanations for the WTA – WTP disparity." *Journal of Economic Behavior and Organization*, 57, 367-379.

Camerer (2005). "Three cheers—psychological, theoretical, empirical—for loss aversion." *Journal of Marketing Research, 42*, 129-133.

Camerer, C.F., T.H. Ho, and J.K. Chong (2004). "A cognitive hierarchy model of games." *Quarterly Journal of Economics, 119,* 861-898.

Carmon, Z., and D. Ariely (2000). "Focusing on the forgone: how value can appear so different to buyers and sellers." *Journal of Consumer Research, 27*, 360-370.

Chapman, G.B. (1998). "Similarity and reluctance to trade." *Journal of Behavioral Decision Making, 11*, 47-58.

Emery, F. (1970). "Some psychological aspects of price." In *Pricing Strategy*, Bernard Taylor and Gordon Wills, eds. Princeton, NJ: Brandon/Systems Press, 98-111.

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

Genesove, D. and C. Mayer (2001). "Loss aversion and seller behavior: Evidence from the housing market." *Quarterly Journal of Economics, 116*, 1233-1260.

Hanemann, W.M. (1991). "Willingness to pay and willingness to accept: How much can they differ?" *American Economic Review, 81*, 635-647.

Horowitz, J.K., and K.E. McConnell (2002). "A review of WTA/WTP studies " *Journal of Environmental Economics and Management, 44*, 426-447.

Kahneman, D., J.L. Knetsch, and R. Thaler (1986). "Fairness as a constraint on profit seeking: Entitlements in the market." *American Economic Review, 76*, 728-741.

Kahneman, D., J.L. Knetsch, and R.H. Thaler (1990). "Experimental tests of the endowment effect and the Coase theorem." *Journal of Political Economy, 98*, 1325-1348.

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

Knetsch, J.L. (1989). "The endowment effect and evidence of nonreversible indifference curves." *American Economic Review, 79,* 1277-1284.

Lichtenstein, D.R., R.G. Netemeyer, and S. Burton (1990). "Distinguishing coupon proneness from value consciousness: An acquisition-transaction utility theory perspective." *Journal of Marketing, 54*, 54-67.

Monroe, K.B. (1973). "Buyers' subjective perceptions of price." *Journal of Marketing Research, 10*, 70-80.

Novemsky, N. and Kahneman, D. (2005). "The boundaries of loss aversion." *Journal of Marketing Research, 42*, 119-128.

Nunes, J. and P. Boatwright (2004). "Incidental prices and their effect on willingness to pay." *Journal of Marketing Research, 41*, 457-466.

Odean (1998). "Are investors reluctant to realize their losses?" *Journal of Finance, 53*, 1775-1798.

Okada, E.M. (2007). "Conservative valuation and WTA versus WTP."

Plott, C.R., and K. Zeiler (2005). "The willingness to pay – willingness to accept gap, the 'endowment effect,' subject misconceptions, and experimental procedures for eliciting valuations." *American Economic Review, 95*, 530-545.

Randall, A., and J.R. Stoll (1980). "Consumer's surplus in commodity space." *American Economic Review, 71*, 449-457.

Sayman, S., and A. Öncülar (2005). "Effect of study design characteristics on the WTA–W TP disparity: A meta analytical framework." *Journal of Economic Psychology, 26*, 289-312.

Shafir, E., P. Diamond, and A. Tversky (1997). "Money illusion." *Quarterly Journal of Economics, 112,* 341-374.

Simonsohn, U., and G. Loewenstein (2006). "Mistake #37: The effect of previously encountered prices on current housing demand." *Economic Journal, 116*, 175-199.

Simonson, I., and A. Drolet (2004). "Anchoring effects on consumers' willingness-to-pay and willingness-to-accept." *Journal of Consumer Research, 31*, 681-690.

Stigler (1966). "The theory of price." New York: Macmillan.

Strahilevitz, M.A., and G. Loewenstein (1998). "The effect of ownership history on the valuation of objects." *Journal of Consumer Research, 25*, 276-289.

Thaler (1980). "Toward a positive theory of consumer choice." *Journal of Economic Behavior and Organization, 1*, 39-60.

Thaler (1985). "Mental accounting and consumer choice." *Marketing Science, 4*, 199-214.

Tversky, A. and D. Kahneman (1991). "Loss aversion in riskless choice: a reference-dependent model." *Quarterly Journal of Economics, 106*, 1039-1061.

Willig, R.D. (1976). "Consumer's surplus without apology." *American Economic Review, 66*, 589-597.

Winer, R. (1986). "A reference price model of brand choice for frequently purchased products," *Journal of Consumer Research, 13*, 250-256.

- Essay 2 -

Preference Reversals Caused by Transaction Disutility

I am grateful to Boris Maciejovsky, Nina Mazar, Daniel Mochon, and Kristina Shampan'ier for their help in collecting experimental data.

ABSTRACT

Reservation prices — the maximum non-owners are willing to pay for goods, or the minimum owners demand to sell them — are commonly believed to reveal consumers' relative preferences among a set of goods: consumers who prefer one good over another presumably will also have a greater reservation price for that good. But reservation prices are influenced by *transaction disutility*: an aversion to bad deals that makes buyers reluctant to pay more than, and sellers reluctant to accept less than, a good's reference price. Alternative measures of preference, such as binary choices between goods, are not regarded by consumers as transactions and therefore are not subject to these distortions. This difference can cause incoherence between *explicit* preferences, and preferences *implied* by stated reservation prices. In three experiments, I find strong evidence for this proposition: manipulations of reference prices have little effect on choice but substantial impact on reservation prices; and preferences can "reverse" when consumers choose a good with a lower reference price over some other good with a higher price. I show that these effects are not caused by quality inferences that participants draw from posted prices, by mere anchoring on reference prices, or by scale compatibility between reference price stimuli and reservation price responses. I also find that preference distortions due to transaction disutility hold not just for binary choice but also preference ranks over many goods, and even when no explicit reference prices are provided.

INTRODUCTION

The theory of the consumer holds that consumers maximize their expected utility by optimizing purchases given a well-ordered preference relation, a set of prices for market goods, and a budget constraint. In other words, a person's purchases (and sales) should reflect his underlying preferences. According to this theory, an individual who prefers one good — considering his current endowment — should not be willing to pay more for a different good. In fact, some theorists have long contended that preferences revealed through transaction behavior are the *only* legitimate basis for inferring what consumers prefer (c.f. Samuelson 1938).

Some evidence, however, suggests that reservation prices — the maximum a non-owner is willing to pay for a good, or the minimum an owner demands to sell it — are influenced by considerations other that the utility consumers expect to get from ownership and use. Thaler (1985) argues that in addition to this *acquisition utility*, consumers also experience *transaction utility* from purchases and sales. Transaction utility is the consequence of considering not only a product's benefits, but also how the terms of the deal stack up against expectations. *Positive* transaction utility results from a better-than-expected deal, and *negative* transaction utility follows from terms that are worse than expected. For example, Thaler found that people who held $10 face tickets to a hockey game they couldn't attend were unwilling to sell at prices that were acceptable to people who held tickets whose face value was $5, but were otherwise identical.

Transaction utility is "real" in the sense that it is a genuine component of consumers' well-being: people derive pleasure from getting good deals, and (psychological) pain from bad ones. But it is sensitive to transient cues, in particular the *reference prices* — posted prices, previously paid prices, etc. — against which actual deal terms are evaluated (Kahneman, Knetsch and

Thaler 1986; Winer 1986). If transaction utility is short-lived, then it is possible that reservation prices reveal preferences which are distorted representations of the utility consumers expect to get. Moreover, if alternative preference measures, such as binary choice between two goods, are not perceived by consumers to be transactions, these measures will not be subject to such distortions. Consequently, we can expect to see incoherence in the preferences indicated by transactional vs. non-transactional measures. In the sense that choice and other non-transactional measures are less vulnerable to fleeting concerns, these measures can be regarded as more faithful indicators of true preference.

Weaver (2008) develops a model of reference prices that assumes that negative transaction utility, or transaction *dis*utility, has much greater influence over judgments and decisions than does positive transaction utility. The theoretical rational for this assumption is loss aversion (Kahneman and Tversky 1979; Tversky and Kahneman 1991): the loss associated with a bad deal is psychologically more prominent than the gain from a good deal. Weaver's experimental evidence supports this assumption.[23] Accordingly, I will propose that preference reversals between *explicit* choices, and the choices *implied* by reservation prices, are caused by transaction disutility — that is, because maximum buying and minimum selling prices can be distorted by a reluctance to trade on unfavorable terms, but choices between goods are not. The model and its predictions are developed in the next section, followed by experimental evidence in support of the model and a concluding section.

[23] Forthcoming evidence presented here is also confirmatory: consumers in Study 2 below appear to be heavily influenced by negative transaction utility, but not at all by positive transaction utility.

THE TRANSACTION DISUTILITY MODEL OF PREFERENCE REVERSALS

Suppose that for any given consumption good, every person has an underlying valuation — an (unobservable) monetary expression of the utility he expects to get from ownership and use. Elicitations of reservation prices attempt to measure valuations, but the reservation prices consumers report are distorted whenever they face a reference price that is unfavorable with respect to their valuations. Here, "unfavorable" means a reference price indicating that trading at one's valuation would be a bad deal. For a buyer (i.e. a non-owner deciding his maximum willingness-to-pay, or WTP), a reference price is unfavorable when it is *less than* his valuation, because bidding his valuation would be overpaying relative to this reference point. The buyer's reluctance to overpay causes him to express a WTP that is less than his underlying valuation. Conversely, for a seller (an owner deciding his minimum willingness-to-accept, or WTA), a reference price is unfavorable when it *exceeds* his valuation, because in this case demanding no more than his valuation would mean receiving less than the good is "worth." The seller's aversion to this bad deal results in a WTA that is greater than his underlying valuation. (Because I assume no positive transaction utility, buyers and sellers simply bid their valuations when the reference prices they face are not unfavorable.)

I assume that choices between goods fully reflect underlying values, and are not distorted by transaction disutility. Therefore, a consumer chooses good A over good B if and only if his valuation for A is higher. The first prediction of this model, then, is that manipulations of reference prices will positively impact stated reservation prices even when they have no influence on choice.

Under specific conditions, transaction disutility also causes preference reversals. Suppose that a consumer values good A more highly than B: $v_A > v_B$. He will therefore choose A over B, regardless of the goods' reference prices. If good A's reference price is also higher, then there should be no reversal; while reservation prices may be distorted, A's will still be higher. But suppose that good B's reference price is higher: $r_B > r_A$. This creates the potential for *differentially* unfavorable reference prices, and consequently for preference reversals. If the consumer is in the role of buyer, he may state $WTP_A < WTP_B$ if v_A is dragged down enough by transaction disutility, which is possible if $r_A < v_A$. Similarly, if he is a seller, he may state $WTA_B > WTA_A$ if v_B is sufficiently inflated, which is possible if $r_B > v_B$. In summary, preference reversals are likely in either of two situations: Buyers are likely to reverse preferences when they choose the good with the lower r and whose valuation for this good exceeds its reference price. Sellers are likely to reverse when they choose the good with the lower r and their valuation for the *alternate* good is *less than* its reference price.

Even under these conditions, preference reversals are not certain. Distortions in valuations of both goods are possible, and the relative distortion in valuations may not be enough to cause preferences to reverse. Moreover, valuations are heterogeneous across consumers, so particular levels of reference prices may induce reversals among some people but not others. Still, conflicts between implicit and explicit preferences should be more frequent under the two conditions specified. We now turn to the experimental evidence collected to test the model.

EXPERIMENTAL EVIDENCE

Study 1: Computer Mouse vs. Trivia Game

In this study, I provided respondents with two different goods and elicited two different measures of preference: explicit choice between the two goods, and reservation prices for both goods — either the most they would pay to purchase, or (conditional on ownership) the least they would demand to sell. The relative reference prices of the two goods were manipulated experimentally to test the proposition that even when choices are not affected by such manipulations, reservation prices are.

Method. The experiment was conducted as a paper-and-pencil survey containing photographs and short descriptions of the two goods: a Logitech computer mouse and a Trivial Pursuit board game. Two groups of participants were recruited. The "buyer" group (N=120) was recruited while waiting for Boston's annual 4th of July fireworks display, and were given ice cream treats for their participation. "Sellers" (N=184) completed the survey in a laboratory, and were paid a flat fee. Incomplete surveys were excluded from analysis. In each group, subjects were randomly assigned to one of two reference price conditions, in which the products' reported retail prices were varied:[24]

M40-T20: The mouse was $40 and the game was $20.

M20-T40: The game was $40 and the game was $40.

Respondents were asked to indicate the good they would prefer to have for free, then to state their reservation price for each good. Buyers stated their maximum willingness to pay (WTP). Sellers were instructed to imagine they had won each good in a raffle, and then state their minimum willingness to accept (WTA).

It is possible that the provided reference prices served as signals about the products' features, popularity or quality (Zeithaml 1988), though the goods were chosen with the intent of minimizing this effect. To test whether subjects made quality inferences from the reference prices, I also asked respondents to rate each product according to how desirable the "average American consumer" would find it, on a scale of 1 (extremely undesirable) to 7 (extremely desirable).

[24] The actual retail price of each good was about $30 on Amazon.com at the time the study was conducted.

Results. Table 1 shows a summary of each group's responses according to reference price condition. In both groups, desirability ratings were not significantly different across conditions for either good. This result suggests that subjects did not make strong quality inferences from the reference prices, but rather that people in different conditions evaluated the "same" products. Explicit preference was also similar in both conditions. In the buyer group, 49% chose the mouse in the M40-T20 condition, vs. 36% in the T40-M20 condition (p = .15). Among sellers, 62% chose the mouse in M40-T20 vs. 55% in T40-M20 (p = .31).[25]

There were, however, large differences in monetary evaluations: in all four cases (two groups × two goods), average reservation prices were higher when the stated retail price for the good was $40 than when it was $20. This finding is consistent with the transaction disutility model. Buyers whose underlying values for either good exceeded $20 were reluctant to pay more than the reference price, reducing average WTP. Similarly, sellers whose underlying values were less than $40 were nevertheless reluctant to accept less, inflating WTA.[26] Consequently, there are significant differences across conditions in implicit preference, i.e. the good for which the subject stated a higher reservation price. The implied preference for the computer mouse among buyers is 73% in M40-T20 vs. 27% in T40-M20 (p < .0001); among sellers it is 94% in M40-T20 vs. 19% in T40-M20 (p < .0001).

This conflict between explicit and implicit choice causes a predictable pattern of preference reversals, as shown in Figure 1. When the computer mouse's reference price was higher (M40-T20 condition), very few of the people who explicitly chose the *mouse* stated a higher

[25] The overall greater preference for the computer mouse in the seller group likely reflects differences in the population from which this sample was selected rather than differences in the buying and selling task, as choice was elicited before roles were revealed.

