Games and Economic Behavior 68 (2010) 443–468

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Preference reversals: The impact of truth-revealing monetary incentives ☆

Joyce E. Berg [a], John W. Dickhaut [b], Thomas A. Rietz [c,*]

[a] Department of Accounting, Henry B. Tippie College of Business, University of Iowa, Iowa City, IA 52242, United States
[b] Economic Science Institute, Chapman University, One University Dr., Orange, CA 92866, United States
[c] Department of Finance, Henry B. Tippie College of Business, University of Iowa, Iowa City, IA 52242, United States

ARTICLE INFO

Article history:
Received 15 August 2005
Available online 13 August 2009

JEL classification:
C91
D81

Keywords:
Preference reversal
Risky choice
Decision making
Uncertainty

ABSTRACT

Researchers vigorously debate the impact of incentives in preference reversal experiments. Do incentives alter behavior and generate economically consistent choices? Lichtenstein and Slovic (1971) document inconsistencies (reversals) in revealed preference in gamble pairs across paired choice and individual pricing tasks. The observed pattern is inconsistent with stable underlying preferences expressed with simple errors. Lichtenstein and Slovic (1973) and Grether and Plott (1979) introduce incentives, but aggregate reversal rates change little. These results fostered numerous replications and assertions that models of non-stable preferences are required to explain reversals. Contrary to this research, we find that incentives can generate more economically consistent behavior. Our reevaluation of existing experimental data shows that incentives have a clear impact by better aligning aggregate choices and prices. The effect is sufficiently large that, with truth-revealing incentives, a stable-preferences-with-error model not only explains behavior, but fits the data as well as any model possibly could.

1. Introduction

An extensive literature documents the robustness of preference reversals, inconsistencies in implied preference orderings when subjects are asked to choose between two gambles versus separately price them. This literature documents that monetary incentives have little effect on overall reversal rates and has been cited as evidence that incentives do not affect behavior in experiments. As a result, several new models of choice have been developed to explain behavior in preference reversal experiments.

We reexamine preference reversal studies where paired comparison (choices) and individual pricing tasks were both used to elicit preferences over monetary gambles. We document a striking incentive effect: though overall rates of reversal are relatively stable across incentive environments, the pattern of responses changes significantly when truth-revealing incentives are used.[1] Without incentives, subjects' revealed preferences appear dependent on the elicitation task. In contrast, when truth-revealing incentives are present, subjects' responses are consistent with stable underlying preferences that are

☆ We thank Colin Camerer, Jim Cox, Robyn Dawes, Philip Dybvig, Ido Erev, Mike Ferguson, John Geweke, William Goldstein, Glenn Harrison, Charles Holt, Joel Horowitz, Jack Hughes, Forrest Nelson, Charles Plott, Gene Savin, Paul Schoemaker, Amos Tversky, Nathaniel Wilcox, and workshop participants at the Economic Science Association, University of Chicago, University of Iowa and Cornell University for thought provoking comments and conversation in the development of this paper.

* Corresponding author.
E-mail address: Thomas-Rietz@uiowa.edu (T.A. Rietz).

[1] By "truth-revealing" we mean incentives that are designed to elicit true preferences under expected utility theory. While there are other possible ways to classify incentives, we show a clear change in the pattern of responses under this particular definition.

0899-8256/$ – see front matter
doi:10.1016/j.geb.2009.07.008

revealed with random errors (e.g., expected utility with random errors). Our tests show that incentive treatments significantly alter the ability of stable preference models to explain the data.

In documenting the effects of incentives, we study preference reversal experiments already in the literature. To avoid confounding incentive effects with other treatment effects in the context of this analysis, we limit our data set to published replications, or near replications, of the classic Lichtenstein and Slovic (1971) preference reversal experiment.[2] We arrive at our conclusion that incentives do matter by using two different analyses of how incentives alter the *pattern* of responses. First, we simply classify experiments by incentive type (no monetary incentives, truth-revealing monetary incentives, and "indeterminate monetary incentives") and examine patterns of aggregate behavior across these treatments.[3] We find that the differences are significant: reversal rates conditional on pricing change dramatically when truth-revealing incentives are present. The conditional reversal rates appear to be consistent with error correction (as defined below) when truth-revealing incentives are present, but not when incentives are absent. We then develop a system for testing between existing formal models of behavior empirically. Test results mirror those found in our initial analysis: when truth-revealing incentives are present, a simple expected-utility-with-random-error model explains the data as well as any model possibly could. However, when incentives are absent, this simple model cannot accommodate the data. Thus, our analysis supports the claim that the behavioral model underlying choice differs depending on the presence or absence of incentives.

In the next two sections, we discuss preference reversal experiments, the selection criteria we use to determine the data we analyze, and the classification of experiments by incentive type. In Section 4, we examine the differences in the patterns of aggregate data without reference to a particular model of behavior. We show that there is a significant change in conditional reversal rates. In Section 5, we develop formal models and tests. In Section 6, we provide our estimation and test results. We also examine the robustness of our results by asking (1) whether relaxing the data selection criteria would affect our results and (2) whether subject self-selection may account for observed differences. We conclude in Section 7.

2. Preference reversal experiments and data

2.1. General description of preference reversal experiments

We isolate those studies that preserve the classic structure of Lichtenstein and Slovic (1971) and do not introduce other factors that might affect the inference on the impact of incentives. Such studies consist of two preference elicitation tasks: paired-choice (indicating a preference while viewing a pair of bets) and individual pricing (assigning a scalar value, generally money or "points," while viewing each bet in isolation). Bet pairs have approximately equal expected value, but large differences in variance.[4] One bet (the P-bet) has a high probability of winning a small amount of money. The other bet (the $-bet) has a low probability of winning a large amount of money. For example, Grether and Plott (1979) contains the following typical preference reversal bet pair: a bet with a 32/36 chance of winning $4.00 and a 4/36 chance of losing $0.50 (the P-bet), and a bet with a 4/36 chance of winning $40.00 and a 32/36 chance of losing $1.00 (the $-bet). While the bets have approximately equal expected values ($3.50 versus $3.56), the P-bet has significantly lower variance (2.00 versus 166.02).

In the paired-choice task, the subject is presented with the two bets simultaneously and asked which one is preferred. In the pricing task, the subject is shown each bet independently and asked to state a scalar, typically the minimum selling price, for the bet. The preference ordering implied by the choice task is compared to the ordering implied by the pricing task to determine whether the subject's orderings are consistent. A preference reversal occurs when the choice and pricing tasks imply different orderings for the two bets.

The data from preference reversal experiments are typically analyzed at the aggregate level rather than at the individual subject level and are presented as frequencies in four response cells (when indifference is not allowed as a response) or nine response cells (when indifference is allowed). We use the same unit of analysis here. Fig. 1 shows a typical four cell data set (from Lichtenstein and Slovic, 1971, experiment 1).[5] Cells a and d represent stable preferences: the bets within a pair are assigned the same ordering by both the choice and pricing decisions. Cells b and c represent reversals: the less risky bet chosen but priced lower (cell b) or the more risky bet chosen but priced lower (cell c).[6] Using the variables a through d to represent the frequencies of observations in cells a through d, the reversal rate is $(b+c)/(a+b+c+d)$.

[2] Of course, while significant treatment differences beyond incentives are confounds for our analysis, they were not necessarily confounding factors in the context of the original research experiments.

[3] Our labels come from observing that "indeterminate" incentives do not necessarily induce expected utility maximizing subjects to truthfully reveal preferences and "truth-revealing" incentives do. By classifying incentives in this way, we are not arguing that subjects are necessarily maximizing expected utility, just that the incentive schemes classified in this manner have a significant impact on behavior.

[4] In fact, many of the experiments we examine use nearly identical sets of bets. Thus, the results we present cannot be attributed to differences in bets across experiments.

[5] In our analysis, we use the four cell analysis and ignore indifference responses. We do this for both data limitation and theoretical reasons.

[6] Cell b reversals, where the P-bet is chosen and the $-bet priced higher, are often called Predicted Reversals. Cell c reversals, where the $-bet is chosen while the P-bet is priced higher are often called Unpredicted Reversals. See Lichtenstein and Slovic (1971, 2006) for further elaboration on these classifications.

	P-bet priced higher	$-bet priced higher
P-bet chosen	Cell a 88 8.48%	Cell b 441 42.49%
$-bet chosen	Cell c 32 3.08%	Cell d 477 45.95%

Fig. 1. Typical pattern of preference responses (from Lichtenstein and Slovic, 1971, experiment 1, 1038 observations).

2.2. Data selection criteria

To select the data included in our analysis, we start with two general searches: (1) an ISI Web of Knowledge/Web of Science (the online version of the Social Science Citation Index) search that included published papers from all of their citation databases, all languages, all file types, from 1979 through 2006 and (2) a Google Scholar search.[7] These two sources are commonly used for citation counts during tenure review. In both searches, we requested all documents containing the words "preference reversal," "preference reversals," "preference-reversal," "preference-reversals," "reversal of preference" or "reversals of preference" in the title. Both searches were conducted on September 3, 2006. The Web of Science search returned 100 items; the Google search returned 215. We next corrected the Google list by filtering out obviously unpublished papers, papers that did not appear in refereed journals or edited volumes, duplicate references and references that we immediately recognized as errors.[8] Then, we concatenated the lists. The result was 129 papers that we analyzed for potential inclusion in our analyses.

Next, we added one paper (Selten et al., 1999) that did not contain "preference reversal" in the title. This paper was brought specifically to our attention by readers in the development of our work. Those readers claimed that excluding this paper would bias results in favor of our hypothesis that incentives change behavior. Indeed, the model we develop below does not explain behavior perfectly in Selten et al. (1999). As a result, we include it to follow a conservative approach in testing our hypothesis. To insure that this does not affect results, we tested all of our results below without including this paper. We found no differences in conclusions. No tests statistics reported in the paper rose above or fell below significance as a result of including or excluding this data.

Finally, as a cross check to make sure that our paper identification method did not miss other papers without preference reversal in the title, we also examined the 141 references cited in Seidl's (2002) review of the preference reversal literature. This led to no additional data sets that could be used for our analysis. In contrast, Seidl (2002) does not reference two papers we included in our main analysis that were published before 2002 and one paper that was published after.

We applied the following criteria to the 130 total papers in sequential order to determine which papers should be included in our analysis:

1. The paper is published in a refereed journal or volume.
2. The paper contains an experiment.
3. At least one experiment in the paper uses human subjects.
4. For the experiment(s) identified in (3), the objects evaluated in at least one experiment are two-prize gambles for real or hypothetical money.
5. For the experiment(s) identified in (4), at least one experiment contains classic preference reversal tasks — a single paired-choice task where subjects see all of the parameters of both bets simultaneously and individual pricing tasks where scalars are chosen for the bets' "prices" or certainty equivalents.[9]
6. For the experiment(s) identified in (5), at least one experiment has no significant design changes to the classic preference reversal experimental design that may confound our analysis.
7. Data for the qualifying experiment(s) is available for cells a, b, c and d used in our analysis.

These seven criteria were used to assure that our sample included experiments where the primary design differences were limited to incentives. Screens 1 and 2 assure we have published papers with original data. Screens 3, 4 and 5 assure consistency of subjects, stimuli and tasks while screen 6 eliminates treatment changes that may obscure, compete as explanations with or otherwise confound the pure effect of incentives. Screen 7 assures that the data needed for our analysis are avail-

[7] The ISI search was limited to papers after 1979 because that was as far back as that database went at the time of our search. The Google Scholar search generated many references that were errors. In many of these, the author names were incorrect or in the wrong order, or the source of the paper or title was in error. To the extent that we recognized these errors immediately, we corrected them before concatenating the two lists.

[8] We eliminate unpublished papers and papers not published in refereed journals because these papers have not undergone a peer review process to verify the quality of the research and are sometimes not available to the research community. In addition, we would have no way of knowing whether we included all, or even a representative sample, of unpublished papers.

[9] This mirrors the definition of a classic preference reversal task as defined in Bostic et al. (1990) and Cox and Grether (1996) (they add that the gambles need to be of approximately equal expected value).

Table 1
Results of filtering papers using selection criteria.

Total papers from initial search	130
1. Paper not published in a refereed journal or volume	16
2. Paper does not contain an experiment	25
3. Experiments do not use human subjects	21
4. Objects evaluated are not two-prize monetary gambles	35
5. Tasks not classic preference reversal tasks	13
6. Substantial change to classic PR design – reported in Section 6.2	8
7. Data not available for cells a, b, c, d	1
Papers included in primary data analysis (Table 3)	11

able. If an experiment met all seven criteria, it is included in our primary analysis. Several papers/experiments introduced design differences in addition to incentive differences. These papers met the first five criteria but failed criterion six. They are included in our Section 6.2.

We were surprised by the small number of pure replications in the preference reversal literature. Though many papers introduced techniques to reduce reversals, only 11 contained experiments replicating the classic Lichtenstein and Slovic study. Applying our seven criteria eliminated papers from our sample as shown in Table 1.[10]

Sixteen of the papers in our initial search were eliminated because the paper was not published in a refereed journal. One paper did not exist — its citation was incorrect. The remaining fifteen were either unpublished or published in a non-refereed volume. Forty-six papers were eliminated because they did not contain original experimental data using human subjects (criteria 2 and 3). Theory papers and discussions of previously published results were eliminated by criterion 2. Experiments that did not use human subjects, for example, rat experiments and chemical analyses, were eliminated by criterion 3.

Another thirty-five papers did not satisfy the requirement that the objects evaluated be two-prize monetary gambles (criterion 4). A detailed list of papers eliminated under criterion 4 appears in Appendix A. Many of the studies eliminated under this criterion were designed to examine whether preference reversals were observable in previously unexplored settings such as contractor bidding, life expectancy choices and restaurant grading.

Papers eliminated under criterion 5 (not a classic preference reversal task) typically introduced alternative valuation or choice elicitation methods designed to eliminate or reduce reversals. Thirteen papers were eliminated under this criterion for reasons such as:

- no pricing task was used or no choice task was used,
- the pricing task required subjects to price bets in a pair simultaneously (this mixes choice and pricing),
- the pricing task involved iterative choices (again, because choice and pricing are mixed),
- all attributes of a bet are not revealed before the subject makes a decision (because we can't tell whether a subject actually saw the binary gamble).

A detailed list of papers eliminated and reasons for elimination under criterion 5 appears in Appendix A.

Of the twenty remaining papers, eight contained substantial design changes (eliminated based on criterion 6) and one did not report data in a format that allowed us to derive the cell frequencies described in Fig. 1. Substantial design changes eliminated papers where:

- bets did not have approximately the same expected value,
- arbitrage was introduced,
- preferences were induced using the binary lottery payoff mechanism,
- an ordinal payoff scheme was used.

All papers eliminated from the main analysis under criterion 6 are discussed in Section 6.2 below. The remaining papers are classified by incentive type in Appendix B and appear in our main analysis.

The screens we use allow us to examine incentive effects in human subjects across similar tasks. Our focus is on transparent, two-outcome monetary gambles of the type first studied in Lichtenstein and Slovic (1971). Table 2 describes the 26 data sets contained in the 11 papers used in our main analysis. These data sets contain a variety of incentive designs conducted by experimenters with a variety of backgrounds and interests. An overview of the experimental designs is contained in Appendix B; design details are in the original papers.

[10] Because papers often contain several experiments, we eliminated data on an experiment by experiment basis. The numbers here represent papers where every experiment in the paper had been eliminated by a particular stage.

Table 2
Data sets analyzed.

Incentive category	Data set number	Paper and experiment	Short name	Data					
				N^a	a	b	c	d	Reversal rate
None	1	Lichtenstein and Slovic (1971), Exp. 1	L&S1	1038	88	441	32	477	46%
	2	Lichtenstein and Slovic (1971), Exp. 2	L&S2	3234	844	876	411	1103	40%
	3	Goldstein and Einhorn (1987), Exp. 2a	G&E2a	127	26	40	13	48	42%
	4	Grether and Plott (1979), Exp. 1a	G&P1a	245	49	71	14	111	35%
Indeterminate	5	Lichtenstein and Slovic (1971), Exp. 3	L&S3	84	21	27	4	32	37%
	6	Lichtenstein and Slovic (1973), positive EV	L&SLV+	484	44	185	25	230	43%
	7	Lichtenstein and Slovic (1973), negative EV	L&SLV−	348	190	46	85	27	38%
	8	Pommerehne et al. (1982), Grp. 1 Run 1	PSZ1.1	160	44	52	9	55	38%
	9	Pommerehne et al. (1982), Grp. 1 Run 2	PSZ1.2	154	31	28	13	82	27%
	10	Pommerehne et al. (1982), Grp 2 Run 1	PSZ2.1	151	39	35	10	67	30%
	11	Pommerehne et al. (1982), Grp 2 Run 2	PSZ2.2	152	27	38	6	81	29%
	12	Mowen and Gentry (1980), warm-up exercises	M&G	57	7	24	7	19	54%
Truth-revealing	13	Grether and Plott (1979), Exp. 1b	G&P1b	262	26	69	22	145	35%
	14	Grether and Plott (1979), Exp. 2b (selling prices)	G&P2SP	209	27	52	25	105	37%
	15	Grether and Plott (1979), Exp. 2b ($ equivalents)	G&P2DE	214	20	61	20	113	38%
	16	Reilly (1982), Stg. 1, Grp. 1	R1.1	343	51	84	57	151	41%
	17	Reilly (1982), Stg. 1, Grp. 2	R1.2	318	42	100	29	147	41%
	18	Berg et al. (1985), Exp. 1, Ses. 1	BDO1.1	124	23	25	18	58	35%
	19	Berg et al. (1985), Exp. 1, Ses. 2	BDO1.2	118	15	23	12	68	30%
	20	Berg et al. (1985), Exp. 2, Ses. 1	BDO2.1	114	35	13	22	44	31%
	21	Berg et al. (1985), Exp. 2, Ses. 2	BDO2.2	110	41	12	20	37	29%
	22	Chu and Chu (1990), psychology students, Exp. 1, Part I	C&CPs	225	50	52	39	84	40%
	23	Chu and Chu (1990), economics students, Exp. 1, Part I	C&CEc	308	60	64	46	138	36%
	24	Selten et al. (1999) Money payments w/o summary statistics	SSA1	186	46	29	11	100	22%
	25	Selten et al. (1999) Money payments w/ summary statistics	SSA2	96	30	18	6	42	25%
	26	Chai (2005), minimum selling price data	C.MSP	558	64	177	47	270	40%

[a] N is the total number of choices in cells a, b, c and d. Indifference choices and prices are ignored.

In addition to our main analysis, we include several omitted papers in Section 6.2 in order to examine whether our results are robust to the selection criteria. The data sets analyzed in Section 6.2 contain significant design changes that make them not directly comparable to the papers in the main analysis. However, they do provide additional evidence about the effects of incentives.

3. Classification by incentives

To begin our analysis, we first classify each experiment according to the form of incentives used: no monetary incentives, truth-revealing incentives, and indeterminate incentives. Details supporting the classification of individual experiments appear in Appendix B.

"No-incentives experiments" use hypothetical bets. We analyze four such experiments, all of which use flat participation fees but have no performance-based rewards in their experimental design. Because no differential reward is given for responding truthfully, any response is optimal for a subject who cares only about the monetary payoffs for the experiment. Included in this category are Lichtenstein and Slovic (1971) experiments 1 and 2, Goldstein and Einhorn (1987) experiment 2a and Grether and Plott (1979) experiment 1a. We label the data sets L&S1, L&S2, G&E2a, and G&P1a, respectively.

"Truth-revealing incentives experiments" incorporate unambiguous incentives for truthfully revealing preferences when subjects are expected utility maximizers.[11] These experiments all use a paired-choice task and a pricing procedure that should elicit truthful revelation of prices (generally the Becker et al., 1964, pricing procedure). We analyze fourteen experiments in this category: three from Grether and Plott (1979), two from Reilly (1982), four from Berg et al. (1985), two from

[11] These "truth-revealing" incentive schemes may not be truth-revealing for non-expected utility maximizing subjects. Without taking a stand on whether subjects are *actually* expected utility maximizers, the data clearly show a significant change in the observed pattern of responses in preference reversal experiments incorporating "truth-revealing" incentives. In Section 6, we show that the data is consistent with stable preferences across gambles and noisy revelation of these preferences. While this is consistent with expected utility maximizing subjects, non-expected utility maximizing subjects may also have been induced to have stable preferences.

Table 3
Preferences, reversal rates and conditional reversal rate asymmetries.