[26] The low reservation prices in the seller group when reference prices were $20 ($13.35 for the mouse, and $13.33 for the trivia game) suggest that many of their underlying values were less than the low reference prices as well. But the gap between reference price and underlying value is greater for the high reference price good, resulting in greater distortion and WTA disparities. Conversely, very few buyers' WTP exceeded $40, so there is little distortion of buyers' valuations for either good when the reference price is high.

reservation price for the game: 12% in the buyer group and 3% in the seller group. But a majority of people who chose the *game* had a higher reservation price for the mouse: 58% among buyers, and 89% among sellers.[27] The pattern is analogous in the T40-M20 condition: very few preference reversals among trivia choosers, and many among mouse choosers. The differences in preference reversal rates are significant in all four cases ($p \leq .005$ by $\chi^2(1)$ tests). Also, in separate logit models for each group, the interaction effects between reference price condition and explicit choice are highly significant ($p < .0001$).

Discussion. These results support the idea that while the choice between two goods is a "pure" measure of preference — that is, it faithfully reflects differences in underlying expected utility — reservation prices are potentially distorted measures. The degree of distortion depends on how "unfavorable" the reference price is with respect to underlying value. For example, as Figure 1 shows, the preference reversal rates are higher among sellers than among buyers, reflecting the fact that most underlying values are substantially less than $40, but relatively few are much more than $20.

The correlations between explicit choice, implicit choice, and reference price provide additional insight into the basis stated reservation prices (Table 2). As expected, choice is essentially uncorrelated with a good's reported retail price. But stated reservation prices are correlated with the reference price ($\rho = .45$ for buyers, $\rho = .74$, both using Spearman rank correlations). Moreover, these correlations are higher than those between reservation price and explicit choice ($\rho = .43$ and $\rho = .31$, respectively). These results suggest that reservation prices are informed both by underlying preference and by reference prices, with the latter carrying at least as much weight, and possibly more.

[27] Subjects who stated equal reservation prices for both goods were excluded from this analysis, but the results are qualitatively unchanged if these are included and counted as non-reversals.

Study 2: Vacation Upgrades

While the results of Study 1 are consistent with the transaction disutility model, two related alternative explanations predict the same overall pattern. The first is anchoring (Tversky and Kahneman 1974): perhaps respondents anchored their reservation prices on the goods' stated retail prices. Strong anchoring effects usually require an explicit comparison between the anchor and the target judgment, e.g., "Would you pay more or less than *x*?" (Brewer and Chapman 2002). I did not ask for such a comparison, but it is possible that respondents made it spontaneously (see, for example, Epley and Gilovich 2001).

A second alternate explanation for the results of Study 1 is the compatibility hypothesis (Slovic, Griffin and Tversky 1990). According to this hypothesis, a stimulus attribute is given greater weight in judgments for which the response shares that attribute's scale. Compatibility has been invoked frequently to account for various kinds of preference reversals. For example, Slovic and Lichtenstein (1971) found that when evaluating risky gambles with similar expected values, many people chose gambles offering a high probability of a relatively low payoff, but demanded more to sell gambles offering a low probability of a much higher payoff. They explained this pattern in terms of compatibility: like the gambles' payoffs, minimum selling prices were expressed in dollars. It is possible that stimulus-response compatibility might have encouraged the participants of Study 1 to weigh reference prices heavily when stating their reservation prices, but not when making binary choices.

The anchoring and response mode compatibility explanations both predict the key results of the mouse vs. trivia game experiment: no effect of reference price on explicit choice, a main effect of reference price on WTP and WTA, and a high probability of a preference reversal among respondents who chose the good with the lower reference price. Study 2 was designed to rule out these alternative models.

Like Study 1, Study 2 uses two stimulus goods. However, in this case all participants received the same reference prices, which were purposefully set high — in excess of most consumers' likely valuations. The role assumed for stating a reservation price was varied experimentally between subjects: either buyer or seller. Both the anchoring and response mode compatibility explanations predict results similar to those found in Study 1, in particular, frequent preference reversals among people who choose the good with the lower reference price. But my model predicts widespread preferences *only among sellers.* Buyers will experience no transaction disutility because their valuations are below the reference price. Therefore, I predict few preference reversals among buyers, regardless of which upgrade they explicitly prefer.

Method. Volunteers (N=110) were solicited from an online community to complete a short web-based survey. They were instructed, "Suppose you are making hotel and flight arrangements for a Caribbean vacation, and are considering upgrading your flight or your hotel room." Two upgrades were presented: a beachfront hotel room, with a "current price: $900 total for 6 nights"; and an upgrade to business class, with a "current price: $600 round trip." [28] Next, subjects assigned to the buyer role were asked to state the most they would pay for each upgrade. Sellers were told, "Imagine you've been given the upgrade for free. What's the least you would sell it for?"

Results and Discussion. The average reservation prices for the two goods, and proportion of people who preferred the hotel upgrade, are shown by experimental condition in Table 3. First, we see that for both goods, owners demanded substantially more (2.5 to 3.2 times as much) to

[28] The order in which the two goods were presented was counterbalanced (with the higher reference price always associated with the hotel room upgrade), but there were no differences according to presentation order. In the results reported, data are collapsed accordingly.

sell than buyers were willing to pay ($p < .0001$). This "endowment effect" is typical (Thaler 1980; Kahneman, Knetsch and Thaler 1990). Note that reservation prices for both goods are well below their respective "current prices," indicating that, as intended, reference prices are considerably higher than underlying valuations.

A large majority of respondents preferred to get the beachfront room upgrade for free; there was no difference among buyers vs. sellers (86% vs. 78%, $p = .28$). There was also no difference in implied preference across condition: 87% of buyers and 88% of sellers stated a higher reservation price for the hotel upgrade ($p = .84$), a finding to which I will return momentarily.

Figure 2 shows the rate of preference reversals by condition and explicit preference: the percentage of participants who explicitly chose one upgrade but stated a higher reservation price for the other.[29] Among sellers, the results closely parallel those of Study 1: almost none (3%) of the hotel choosers had inconsistent preferences, but fully half (50%) of the flight choosers did ($\chi^2(1) = 10.74$, $p < .0001$). Many sellers who chose the business class upgrade demanded more for the hotel upgrade because its reference price was $300 higher. Among buyers, however, *none* of the subjects' preferences were in conflict — neither hotel nor flight choosers'.[30] This result is predicted by the transaction disutility model, but is inconsistent with the anchoring and response mode compatibility explanations. Because consumer valuations tended to be well below reference prices in this experiment, WTPs were not distorted, and faithfully reflected underlying preferences.

[29] As in Study 1, subjects whose reservation prices for the two goods are equal were excluded from the analysis, but the results are qualitatively unchanged if these subjects are counted as having consistent preferences.

[30] There is clearly an interaction effect between explicit choice and role, but because of the absence of any preference reversals among buyers, it was not possible to estimate a logit model.

As noted above, we observe no difference in implicit preference between buyers and sellers. This result is not surprising. In the case of buyers, implied preferences should (and do) mirror explicit choices because of the absence of transaction disutility. Among sellers, we expect some flight upgrade choosers to demand more for the hotel upgrade, increasing the implied preference for the latter. We do in fact observe this (88% implicit vs. 78% explicit preference for the hotel), but since most people chose the beachfront room upgrade, there is probably a ceiling effect.

In summary, the results of Study 2 are consistent with the transaction disutility model and counter to the competing explanations of response mode compatibility and anchoring. A third study explores the robustness of the model.

Study 3: Four Goods

Study 3 presents participants with a somewhat different and more difficult task: evaluation of four goods rather than two. I manipulated the reference prices of these goods between subjects, and elicited preference ranks, desirability ratings, and reservation prices. Like choosing between two goods, ranking four goods by preference should not invoke transaction disutility. Therefore, I predict that average ranks and desirability ratings will be similar across conditions, whereas reservation prices will vary with reference prices.

I also included two novel reference price conditions, to test additional implications of the model. In one condition, all four goods bore the same retail price. This should eliminate situations where there is enough differential transaction disutility that consumers will pay more for a lesser-preferred good, so we should see fewer preference reversals. In another condition, I provided no reference prices, but asked respondents to estimate each good's retail price. Weaver (2008) shows that in the absence of explicitly provided reference prices, consumers spontaneously generate their own estimates based on their beliefs about market prices, and use these estimates to assess whether potential transactions are favorable (see also Brown 2005). If so, subjects' retail price estimates should predict reservation prices and preference reversals in the same way that explicit reference prices do.

Method. Respondents were given a survey containing photographs and descriptions of four goods: a pair of Motorola walkie-talkie radios, a Leatherman utility tool, a Trivial Pursuit board game, and a six-month subscription to People magazine. They were asked to rank the products from 1 to 4 to in order of their preference to receive each for free, to indicate their maximum willingness to pay for each, and to rate each from 1 to 7 according to how desirable they guessed the average American consumer would consider it to be. As explained above, participants in one condition also estimated the products' retail prices.

Participants were students at Princeton and MIT, who used paper-and-pencil and were paid a flat fee of $2 or $3; and an online panel, who were given a 10% chance to win $20. Forty-five surveys were discarded because of incomplete responses or bad data, almost all of which were rank violations: I required each good to have a unique rank. Five responses (two WTP reports and three price estimates) were significant outliers, and so were trimmed to the maximum value among other respondents in the same condition. A net total of 319 participants were included in the analysis.

Each participant was randomly assigned to one of four conditions, in which I varied the reported retail prices of the goods:

Actual: True retail prices from Amazon.com: radios $39, utility tool $63, game $30, and magazine $57.

Jumbled: Randomly reassigned prices: radios $63, utility tool $57, trivia $39, and magazine $30.

Equal: Each good was reported to have a price of $40.

Elicited: No prices were reported. Instead, participants were asked to guess each good's retail price. These estimates were elicited after WTP judgments, so that they would not influence WTP unless they occurred to participants spontaneously.

Results. Table 4 shows the mean ranks, desirability ratings, and WTP for the four goods. To make it easier to interpret the results, for each product I have arranged the experimental conditions in increasing order of reference price. (For the elicited prices condition, I use and report the average estimated retail prices.) To the extent that reference prices have a positive impact on the dependent variables, then from left to right we will see an increase in desirability, a decrease in average rank (because lower numbers indicate greater preference), and an increase in willingness to pay. The rightmost column shows the results and ANOVA tests for differences across the four conditions. The transaction disutility model predicts that there will be no such trend for desirability or (explicit) rank, but there will be for WTP.

The overall pattern of results strongly supports the model. For three of the four goods, there are no significant differences in desirability rating. The exception is the radios, for which rated desirability is lower in the equal prices condition. But even here, there is no clear pattern of desirability increasing with reference price; and if the outlier were discarded, the differences across the remaining three conditions would not be significant ($p = .29$). As in Study 1, participants do not appear to have inferred product quality from the reference prices they were given. Average preference rank is also generally stable as reference prices are varied. For three of the four goods, there are no significant differences across conditions, and for the utility tool, differences are marginally significant (p = .10). Again, however, there is no apparent pattern of increasing preference for the tool as its stated retail price increases.

Differences in willingness to pay between conditions, however, are highly significant (p < .0001 for all four goods). Specifically, there is a clear trend of increasing WTP as reference prices increase. Roughly speaking, small differences in reference prices cause small differences in WTP, but large "jumps" in the reference price have a large impact. In fact, across all four goods, the correlation between reference price and mean WTP is .72. There is one notable exception to this pattern: WTP for the trivia game is *lowest* in the condition with the highest reference price (equal prices). This result, which is counter to the proposed model, may be due in part to a weak reference price manipulation: the range of retail prices for the trivia game is just $14 (=$40–$26), as opposed to $24 for the radios and $35 for both the utility tool and the magazine.

In any case, despite this one anomaly, the overall picture that emerges is this: as in Study 1, reference prices do not influence either consumers' explicit preferences nor their beliefs about the preferences of others, but lower prices do compel consumers to pay less..

In this study, preference reversals are indicated by a conflict between a consumer's explicit rankings and his reservation prices.[31] Overall, the rate of preference reversals is quite high, reflecting the difficulty of the task — maintaining cohesion among four goods rather than just two as in Study 1. The model predicts that the frequency of reversals will be lowest in the Equal

[31] In cases where a participated stated equal WTPs for two or more of the goods, I judged preferences as cohesive if and only if the goods were ranked consecutively. For example, ranks 1-2-3-4 and respective WTPs $40-$30-$30-$20 were judged cohesive, but ranks 1-2-3-4 and respective WTPs $40-$30-$20-$30 were judged a preference reversal.

prices condition, because there are no price disparities to distort inherent preferences. Therefore, in this condition WTP should be a more "pure" reflection of true preference. This is indeed the case: 74% of participants' preferences are incoherent in the actual prices condition, 71% for jumbled, and 71% for elicited, but just 43% for equal prices ($p \leq .02$ by pairwise Pearson chi-square tests). Note also the high reversal rate for elicited prices: the reference prices that subjects presumably generate themselves without prompting appear to have just as strong an impact on reservation prices as salient, externally-provided retail prices do.

Additional evidence of the mediating role of reference prices in preference reversals can be found by focusing on each participant's most-preferred good, i.e. the product personally ranked #1. Preferences are cohesive if this highest-ranked is also the good for which he expresses the highest willingness-to-pay.[32] The model predicts that for each condition (except Equal, in which there are no price differences), preference reversals will occur *less* frequently when the top-ranked good happens to be relatively expensive, and *more* frequently when the top-ranked good is relatively inexpensive. Figure 3 shows the reversal rates for these top-ranked goods. For each condition, I have arranged the goods are according to their price ranks (e.g., the utility tool is left-most in the actual prices condition, because its retail price of $63 was the highest among all goods in that condition). For the elicited prices condition, the price ranks are defined according to average price estimates; for equal prices, the goods are not ordered because there are no price ranks.

When the goods are ordered this way, the model predicts that for the actual, jumbled, and elicited prices conditions, preference reversals will increase from left to right. We observe this trend in all three conditions: reversals are least frequent for the top-ranked good, and increase

[32] Again, ties were judged cohesive, so technically the test is: preferences were judged cohesive if and only if the participant stated a WTP for no good higher than that for his top-ranked good.

monotonically for all other goods. Note that for elicited prices, the pattern holds despite the fact that price ranks based on average price guesses are a crude proxy for individual-level subjective price ranks, which vary from consumer to consumer. In fact, when we use each participants' individual retail price estimates in the elicited prices condition (as opposed to averages across all participants), perceived price rank becomes a near-perfect predictor of preference reversals. Aggregating across conditions and assigning the rank 2.5 (=[1+2+3+4]/4) to every goods in the equal prices condition, the correlation between the price rank of the top-ranked product and the average reversal rate is .94.

The model also predicts that preference reversal rates will be similar in the equal prices condition, regardless of which good is highest-ranked. Although there are significantly fewer reversals for radios than for the other products in this condition, the range of reversal rates across goods is far smaller than in the other conditions (.32, vs. an average of .72 in the other three conditions).