Incentive category	Data set	Average preference for the P-bet according to		Abs. diff. between P-bet preference measures	Reversal rate	Conditional (on choice) reversal rates			Conditional (on pricing) reversal rates		
		Choices $(a+b)/(a+b+c+d)$	Prices $(a+c)/(a+b+c+d)$		$(b+c)/(a+b+c+d)$	P-bet $(b/(a+b))$	\$-bet $(c/(c+d))$	Difference	P-bet $(c/(a+c))$	\$-bet $(b/(b+d))$	Difference
None	L&S1	0.51	0.12	0.39	46%	0.83	0.06	0.77	0.27	0.48	−0.21
	L&S2	0.53	0.39	0.14	40%	0.51	0.27	0.24	0.33	0.44	−0.12
	G&E2a	0.52	0.31	0.21	42%	0.61	0.21	0.39	0.33	0.45	−0.12
	G&P1a	0.49	0.26	0.23	35%	0.59	0.11	0.48	0.22	0.39	−0.17
Indeterminate	L&S3	0.57	0.30	0.27	37%	0.56	0.11	0.45	0.16	0.46	−0.30
	L&SLV+	0.47	0.14	0.33	43%	0.81	0.10	0.71	0.36	0.45	−0.08
	L&SLV−	0.68	0.79	0.11	38%	0.19	0.76	−0.56	0.31	0.63	−0.32
	PSZ1.1	0.60	0.33	0.27	38%	0.54	0.14	0.40	0.17	0.49	−0.32
	PSZ1.2	0.38	0.29	0.10	27%	0.47	0.14	0.34	0.30	0.25	0.04
	PSZ2.1	0.49	0.33	0.17	30%	0.47	0.13	0.34	0.20	0.34	−0.14
	PSZ2.2	0.43	0.22	0.21	29%	0.58	0.07	0.52	0.18	0.32	−0.14
	M&G	0.54	0.25	0.30	54%	0.77	0.27	0.50	0.50	0.56	−0.06
Truth-revealing	G&P1b	0.36	0.18	0.18	35%	0.73	0.13	0.59	0.46	0.32	0.14
	G&P2SP	0.38	0.25	0.13	37%	0.66	0.19	0.47	0.48	0.33	0.15
	G&P2DE	0.38	0.19	0.19	38%	0.75	0.15	0.60	0.50	0.35	0.15
	R1.1	0.39	0.31	0.08	41%	0.62	0.27	0.35	0.53	0.36	0.17
	R1.2	0.45	0.22	0.22	41%	0.30	0.16	0.13	0.41	0.40	0.00
	BDO1.1	0.39	0.33	0.06	35%	0.48	0.24	0.24	0.44	0.30	0.14
	BDO1.2	0.32	0.23	0.09	30%	0.39	0.15	0.24	0.44	0.25	0.19
	BDO2.1	0.42	0.50	0.08	31%	0.73	0.33	0.40	0.39	0.23	0.16
	BDO2.2	0.48	0.55	0.07	29%	0.77	0.35	0.42	0.33	0.24	0.08
	C&CPs	0.45	0.40	0.06	40%	0.49	0.32	0.17	0.44	0.38	0.06
	C&CEc	0.40	0.34	0.06	36%	0.48	0.25	0.23	0.43	0.32	0.12
	SSA1	0.40	0.31	0.10	22%	0.61	0.10	0.51	0.19	0.22	−0.03
	SSA2	0.50	0.38	0.13	25%	0.63	0.13	0.50	0.17	0.30	−0.13
	C.MSP	0.43	0.20	0.23	40%	0.27	0.15	0.12	0.42	0.40	0.03

Chu and Chu (1990),[12] two from Selten et al. (1999) and one of the data sets in Chai (2005).[13] These are denoted G&P1b, G&P2SP, G&P2DE, R1.1, R1.2, BDO1.1, BDO1.2, BDO2.1, BDO2.2, C&CPs, C&CEc, SSA1, SSA2 and C.MSP, respectively.

"Indeterminate-incentives experiments" use bets that have real monetary payoffs, but the designs do not strictly induce truthful revelation for utility maximizing subjects. We include these experiments to determine whether monetary incentives alone affect behavior in preference reversal experiments or whether the more restrictive definition of truth-revealing incentives is necessary. These experiments include Lichtenstein and Slovic (1971) experiment 3, two Lichtenstein and Slovic (1973) experiments conducted in Las Vegas, four experiments from Pommerehne et al. (1982) and the "warm up" exercises from Mowen and Gentry (1980). We label the data sets L&S3, L&SLV+ (this session used positive expected value bets), L&SLV− (this session used negative expected value bets) and PSZ1.1, PSZ1.2, PSZ2.1, PSZ2.2 and M&G, respectively.

4. Model free effects of incentives in preference reversal data patterns

Table 3 presents summary statistics for each of the data sets in our main analysis. In the "preferences" columns, we show the percentage of instances in which the P-bet is revealed as preferred in each task and the difference in these two percentages. The next column repeats the overall reversal rate (from Table 2) for convenience. The last six columns present conditional reversal rates and the differences in these rates. We discuss each item in detail in the following subsections.

4.1. Effects on preference over bets

The effects of truth-revealing monetary incentives are immediately apparent in the percentage of P-bet choices. P-bet choices ranged from 49% to 53% and averaged 51% in experiments without incentives. However, when truth-revealing incentives are introduced, P-bet choices ranged from 32% to 50%, with an average of 41%. Choices are decidedly more risk seeking on average under truth-revealing incentives. The difference across treatments is significant. A Wilcoxon rank-sum

[12] Both data sets are from experiment 1, part I and consist only of the Reilly (1982) replication data (before subjects knew that there would be a part II to the experiment, where new choices were given to subjects and inconsistencies were arbitraged).

[13] Chai (2005) presents four "sets" of data that are not actually independent. All correspond to the same choices but different value elicitation procedures. We use the set from Chai's minimum selling price task data set here.

statistic comparing the rates in no-incentives experiments to rates in truth-revealing-incentives experiments using each experiment as a data point is 2.867 (p-value = 0.0041).[14]

Truth-revealing incentives also align revealed preferences. Table 3 shows the absolute difference between the percentage of instances in which subjects prefer the P-bet according to the choice task and the percentage of instances in which subjects prefer the P-bet according to the pricing task. These statistics represent the degree of inconsistency of preferences across tasks at an aggregate level. The average absolute difference without incentives is 25%. Under indeterminate incentives it is 22%. Under truth-revealing incentives it drops to 12%. Again, the difference across treatments is significant. A Wilcoxon rank-sum statistic comparing the absolute differences in no-incentives experiments to rates in truth-revealing-incentives experiments using each experiment as a data point is 2.230 (p-value = 0.0257). Thus, incentives result in clear, consistently risk-seeking revealed preferences across the two tasks.

4.2. Effects on reversal rates

While responses are more aligned, reversal rates do not necessarily fall. Rates range from 35% to 46% (average 40%) in experiments without incentives, 27% to 54% (average 40%) in experiments with indeterminate incentives, and 22% to 41% (average 34%) in experiments with truth-revealing incentives. While some truth-revealing incentives experiments exhibit lower reversal rates and rates are lower on average, many reversal rates are higher than those observed under no-incentives or indeterminate incentives. The difference across treatments is marginally significant. A Wilcoxon rank-sum statistic comparing the rates under no-incentives to rates under truth-revealing incentives using each experiment as a data point is 1.699 (p-value = 0.0893). Such data leads some researchers (e.g., Camerer and Hogarth, 1999, Table 1) to conclude that incentives do not affect behavior (or may even make behavior less rational) in preference reversal experiments.

4.3. Effects on conditional reversal rates

Conditional reversal rates and their signed differences are also presented in Table 3. Reversal rates conditional on the P-bet or $-bet being chosen are shown in columns 7–9; rates conditional on the P-bet or $-bet being priced higher are shown in columns 10–12. Truth-revealing incentives have a clear impact, primarily on the asymmetry in reversal rates conditional on the pricing task ranking. While the asymmetry in reversal rates conditional on choice differs little across treatments, incentives change the sign of the asymmetry conditional on the price-ranking. The Wilcoxon rank-sum statistic comparing the latter asymmetry under no-incentives to the asymmetry under truth-revealing incentives using each experiment as a data point is 2.761 (p-value = 0.0058).

In summary, analysis of the aggregate data shows significant effects of incentive type. Without incentives, the average frequencies of bet choices are near 50:50, but the $-bets are usually priced higher. When truth-revealing incentives are present, subjects choose the $-bets significantly more often while continuing to price them higher, creating more coherence across decisions. While the overall reversal rates seem unaffected by the incentive schemes, reversal rates conditional on prices change dramatically. As we explain below, the conditional reversal pattern under truth-revealing incentives is consistent with a stable preference across the bets expressed with uncorrelated errors. The data also show that it is not simply the existence of incentives that creates these differences. It is the truth-revealing nature of the incentives.[15]

5. Modeling the effects of incentives in preference reversal experiments

In this section, we discuss three types of models of individual behavior in preference reversal experiments and show how they are related to aggregate choice patterns. One type assumes consistent preference across bets that may be revealed with uncorrelated random errors. This includes the expected utility with noise (two-error-rate) model of Lichtenstein and Slovic (1971) and restricted versions of this model that make it a one-error-rate model and a zero-error model (strict expected utility). A second type is the anchor-and-adjust based expression theory developed by Goldstein and Einhorn (1987). In this model, aspects of the way the choices are made induce systematic misrepresentation of preferences that differ depending on the preferences of the subject. A third type is the task dependent preferences model of Tversky and Thaler (1990). In this model, preferences are constructed not from the gambles alone but also from the tasks themselves. Our contribution is to provide a nested modeling structure that allows us to distinguish empirically between the various models posited in the literature to explain behavior. With this structure, we can show when particular models are consistent with the patterns of behavior actually observed and whether incentives affect which models can explain the data.

We analyze behavior in preference reversal experiments in three steps. First, we show how cell frequencies and maximum likelihood estimation can be used to estimate parameters for underlying models of behavior in each data set. Then, for any given data set, we define a "best-fit" benchmark model that defines the maximum likelihood attainable by any

[14] P-bet choices ranged from 38% to 68% and averaged 52% in experiments with indeterminate incentives. Across the varying incentive mechanisms in the indeterminate incentive group, preferences seem to change, but not in a systematic way.

[15] None of the Wilcoxon statistics discussed above would be significant at the 95% level of confidence if we were to group the indeterminate incentives with the truth-revealing data sets.

conceivable underlying model of behavior. By construction, no model can ever fit the data better than this benchmark.[16] Though this model is only a statistical representation of fit, in context, restrictions on this model yield the types of behavioral models discussed above. Finally, we test the behavioral models against the benchmark to determine whether each of these behavioral models attains the best possible fit. Attaining the best possible fit means that we can never find a model that explains the data better. A significantly worse fit means that there may be models that can explain the data better.

5.1. Maximum likelihood estimation of behavioral models of preference reversal

We use maximum likelihood estimation to find parameters that maximize the likelihood of the observed data in cells a, b, c, and d in Fig. 1 subject to restrictions imposed by the model under investigation.[17] In fitting models, note that the cell frequencies in Fig. 1 represent a multinomial distribution. The joint log likelihood function based on predicted cell frequencies is:

$$L = \ln(n!) - \ln(A!) - \ln(B!) - \ln(C!) - \ln(D!) + A\ln\big[m_a(\theta)\big] + B\ln\big[m_b(\theta)\big] + C\ln\big[m_c(\theta)\big] + D\ln\big[m_d(\theta)\big],$$

where n is the number of observations in the data set; A, B, C and D are the total numbers of observations in cells a, b, c and d; $m_a(\theta)$, $m_b(\theta)$, $m_c(\theta)$ and $m_d(\theta)$ are the model's predictions of frequencies for cells a, b, c and d based on the model's underlying parameters, θ. Estimates and their variances are found using standard maximum likelihood techniques (see Judge et al., 1982). The value of the likelihood function given the estimated parameters indicates the model's ability to explain the data. We use likelihood ratio tests to distinguish between the behavioral models that we study.

5.2. The "best-fit" benchmark model

The multinomial nature of reported preference reversal data allows us to define a "best-fit" benchmark. If the predicted cell frequencies could be set freely, then the (global) maximum likelihood is attained by matching the predicted cell frequencies to the observed frequencies. That is, $a = m_a(\hat{\theta})$, $b = m_b(\hat{\theta})$, $c = m_c(\hat{\theta})$, and $d = m_d(\hat{\theta})$. Because this solution results in the *global* maximum of the likelihood function, it provides the best possible fit that any model can possibly attain in explaining these cell frequencies.

We call the model corresponding to this best fit, the Best-Fit Benchmark Model. Though it need not correspond to actual behavior, it serves two important purposes. First, it always achieves the global best fit to the data. As a result, it can be used as a benchmark model against which the explanatory power of more restrictive models can be measured. This is similar in spirit to developing "fully saturated" models for testing against in the sense that we find and test against models that can always be parameterized to predict exactly the observed cell frequencies.[18] Second, the Best-Fit Benchmark Model nests the behavioral models of preference reversal in the existing literature. This allows us to test the restrictions implied by various explanatory models using simple likelihood ratio tests. The way in which these models nest is shown in Fig. 2 and described in the rest of this section.

The parameters of the benchmark model are q, r, s_P and $s_\$$. There are many ways to interpret these parameters and we provide one here to help clarify how several behavioral models can be nested in this model. The parameter q corresponds to the percentage of subjects who have an underlying preference for the P-bet.[19] In the choice task, subjects make errors and, at the rate r, report the wrong preference ordering. Reported prices also sometimes change the apparent ordering of the bets. Denote by s_P the rate at which orderings are reversed for subjects who actually prefer the P-bet. Denote by $s_\$$ the rate at which orderings are reversed for subjects who actually prefer the \$-bet.[20]

[16] Note, we are fitting a model to the aggregate-level data, not each individual subject's responses. This benchmark model is a statistical representation of fit, not a specific behavioral model.

[17] Allowing for errors and applying maximum likelihood to test between competing models of behavior is similar to Camerer (1989), Starmer and Sugden (1989), Harless and Camerer (1994) and Hey and Orme (1994). We focus on only the four cells a, b, c, and d for several reasons. First, not all experiments permitted indifference as a response. Second, indifference responses do not have unambiguous implications for preference orderings. For example, a subject may rationally state the same price for two gambles while choosing one gamble over the other if the difference in preference is not sufficiently high to be observed in the price grid used in the experiment. Third, the models generally predict continuous preferences over the gambles, so the chances of being truly indifferent are vanishingly small according to theory. Finally, we note, that by using aggregate data we are making implicit assumptions about the similarity of behavior across subjects and bets. In the language of Ballinger and Wilcox (1997), we are investigating "representative" decisions across agents and bets. Ballinger and Wilcox (1997) argue that this aggregation form "should be tested; and its rejection motivates models that allow for various kinds of heterogeneity." Allowing the preference-dependent error rates that lead to expression theory (see below) is a type of heterogeneity and it appears to be required to explain the data without incentives. However, with incentives, we never reject the representative form of the two-error-rate model (again, see below), so we do not investigate heterogeneity further.

[18] However, the intuition from the typical log linear case does not always apply here because of the non-linearities in the models. For example, the two-error-rate model developed below might appear fully saturated because it has three parameters and must fit three free cell frequencies. In fact it cannot always fit all three observed cell frequencies because of its quadratic form.

[19] While one can develop different interpretations of q and the parameters that follow for each model, we will follow interpretations that are natural extensions of models already in the existing literature. In the end, it is not the interpretation of the parameters that matters. What matters is the fact that the models that can fit the data change as a result of the incentive treatment.

[20] This is the extension of Lichtenstein and Slovic's (1971) two-error-rate model, which assumes $s_\$ = s_P = s$.

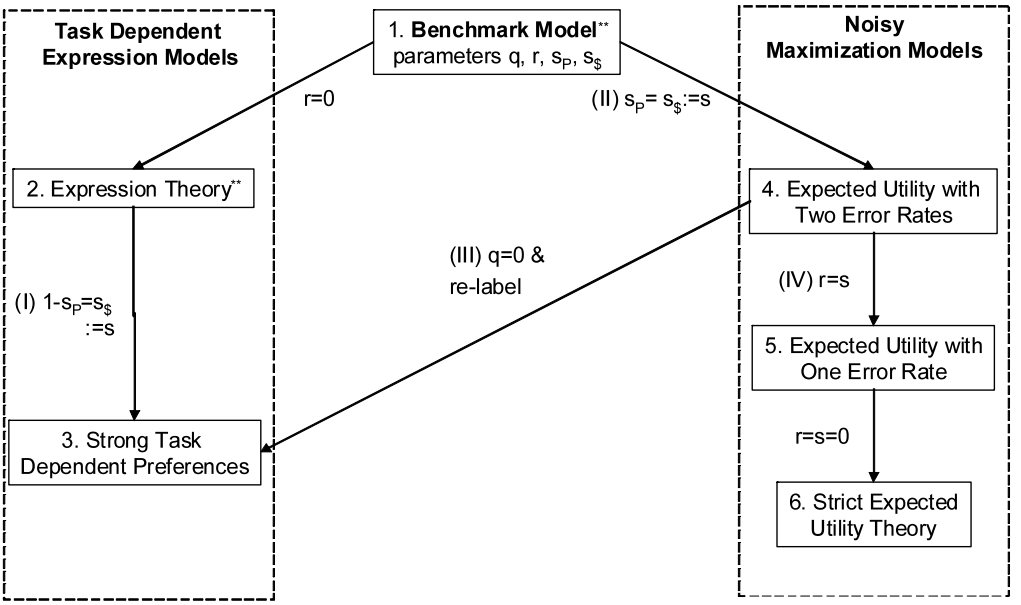

Fig. 2. Relationships between models studied. Numbered models correspond to model numbers in the text. Numbered restrictions correspond to likelihood ratio tests in Table 5. **The benchmark model and expression theory are observationally equivalent models that always attain the best-fit benchmark.

	P-bet priced higher	$-bet priced higher
P-bet chosen	$a = (q)(1-r)(1-s_P) + (1-q)(r)(s_\$)$	$b = (q)(1-r)(s_P) + (1-q)(r)(1-s_\$)$
$-bet chosen	$c = (q)(r)(1-s_P) + (1-q)(1-r)(s_\$)$	$d = (q)(r)(s_P) + (1-q)(1-r)(1-s_\$)$

Fig. 3. Pattern of responses for the benchmark model, where: $q =$ percentage of subjects whose underlying preference ordering ranks the P-bet higher; $r =$ error rate in the paired-choice task; $s_P =$ error rate in the pricing task when P-bet preferred; $s_\$ =$ error rate in the pricing task when $-bet preferred.

Under the benchmark model, data will conform to the pattern shown in Fig. 3. This model can always be parameterized to attain the best-fit benchmark likelihood. This is true because for any observed a, b, c and d, one can find valid parameters for the model such that $a = m_a(\hat{\theta})$, $b = m_b(\hat{\theta})$, $c = m_c(\hat{\theta})$ and $d = m_d(\hat{\theta})$. To see this, set $r = 0$ and maximize the (global) likelihood function by matching observed frequencies to predictions. Setting the model predictions equal to the observed frequencies shows that the estimated parameters must solve $a = \hat{q}(1-\hat{s}_P)$, $b = \hat{q}\hat{s}_P$, $c = (1-\hat{q})\hat{s}_\$$ and $d = (1-\hat{q})(1-\hat{s}_\$)$. Simple algebra shows that $\hat{q} = a + b, \hat{s}_P = b/(a+b)$ and $\hat{s}_\$ = c/(c+d)$ solve the system and represent valid fractions and probabilities. Hence, the benchmark model always achieves the (global) best-fit benchmark. By definition, no model can ever fit the aggregate data better than the benchmark model.[21]

5.3. Task dependent evaluation models

A single restriction on the benchmark model produces Goldstein and Einhorn's (1987) expression theory (Model 2 in Fig. 2). Expression theory argues that subjects have a preference ordering over bets that is accurately reported in the choice task. However, in the pricing task, valuations are influenced by an anchor and adjust mechanism based on the compatibility of the units of measure used in prices and bet outcomes. Because of this, price evaluations of bets may differ from choice evaluations. Further, Goldstein and Einhorn (1987) conjecture that the rate of inconsistencies will depend on whether the subject prefers the P-bet or the $-bet in the choice task.

Starting with the benchmark model let q represent the fraction of subjects who prefer the P-bet in the choice task. Because this preference is accurately reported in the choice task, set $r = 0$. Denote the conditional reversal rates by the fractions s_P for reversals conditional on the P-bet being chosen and $s_\$$ for reversals conditional on the $-bet being chosen. Given this notation, expression theory predicts the cell frequencies given in Fig. 4. This pattern results from applying the restriction that $r = 0$ to the response pattern in Fig. 3.

Like the benchmark model, expression theory can explain any pattern of aggregate preference reversal behavior. To see this, use exactly the same argument as used to show that the benchmark model is a best-fit model. Both the benchmark

[21] We note that in this aggregation and the others that we study later, there is an implicit assumption of independence. When we do estimation on these models, we are effectively estimating parameters for a "representative decision maker" in the language of Ballinger and Wilcox (1997).

	P-bet priced higher	$-bet priced higher
P-bet chosen	$a = (q)(1 - s_P)$	$b = (q)(s_P)$
$-bet chosen	$c = (1 - q)(s_\$)$	$d = (1 - q)(1 - s_\$)$

Fig. 4. Pattern of responses according to expression theory, where: $q =$ percentage of subjects whose underlying preference ordering ranks the P-bet higher; $s_P =$ the fraction of the outcomes in which pricing biases reverse apparent preference orderings when the P-bet is actually preferred; $s_\$ =$ the fraction of the outcomes in which pricing biases reverse apparent preference orderings when the $-bet is actually preferred.

	P-bet priced higher	$-bet priced higher
P-bet chosen	$a = (q)(1 - r)(1 - s) + (1 - q)(r)(s)$	$b = (q)(1 - r)(s) + (1 - q)(r)(1 - s)$
$-bet chosen	$c = (q)(r)(1 - s) + (1 - q)(1 - r)(s)$	$d = (q)(r)(s) + (1 - q)(1 - r)(1 - s)$

Fig. 5. Pattern of responses generated by the two-error-rate model, where: $q =$ percentage of subjects whose underlying preference ordering ranks the P-bet higher; $r =$ error rate in the paired-choice task; $s =$ error rate in the pricing task.

model and expression theory always achieve the best-fit benchmark and are observationally equivalent on aggregate preference reversal data. Neither model has testable implications for aggregate data.[22]

An alternative interpretation of parameters in this model produces task dependent preferences (Tversky and Thaler, 1990) where there is some systematic preference, but the elicitation task may change the preferences in some cases. Define q as the percentage of instances where subjects prefer the P-bet when preferences are elicited using the choice task. Some percentage of them, s_P, changes their preference in the pricing task. A (potentially different) percentage of subjects who prefer the $-bet change their preference in the pricing task. Denote this by $s_\$$. This model differs from expression theory only in interpretation. It is identical mathematically and cannot be distinguished by the aggregate data.