Discussion. The results of this study confirm that preference ranking is psychologically similar to binary choice, in that it does not induce concerns about transactional concerns. Moreover, subjects' (subsequently) elicited retail price estimates appear to function just as explicit reference prices do: they do not influence ranks or desirability ratings, but do influence WTP and are a strong predictor of preference reversals. This finding provides further evidence that consumers do in fact spontaneously consider expected market prices before transacting. Like Study 2, it also helps reject the response mode compatibility hypothesis, as there were no stated prices to facilitate compatibility.

One surprising result of this experiment, however, is the high percentage of respondents (43%) who showed some incoherence between their implicit and explicit ranks in the equal prices condition. Although this is significantly lower than in any other condition, it is still a non-trivial proportion. As noted above, task complexity may be one reason for this result. But it is also possible that some respondents did not find the claim that all of these products have the same prices to be credible, and so replaced them with their own estimates. Figure 3 provides evidence in favor of this explanation. This figure shows that the patterns of preference reversals are very similar in the equal and elicited price conditions: low when radios are the top-ranked good, and higher for the other goods, whose rates are roughly equal to each other. But the overall reversal rate is lower in the equal prices condition. This is exactly the pattern that would emerge if some respondents adopted the $40 reference price for each good, but others rejected these reference prices and spontaneously estimated their own (as subjects did in the elicited condition).

Aggregating across goods and conditions, I computed pairwise Spearman rank correlations between rank, WTP, and stated retail price: ρ(rank, retail price) = –.09; ρ(WTP, retail price) = .40; ρ(WTP, rank) = –.42 (correlations with ranks are negative because lower numbers indicate greater preference). These results are nearly identical to those from Study 1, again suggesting that reservation prices are about equally informed by true preferences and by reference prices.

CONCLUDING DISCUSSION

The three experiments reported here were designed to test the hypothesis that transaction disutility can create influence monetary measures of value in a way that distorts underlying preferences. Evaluating a computer mouse and a board game, buyers showed reluctance to pay more than reference prices, and sellers were reluctant to pay less than reference prices. But reference prices had very little impact on explicit choice between goods. Consequently, consumers who chose the product with a lower reference price often nevertheless expressed higher maximum buying or minimum selling prices for the alternative product. A subsequent study involving evaluations of four different goods revealed analogous conflicts between stated reservation prices and preference ranks of the goods. Somewhat surprisingly, this study also showed transaction disutility to be just as strong even when no reference prices were provided. In this case, participants appeared to spontaneously estimate the goods' market prices, and their maximum buying prices (but not preference ranks) were influenced in accordance with these estimates.

When different measures of consumers' preferences diverge, it is difficult (impossible?) to say with certainty which measures most faithfully reflect true preferences, or indeed if "true" preferences exist. But in the data presented here, explicit preferences are remarkably insensitive to reference prices, which arguably (aside from their value as quality signals, as discussed further below) *should* have no influence on utility, particularly for products that are not consumed conspicuously. It is therefore reasonable to declare that explicit choices and preference ranks more accurately reflect expected utility than do reservation prices (either buyers' WTP or sellers' WTA). Given this position, it is useful to think of monetary assessments as *constructed* preferences. Correlations between explicit choices, reference prices and reservation prices can then shed light on the nature of these constructions. My data show that, roughly speaking, reservation prices are informed equally by true preferences (as indicated by choice) and reference prices. However, when underlying valuations and unfavorable reference prices diverge sharply — as was the case, for example, among sellers in Study 1 — the distortion of underlying values can be so great that the influence of reference prices dominates.

It should be emphasized that there is no claim that reference prices *cannot* influence expected utility. Prices can serve as useful signals of quality, especially so when quality is uncertain (Bagwell and Riordan 1991).[33] The inferences consumers make from these signals presumably would be reflected in explicit choices and preference rankings.[34] The claim, rather, is that quality inferences do not explain the results presented *here*: that reservation prices vary substantially even when explicit choices and desirability ratings do not, and that preference reversals are far more likely when the consumer chooses one good while facing a significantly unfavorable reference price for an alternative good.

In Weaver (2008), I argued that negative transaction utility (distortion of WTP and WTA due to unfavorable reference prices) is stronger than positive transaction utility (distortion due to favorable reference price). The results reported here provide additional evidence that this is the case. Both kinds of distortions have the potential to cause preference reversals. But in Study 2, in which consumers evaluated two different vacation upgrades, we see evidence only of the former. If consumers experience positive transaction utility, then the resulting distortions in WTP should be enough to cause preference reversals among some buyers. Specifically, some buyers who chose the business class flight upgrade should have been willing to pay more for the beachfront hotel, because its reference price was 50% (and $300) higher. But there are no such preference reversals, confirming that the describing the phenomenon of interest as transaction *dis*utility is indeed appropriate.

[33] In fact, prices can even influence *experienced* utility. Shiv, Carmon and Ariely (2005) found that people solved more mental puzzles when they consumed a full-priced energy drink than when the same energy drink bore a discounted price.

[34] In fact, we see some directional (though non-significant) evidence of prices as signals in our data. In Study 1, for example, slightly more people chose the mouse when its reference price was $40 (55%) than when it was $20 (45%).

Alternative explanations for preference reversals

The aforementioned *compatibility hypothesis* has been proposed as the major reason for a number of classes of preference reversals — not only in the cases of risky gambles as Lichtenstein and Slovic (1971) and Schkade and Johnson (1989) have found, but also for judgments of riskless options (Tversky, Slovic and Kahneman 1990). As Tversky and Thaler (1990) explain, the reasoning behind scale compatibility is two-fold. First, if the stimulus and response do not match, mapping one onto the either requires additional cognitive effort, which may reduce the weight given to the stimulus or introduce error. Second, a stimulus that is compatible with the response mode may focus the respondent's attention on that stimulus. But we find little support for the compatibility hypothesis here. For one thing, in Study 2 *no one* who chose the business class flight was willing to pay more for the beachfront upgrade, despite the fact that the "compatible" stimuli indicated a much higher price for the latter. Moreover, in Study 3 we observe similar patterns of preference reversals *regardless of whether reference prices are explicitly provided* — that is, even in the absence of compatible stimuli. Transaction disutility appears to be a different source of preference conflict than compatibility.

Findings in accordance with the transaction disutility model, however, are reported by Amir, Ariely and Carmon (2008). They show that monetary assessments (i.e. reservation prices) for stimulus goods are poorly correlated with predicted utility (e.g. ratings of expected enjoyment). The authors propose that monetary assessments are informed by transaction cues, whereas predicted utility judgments are informed by experience cues. To the extent that these cues differ, they can lead to discordant judgments. Although this research studies predicted utility and not choice, if we assume that explicit choices are informed primarily by expected utility, these results are broadly consistent with my model. There are, however, important differences. I find that “transaction considerations” (in my case, reference prices) only matter to the extent that they create *unfavorable* comparisons to underlying valuations — high reference prices are largely ignored by buyers, and low reference prices by sellers. Moreover, Amir et al. argue that the relative salience of different cues determines their weight in judgments, but my Study 3 shows that the influence of reference prices is consistently strong even when they are altogether absent, and therefore certainly not salient.

Implications for managers

The experimental results presented here obviously imply that response mode is important in judgments about value and preference. But they also suggest the essential role of context: the value people place on things can depend considerably, sometimes greatly, on circumstance. Suppose we ask a bar patron his reservation price for a mojito. His answer would surely depend on the market environment: the price of a margarita, the going rate for a mojito at nearby bars, the attractiveness of the woman sitting next to him, etc. In an alternate world in which only one bar has access to mint leaves (an essential ingredient for mojitos), many consumers’ reservation prices would exceed the current market price. But in that case, the bar’s market (reference) price for a mojito would also be higher! The marketplace holds so much sway that the very idea of reservation prices may hold little meaning without reference to a particular market. This influence is not limited to market goods. I cannot buy a trip to the moon (yet), but my hypothetical reservation price is influenced by my subjective reference prices for similar experiences: a zero-gravity airplane ride, a hot air balloon trip, a visit to the Kennedy Space Center, whitewater rafting.

Unfortunately, our results suggest that attempts to elicit "context-free" judgments may be quixotic: when no context is provided (in the form of suggested retail prices, for example), consumers appear to adopt whatever context they deem appropriate. Again, this strategy may not be limited to market goods. Rather, it's possible that people spontaneously adopt reference prices as part of the cognitive process to make contingent valuations. Such judgments tend to result in large disparities between buyers and sellers (e.g., Hammack and Brown 1974), suggesting that the reference prices people adopt for public goods and other non-market goods are very high. This is not surprising: the (real or perceived) uniqueness of goods like the Grand Canyon or personal health readily conveys high, even exorbitant, "market" prices.

A very desirable feature of reservation prices is that they provide a continuous measure of preference, such that managers can readily construct demand curves, quantify part-worths in conjoint studies, etc. My results indicate that such measures are capricious, but they also suggest some strategies for more reliably assessing expected utility. A brand manager about to launch a new product might, for example, present test consumers with binary choices between the test product and competing products already on the market, and use the results — the proportion of consumers who express a preference for the novel offering — to determine an appropriate price premium (or discount!) for his new product. Another possibility is to ask consumers for a series of choices between the target good and an increasing number of some other product: would you rather have this good or x widgets? x+1 widgets? etc.[35] The resulting data would still be ordinal, but would be more precise than percentages from one-time binary choices. The potential danger of this approach is that subjects might come to view the widgets as a proxy for money, which could create transaction disutility.

[35] Obviously, this technique would require that the alternate good not satiate rapidly.

In the end, *some* market context will almost certainly inform the judgments of market research participants. The corollary, however, is that the *particular* market context adopted by consumers adopt is fungible. Astute managers can suggest a context in which their own goods appear relatively attractive, increasing their revenue and/or market share.

REFERENCES

Amir, O., D. Ariely, and Z. Carmon (2008). "The dissociation between monetary assessments and predicted utility" *Marketing Science*, in press.

Bagwell, K. and M.H. Riordan (1991). "High and declining prices signal product quality." *American Economic Review, 81*, 224-239.

Brewer, N.T. and G.B. Chapman (2002). "The fragile basic anchoring effect." *Journal of Behavioral Decision Making, 15*, 65-77.

Brown, T.C. (2005). "Loss aversion without the endowment effect, and other explanations for the WTA – WTP disparity." Journal of Economic Behavior and Organization, 57, 367-379.

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

Hammack, J., and G.M. Brown (1974), *Waterfowl and Wetlands: Toward Bioeconomic Analysis*, Baltimore: Johns Hopkins Press for Resources for the Future.

Kahneman, D., J.L. Knetsch, and R. Thaler (1986). "Fairness as a constraint on profit seeking: Entitlements in the market." *American Economic Review, 76*, 728-741.

Kahneman, D., J.L. Knetsch, and R.H. Thaler (1990). "Experimental tests of the endowment effect and the Coase theorem." *Journal of Political Economy, 98*, 1325-1348.

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

Lichtenstein, S., and P. Slovic (1971). "Reversals of preference between bids and choices in gambling decisions." *Journal of Experimental Psychology, 89*, 46-55.

Samuelson, P.A. (1938). "A note on the pure theory of consumer's behaviour." *Economica, 5*, 61-71.

Schkade, D.A. and E.J. Johnson (1989). "Cognitive processes in preference reversals." *Organization Behavior and Human Performance, 44*, 203-231.

Shiv, B., Z. Carmon, and D. Ariely (2005). "Placebo effects of marketing actions: Consumers may get what they pay for." *Journal of Marketing Research, 42*, 383-393.

Slovic, P., D. Griffin, and A. Tversky (1990). "Compatibility effects in judgment and choice." In Hogarth, Robin M., ed., Insights in Decision Making: Theory and Applications. Chicago: The University of Chicago Press.

Thaler, R.H. (1980). "Toward a positive theory of consumer choice." *Journal of Economic Behavior and Organization, 1*, 39-60.

Thaler, R.H. (1985). "Mental accounting and consumer choice." *Marketing Science, 4*, 199-214.

Tversky, A. and D. Kahneman (1974). "Judgment under uncertainty: Heuristics and biases." *Science, 185*, 1124-1131.

Tversky, A. and D. Kahneman (1991). "Loss aversion in riskless choice: a reference-dependent model." *Quarterly Journal of Economics, 106*, 1039-1061.

Tversky, A. P. Slovic, and D. Kahneman (1990). "The causes of preference reversal." *American Economic Review, 80*, 204-217.

Tversky, A. and R.H. Thaler (1990). "Anomalies: Preference reversals." *Journal of Economic Perspectives, 4*, 201-211.

Weaver, R. (2008). "Transaction disutility and the endowment effect." Working paper.

Winer, R. (1986). "A reference price model of brand choice for frequently purchased products," *Journal of Consumer Research, 13*, 250-256.

Zeithaml, V.A. (1988). "Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence." *Journal of Marketing, 52*, 2-22.

TABLES AND FIGURES

MOUSE AND GAME: BUYER GROUP				
		Mouse $40 Game $20	Game $40 Mouse $20	Significance Test
Desirability rating	Computer mouse	4.4	4.2	p = .28
	Trivia game	4.5	4.4	p = .47
Willingness to pay	Computer mouse	$27.22	$20.21	p < .003
	Trivia game	$19.25	$25.93	p < .0004
Preference (mouse %)	Explicit choice	49%	36%	p = .15
	Implied by WTP	73%	27%	p < .0001

MOUSE AND GAME: SELLER GROUP				
		Mouse $40 Game $20	Game $40 Mouse $20	Significance Test
Desirability rating	Computer mouse	4.8	4.5	p = .21
	Trivia game	4.7	4.5	p = .48
Willingness to accept	Computer mouse	$24.02	$13.35	p < .0001
	Trivia game	$13.33	$19.88	p < .0001
Preference (mouse %)	Explicit choice	62%	55%	p = .31
	Implied by WTA	94%	19%	p < .0001

Table 1. Measures of preference for a computer mouse and trivia board game, for buyer and seller groups (Study 1).

MOUSE AND GAME
SPEARMAN RANK CORRELATIONS

	Buyers	Sellers
p(choice, retail price)	.13	.08
p(reservation price, retail price)	.45 *	.74 *
p(reservation price, choice)	.43	.31 *

Table 2. Correlations between measures of value and reference price (Study 1).

VACATION UPGRADES: CHOICE VS. RESERVATION PRICE

		Buyers	**Sellers**	**Significance Test**
Preference (hotel %)	Explicit choice	86%	78%	p = .28
	Implied by reservation price	87%	88%	p = .84
Reservation price	Business class flight upgrade	$106	$343	p < .0001
	Beachfront hotel upgrade	$220	$543	p < .0001

Table 3. Measures of preference for upgrades to a business class flight and a beachfront hotel room (Study 2).