However, another restriction (Restriction I in Fig. 2) leads to a strong form of task dependent preferences where preferences are constructed purely through the elicitation task and preference orderings implied by responses are entirely task specific (Model 3 in Fig. 2). This implies independence of the rows and columns in Fig. 1. To achieve this, simply restrict $1 - s_P = s_\$ = s$. This is a two parameter model with q representing the fraction of choices for the P-bet and s representing the fraction of instances in which the P-bet is priced higher. Testing whether strong task dependent preferences explains the data significantly worse than the benchmark is a simple χ^2 test of independence of the rows and columns of Fig. 1.[23]

5.4. Noisy maximization models

We define noisy maximization in the same manner as Berg et al. (2003). Expected utility maximization would lead subjects to have invariant preferences over the bets. Camerer and Hogarth (1999) argue that even if this was so, the "labor cost" of optimization and accurate reporting may result in errors. The error rates should depend on the difficulty of optimization relative to the payoffs for optimizing. In the case of preference reversal experiments, the choice and pricing tasks may have different difficulties and different error costs, so the two tasks could have different error rates. Ballinger and Wilcox (1997) also argue that error rates should differ by task.[24] Lichtenstein and Slovic's (1971) two-error-rate model (Model 4 in Fig. 2) has exactly these properties. It results from a single restriction on the benchmark model (Restriction II in Fig. 2 that $s_P = s_\$ = s$) and results in the response pattern given in Fig. 5.

Unlike the restriction that yields expression theory, the restriction in the two-error-rate model is meaningful and makes the model refutable. In particular, matching the frequencies and solving for q, r and s shows that, when a solution exists, the following relationships maximize the global likelihood function:

$$\hat{q}(1 - \hat{q}) = \frac{ad - bc}{(a + d) - (b + c)}, \tag{1}$$

$$\hat{r} = (a + b - \hat{q})/(1 - 2\hat{q}) \tag{2}$$

and

$$\hat{s} = (a + c - \hat{q})/(1 - 2\hat{q}).^{25} \tag{3}$$

[22] Of course, additional restrictions on the expression theory model can make it testable. As a referee pointed out, one interpretation of the anchor and adjust theory could be that only subjects who prefer the P-bet make errors in pricing. As a result, $s_\$ = 0$. However, because this model predicts $c = 0$, it is refuted whenever there are observations in cell c (as is true in all observed data sets). A weaker restriction that $s_\$ < s_P$ is consistent with the data in all but three cases (see Table 4 below). In the first, L&SLV−, we expect the relationship to be reversed. In the other two cases, the difference between $s_\$$ and s_P is not significant (see test statistics in Table 5, column II). As a result, whether expression theory explains the data appears unaffected by incentives.

[23] In the remainder of the text, we will typically drop the "strong" adjective for brevity.

[24] Their language for this is "set conditional error rates." We thank an anonymous referee for pointing this out.

[25] Due to the quadratic form, there are two equivalent sets of parameters that satisfy these equations because q and $1 - q$ are interchangeable. The resulting estimates of r and s are each one minus the original estimate. We do not take a stand on which set of estimates is "correct" because it is irrelevant to the likelihood function (both sets give the same likelihood) and, hence, to the likelihood ratio tests discussed below. We let the data choose which set we display in the tables by minimizing the sum of the error rates r and s. This also gives P-bet preference measures (q's) that accord with Table 3.

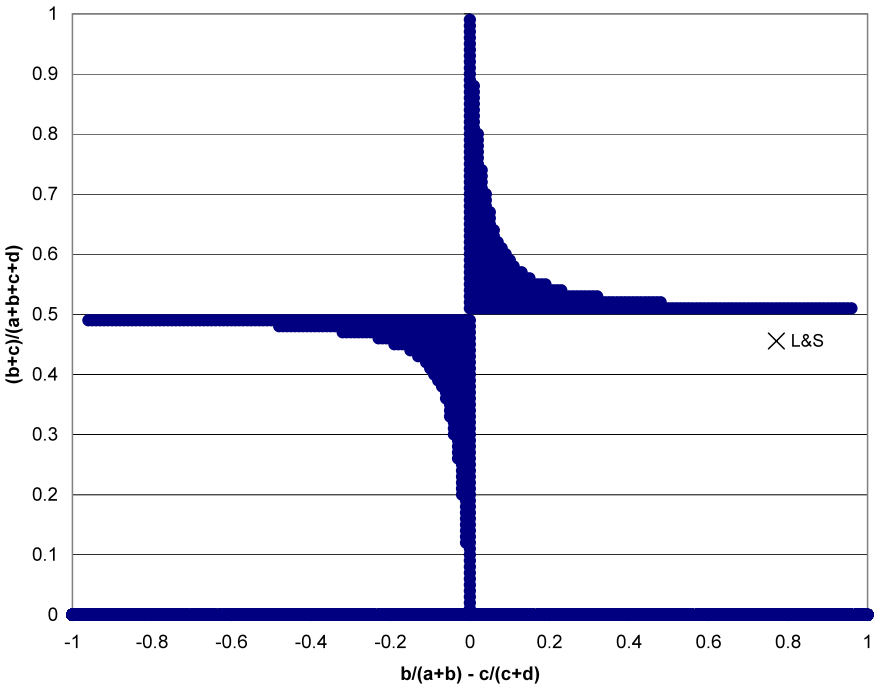

Fig. 6. Reversal configurations that can be explained by the two-error-rate model when subjects choose the P-bet slightly more often than the $-bet ($(a+b)/(a+b+c+d) = 0.509633911$).

When a solution exists, the two-error-rate model achieves the (global) best-fit benchmark. Thus, it explains the data as well as the benchmark model, as well as expression theory and, in fact, as well as any model possibly could. Note however that the two-error-rate model does impose testable restrictions on the parameter estimates. In particular, this model will not achieve the best-fit benchmark when $(ad-bc)/(a-b-c+d) < 0$ or $(ad-bc)/(a-b-c+d) > 0.25$ because Eq. (1) cannot be solved. Further, Eqs. (2) and (3) may not result in valid error rates in the $[0,1]$ range even when (1) can be solved.

The restrictions imposed by the two-error-rate model are stronger than one might think. Because there are three free cell frequencies to fit in Fig. 5, one is tempted to conjecture that any three parameter model should explain the data. However, this intuition fails for the two-error-rate model because cell frequencies are not simple linear functions of q, r, and s. Using Monte Carlo simulations, we find that, out of all possible cell frequencies that can be explained by the benchmark model, only about one third can be explained as well by the two-error-rate model. In the rest of the cases, one or more restrictions keep the two-error-rate model from achieving the best-fit benchmark.[26]

While simulation results are informative about the general performance of the two-error-rate model, what is more important is to consider whether the two-error-rate model is likely to fit the patterns of data typically seen in preference reversal experiments. We find that the two-error-rate model's ability to fit the data is primarily determined by the *same two factors* that appear to be affected by incentives: (1) whether subjects have a distinct preference for one bet-type over another and (2) the asymmetries in conditional reversal rates. In particular, the two-error-rate model cannot fit patterns of data when subjects are approximately indifferent between the bets and there are asymmetries in the conditional reversal rates. Further, it cannot fit the data when subjects have distinct preferences for one bet-type and the asymmetries in conditional reversal rates are inconsistent with those preferences.

Consider the case in which subjects do not exhibit a clear preference for one bet-type over another. Empirically, this is typical of data from preference reversal experiments conducted *without* incentives: P-bets and $-bets are chosen with approximately equal frequency in the choice task. Fig. 6 shows the two-error-rate model's ability to fit the data in such a case. We graph the asymmetry in conditional (on choices) reversal rates $(b/(a+b)-c/(c+d))$ and the overall reversal rate $(b+c)/(a+b+c+d)$ for choices that show only a slight preference for the P-bet $((a+b)/(a+b+c+d) = 0.509633911$, corresponding to Lichtenstein and Slovic's, 1971, experiment 1). The shaded area shows combinations of reversal rates and asymmetries that can be explained by the two-error-rate model. The white areas are instances in which the two-error-rate model fails.

Notice that the two-error-rate model can explain data only under quite limited conditions: (1) when there are no reversals (the bottom axis), (2) when reversals are about 50% (horizontally across the middle of the diagram, indicating true

[26] We arrive at this number through 50,000 simulations, drawing cell frequencies randomly from the feasible set. Other simulations show that the two-error-rate model also frequently fails to fit the data when it is simulated from expression theory and task dependent preferences.

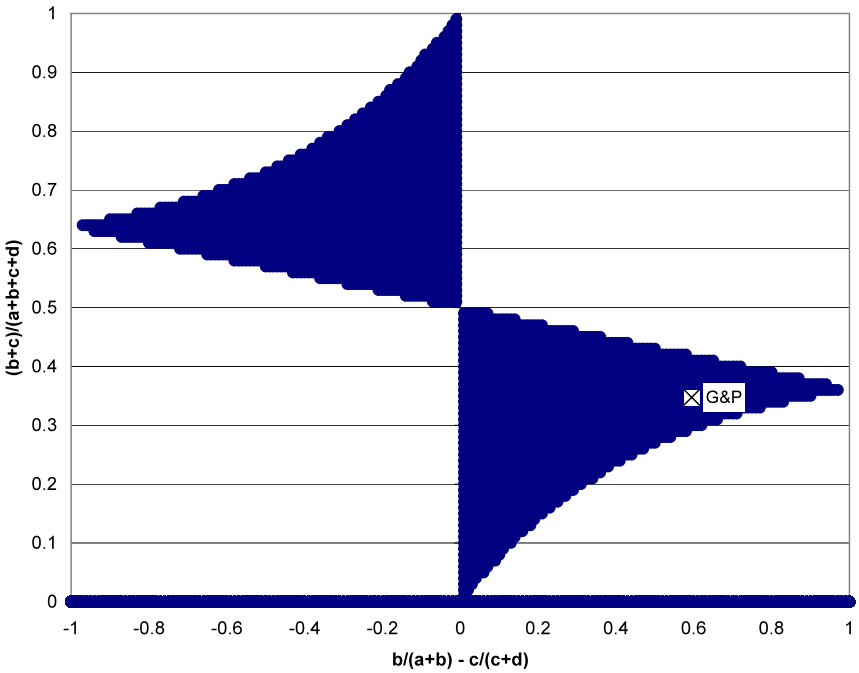

Fig. 7. Reversal configurations that can be explained by the two-error-rate model when subjects strongly prefer the $-bet according to the choice task $((a+b)/(a+b+c+d) = 0.36259542)$.

indifference between the bets) or (3) when the conditional reversal rates are roughly symmetric (vertically up the middle of the diagram, indicating that reversals for either choice are approximately the same). Simply put, when there are not strong preferences across the bet-types, asymmetric conditional reversal rates cannot be accommodated by the two-error-rate model. Instead, a model such as the anchor and adjust hypothesis of expression theory is needed to accommodate the asymmetries in conditional reversal rates.

Now consider the case where subjects have a distinct preference for one bet-type over another. Empirically, this is typically the case in preference reversal experiments *with* truth-revealing incentives. Fig. 7 shows the two-error-rate model's ability to fit the data in such a case. As in Fig. 6, we graph the asymmetry in conditional reversal rates and the overall reversal rate for a particular choice pattern. In this case, we examine subjects who prefer the $-bet on average according to the choice task $((a+b)/(a+b+c+d) = 0.36259542$, corresponding to Grether and Plott's (1979), experiment 1b). Again, the shaded area shows combinations of reversal rates and conditional asymmetries that can be explained by the two-error-rate model. Notice that the two-error-rate model can only explain data when (1) there are no reversals (the bottom axis) or (2) when the asymmetries are in the direction of error correction. Patterns of reversal that are inconsistent with simple error correction (the white areas of the graph) are inconsistent with the two-error-rate model.[27]

Two other maximization models of choice can be generated by restrictions on the two-error-rate model. First, restricting $r = s$ (Restriction IV in Fig. 2) leads to a one-error-rate model (Model 5 in Fig. 2). Though Camerer and Hogarth's (1999) capital-labor-production theory would argue that this is unlikely, we examine the one-error-rate model in order to determine whether the less restrictive two-error-rate model proposed by Lichtenstein and Slovic (1971) is necessary to explain the data. Further restricting $r = s = 0$ leads to an expected utility model without errors (Model 6 in Fig. 2). Expected utility without errors would predict no reversals at all. As a result, the predicted frequencies in cells b and c are zero. Any observations in these cells contradict the model and likelihood ratio tests are not meaningful. Indeed, it is the existence of observations in cells b and c that has led prior researchers to claim that preference reversals are inconsistent with expected utility theory. However, our modeling framework allows us to ask whether the inconsistencies are statistically significant (using χ^2 test statistics).

5.5. Summary of models

In summary, expression theory (Model 2 in Fig. 2) can explain any observed data set and is observationally equivalent to the benchmark model (Model 1) because it can always be parameterized to attain the best-fit, global maximum likelihood. The two-error-rate expected utility model is also a restricted form of the benchmark model, but in this case the restriction

[27] Graphs based on the asymmetry in reversal rates conditional on price-ranking lead to the same conclusion: the two-error-rate model can only explain data when asymmetries are consistent with an error correction explanation.

Table 4
Model estimates.

Incentive category	Data set	Benchmark/expression theory				Two-error-rate model				Task dependent preferences		
		Estimates			Log likelihood	Estimates			Log likelihood	Estimates		Log likelihood
		q	s_P	$s_\$$		q	r	s		q	s	
None	L&S1	0.51	0.83	0.06	−9.39	0.12	0.46	0.00	−19.60	0.51	0.12	−23.49
	L&S2	0.53	0.51	0.27	−11.99	0.39	0.40	0.00	−33.46	0.53	0.39	−94.72
	G&E2a	0.52	0.61	0.21	−7.04	0.31	0.42	0.00	−7.87	0.52	0.31	−9.51
	G&P1a	0.49	0.59	0.11	−7.77	0.26	0.35	0.00	−10.83	0.49	0.26	−22.44
	ON	0.52	0.59	0.21	−12.43	0.32	0.41	0.00	−49.76	0.52	0.32	−122.12
Indeterminate	L&S3	0.57	0.56	0.11	−6.17	0.30	0.37	0.00	−9.81	0.57	0.30	−11.86
	L&SLV+	0.47	0.81	0.10	−8.50	0.14	0.43	0.00	−9.35	0.47	0.14	−12.90
	L&SLV−	0.68	0.19	0.76	−8.24	0.96	0.31	0.18	−8.24	0.68	0.79	−8.72
	PSZ1.1	0.60	0.54	0.14	−7.20	0.33	0.38	0.00	−15.28	0.60	0.33	−16.61
	PSZ1.2	0.38	0.47	0.14	−7.12	0.27	0.25	0.04	−7.12	0.38	0.29	−20.51
	PSZ2.1	0.49	0.47	0.13	−7.12	0.32	0.30	0.00	−8.72	0.49	0.32	−21.37
	PSZ2.2	0.43	0.58	0.07	−6.82	0.22	0.29	0.00	−8.10	0.43	0.22	−20.40
	M&G	0.54	0.77	0.27	−5.77	0.14	0.56	0.14	−5.77	0.54	0.25	−5.84
	OIND	0.53	0.52	0.21	−10.84	0.35	0.37	0.00	−26.42	0.53	0.35	−75.47
Truth-revealing	G&P1b	0.36	0.73	0.13	−7.76	0.12	0.32	0.08	−7.76	0.36	0.18	−11.70
	G&P2SP	0.38	0.66	0.19	−7.65	0.16	0.32	0.13	−7.65	0.38	0.25	−10.53
	G&P2DE	0.38	0.75	0.15	−7.50	0.10	0.35	0.10	−7.50	0.38	0.19	−9.01
	R1.1	0.39	0.62	0.27	−8.55	0.17	0.34	0.22	−8.55	0.39	0.31	−10.58
	R1.2	0.45	0.70	0.16	−8.23	0.22	0.40	0.01	−8.23	0.45	0.22	−12.11
	BDO1.1	0.39	0.52	0.24	−7.01	0.25	0.27	0.16	−7.01	0.39	0.33	−10.87
	BDO1.2	0.32	0.61	0.15	−6.66	0.16	0.24	0.11	−6.66	0.32	0.23	−10.82
	BDO2.1	0.42	0.27	0.33	−6.90	0.42	0.00	0.31	−7.16	0.42	0.50	−15.87
	BDO2.2	0.48	0.23	0.35	−6.82	0.55	0.29	0.00	−7.28	0.48	0.55	−17.13
	C&CPs	0.45	0.51	0.32	−8.03	0.34	0.35	0.17	−8.03	0.45	0.40	−11.53
	C&CEc	0.40	0.52	0.25	−8.40	0.27	0.29	0.16	−8.40	0.40	0.34	−17.32
	SSA1	0.40	0.39	0.10	−7.26	0.31	0.22	0.00	−7.38	0.40	0.31	−35.97
	SSA2	0.50	0.38	0.13	−6.41	0.38	0.25	0.00	−7.52	0.50	0.38	−20.08
	C.MSP	0.43	0.73	0.15	−8.99	0.18	0.39	0.03	−8.99	0.43	0.20	−14.85
	OTR	0.41	0.60	0.20	−11.81	0.24	0.33	0.09	−11.81	0.41	0.28	−91.18

is meaningful. Many patterns of data are inconsistent with this model. Only those patterns representing stable preferences, random errors and error corrections can be explained by the two-error-rate model. A likelihood ratio test between it and the benchmark model is equivalent to testing it against expression theory because of the likelihood equivalence of the benchmark model and expression theory. Further restrictions result in a one-error-rate expected utility model (Model 5) and strict expected utility without errors (Model 6). As we note below, an extreme form of task dependent preferences (Model 3) is also a restricted form of the two-error-rate model. As a result, simple likelihood ratio tests can distinguish between it and the two-error-rate model.

6. Results of behavioral model estimation and comparison

6.1. Main results

Tables 4 and 5 contain our maximum likelihood results. Table 4 contains estimated parameters and log likelihoods for each model.[28] For each data set, we show the log likelihood and parameter estimates for expression theory (recall that this is equivalent in likelihood to the best-fit benchmark). Next, we show the parameter estimates and log likelihood of the two-error-rate model.[29] Then, we show the parameter estimates and log likelihood for the task dependent preferences

[28] For the task dependent preference model, we present the estimates from the derivation in Fig. 2 where we restrict r to zero and estimate q and s. With some re-labeling, this is equivalent to the other derivation in Fig. 2 where q is restricted to zero. Either method gives independent preference parameters across the choice and pricing tasks and it is these preference parameters we are estimating.

[29] Readers may wonder why the estimated pricing task error rates are lower than the estimated choice task error rates. At first blush, this seems at odds with the conventional wisdom that a paired choice task is the least error-prone method of revealing preferences. However, in the paired choice tasks, the subjects must simultaneously compare multiple attributes of the gambles. If they focus on one attribute, say the high payoff, for example, then they may not make a choice that reflects their true overall assessment of the gambles. In the language of Hogarth (1980), such decision makers are using "non-compensatory" evaluations. They are not incorporating all aspects of the gambles in their choice. However, the pricing task forces subjects to rank each gamble on a single common scale (dollar values). This may force them to consider all aspects of each gamble and use (potentially non-linear) "compensatory" evaluation to integrate the different attributes. Hogarth (1980) argues the compensatory evaluation may lead to better choices. That is, forcing the integration of attributes through the pricing task may result in rankings that more closely reflect subject preferences.

Table 5
Test statistics.

Incentive category	Data set	Likelihood ratio test statistics				χ^2 test statistics	
		(I)[a] expression vs. task dep. prefs. (1 dof)	(II)[a] expression vs. two-error-rate model (1 dof[b])	(III)[a] two-error-rate model vs. task dep. prefs. (1 dof)	(IV)[a] two-error-rate model vs. one-error-rate model (1 dof)	Task dep. pref. (1 dof[c])	Strict expected utility (1 dof[d])
None	L&S1	28.20*	20.42*	7.78*	404.96*	27.17*	868.98*
	L&S2	165.47*	42.96*	122.51*	128.91*	162.97*	2,137.73*
	G&E2a	4.95*	1.66	3.29	12.76*	4.87*	90.96*
	G&P1a	29.35*	6.13*	23.23*	35.65*	28.15*	130.16*
	ON	219.37*	74.66*	144.72*	431.85*	215.29*	3,209.87*
Indeterminate	L&S3	11.38*	7.28*	4.10*	11.85*	10.48*	49.13*
	L&SLV+	8.79*	1.70	7.09*	136.27*	8.74*	370.95*
	L&SLV−	0.96	Equal	0.96	11.79*	0.98	210.08*
	PSZ1.1	18.82*	16.15*	2.67	17.36*	17.50*	98.59*
	PSZ1.2	26.78*	Equal	26.78*	5.62*	26.93*	55.88*
	PSZ2.1	28.49*	3.19	25.30*	11.52*	27.15*	64.10*
	PSZ2.2	27.15*	2.54	24.61*	23.40*	26.27*	61.93*
	M&G	0.14	Equal	0.14	10.96*	0.14	67.96*
	OIND	129.27*	31.18*	98.10*	102.13*	125.94*	948.25*
Truth-revealing	G&P1b	7.88*	Equal	7.88*	25.49*	8.15*	139.43*
	G&P2SP	5.75*	Equal	5.75*	9.67*	5.87*	121.92*
	G&P2DE	3.02	Equal	3.02	21.74*	3.09	130.33*
	R1.1	4.05*	Equal	4.05*	5.20*	4.08*	242.25*
	R1.2	7.75*	Equal	7.75*	41.34*	7.78*	217.05*
	BDO1.1	7.72*	Equal	7.72*	1.14	7.81*	65.83*
	BDO1.2	8.31*	Equal	8.31*	3.52	8.74*	49.76*
	BDO2.1	17.95*	0.51	17.43*	1.83	17.42*	50.51*
	BDO2.2	20.61*	0.91	19.69*	1.11	19.87*	45.13*
	C&CPs	7.00*	Equal	7.00*	1.86	6.99*	152.80*
	C&CEc	17.84*	Equal	17.84*	2.96	17.95*	171.11*
	SSA1	57.43*	0.24	57.18*	8.16*	55.69*	166.10*
	SSA2	27.34*	2.22	25.12*	4.05*	25.60*	127.32*
	C.MSP	11.72*	Equal	11.72*	80.38*	11.82*	374.23*
	OTR	158.74*	Equal	158.74*	145.34*	160.23*	1,809.07*

* Significant at the 95% level of confidence under the assumption that the statistic is distributed χ^2 with the given degrees of freedom (the cutoff level for 1 degree of freedom is 3.841).