FOUR GOODS: RANK VS. MAXIMUM BUYING PRICE

RADIOS	$39 Actual	$40 Equal	$56 Elicited	$63 Jumbled	Prob > F
Desirability	5.0_{a}	4.0_{b}	4.8_{a}	5.1_{a}	< .0001
Average explicit rank	2.2	2.0	2.0	2.1	.78
WTP	$\$28.65_{a}$	$\$26.40_{a}$	$\$40.41_{b}$	$\$40.28_{b}$	< .0001
Average implicit rank	2.0_{a}	1.7_{a}	1.4_{b}	1.3_{b}	< .0001

UTILITY TOOL	$28 Elicited	$40 Equal	$57 Jumbled	$63 Actual	Prob > F
Desirability	4.1	4.3	4.4	4.1	.43
Average explicit rank	2.8	2.4	2.4	2.6	.10
WTP	$\$18.67_{a}$	$\$16.96_{a}$	$\$28.10_{b}$	$\$28.20_{b}$	< .0001
Average implicit rank	2.7_{a}	2.5_{ab}	2.2_{bc}	2.1_{c}	< .0001

GAME	$26 Elicited	$30 Actual	$39 Jumbled	$40 Equal	Prob > F
Desirability	4.2	4.2	4.2	3.8	.21
Average explicit rank	2.2	2.3	2.5	2.4	.22
WTP	$\$19.95_{ab}$	$\$19.54_{ab}$	$\$21.56_{a}$	$\$16.86_{b}$	.02
Average implicit rank	2.5_{a}	3.1_{c}	2.8_{bc}	2.5_{ab}	< .0001

MAGAZINE	$22 Elicited	$30 Jumbled	$40 Equal	$57 Actual	Prob > F
Desirability	4.6	4.5	4.8	4.4	.46
Average explicit rank	3.0	3.0	3.1	2.9	.65
WTP	$\$12.19_{a}$	$\$13.40_{a}$	$\$10.49_{a}$	$\$22.20_{b}$	< .0001
Average implicit rank	3.4_{ab}	3.7_{a}	3.2_{b}	2.8_{c}	< .0001

Table 4. Measures of preference for walkie-talkie radios, a utility tool, a trivia game, and a magazine subscription (Study 3). For each good, the experimental conditions are ordered differently such that the good's reference prices increase from left to right. In the Elicited reference price condition, the average estimated retail price is listed. The rightmost column of each row shows p-values for ANOVA tests. Within each row, numbers with different alphabetical subscripts are significantly different ($p < .05$ by Tukey HSD tests).

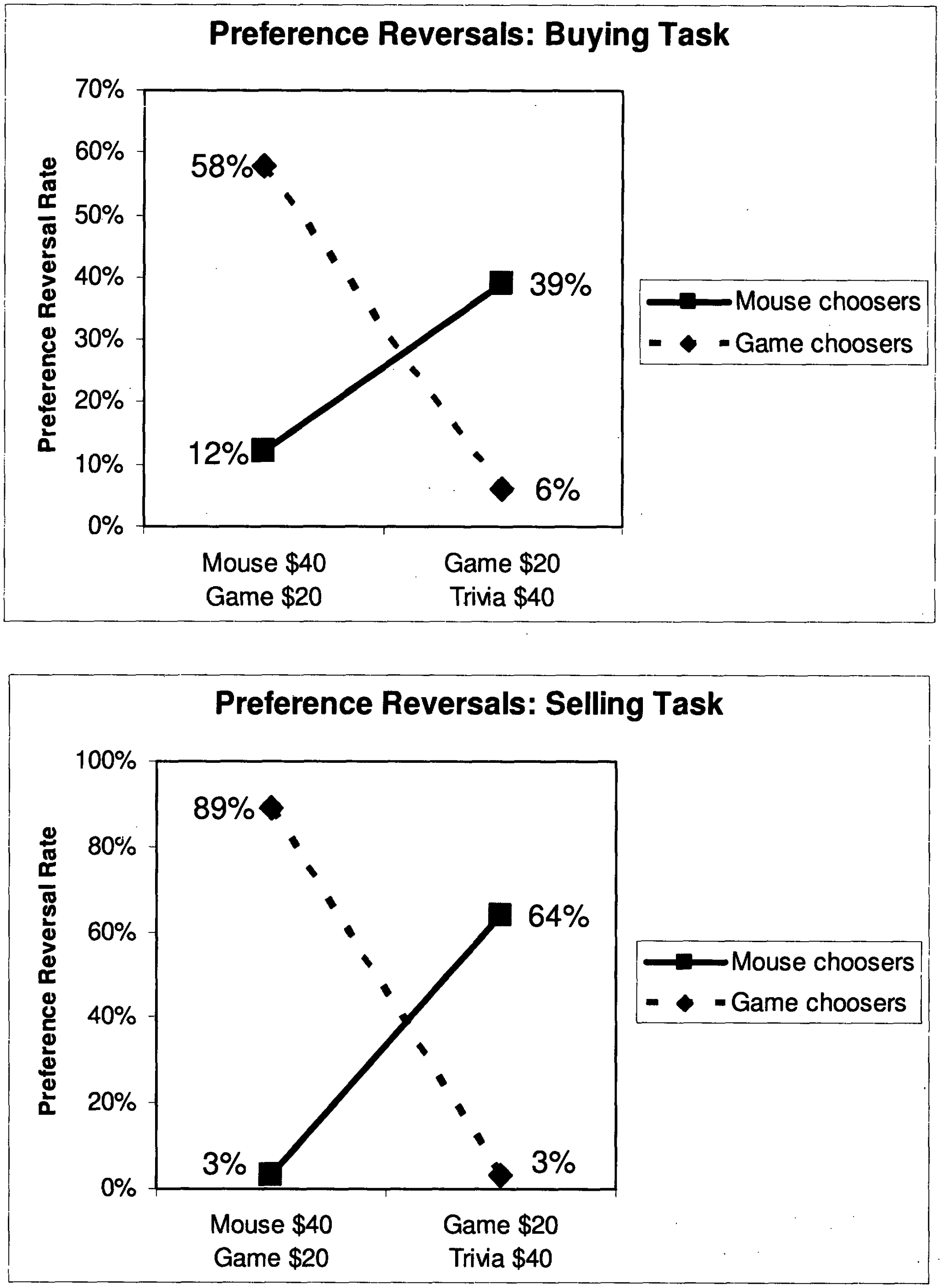


Figure 1. Proportion of subjects whose explicit preferences (choice between the two products) and implicit preferences (as indicated by reservation prices) are in conflict (Study 1).

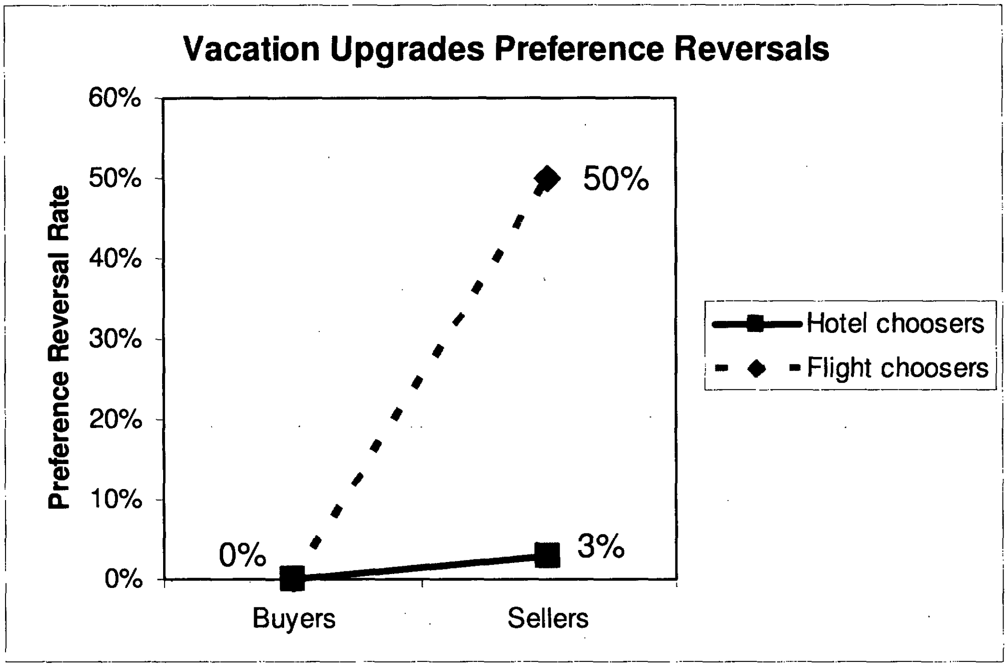


Figure 2. Proportion of subjects whose explicit preferences (choice between the two upgrades) and implicit preferences (as indicated by reservation prices) are in conflict (Study 2).

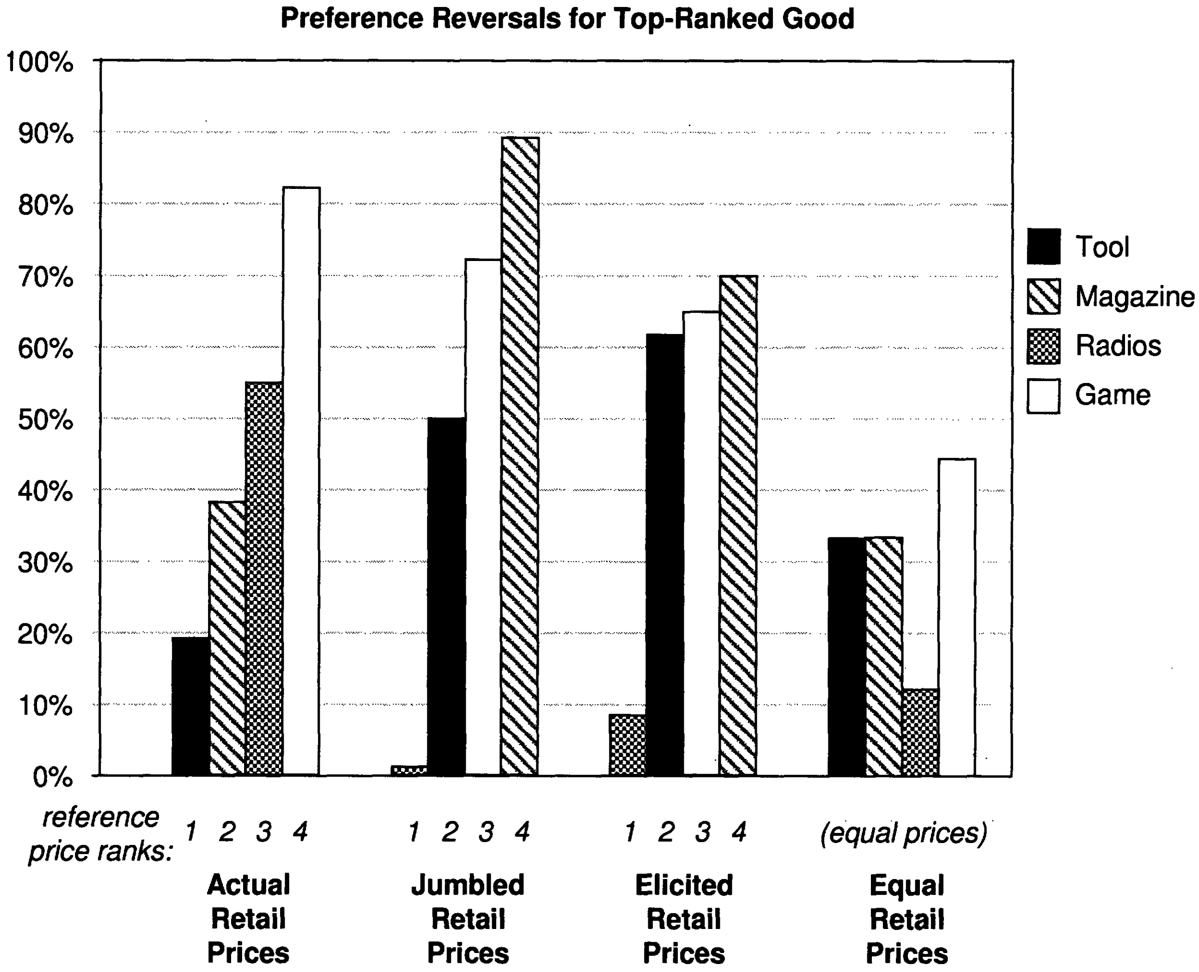


Figure 3. Proportion of subjects whose most-preferred good (i.e. the good they rank #1 to receive for free) are not also the good they would pay the most for, by experimental condition and reference price rank (Study 3).

- Essay 3 -

Using the Bayesian Truth Serum to Obtain and Reward Truthful Answers

I thank Shane Frederick, John Hauser, and especially Drazen Prelec for helpful discussion and comments, and gratefully acknowledge the support of MIT's Center for Innovation in Product Development.

ABSTRACT

The "Bayesian Truth Serum" (BTS) is a survey scoring method that provides truthtelling incentives for respondents answering multiple-choice questions about intrinsically private matters: opinions, tastes, past behavior. The method requires respondents to supply not only their own answers, but also percentage estimates of others' answers. The formula then assigns high scores to answers that are surprisingly common, i.e. whose actual frequency exceeds their predicted frequency. Two studies demonstrate that this method both encourages and rewards truthful responses. First, we simulated various deception strategies and compared them to respondents' actual answers on four surveys with diverse content: personality, humor, attractiveness, and purchase intent. For most respondents, BTS penalized every deception policy we tested. Second, we conducted a general knowledge questionnaire in which we listed items such as brand names, famous people, and scientific terms. One-third of the items were nonexistent foils. When survey takers knew they would be paid for higher BTS scores, they reported recognition of fewer foils, verifying that the method created a credible and persuasive incentive to tell the truth. Moreover, our assertion that truthtelling would be rewarded was shown to be valid: respondents scored higher, and earned more money, when they claimed ignorance of foils.

INTRODUCTION

In opinion research as traditionally conducted, respondents are given no incentives for performance — for the quality or usefulness of their answers. They may, of course, be compensated for time and effort, but the level of compensation does not hinge on the particular answers that they provide. There is, in other words, no hidden answer key by which the survey administrator judges some answers to a question as more worthy of compensation than others.

The reason for this is straightforward: when questions deal with intrinsically private matters — a respondent's opinions, preferences, intentions, or past behaviors — then the correct answer for a particular individual is simply the answer that best matches his private opinions or preferences, and the survey administrator is in no position to judge whether any given answer really does reflect the respondent's tastes. To evaluate answers, the administrator would apparently need to know which answer is personally correct for each respondent. But such an omniscient administrator would not need to conduct a survey in the first place.

In Prelec (2004), one of us proposed a "Bayesian Truth Serum" (BTS) scoring method that provides incentives for providing truthful — in the dual sense of honest and carefully considered — answers to questions dealing with personal information. The key idea behind BTS is to assign a high score to an answer whose actual frequency is greater than its predicted frequency, with predictions drawn from the same population that supplies the answers. To implement this method, the administrator asks each respondent to provide not only a personal answer, but also estimates in percentage terms how other respondents will answer the same question. With this additional input, a numerical answer key can be generated in which the "information score" for any answer is the log-ratio:

$$\textit{Information score for an answer} = \log \frac{\textit{actual relative frequency of the answer}}{\textit{(geometric) mean predicted frequency of the answer}} \qquad (1)$$

Although (1) identifies some answers as "winners" and other answers as "losers" after the survey is analyzed, each respondent has reason to believe that the answer that matches his own private opinion has the best chance of achieving a high score. Specifically, the BTS theorem states that under certain general conditions, personally truthful answers maximize the expected information score for any respondent who believes that others are answering truthfully — i.e., that they are giving truthful answers and optimal Bayesian predictions of the distribution of answers. The scoring system transforms a survey into a competitive, zero-sum contest, in which truthtelling is a strict Bayesian Nash equilibrium (Prelec, 2004). Like other Bayesian mechanisms, BTS exploits the subjective correlation between one's opinion and the opinions of others (Cremer & McLean, 1988; d'Aspremont & Gerard-Varet, 1979; Johnson, Pratt, & Zeckhauser, 1990; McAfee & Reny, 1992; McLean & Postlewaite, 2002; Miller, Resnick, & Zeckhauser, forthcoming). However, unlike previous mechanisms, BTS does not incorporate assumptions about this correlation into the scoring function. Here, the function is generic, not requiring any input from the survey administrator.

We may contrast (1) with consensus scoring, which would assign a high score to the most popular answer. Consensus scoring creates incentives for deception by respondents who suspect that their opinion is in the minority. The information scoring criterion, however, creates no such incentives: untruthful answers have lower expected scores, irrespective of whether the respondent believes that his opinion is common or rare. In this sense, the BTS system levels the playing field between typical and atypical opinions.

Here we provide the first experimental evidence that BTS both rewards and encourages truthtelling. First, we verify that the assumptions on which Prelec's (2004) theorem rest sufficiently hold in empirical settings such that truthful answers outscore deceptive ones. In four surveys, with content chosen to be neutral enough that we can plausibly treat actual answers as truthful, we test whether respondents could have scored higher by engaging in a systematic deception policy, i.e., given non-truthful answers according to some algorithm. For example, we test whether respondents would have achieved higher scores if they had given the answers they believed would be most popular instead of their actual answers. We also examine whether respondents would have done better by misrepresenting their demographic characteristics (gender), or by simulating the answers of some other person whom they "know well." We can find no deception policy that reliably benefits survey takers or any identifiable subgroup of them.

Our second study demonstrates that when incentives to deceive are present, the prospect of payment based on one's information score can create an opposing incentive that is credible and persuasive enough to induce more truthtelling, improving the quality of survey data. We conduct a general knowledge questionnaire in which we ask respondents if they recognize various items: electronics brand names, historical figures, philosophy terms, etc. Paulhus and his colleagues (2003) find that people with a need for self-enhancement tend to overclaim their knowledge on such questionnaires, and because one-third of the items are nonexistent foils, we can measure the degree of deception. We also give some respondents an extra incentive to lie by giving hem an extra payment for each item they claim to recognize. When survey takers know that those with the highest BTS scores will be paid a significant bonus, they claim recognition of fewer of the foils than when bonuses are awarded randomly. We also confirm the findings of our first study:

respondents do in fact achieve higher scores, and earn more money, when they respond truthfully.

We begin by providing a simple example of how the BTS method rewards truthtelling when the truth is unknowable to the administrator, and by highlighting the assumptions on which Prelec's (2004) results relies. Following this, we describe the studies we conducted to verify the validity of these assumptions in practice and to test whether BTS can successfully overcome the temptation to deceive. The final section concludes.

BTS SCORING THEORY

Intuition behind the "surprisingly common" criterion

BTS scoring works at the level of a single question. For example, we might ask: "Imagine that your host offers a glass of red or white wine before dinner. Which would you take: red or white?" To implement BTS scoring, each respondent must both give his own answer and predict the fraction of people who will endorse each answer. *Both* components of his response are scored. Predictions are scored for accuracy — for how well they match the empirical frequencies. Personal answers are scored for being surprisingly common — more common than collectively predicted. An answer endorsed by 40% of respondents would be surprisingly common and would receive a high information score if the group's collective prediction were 20%, but would be surprisingly *un*common and hence low-scoring if predictions averaged 70%.

Prelec (2004) proves that we truthful answers can be expected to be surprisingly common (and therefore high scoring). Why should this be so? Developing our wine example more fully will provide the intuition. Consider "Sarah," an imagined survey respondent asked to choose between Red and White wine before dinner. Suppose that Sarah personally prefers Red, and predicts that a majority of wine drinkers share her preference, as shown in panel (a) of Figure 1.

We can think of these predictions as her estimates of the numerator of equation (1) for the answers Red and White, respectively.

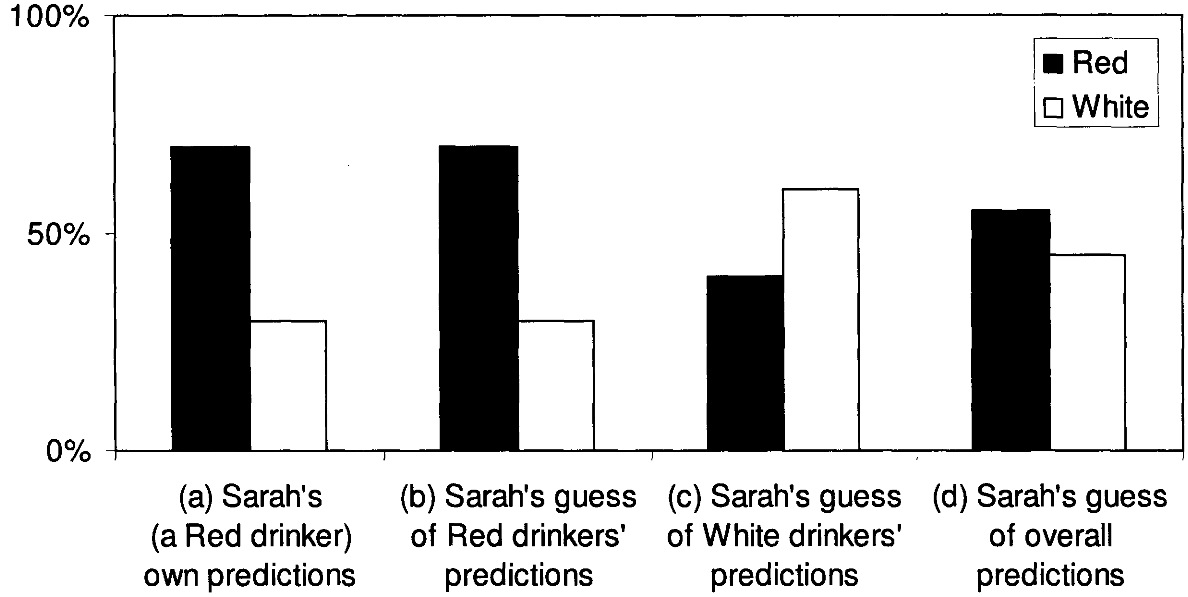


Figure 1: Truthful answers are more likely to be surprisingly common because each person's own preference is a (rational) signal about preferences in general.

To determine which answer is most likely to be surprisingly common, Sarah must also estimate the denominator of (1) — that is, she must predict others' predictions. To construct her estimate, she can usefully divide the survey group into two types of people: those who prefer Red like her, and those who prefer White. Let us suppose that she believes that in general, a person's own preference positively influences his estimate of the overall popularity of that preference. She might then guess that other Red drinkers will have predictions similar to her own, as in panel (b), because their predictions are conditioned on the same private signal — i.e., that they personally prefer Red. She infers that White drinkers, on the other hand, are likely to estimate a smaller degree of preference for Red, because their private signal favors White (panel c). Because the overall group's prediction is a weighted average of the these two types, Sarah estimates that this prediction — the denominator of (1) — will fall somewhere between the predictions made by Red drinkers and those made by White drinkers (panel d). Comparing panels (a) and (d) of the figure, we see that by this logic, Sarah's best guess is that Red — her true preference — will be surprisingly common, and that White will be surprisingly uncommon.

We emphasize three points about this example. First, its logic holds regardless of whether Sarah's opinion is (or she expects it to be) in the majority as in our example, or in the minority. Second, the same reasoning that leads Sarah to expect that her true answer will be surprisingly common applies to White wine drinkers, and more generally to any personally true answer on a multiple choice opinion survey. In *outcome*, of course, both answers will not be correct — either Red or White, but not both, will be more common than predicted. But in *expectation*, one's own preference is his best guess of the most surprisingly common response. Finally, our claim that rewarding surprisingly common answers will reward the truth does not depend on survey takers actually mimicking Sarah's elaborate thought process — of course they don't.

A key assumption that BTS *does* rely on, however, is that respondents, either consciously or unconsciously, use their own tastes as information about the popularity of those tastes among others. The method's validity, then, rests on a testable empirical proposition: that there is a positive relationship between individual opinions and population estimates. In the context of the wine example, the question is whether Red drinkers really do give higher estimates of Red's popularity than do White drinkers, as depicted in the hypothetical data of Figure 1. Fortunately, a great deal of experimental evidence supports this proposition. The seminal experiment was done by Ross, Greene & House (1977), who asked students whether they would be willing to walk around campus wearing a sign that read "Repent." Students who were themselves willing to wear the sign tended to give higher estimates of the proportion of others who would also oblige. This result has been replicated in dozens of studies (Marks & Miller, 1987), but some time passed before its normative status was properly addressed. Ross, Greene & House declared the effect a "false" consensus — an egocentric assumption that others are similar to ourselves. Dawes (1989, 1990), however, argued convincingly in favor of a Bayesian interpretation of the finding: predictions of behavior are correlated with one's own behavior because people rationally update a prior belief based on a "sample of one." Some debate remains about whether experimental subjects use sample information efficiently (Engelmann & Strobel, 2000; Krueger & Clement, 1994), but the evidence is overwhelmingly clear that people who hold a particular opinion or preference give higher than average estimates of the prevalence of that opinion or preference.[36]

The BTS scoring formula

To introduce the formal BTS scoring rule and the highlight the assumptions on which it rests, we begin with some notation. We index respondents by $r \in \{1,2,\ldots,\}$, and their answer and prediction to an m-multiple choice question as x^r and y^r, respectively: $x^r \in \{1,..,m\}$, and $y^r = (y_1^r,.., y_m^r)$ $(y_k^s \geq 0, \sum_k y_k^s = 1)$. (In the informal wine example above, m=2.) We can then calculate the sample frequencies, $\bar{x}_k$, and the (geometric) average of predicted frequencies, $\bar{y}_k$,

$$\begin{aligned} \bar{x}_k &= \frac{1}{n}\sum_{r=1}^{n} I(x^r = k) \\ \log \bar{y}_k &= \frac{1}{n}\sum_{r=1}^{n} \log y_k^r, \end{aligned} \tag{2}$$

[36] Engelman and Strobel (2000) explored whether the use of own opinions to infer population preferences is egocentric — i.e., whether people weight their own opinions more heavily than others'. They first informed participants of the (factual) preferences of a random subset of peers, then elicited predictions for the population. Although estimates were biased in favor of the sample data (a "consensus effect"), participants' private signals had no special status and indeed were often underweighted relative to signals from random, anonymous others (no "false consensus effect").

where *I(·)* is the zero-one indicator function, and *n* is the sample size. Answers are evaluated according to their *information score*, which is the log-ratio of actual $\bar{x}_k$ to predicted $\bar{y}_k$ endorsement frequencies. The total BTS score then combines the information-score with a separate score for the accuracy of predictions:

$$\text{BTS score for } r \equiv u(x^r = j, y^r) = \log\frac{\bar{x}_j}{\bar{y}_j} + \sum_{k=1}^{m} \bar{x}_k \log\frac{y_k^r}{\bar{x}_k} \qquad (3)$$

$$= \text{Information score} + \text{Prediction score.}$$

Prelec (2004) proves that for the game defined by (3), truthtelling is a strict Bayesian Nash equilibrium. We do not repeat the details of this proof here, but rather highlight the assumptions on which it relies, whose validity we aim to verify.

The proof first assumes that the number of survey takers is sufficiently large that each individual response has a negligible impact on the sample frequencies. Then truthtelling is a Nash equilibrium if truthful predictions and truthful answers maximize expected score. In the case of the prediction score component of (3), it is a standard result that truthful predictions are optimal with a log rule (Cooke, 1991).

The result that truthful *answers* are optimal rests on the assumption that people reason like Bayesian statisticians to construct their guesses about the distribution of responses. Let each respondent's truthful answer to a question with *m* possible answers be indexed by a random variable $t^r \in \{1,..,m\}$ (as distinguished from his *reported* answer x^r). The distribution of opinions in the infinite population is given by an *m*-dimensional vector, $\omega = (\omega_1,..,\omega_m) \in \Omega = \Delta^m$. We distinguish between respondents' posterior, which is to say their actual beliefs about this parameter, $p(\omega \mid t^r)$, and the hypothetical common prior distribution $p(\omega)$. We further assume that these densities have three characteristics: *Common prior.* It is common knowledge the posterior beliefs, $p(\omega \mid t^r)$, are consistent with Bayesian updating from a common prior distribution, $p(\omega)$. *Conditional independence.* Opinions are independent, conditional on the actual distribution: $p(t^r=k, t^s=i \mid \omega) = p(t^r=k \mid \omega)\, p(t^s=i \mid \omega)$. *Stochastic relevance.* Respondents with different opinions have different posterior beliefs: $k \neq i$ implies $p(\omega \mid t^r=k) \neq p(\omega \mid t^s = i)$.

Essentially, we assume that in constructing their guesses people begin with some "prior belief" about the distribution of answers to the question under consideration, then update this prior with their own opinion, which they regard as a draw of a ball from an urn containing an unknown mixture ω of differently colored balls. Before drawing the ball, they share the same beliefs $p(\omega)$ about possible mixtures. After drawing, they update these beliefs using Bayes' rule: $p(\omega \mid t^r=k) = p(t^r=k \mid \omega)p(\omega)/p(t^r=k)$, where $p(t^r=k)$ is obtained by taking the expectation of $p(\omega_1,..,\omega_m)$ on the k-th coordinate. Conditional independence implies that respondents who draw the same color also form the same estimate of the proportions in the urn (hence the drawn ball represents the only source of information about these proportions, apart from the common prior). Under these assumptions, the answer that maximizes a respondent's expected information score is his true opinion: $x^r = t^r$.

How strong are these assumptions? First, we note that there is little evidence that a "prior" belief — uninformed by one's own preference — has any psychological reality. Nevertheless, as the empirical literature on false consensus attests, the Bayesian model seems to be a good as-if description of judgment. Second, for the common prior and conditional independence assumptions to hold strictly, all respondents who share a given opinion must make the same predictions (assuming they truthfully report expected frequencies). Of course we don't observe such unanimity in practice. Instead, there is considerable variation in predictions, which can be interpreted as a shared Bayesian posterior among like-minded people, plus a noise term. One purpose of our studies is to test whether information scoring is robust to this noise.