[a] The test numbers correspond to the numbered restrictions in Fig. 2.

[b] If the log likelihoods are equal, the models explain the data equally well and there is no meaningful likelihood ratio test.

[c] This is a simple Pearson's χ^2 test of independence.

[d] This is a χ^2 test of the actual cell frequencies versus the closest fit frequencies under the restrictions that $b = 0$ and $c = 0$.

model.[30] Table 5 contains the likelihood ratio and χ^2 test statistics that distinguish models. Note that occasionally a likelihood ratio cell contains the word "equal." This denotes equality in the log likelihoods for the two models and indicates that the two models explain the data equally well.[31]

To provide a more stringent test, we also fit the models by requiring a single set of parameters for each incentive condition. To do this, we aggregate across data sets within each incentive condition before fitting the models. These tests are indicated by rows at the bottom of each incentive condition. These rows are labeled ON for the Overall No-Incentives treatment data, OIND for the Overall INDeterminate incentives treatment data and OTR for the Overall Truth-Revealing incentives treatment data.

[30] Note that, while maximum likelihood always allows us to estimate the parameters of each model, one cannot interpret the estimates directly in the context of the model when the model itself is rejected. For example, we cannot make any assertions about how incentives affect error rates because the two-error-rate model is typically rejected when there are no incentives.

[31] We use χ^2 statistics because they are an obvious measure of distance from each model's achieved fit to the achieved fit of other models, including the best-fit benchmark. These statistics will be distributed according to a χ^2 distribution when testing equality restrictions in the parameter space. However, to be technically correct, the restriction of the two-error-rate model is an inequality restriction (see the discussion of Eqs. (1) through (3) above). While working out the exact distribution of the statistic in this case is beyond the scope of this paper, the χ^2 statistic remains a valid measure of distance and, to place each one in context, we will report significance levels according to the χ^2 distribution. Extensive simulation of these models leads to the same conclusions.

Result 1. Expression theory (or a weak form of task dependent evaluation) is generally necessary to explain the data without incentives.

This result is based on the likelihood ratio test statistics numbered I and II in Table 5. The statistics in the column labeled I determine whether the task dependent preferences model (Model 3 in Fig. 2) explains the data significantly worse than expression theory (Model 2). In all four cases with no monetary incentives, expression theory fits the data better than extreme task dependent preferences. This is true using the aggregated data test statistic as well. The statistics in column II show that in three of the four data sets without monetary incentives, expression theory (Model 2) explains the data significantly better than the two-error-rate model (Model 3). It fits the data significantly better in the aggregated data as well.

Result 2. Results are mixed under indeterminate incentives.

First we compare expression theory (Model 2 in Fig. 2) and task dependent preferences (Model 3) when incentives are indeterminate. According to the likelihood ratio statistics in column I, in six of eight cases (75%), the task dependent preference model fits the data significantly worse than expression theory. In the other cases, the differences are insignificant. In the aggregate data, expression theory performs significantly worse in fitting the data. Now consider the two-error-rate model's (Model 4) fit. Based on the likelihood ratio test statistics in column II, the two-error-rate model explains the data significantly worse than expression theory in two of eight cases (25%). In three cases (37.5%), the fits are identical (resulting in identical likelihoods and no meaningful likelihood ratio test). In the other three cases (37.5%), the differences are not significant. However, in the aggregated data, the two-error-rate model performs significantly worse than expression theory.

Result 3. When truth-revealing incentives exist, *no* model could do a better job of fitting the existing data than noisy maximization in the form of the two-error-rate model.

Again, this result is based on the likelihood ratio tests in Table 5. Column II shows that in ten of fourteen cases (71.4%), the two-error-rate model (Model 4 in Fig. 2) achieves the best-fit benchmark and fits the data exactly as well as the benchmark model (Model 1) or expression theory (Model 2). This means that for these cases, no model can fit the data better than the two-error-rate model. In the remaining four cases, the differences are insignificant. Thus, even in these cases, no model with the same degrees of freedom can fit the data *significantly* better than the two-error-rate model. Strikingly, when aggregated across all the data (effectively restricting subjects to common task error rates across experiments), the two-error-rate model fits as well as the best-fit model.

Result 4. The two-error-rate specification is usually necessary for explaining the data under truth-revealing incentives. Further restrictions on the two-error-rate model make it perform significantly worse in explaining the data.

Column III tests the restriction that leads to task dependent preferences.[32] In all but five cases, the two-error-rate model fits the data significantly better than task dependent preferences. Only one of these is under truth-revealing incentives. Thus, the two-error-rate model fits better than task dependent preferences, which has no errors, but allows for inconsistent preference orderings. Column IV tests the significance of the restriction that leads to the one-error-rate model. In all but the Berg et al. (1985) and Chu and Chu (1990) sessions, this is a significant restriction. In the aggregate data, the one-error-rate model performs significantly worse under each incentive treatment. Strict expected utility is always rejected. This implies that error rates that differ across tasks are necessary for explaining the data. This is consistent with Camerer and Hogarth's (1999) capital–labor–production theory and the assertion that the two tasks have different degrees of difficulty relative to the payoffs to error reduction.

Result 5. The model that best explains subject behavior changes when truth-revealing incentives are incorporated in the experiments, shifting from expression theory (or an observationally equivalent model) to a model of stable preferences with errors (i.e., noisy maximization).

This result is simply stating the significance of Results 2 through 4 jointly. Without incentives, expression theory is necessary to explain the data. With truth-revealing incentives, expected utility with errors explains the data as well. The result is obvious in the aggregated data. We test joint significance in two ways: by using χ^2-tests and by using simulation data.

To perform the χ^2-tests, begin by counting the number of instances in which the two-error-rate model performs significantly worse than expression theory in explaining the data under each incentive condition versus the instances in which it does not: 3 instances versus 1 instance under no-incentives, 2 versus 6 instances under indeterminate incentives and 0 versus 14 instances under truth-revealing incentives. Treating each data set as an observation, these frequencies form the basis

[32] This statistic makes sense because restricting $q = 0$ in the two-error-rate model (Restriction III in Fig. 2) makes it equivalent to task dependent preferences (Model 3 in Fig. 2). This restriction leads to independence of the rows and columns, though the notation differs from the restriction that gives task dependent preferences from expression theory. This nests the two models to make the likelihood ratio test valid.

Table 6
Frequencies of likelihood ratio test statistic rejection rates and χ^2 tests of incentives effects.

	Expression vs. two-error-rate	Expression vs. task dep. prefs.	Two-error-rate vs. one-error-rate	Two-error-rate vs. task dep. prefs.
None	3/4 (75%)	4/4 (100%)	4/4 (100%)	3/4 (75%)
Indeterminate	2/8 (25%) 3 equal	6/8 (75%)	8/8 (100%)	5/8 (62.5%)
Truth-revealing	0/14 (0%) 9 equal	13/14 (93%)	8/14 (67%)	13/14 (93%)
$\chi^2(2)$ tests for treatment effects[a]	11.51[b] (0.003)	2.21 (0.332)	6.69[b] (0.035)	3.12 (0.210)

[a] Pearson's χ^2 test for independence of the frequencies of rejection and the incentive category in a given column of the table.
[b] Significant at the 95% level of confidence.

for χ^2-tests of independence between the incentive condition and the likelihood that the two-error-rate model explains the data as well as any model possibly could. This test appears in Table 6 along with similar χ^2-tests for incentive effects on the ability of other models to explain the data as well as expression theory or observationally equivalent models.

There are two significant effects of incentives.[33] The two-error-rate model cannot explain the data when there are no incentives and generally explains the data as well as any model possibly could explain it when there are truth-revealing incentives. In addition, under truth-revealing incentives, the performance of the one-error-rate model improves significantly relative to the two-error-rate model. Task-dependent preferences cannot explain the data in any of the incentive conditions. The two-error-rate model nearly always explains the data better than a one-error-rate model and better than task-dependent preferences, independent of the incentive condition.

We also generate a test based on Monte Carlo simulation of data patterns in preference reversal experiments. If the cell frequencies were drawn randomly from a uniform distribution over the feasible set, expression theory achieves the best-fit benchmark likelihood for all cases, but the two-error-rate model achieves the best-fit for only 37.8% of the cases.[34] Label these cases "success" for the two-error-rate model. This forms the basis for a binomial test of the two-error-rate model's success in best-fitting the data. If the two-error-rate model were not generating the data and, instead the observed frequencies were drawn uniformly from the feasible set, the predicted number of successes for the two-error-rate model under no incentives would be $0.378 \times 4 = 1.51$ and the actual is zero. The difference is not significant. With indeterminate incentives, the predicted number of successes would be $0.378 \times 8 = 3.024$ and the actual is 3. Again, the difference is not significant. With truth-revealing incentives, the predicted number of successes would be $0.378 \times 14 = 5.30$ and the actual is 10. The p-value of the one-sided binomial test statistic is 0.011 and the two-sided is 0.012. As a result, we see that the two-error-rate model fits significantly more often than we would expect with random behavior. The difference between the frequencies under no-incentives and truth-revealing incentives is significant ($\chi^2(1) = 6.43$, p-value $= 0.011$). These results show that the underlying model consistent with behavior shifts from one that is inconsistent with noisy maximization (under no-incentives) to one that is consistent with noisy maximization (under truth-revealing incentives).

6.2. Additional evidence

Here, we ask whether relaxing our "near replication" criterion (criterion 6) would have affected our conclusions. The discussion here includes all of the papers that failed criterion 6 where sufficient data was available to indicate the potential impact of incentives. We also discuss additional experiments from papers in our main analysis where the additional experiments alone would have failed criterion 6 and one paper that failed criterion 5, but studied the impact of incentives directly. Thus, we examine the robustness of our results to expanding the data sets to include a variety of treatments that have potentially confounding effects for our analysis (though they would not have been confounding relative to the original research questions).