Although there are other equilibria, in which respondents either randomize or give the same answer irrespective of opinion, these equilibria tend to be strategically implausible. For example, if the population contains two subgroups, experts and nonexperts, such that experts know the nonexperts' opinions but not vice versa, then the nonexperts would receive a negative total BTS score in the truthtelling equilibrium, and would consequently prefer the uninformative equilibrium, where everyone chooses answers at random and receives zero total score. However, there is nothing the nonexperts can do strategically to induce the experts to join them in randomizing their responses. In a "struggle of wills," if the nonexperts randomize and the experts respond truthfully, this will only increase the advantage of the experts in total score. Hence, the experts — who prefer the truthtelling equilibrium — have nothing to fear by answering truthfully.

In concluding this introduction to BTS scoring theory, we emphasize that someone who supplies a "losing" answer is not thereby convicted of dishonesty or incompetence. A low score does not automatically indicate deception at the level of an individual answer; rather, BTS scoring penalizes the *intention* to deceive or respond carelessly by linking such intentions with a lower *expected* score.[37]

COMPARING TRUTHTELLING TO SYNTHETIC DECEPTIVE RESPONSE STRATEGIES

Approach

We now turn to our first experimental study, whose objective is to test whether the scoring method does in fact reward truthtelling in the context of real data. The truthtelling theorem relies

[37] The situation is analogous to tests like the SAT, except that for aptitude tests the incentives for truthfulness are self-evident, and the "winning" answer is also defined as objectively correct. For both aptitude tests scored against objective knowledge and opinion surveys scored using BTS, one cannot determine whether a false answer reflects the test-taker's honest opinion or is an instance of deception. What the answer key does ensure is that the test-takers' incentives are properly aligned: giving an answer that one believes is incorrect always reduces expected score.

on sufficient conditions that are unlikely to be satisfied in any actual data set. Conditional independence may fail, because of hidden segments or other reasons. There is little evidence in favor of the common prior assumption; even if this assumption is correct, the literature on heuristics and biases clearly demonstrates that the updating from prior to posterior probabilities does not fully conform to Bayes' rule. And although the theorem requires that respondents who share an opinion will have identical posterior probabilities, reported probability estimates will surely vary. The question we now take up is whether the result that truthful responses outscore deceptive ones holds up in practice, or rather the discrepancies between actual and ideal data can be exploited in some systematic way.

Our approach is this: First we administered a series of opinion and market research surveys to different sample groups. The content of these surveys was sufficiently benign that we could reasonably take the respondents' actual answers as truthful. Next, for each survey we defined a set of deception strategies: policies according to which answers deviate from the actual survey data. We then computed BTS scores for actual responses to the scores that would have been earned under each deception strategy. We assess vulnerability to deception by comparing these results — i.e., by testing whether the respondent group, or some identifiable segment thereof, would have attained a higher score by applying some systematic non-truthful policy. Because truthful predictions are well-known to be optimal with a logarithmic scoring rule, we discuss prediction scores no further and confine our attention to verifying the information scoring component of the BTS scoring rule.

Our assumption that in these data sets actual answers are really truthful is worth examining. Indeed, although the questions in our surveys do not give respondents strong reasons to deceive, it is likely that some fraction of answers do not correspond to true opinions, because of carelessness or self-impression management. Even if so, however, the logic of testing actual responses against deception policies remains valid. Since we do not know which actual answers are truthful and which are not, superimposing a synthetic strategy on actual data will bear equally on both types of answers and will only further reduce the fraction of truthful ones. Essentially, our methodology tests whether additional corrupting of answers that may already be somewhat corrupted reduces information scores.

Content of the four surveys

We conducted four separate surveys with varied content to test the effectiveness of BTS across multiple domains and respondent pools. All surveys were administered with paper and pencil, and the respondents' anonymity was strictly preserved. Brief descriptions of each survey are below, and details summarized in Table 1:

Personality. The survey contained 13 statements from an assessment exercise used by an executive recruitment ("headhunter") firm to evaluate personality traits of job seekers. The items are "difficult" in the sense that the answer potential employers would prefer is not obvious. One statement, for example, asserts: *When I'm under a lot of stress I would rather relax by myself than relax with my family.* Respondents indicated whether they personally agreed with each statement, and also estimated the percentage of their peers who would agree. Survey takers were MBA students who volunteered to participate. They were encouraged to respond truthfully and accurately, and were told that they would receive (anonymous) feedback in the form of a factor analytic interpretation of their answers.

Faces. For the second survey, we showed participants a set of 24 color photographs of young adults (13 women and 11 men). As with the Personality task, they made a binary judgment — in this case, whether they regarded the person in each photo as attractive. However, we extended the survey design in two ways. First, whereas for the first survey all responses were pooled, here we computed information scores separately for men and women to account for the possibility that preferences vary systematically by gender. To facilitate this split, respondents reported their own sex, and gave separate predictions about the percentage of male and female college students who would find each photo attractive. Second, to test the performance of "impersonation" as a deception strategy, we also asked each respondent to consider the tastes of someone else of their choosing, and to rate each photo as if in the shoes and mind of that other person.[38] The participants were MIT students approached in a student center, who were given a $2 coupon for a local convenience store as compensation.

Humor. This survey is identical in design to Faces: two answer choices for each item, segmentation by gender, and the elicitation of impersonated responses in addition to the subjects' own opinions. Again participants were recruited at the MIT student center and compensated with a convenience store coupon. Here, however, we chose a domain with greater ambiguity and variety in tastes. Specifically, we presented a series of 13 "Deep Thoughts by Jack Handey," which are mock-profound observations that originated as a recurring segment on the television show *Saturday Night Live*. Deep Thoughts can be hilarious, offensive, or utterly obscure, depending on one's taste in humor and counter-cultural sophistication. Respondents read the thoughts and judged them as funny or unfunny. They knew that all items were bona fide Deep Thoughts.[39]

Purchase Intentions. This final survey uses content directly relevant to product development: estimates of purchase intent for new products. We presented to participants

[38] To make the impersonation target more concrete, we encouraged subjects to select someone they know well, and asked them to report his or her first name and gender.

[39] An example Deep Thought we used in our survey: *If you go to a costume party at your boss's house, wouldn't you think a good costume would be to dress up like the boss's wife? Trust me, it's not.*

photographs, descriptions, and retail prices of six novel products from a Sharper Image catalog: a portable exercise cycle, a motorized DVD storage tower, electronic drum sticks, a wearable air purifier, and an automatic eyeglass cleaner. As for Faces and Humor, we segmented the respondents by sex. Moreover, whereas previously the response options were binary, here we offered four choices: definitely will buy, probably will buy, probably will not buy, and definitely will not buy. Respondents, who were MBA student volunteers, indicated their own purchase intent and the estimated percentage of their peers who would select each category. Because we segmented the sample pool, it was therefore necessary to elicit *eight* separate predictions for each product: four response options x two genders.

Survey Content	N	Number of Items	Demographic Split	Response Options	Deception Strategies Tested
Personality	104	13	n/a	agree disagree	10
Faces	41	24	Gender	attractive not attractive	18
Humor	46	13	Gender	funny not funny	18
Purchase Intent	106	6	Gender	definitely will buy probably will buy probably will not buy definitely will not buy	12

Table 1: Summary of the design of the four surveys conducted for Study 1.[40]

Development of deception strategies

The space of possible deception strategies is in principle very large; we will evaluate strategies that bear on the following questions:

1. Is the scoring system neutral (as theory predicts) between opinions that are expected to be typical or atypical? Thinking informally, a respondent might be tempted to

[40] The sample sizes reflect the exclusion of participants who did not report their gender or failed to complete large portions of the survey: 16 for Faces, 11 for Humor, and 3 for Purchase Intent.

game the scoring system in one of two ways: either by boosting the numerator of the information score ratio by choosing the most "typical-looking" answers, or by reducing the denominator by choice of unusual, off responses. Neither strategy should pay off.

2. Does the scoring system favor people who believe they are typical or atypical across an entire battery of items? More generally, are there personality traits or demographic characteristics that might justify the use of some deception strategy?
3. Are the penalties for deception "sufficiently large," and do they extend to false reporting of demographic characteristics?
4. Can the scoring system discriminate between authentic personal answers and answers contrived to mimic the opinions of another individual? In other words, can a person expect to achieve a higher score by pretending to be someone else?

Broadly, we test two types of strategies: synthetic deception, in which we transform the survey takers' actual responses according to algorithms designed to mimic various styles of untruthfulness; and directly instructed deception, in which we ask respondents to lie in some specific way — in particular, to impersonate someone else. Some synthetic strategies are very simple, such as *always affirm*: rate all items as funny, attractive, etc. depending on the survey content. Others are more complex functions of the respondents' predictions, like *consensus for other sex*: pretend to be a member of the other sex, and give the answer you expect to be in the majority for that sex. The Appendix describes in more detail all 20 deception strategies we test.

We do not test any strategies generated spontaneously by respondents. The advantage of this approach is that there is no ambiguity about the strategy that is being evaluated. Note also that not all strategies are applicable to all surveys. For example, none of the algorithms that use respondent sex as an input apply to Personality, because we did not collect gender information in that survey.

Results and discussion

Using equation (1), we computed total information scores for all respondents based on their actual responses, and then on the data sets results from the application of the deception policies. We averaged these scores across respondents, and to enable more meaningful comparisons across surveys, we divided these averages by the number of survey questions, yielding average information scores per item. Table 2 summarizes our findings. Taking actual answers as the respondents' true opinions, these results very strongly confirm Prelec's (2004) theorem that information scoring rewards truthtelling over deception. Of the 58 total strategies tested across the four surveys, 53 score significantly lower than actual answers, three are not significantly different, and two score significantly higher. On three of the four surveys, truthful responses outscored every deception strategy we devised.

Strategy	Personality		Faces		Humor		Purchase Intent	
Actual answers	**.17**		**.19**		**.31**		**.42**	
Always affirm (agree, attractive, funny, will buy)	.04	****	-.05	****	**.60**	****	-.70	****
Always reject (disagree, not attr, not funny, won't buy)	.11	****	.03	****	-.19	****	.38	**
Reverse answers	-.02	****	-.21	****	.10	***	-.79	****
Claim to be a member of the other sex			.09	****	.25	**	.36	***
Reverse answers and claim to be other sex			-.20	****	.13	**	-.75	****
Consensus: try to be in the majority	.13	****	.13	****	.17	****	.31	****
Contrarian: try to be in the minority	.03	****	-.12	****	**.31**		-.89	****
Mild consensus: try to ensure mild agreement	.16	*	.17	***	.25	***		
Mild contrarian: try to ensure mild disagreement	.10	****	-.05	****	**.39**			
Mostly consensus, but retain very atypical answers	.12	****						
Mostly contrarian, but retain very typical answers	.11	****						
Consensus for other sex (and claim to be that sex)			.10	****	.08	****	.22	****
Contrarian for other sex			-.18	****	**.36**		-.89	****
Mild consensus for other sex			.09	****	.19	****		
Mild contrarian for other sex			-.14	****	**.39**	*		
Consensus for own sex but claim to be other sex			.10	****	.14	****	.31	****
Consensus for other sex but truthfully report own sex			.14	***	.11	****	.34	***
Impersonate a well-known other			.13	****	.20	**		
Counter-impersonate: reverse impersonated answers			-.17	****	.20	*		
Answer randomly	.07	****	.01	****	.20	**	-.15	****

Table 2: Average information score per survey item across all respondents, for respondents' actual answers and the data sets resulting from various deception strategies. Asterisks indicate results of two-tailed paired t-tests comparing scores under truth (actual answers) to deception: * $p < .05$, ** $p < .01$, *** $p < .001$, and **** $p < .0001$.

On the Humor survey, however, five strategies were more successful (two significantly so) than the respondents' actual answers. As the table shows, these strategies were *always funny* and variations of *contrarian*. Why did these policies outperform actual responses? The explanation is straightforward: it turns out that Deep Thoughts are surprisingly funny as a category. For nearly every Deep Thought — 12 out of 13 among both men and women — more people found it funny than the group collectively expected. Applying these strategies changed many responses from *not funny* to *funny*. This finding, however, does not automatically disprove the truth-rewarding properties of the scoring rule. When a series of questions tap similar content, as is the case here, it is possible that the same answer will emerge as the high-scoring answer for all, or most, of the items. However, to exploit this regularity, respondents would have to know the direction of surprise, and we cannot infer that they are able to do so here. It might have turned out that Deep Thoughts are surprisingly *un*funny, or as in the Faces survey, which has an identical design, that there is no directional regularity to exploit.

In situations like this, when the questionnaire presents a homogeneous set of items, a respondent's opinion combines two stochastic signals. The first is the overall appeal of Deep Thoughts for that respondent, and does not change across the 13 items. The second signal captures how funny he finds a particular Deep Thought relative to the others presented. This signal is resampled for each item.

We can eliminate the effect of the common signal by constraining strategies to the same number of funny ratings per respondent as in the actual data. For example, a *constrained consensus* strategy for a survey taker who rated five Deep Thoughts as funny would first sort the items by predicted popularity (i.e., by the survey takers' predictions of how many others would find them funny), then assign the rating *funny* to the top five. As expected, this quota system eliminates the advantage of the contrarian strategies. Under this constraint, both *consensus* (average information score +.27, $p < .001$) and *contrarian* (+.20, $p < .0001$) fall short of actual scores, and consensus outperforms contrarian, fully replicating the pattern of the other three surveys.

Can some subgroup of people benefit from deception? The analysis so far shows that respondents as a group cannot reliably benefit from deception. But it is possible that averaging across respondents conceals the existence of segments that might profitably deceive. To so profit, people in the relevant segment would need to be able to identify themselves beforehand.

We test two segmentation variables: gender and "subjective typicality," which we will define momentarily.

Conducting the analysis of Table 1 separately for men and women broadly confirms the results of the groups as a whole. Eliminating the Personality survey, for which we did not collect gender data, we tested 48 deception strategies across the three remaining surveys. For men, actual responses outscored deception 43 times (40 of which are significantly higher at $p < .05$ by two-tailed paired t-tests), and deception was higher in only 5 cases (2 significant at $p < .05$). The results are similar among women: actual answers scored higher in 44 comparisons (37 significantly so) vs. 4 cases where deception scored higher (1 significantly so).[41] Moreover, eight of the nine cases in which a deception strategy outperformed actual responses were, as with the overall data set, policies resulting in rating more Deep Thoughts funny on the Humor survey: *always funny* and variants of *contrarian.* For both men and women, actual responses again significantly ($p < .03$) outscore all of these approaches when the number of funny ratings is held constant for the *constrained consensus* and *constrained contrarian* strategies. The only other instance of a successful deception strategy was *always won't buy* on the Purchase Intent survey among men, which yielded a trivial information score improvement of .004 per item over actual responses ($p > .80$). In summary, then, it is very unlikely that respondents could expect to improve over truthful responses by lying in some gender-tailored way.

Another plausible way of discriminating individuals who might gain from deception is subjective typicality: the degree to which a respondent expects his judgments to coincide with others'. This is the proportion of the sample that the respondent expects will share his answer,

[41] We see convincing results among women despite very small sample sizes: just 12 women took the Faces survey, and 14 completed Humor.

averaged across all questions (i.e., it is the average estimated consensus).[42] To test the usefulness of this screening variable, we selected the ten most subjectively typical respondents on each survey and again tested our deception strategies against actual responses, but we found little reason to believe that even these elite few can reliably score higher by lying than telling the truth. Among this subset of respondents, actual answers outscored deception on 46 out of 58 tests across the four surveys, with deception winning 11 times and one tie. Nine of the cases in which deception won were for the Humor survey, and again can be explained by the main effect of switching answers to the surprisingly common (and hence high-scoring) response *funny*. Even when using the consensus strategy, for which the correlation between subjective typicality and the advantage of deception is highest (about .50), these "top ten" respondents *lose* an average of .02 points per item.[43]

Segmenting on gender or subjectively typicality fails for two reasons. First, the discriminatory power of these segmenting variables is weak. Correlations between typicality and the gains from deception over truth, for example, are generally modest, averaging r = .21 in our data. Second, actual answers strongly outperform deception *in general.* Therefore, even if we could reliably identify the individuals with the most to gain, the expected improvement in their information scores would still be little or none, and there is considerable risk of a large penalty. While we cannot rule out the possibility that respondents who might profit from deception have some other way of identifying themselves, it seems that the two segmentation variables available to us — gender and a person's belief about how closely her opinions conform to the group's — provide little useful information.

[42] Consistent with a consensus bias, the mean subjective typicality was .66: respondents predicted that their peers would agree with two-thirds of their judgments. One person predicted a consensus of 94%!

[43] Incidentally, deception fares even worse among the "bottom ten," i.e. the most subjectively *a*typical respondents. For this group, actual answers outscore deception in 52 of 58 cases.

Information score penalty as a function of information loss. Some deception strategies appear ex ante more likely to be successful than others. For example, trying to anticipate and join (or oppose) the majority opinion seems a more promising approach than responding at random. And a few of the strategies we developed, such as generally trying to be contrarian but retaining answers expected to be very typical, are almost bizarre. Despite these differences in face validity, it turns out that all deception strategies are equally bad in the sense that the size of the penalty incurred for lying depends almost entirely on the extent to which truthful responses are changed. The strong linear relationship between the deviation from actual responses and resulting information score is shown in Figure 2. In fact, the correlations are .93, .94, and .98 for the Personality, Faces, and Purchase Intent surveys, respectively. For Humor, this correlation is only .10, reflecting the disruption in scores caused by the fact that nearly all Deep Thoughts are surprisingly funny.

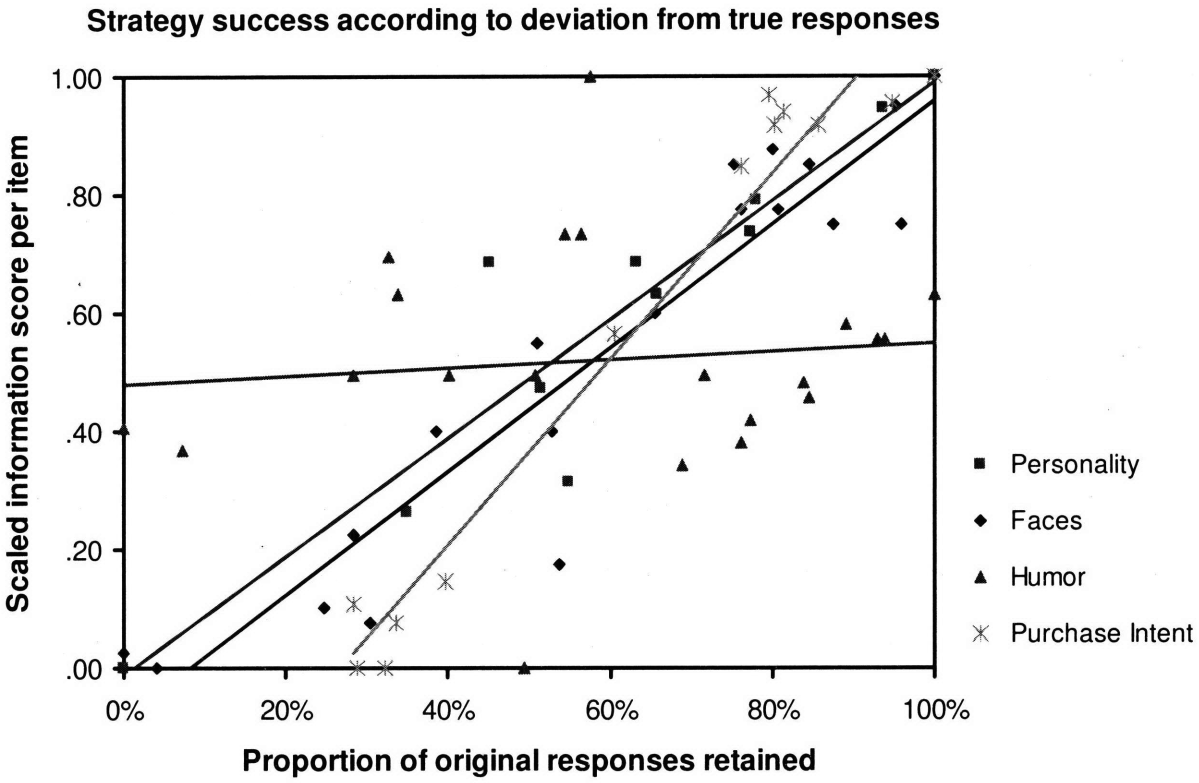


Figure 2: Information score under actual responses and deception strategies as a function of the proportion of original responses retained under each strategy. Within each survey, information scores are scaled such that the best-performing strategy (usually actual responses) is 1, and the worst is 0.

The finding that the reduction in information scores is a linear function of the "information loss" relative to actual responses is striking, particularly when considering the impersonation strategies respondents used on the Faces and Humor surveys. Apparently, we understand our own preferences in a much richer way than those of even well-known others, and our mimicry is somehow one-dimensional in a way that information scoring can easily discern as less than an authentic person.

USING BTS TO ELICIT TRUTHTELLING WHEN THERE ARE INCENTIVES TO DECEIVE

Because respondents had no particular reason to misrepresent their preferences about the benign content of Study 1's surveys, their responses were ostensibly truthful. Under this assumption, we demonstrated that the BTS method functions on real data as Prelec's (2004) theorem predicts: information scores are in fact highest for truthful answers. Two questions remain, however. First, when we elicit information that is by nature subjective, how can we really know when people are telling the truth? To fully establish that BTS rewards honesty, we must be able to distinguish truthful responses from untruthful ones. Second, when the circumstances are such that people *do* have incentives to deceive, is BTS effective at overcoming these incentives and getting respondents to tell the truth? Study 2 addresses these questions.

The over-claiming questionnaire

Over-claiming is the tendency to claim knowledge about or awareness of non-existent items. Phillips and Clancy (1972) developed an index of over-claiming for use in consumer surveys by asking respondents to rate their familiarity with various consumer goods. None of the goods on the survey actually existed, so any claims of familiarity suggested a predisposition towards exaggerating one's knowledge.

The over-claiming technique provides a useful test of the Bayesian Truth Serum because it provides an objective criterion for measuring truthtelling: if survey takers are completely truthful, they should not claim recognition of any bogus items. We emphasize that by "truthful," we mean not only that respondents honestly report their knowledge, but also that they consider their answers thoughtfully and make efforts to avoid careless mistakes. The technique also has a built-in incentive to deceive. Paulhus et a. (2003) showed that individuals with a psychological need for self-enhancement tend to over-claim. We can therefore test whether compensation tied to BTS scores can overcome this tendency.

For our survey we used a subset of the over-claiming questionnaire developed by Paulhus and Bruce (1990), presenting 12 items from each of six categories: arts, historic names, authors and characters, computers and electronics, life sciences, and philosophy. We asked subjects to indicate whether they personally recognized each item, and to estimate the percentage of others taking the questionnaire who would report recognizing it.[44] Within each category, one-third of the items were non-existent foils; subjects were not warned of the presence of foils.

Experimental design and predictions

To test for the effects of BTS-based truthtelling incentives vs. competing incentives, we used a 2 (BTS-based truthtelling incentives: yes, no) × 2 (explicit deception incentives: yes, no) between subjects experimental design. Subjects assigned to the BTS incentives condition were given the following instructions:

> One-third of the participants in your session will receive a $25 bonus. The winners will be the one-third of people with the highest "BTS Score," and will be paid at the end of your session. BTS Scoring is a method recently invented by an MIT professor, and published in the academic journal *Science*. The method scores responses to surveys and questionnaires, similar to the way tests like the SAT are scored. The difference is that while SAT scoring rewards you for choosing the objectively correct answer, there is no "objectively correct" answer on opinion surveys. Instead, BTS scoring rewards you for answering honestly. Even though there is no way for anyone to know if your answers are

[44] In Paulhus and Bruce's original questionnaire, respondents indicate their degree of familiarity with each item on a 7-point scale. We used a binary measure to avoid imposing on subjects the difficult task of estimating the distribution of responses across seven response options.

> truthful — they're your personal opinions and beliefs — your score will be higher on average if you tell the truth. Because only the top one-third of BTS scorers in your session will receive the bonus, you are most likely to earn the $25 bonus if you answer as truthfully as you can. By "truthfully," we mean not only that you are honest, but also that you consider each question thoroughly before deciding your answer, and that you take care to avoid mistakes (such as clicking option A when you meant to click option B).

Subjects in the control (no truthtelling incentives) condition were told:

> One-third of the participants in your session will receive a $25 bonus. The winners will be chosen randomly and paid when the experiment is complete. Please answer as honestly as you can.

Although we expected that some survey takers would be predisposed to self-enhancement and would therefore be motivated to over-claim, we also wanted to introduce an explicit incentive to deceive in order to clearly compare the impact of competing encouragements to respond honestly and dishonestly. Accordingly, we promised some participants an extra payment in proportion to the number of survey items they claimed to recognize. These respondents were given additional instructions that varied depending on the condition to which they were assigned:

> *Respondents in the BTS condition:*
>
> We will pay you an extra 10 cents for each item that you recognize. However, remember that your BTS score will be lower if you do not respond truthfully, so you will be more likely to earn the $25 bonus if you answer as honestly as possible, and only claim recognition of the items you actually recognize.
>
> *Respondents in the control condition:*

> We will pay you an extra 10 cents for each item that you recognize. However, we still want you to answer as honestly as possible, and only claim recognition of the items you actually recognize.

The remaining respondents (i.e., those in the no explicit deception incentives condition) were given no additional instructions.

A total of N=133 people participated in the study, so there were about 33 subjects in each of the four conditions. The study was conducted in four separate sessions with students recruited at Harvard and MIT. The questionnaire was administered by computer, which enabled us to compute the respondents' information scores as they finished the survey, and to pay the subjects according to their experimental conditions at the end of each session as promised.[45] (A random number generator was used to identify the $25 bonus recipients in the control condition.)

If, based on the instructions provided, participants believe that BTS scoring really rewards truthtelling, we expect to them to claim recognition of fewer foils in the BTS condition than in the control condition. The experimental cell in which respondents faced competing incentives — a BTS truthtelling incentive as well as an explicit deception incentive — provides a stronger test. If belief in BTS scoring is strong, respondents should ignore the offer of 10¢ per item recognized and respond honestly. But if respondents are skeptical about our ability to reward honesty, they will reject the BTS incentives and over-claim their recognition of items on the questionnaire.

Results

Truth*-rewarding *property of BTS. With six categories and 12 categories per item, our questionnaire contained 72 total items. One of the authentic items was mislabeled and was

[45] To simplify the real-time computations, we based each respondent's score on information scores computed using the responses of the previous sessions' participants. We used scores computed from a pilot study for the first session. Information scores become more stable as N increases, but in expectation truthtelling is optimal even for very small N.

excluded, leaving 47 reals and 24 foils for analysis. We first test our claim that, under conditions when we can reliably identify which responses are "truthful," that information scores are higher for these responses. Table 3 shows the average information score per item for each response option: *I recognize* or *I do not recognize* the item.

	Control	Control plus 10¢ per item recognized	BTS	BTS plus 10¢ per item recognized
Reals				
I do recognize	+.28	+.38	+.16	+.22
I don't recognize	−.11	−.28	+.08	−.02
Foils				
I do recognize	−.70	+.23	−.99	−.93
I don't recognize	+.21	+.08	+.34	+.27

Table 3: Average information score per survey item, for reals and foils, according to experimental condition.

The most important results are those for the two BTS conditions, because only under BTS incentives does Prelec's (2004) theorem predict that truth will necessarily outscore deception; and for foils, because only for these items is there an objectively truthful response. These results, highlighted in the table, show that the truthful response of *don't recognize* does indeed convincingly outscore *recognize* in both conditions where such honesty is incentive-compatible (BTS: $t_{23} = -7.52, p < .0001$; BTS+10¢: $t_{23} = -6.62, p < .0001$). *Don't recognize* scored higher for 21 of the 24 foils in the BTS condition, and 22 of 24 in the BTS+10¢ condition. These findings are consistent with Study 1: truthful responses are surprisingly common. Moreover, information scores for the two BTS conditions are very similar; evidently, the presence of a countervailing incentive to over-claim does not significantly distort the truth-rewarding property of information scoring.

The remaining information scores are more difficult to interpret. On real items, recognition mildly outscores non-recognition in the BTS conditions (significantly so only for BTS+10¢), suggesting that respondents know somewhat more about the questionnaire topics that the group expected. There is no truth-telling equilibrium in the control conditions. Overall, however, these broader results indicate that the truth-telling rewards are very decisive, as the difference between scores for don't recognize and recognize is largest under the BTS conditions' foil items. The results also show that higher information scores for non-recognition are not simply an artifact of the domain, because most of the other scores deviate from this pattern.

Truth*-inducing *property of BTS. We used signal detection analysis (Swets, 1964) to analyze whether BTS induced survey takers to answer the questionnaire more truthfully. Accordingly, responses fall into one of four categories: *hits*, or claiming to recognize items that exist (i.e.); *misses*, failure to recognize items that exist; *false alarms*, claiming to recognize non-existent items (i.e. foils); and *correct rejections*, claiming to not recognize non-existent items. Table 4 shows the average proportion across subjects of hits and false alarms. It also shows *accuracy*, the degree of discrimination between real and foil items, which we define as hits minus false alarms; and *bias*, a measure of the general tendency to over-claim, defined as hits plus false alarms.

	Control	Control+10¢	BTS	BTS+10¢
Hits	.58	.71	.57	.57
False alarms	.20	.42	.14	.14
Accuracy	.38	.29	.43	.43
Bias	.79	1.12	.71	.72

Table 4: Signal detection analysis results for the over-claiming questionnaire.

Of greatest interest are the false alarms — the tendency to claim knowledge about items that are not real. An ANOVA test shows a main effect of the truthtelling incentive ($F_{1,129} = 24.2, p < .0001$), indicating fewer false alarms in the presence of BTS incentives; and of the deception incentive ($F_{1,129} = 9.4, p < .003$), indicating more false alarms when subjects were paid 10¢ per item recognized. The interaction was also significant ($F_{1,129} = 9.8, p < .003$): the deception incentive had greater influence in the absence of BTS incentives. Moreover, pooling the two BTS conditions and disregarding the Control+10¢ group, we find that subjects claim to recognize fewer foils under BTS as compared to the control condition absent imposed inducements to over-claim ($t_{98} = -2.1$,

$p < .04$). The bias results are similar: an ANOVA test confirms that bias is lower under the BTS truthtelling incentive, higher under the deception incentive; their interaction is again significant.

The results for hits are analogous, but driven entirely by a higher hit rate in the Control+10¢ group than the other groups ($p < .0001$). This is not surprising, as the payment per recognized item creates an incentive to over-claim on reals as well as foils, and of course foils and unfamiliar reals are indistinguishable. Finally, an ANOVA for accuracy shows a main effect of the truthtelling incentive, but not for the deception incentive or the interaction term. Together, these results suggest that respondents were more truthful when facing BTS incentives, even in the presence of competing incentives to over-claim.

Subjects in the BTS condition also spent nearly a minute longer on the questionnaire (7:23 vs. 6:28,

$t_{131} = 2.87, p < .005$), suggesting that they responded more carefully and thoughtfully than did those in the control condition.

Discussion

These results of Study 2 confirm that compensating people according to their BTS scores creates a compelling incentive to respond more truthfully. Evidently, respondents found our explanation of the scoring system credible, even though the idea of a "truth serum" might reasonably arouse skepticism. The signal detection analysis results are virtually identical for the BTS groups with and without the additional payment per recognized item, suggesting that survey takers had enough faith that veracity would be rewarded that they ignored the financial inducement to exaggerate their knowledge. This assessment of the relative incentives is rational for risk neutral people, as forgoing a sure 10¢ per item recognized in hopes of earning a BTS score in the top third generated higher expected earnings. Respondents in the BTS+10¢ group claimed to recognize 30.3 items (presumably reflecting their actual knowledge), earning $3.03 per person on average. Had they claimed to recognize all 72 items, they could have earned $7.20 – $3.03 = $4.17 more. However, in expectation their reward for truthtelling was 1/3 of $25, or $8.33.[46]

It is noteworthy that even in the BTS condition, respondents claimed to recognize 14% of the nonexistent items. This may be a floor effect, and that even when facing incentives survey takers will inevitably make some mistakes, or that larger rewards are necessary to further reduce "recognition" of foils. Another possibility is that these residual false alarms are the result of psychological processes that do not respond to external incentives. Paulhus (1984) distinguishes two types of socially desirable responding: *impression management*, the purposeful manipulation of answers in order to create a positive social image; and *self-deceptive positivity*, which may be unconscious and is used to help maintain self-esteem, optimism, and related personality constructs. Although variations in demand for social desirability (such as an expectation that survey responses will be made public) moderate impression, they have no influence on self-deception. In fact, a subset of people who score high on measures of narcissism appear unfazed by attempts to confront or embarrass them with evidence of their exaggerated self-presentation.

[46] Interestingly, even in the Control+10¢ group people claimed to recognize only 43 items on average, thus forgoing $2.89 (= $7.20 – $4.31) per person despite the absence of any financial incentive to tell the truth. Similar restraint is explored by Mazar, Amir and Ariely (2005), who argue that an internal motivation to perceive oneself as honest limits the exploitation of rewards for dishonesty.

Financial incentives for candor, even when credible and meaningful, may have limited influence on self-deception.

CONCLUDING DISCUSSION

This article has had two primary goals. Our first goal was to empirically verify Prelec's (2004) theoretical result that the Bayesian Truth Serum rewards truthful responses over deceptive responses, despite the fact that for subjective content, the researcher cannot know whether any particular response is truthful or not. Second, we wanted to test whether we could effectively use the truth-rewarding property of BTS to create credible incentives for survey takers to respond more truthfully.

Two experimental studies confirm these propositions. Study 1 comprised four different surveys with varied content and question types. Taking actual responses as truthful, we developed a set of policies according to which people wishing to deceive might deviate from the truth, e.g. always trying to be in the majority. Comparing information scores using actual responses to the those for these various policies, we found that truthful responses reliably outperformed every synthetic deception strategy we tested. Moreover, the outward merits of any particular type of deception were a poor predictor of its performance. Rather, the penalty imposed by the scoring system depended almost entirely on the number of original responses that were changed. Strategies resulting in few changes scored relatively well — though not as well as the truth — whereas strategies that radically changed the original responses scored very poorly.

Study 2 used the over-claiming technique developed by Paulhus et al. (2003) to provide an objective basis for judging whether subjective responses are true. Respondents are asked to indicate whether they are familiar with different items of various types (such as consumer brands), but some of the items on the list aren't real. Therefore, any claims of familiarity with such foils are cannot be true. We gave a familiarity survey to a group of people after explaining the BTS scoring system and informing them that the top one-third of scorers would be paid a $25 bonus. This group claimed to recognize fewer foils than did people in a control group who were ignorant of BTS and whose bonus recipients were determined randomly. Moreover, our explanation of a "truth serum" scoring system was credible enough that participants ignored specific incentives to deceive. Half of the respondents were given an extra ten cents for each item recognized. In the control payment condition, these respondents claimed to recognize many more foils (as well as reals). But in the BTS payment group, people apparently rejected the small sure gain from exaggerating their knowledge in favor of the larger expected gain from responding truthfully.

We conclude with a discussion of the ability of BTS to reduce different kinds of deviation from the truth, and of some issues of practical importance when considering using the method.

Different Types of Untruthfulness

Broadly, we can divide untrue responses into three categories: intentional deception; carelessness; and inauthentic responding, where, for reasons that may not be fully conscious, a respondent gives answers that are biased by social norms or the opinions of others. It is useful to consider the effect of BTS on these different types of deviating from the truth.

It is clear from our results that the truth serum method *penalizes* all three kinds of untruthfulness, which were represented in Study 1 by various synthetic strategies. Intent to deceive was simulated by such strategies as reversing truthful responses and misrepresenting gender. Carelessness was represented by randomized responses.[47] Inauthentic response strategies included trying to join the majority (or minority) and directed impersonation — policies that mimic biases that might degrade truthfulness. As we have seen, all of these departures from truthfulness reduced information scores, and were equally bad in the sense that the reduction in score was a linear function of the number of responses changed, without regard for the style of untruthfulness that led to the change.

Although we also found convincing evidence that BTS *motivates* respondents to tell the truth, whether it can reduce all three types of untruthfulness is less clear. The results of Study 2 show that BTS can certainly reduce intentional deception. When participants were paid to exaggerate their knowledge (10¢ per item recognized), they claimed knowledge of two-thirds fewer foil items when also given BTS incentives as when not (3.4 vs. 10.1 foils out of 24 total on the questionnaire). They also claimed to recognize fewer real items. Disentangling carelessness and inauthenticity, however, is harder. Subjects in the control condition with no financial incentive to deceive claimed to recognize 20% of the foils. BTS reduced this rate by about one-third (to 14%). But we do not know if this improvement was due to greater care, a suppression of the (unconscious?) desire for self-enhancement, or some combination of the two. The fact that people in the BTS condition spent more time on the survey, however, suggests that at least some of the improvement results from more careful responding.

Practical Considerations

To successfully implement BTS, two conditions must be met. First, skeptical respondents must be convinced that the method really can reward the truth. This does not mean that researchers

[47] Randomization also captures the notion of stereotyped responding, where the same answer is given across a block of questions, except that in this case the randomizing "coin-toss" is performed only once, at the start of the block.

should necessarily explain the specific scoring formula or the intuition behind it; indeed, providing these details may be counter-productive, because respondents might be confused by them or falsely conclude they can devise a strategy to beat the system. Our explanation emphasized the reputation of the source (an MIT professor, whose research was published in the journal *Science*) and drew an analogy to scoring for tests of objective knowledge. Debriefing interviews suggested that respondents found our claim credible, but other instructions may be equally or more effective. Future research can help determine this. Success also requires that BTS-based incentives be large enough to overcome any competing incentives to dissemble. If the survey questions are sensitive — about drug use or sexual behavior, for example — respondents may prefer to forgo a small financial reward rather than reveal socially stigmatized conduct. It is also possible that at least some sources of inauthenticity are fully unconscious and stubbornly resistant to even large financial rewards for telling the truth.

We have not dealt here with the broader issue of when it is appropriate to introduce performance-based incentives. We take it as uncontroversial that incentives may be useful in some circumstances and counterproductive in others (Camerer & Hogarth, 1999). Because scoring transforms a survey of opinions into something that feels like a test of knowledge, it fundamentally changes the relationship between the respondent and survey sponsor. The sponsor, for example, can ask respondents to prepare in advance, such as by trying out a product or service relevant to the questionnaire. In that case, respondents who do their homework have a better chance of doing well in the survey, just as they would on an SAT test. There are also other potential advantages of scoring. Competition creates reputational stakes that can spice up an otherwise dull survey experience; scores can be used to filter more careful or thoughtful respondents, who can then be retained for future studies; scores can also function as performance feedback, teaching respondents how to provide better information. Collectively, the benefits of the Bayesian Truth Serum are substantial and varied enough to make the method a useful tool in many circumstances when a researcher wants to learn about a target group's opinions or beliefs.

APPENDIX: DECEPTION STRATEGIES USED IN STUDY 1

The deception strategies used to test the ability of information scoring to reward truthtelling in the four surveys of Study 1 are described below. Note that in all cases, we only transformed the survey takers' judgments about the statements, not their estimates of the fraction of people endorsing each answer. With the sample sizes in our studies, changing a single respondent's predictions has a negligible impact on overall scores.

Always affirm: Affirmative response for all items: *agree* (Personality), *attractive* (Faces), *funny* (Humor), and *will buy* (Purchase Intent). Because Purchase Intent has two assenting responses — *probably will buy* and *definitely will buy* — we took the average of the information scores for these two answers.

Always reject: Negative response for all items: *disagree*, *not attractive*, *not funny*, and *will not buy*. Again, we averaged Purchase Intent's two dissenting responses.

Reverse answers: The opposite of each actual answer. For Purchase Intent, which has four possible responses, we changed actual answers to their "mirror image," e.g. *probably will buy* became *probably will not buy*.

Claim to be a member of the other sex: Retain actual judgments, but apply information scores for the subject's opposite sex.

Reverse answers and claim to be other sex: Apply both strategies 3 and 4.

Consensus: Change actual answers to the response the survey taker expects to be in the majority (or, for Purchase Intent, the expected mode) based on his predictions. For the three binary response choice surveys, ties — predictions that exactly half the group would give each response — were handled by retaining the original response. For Purchase Intent ties (multimodal predictions), we averaged the information scores of the modes.

Contrarian: Changed actual answers to the expected minority response. Ties were dealt with in the same way as for consensus.

Mild consensus: Similar to strategy 6, but only change actual answers when the survey taker predicts that less than 30% of others will agree with him. For example, if a person's actual response is *funny* and prediction is that 20% of others will agree, his response becomes *not funny*.

Mild contrarian: Similar to strategy 7, but only change actual answers when the survey taker predicts that more than 70% of others will agree with him.

Mostly consensus, but retain very atypical answers: If the survey taker expects 30-50% of others to agree with him, switch to the expected majority response. Otherwise, keep original response.

Mostly contrarian, but retain very typical answers: If the survey taker expects 50-70% of others to agree with him, switch to the expected minority response. Otherwise, keep original response.

Consensus for other sex: Switch the survey taker's stated sex, and change to the responses he expects to be in the majority for that sex.

Contrarian for other sex: Switch the survey taker's stated sex, and change to the responses he expects to be in the minority for that sex.

Mild consensus for other sex: Switch the survey taker's stated sex, and change to the expected majority response for that sex when predicting less than 30% agreement.

Mild contrarian for other sex: Switch the survey taker's stated sex, and change to the expected minority response for that sex when predicting more than 70% agreement.

Consensus for own sex but claim to be other sex: Apply both strategies 4 and 6.

Consensus for other sex but claim own sex: Switch answers to those expected to be in the majority for the other sex, but retain the survey taker's true sex.

Impersonate another: Use the reported sex and responses that the survey taker gave when impersonating the preferences of some other specific person he knows well. Unlike other strategies, which transform responses according to some algorithm meant to mimic deception, impersonation is actual deception, based on real (non-truthful) responses from our survey takers.

Counter-impersonate: Reverse all responses, including the reported sex, that the survey taker gave when impersonating a well-known other.

Random: Retain actual sex, but generate random responses from a uniform distribution.

REFERENCES

Camerer, C. F., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor production framework. *J. Risk Uncert., 18*, 7-42.

Cooke, R. M. (1991). *Experts in Uncertainty.* New York: Oxford University Press.

Cremer, J., & McLean, R. P. (1988). Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica, 56*, 1247-1257.

d'Aspremont, C., & Gerard-Varet, L.-A. (1979). Incentives and Incomplete Information. *J. Public Econ., 11*, 25-45.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *J. Exp. Soc. Psychol., 25*, 1-17.

Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect. In R. Hogarth (Ed.), *Insights in Decision Making* (pp. 179-199).

Engelmann, D., & Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics, 3*(3), 241-260.

Johnson, S. J., Pratt, J., & Zeckhauser, R. J. (1990). Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case. *Econometrica, 58*, 873-900.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *J. Pers. Soc. Psychol., 67*, 596-610.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychol. Bull., 102*, 72-90.

Mazar, N., Amir, O., & Ariely, D. (2005). (Dis)Honesty: A combination of internal and external rewards. Working paper, Sloan School of Management, Massachusetts Institute of Technology.

McAfee, P., & Reny, P. (1992). Correlated information and mechanism design. *Econometrica, 60*, 395-421.

McLean, R., & Postlewaite, A. (2002). Informational Size and Incentive Compatibility. *Econometrica, 70*, 2421-2454.

Miller, N. H., Resnick, P., & Zeckhauser, R. J. (forthcoming). Eliciting Informative Feedback: The peer-prediction method. *Management Science.*

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609.

Paulhus, D. L., & Bruce, N. (1990, June). Validation of the OCQ: A preliminary study. Paper presented at the annual convention of the Canadian Psychological Association, Ottawa, Ontario, Canada.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*, 890-904.

Phillips, D. L., & Clancy, K. J. (1972). Some effects of 'social desirability' in survey studies. *American Journal of Sociology, 77*, 921–940.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science, 306*, 462-466.

Robins, R. W., & John, O. P. (1997). Effects of visual perspective and narcissism on self-perception: Is seeing believing? *Psychological Science, 8*, 37-42.

Ross, L., Greene, D., & House, P. (1977). The "False Consensus Effect:" An Egocentric Bias in Social Perception and Attributional Processes. *J. Exp. Soc. Psychol., 13*, 279-301.

Swets, J. A. (1964). *Signal detection and recognition by human observers.* New York: Wiley.