In attempts to create more consistent choices, Tversky et al. (1990)[35] and Cubitt et al. (2004) both change the value elicitation procedure by using an ordinal payoff scheme. In these papers, there are six otherwise comparable data sets, three with incentives and three without: Tversky et al. (1990), Study 1, Data Sets I, II, III and I$_\$$; and Cubitt et al. (2004), Monetary Valuation Data Set and Probabilistic Valuation Data Set. The ordinal valuation procedure is a significant enough design change to make it difficult to compare these data sets to those included in the main analysis. However, were we to include them it would not change our results. Reversal rates range from 37% to 51% and show no apparent effect of incentives. However, as in the main analysis, the choices and valuations are more coherent under incentives than not (the average

[33] The statistic on the two-error-rate model versus the one-error-rate model is marginally significant. If one used only the no-incentives and truth-revealing incentives data, the only significant statistic is the one comparing expression theory and the two-error-rate model.
[34] This percentage does not change if you restrict the space to cases where there are more consistent than inconsistent price/choice combinations ($a + d > b + c$), which characterizes most of the experiments.
[35] We note that the expected values of the bets in Tversky et al. (1990) also differed somewhat more than in typical preference reversal experiments.

absolute difference in P-bet preference measures according to the two tasks was 0.40 without incentives and 0.24 with incentives). The reversal rates conditional on choice remain asymmetric with much lower reversal rates observed conditional on choosing the $-bet (13.5% overall) than conditional on choosing the P-bet (60% overall). The absolute difference in reversal rates conditional on their valuation tasks (their replacement for the pricing tasks in the main analysis) is much lower under incentives than not (13% versus 60%) and much lower than the difference conditional on the choice task when subjects are paid (13% versus 47%). The two-error-rate model performs significantly worse than expression theory for all three of the data sets without incentives, but only in one of the three with incentives.[36]

In a different attempt to create more consistent choices, Berg et al. (1985) present four additional data sets where subjects were paid in a truth-revealing manner. However, between two sequential sessions, the researchers arbitraged subjects who exhibited inconsistencies in prices and choices. The subjects knew that the experimenters would engage in transactions with subjects between the two sessions. They did not know in advance that the only transactions would result from arbitrage opportunities where the subjects lost money for inconsistencies. This does not change the truth-revealing nature of the incentives, but we thought this was a significant enough design change that we included neither the pre- nor post-arbitrage data in our main analysis. However, the data is consistent with our main analysis. The pattern of reversal and conditional reversal rates mirrors the pattern in other truth-revealing-incentives sessions. In three of the four data sets, the two-error-rate model explains the data as well as the best-fit benchmark. In the fourth, it is not significantly different.

In an attempt to equalize the choice frequency across gamble types, Casey (1991, 1994) created significant differences in the expected values of the bets in each pair. In these papers, there are four otherwise comparable data sets, two with incentives and two without: Casey (1991) experiment 2, large bets and experiment 2, small bets; and Casey (1994) large stakes bets and small stakes bets. Again, this is a significant enough design change to make it difficult to compare these data sets to those included in the main analysis. In fact, if it creates a clear difference between the gambles in the eyes of the subjects, the two-error-rate model should perform better regardless of incentives. Indeed, the two-error-rate model achieves the best fit in three of the four data sets and is not significantly worse in the fourth. This is consistent with the two-error-rate model explaining the data when there are clear preferences and a difference in expected values generating clear preferences with or without incentives. It also accords well with the Berg et al. (2003) attempt to create a clear preference across gambles by inducing risk preferences, which we discuss shortly.

MacDonald and Huth (1989) use monetary payments and a series of auctions to determine valuations of gambles. For each gamble, the subjects engage in at least 2 sequential $n+1$ auctions to determine whether or not they sold the gambles. After the first round of bidding, the researchers gave the subjects feedback about the auction and required at least one more round of bidding before closing the auction. After the second round, subjects received feedback and voted on whether to make the second round binding. Unanimity was required. If the vote was not unanimous, they had a third round, fourth round, or more until the subjects voted to make a round binding (or a limit on the number of rounds was reached). While bids would be truth-revealing if a single $n+1$ auction was used, the potential for inter-period gaming may disrupt the (otherwise) truth-revealing nature. As a result, we do not include this in the main analysis. However, the results would support our conclusion. The two-error-rate model explains the data as well as the best-fit benchmark whether the initial auction bids or final auction bids are used to measure values for subjects.

In attempts to control preferences across gambles, Selten et al. (1999) and Berg et al. (2003) use binary lottery payoff mechanisms intended to induce risk preferences. In these papers, there are five otherwise comparable data sets, all with truth-revealing monetary incentives: Selten et al. (1999) binary payoff data sets with and without feedback statistics (both inducing risk neutrality) and Berg et al. (2003) data sets with induced risk aversion, risk neutrality and risk seeking preferences. Again, this is a significant enough design change to make it difficult to compare these data sets to those included in the main analysis. However, again, the preference and reversal patterns mirror those in the main analysis. The two-error-rate model performs well in these data sets. It achieves the best fit in three of the five data sets and is not significantly different in a fourth. The lone failure arises when subjects are induced to be risk neutral and, hence, indifferent between the gambles (Selten et al., 1999, binary payoffs with feedback statistics). If subjects are truly indifferent, any pattern of responses is optimal. The data are always consistent with the two-error-rate model when there are clear preferences, in this case induced by the binary lottery procedure.

Finally, one paper that fails criterion 5 also sheds light on the impact of incentives. Ordonez et al. (1995) present gambles simultaneously for choice and pricing tasks, allowing subjects to revise responses if desired, so that choices and prices are not obtained independently as required by our criterion 5. However, they also manipulate incentives directly. Without incentives, their treatment had little effect on reversals relative to typical rates in preference reversal experiments without incentives. When they used monetary incentives, choice inconsistencies declined significantly.[37]

Overall, the additional evidence is strongly consistent with the idea that incentives or other mechanisms can create clear and consistent preferences across the gambles. When this is the case, the two-error-rate model of consistent preferences with error generally does as well at explaining the data as any model could.

[36] We note that Cox and Grether (1996) also use an ordinal payoff scheme. While their data is not presented in a format that allows estimation of the two-error-rate model, they do show that the reversal rate goes down as subjects engage in repeated auctions to value the gambles.

[37] This is consistent with our argument. However, the data was not presented in a format that allows estimation of the two-error-rate model.

6.3. Notes on subject self-selection

Our results show previously undocumented effects on outcomes in preference reversal experiments when classifying the incentive structure of the experiments as a treatment variable. Under truth-revealing incentives, subject choices are more closely aligned than without such incentives. There are clear effects on the patterns of reversals. Finally, as we show in the prior section, the incentive treatment changes whether the two-error-rate model fits the data.

However, there is a potential confound for our analysis: contrasts in this study do not arise specifically from subjects being randomly assigned to three incentive conditions. In fact, subjects made a sequence of choices that led them to participate in a particular experiment and incentive condition. There are a number of factors that may lead to self-selection of subjects to incentive treatments:

1. Different types of students may choose different types of classes. As a result, recruiting from different types of classes or subject pools (e.g., psychology versus economics classes or pools) may result in different types of subjects in experimental sessions.
2. The culture of research in a discipline (e.g., psychology versus economics) may lead potential subjects to expect different payment mechanisms. As a result, different types of subjects may come to psychology versus economics experiments.
3. Knowledge of the incentive scheme used (paid versus unpaid or contingent on performance versus not) may affect the types of subjects who show up for a session.

To address self-selection issues, we asked the researchers who conducted the studies referenced in Table 2 to provide us specific information about their recruiting procedures and information conveyed to subjects about incentive mechanisms used. We received responses from at least one researcher on all but one paper.

Two pieces of evidence allow us to address partially the issue of the type of subjects in the recruiting pools. First, with one notable exception, all researchers used a mixed subject pool. As a result, differences across experimental sessions cannot arise from the fact that some researchers used economics students and others psychology. The exception was Chu and Chu (1990), leading to the second piece of evidence. Their "Group A" subjects were drawn from psychology classes, while "Group B" subjects were drawn from economics classes. Table 3 through Table 5 show no significant differences between their groups. If the subject pool alone explained the differences, we would expect differences to arise across Groups A and B.

Two pieces of evidence allow us to address partially the issues of the cultures of research across the disciplines. First, Grether and Plott (1979), both trained in economics, have sessions run under no-incentives and truth-revealing-incentives conditions. Similarly, Lichtenstein and Slovic (1971, 1973), both trained in psychology, have sessions under no-incentives and indeterminate-incentives conditions. The pattern in this subset of data mirrors the pattern in the overall data. To get a more complete picture, Table 7 generates test statistics similar to those in Table 6. However, here, we use the disciplines of the researchers as the treatment variable to control for any potential discipline specific expectations on the parts of subjects. The significant effects found in Table 6 disappear when we divide the data according to the disciplines of the researchers.

Finally, three pieces of evidence allow us to address partially whether knowledge of the payment scheme being used drives the results. First, while one researcher did not respond to our inquiry, all other researchers responded that they told subjects during recruiting that they would be paid (no-incentives treatments paid flat fees to subjects). So, being paid by itself cannot be causing the significant differences we observe. Second, Grether and Plott (1979) recruited all subjects in exactly the same way. Only after arriving were the subjects randomly spit into incentive treatments. The pattern in this subset of data mirrors the pattern in the overall data. Finally, to get a more complete picture, Table 8 generates test statistics similar to those in Table 6. However, here, we use as the treatment condition whether the subjects were told in advance there would be performance based (contingent) payoffs.[38] The significant effects found in Table 6 disappear when we divide the data according to subjects' advance knowledge of the payoff scheme.

7. Conclusions

Decades after it was first documented, preference reversal continues to be a topic of research and debate. The consistency of preferences, or lack thereof, lies at the core of decision theory and matters in applications as well. For example, if institutions and preferences are interdependent (as Tversky et al., 1990, suggest), then the institutions we design for allocating resources (e.g., sealed bid auctions versus oral double auctions) actually feed back into the preference orderings of the participants. This would complicate greatly optimal institutional design, if not make it impossible. In contrast, if economic agents have stable preference orderings, and observed anomalies such as preference reversals are merely the result of simple errors, then designing efficient institutions becomes simpler.

We conduct a literature search covering more than 35 years of published papers, to find near replications of Lichtenstein and Slovic's (1971) original preference reversal experiments where the main treatment difference is the type of incentives used in the experiments. The result of this search was a large number of papers, but surprisingly few that replicated or

[38] While this is not documented in any of the papers, we classify the sessions using the best recollections of the researchers according to our private communication with them.

Table 7
Frequencies of likelihood ratio test statistic rejection rates and χ^2 tests of discipline effects.

Disciplines of researchers	Expression vs. two-error-rate	Expression vs. task dep. prefs.	Two-error-rate vs. one-error-rate	Two-error-rate vs. task dep. prefs.
Psychology	3/7 (43%)	6/7 (86%)	6/7 (86%)	5/7 (71%)
Other	0/2 (0%)	1/2 (50%)	2/2 (100%)	1/2 (50%)
Economics	2/17 (12%)	16/17 (94%)	12/17 (71%)	15/17 (88%)
$\chi^2(2)$ tests for treatment effects[a]	3.60 (0.165)	3.48 (0.175)	1.29 (0.525)	0.22 (0.329)

[a] Pearson's χ^2 test for independence of the frequencies of rejection and the incentive category in a given column of the table.
[b] Significant at the 95% level of confidence.

Table 8
Frequencies of likelihood ratio test statistic rejection rates and χ^2 tests of recruiting effects.

Information about incentives that was communicated to subjects during recruiting	Expression vs. two-error-rate	Expression vs. task dep. prefs.	Two-error-rate vs. one-error-rate	Two-error-rate vs. task dep. prefs.
Contingent incentives NOT communicated during recruiting	4/11 (36%)	10/11 (91%)	11/11 (100%)	8/11 (73%)
Unknown or unsure	0/2 (0%)	1/2 (50%)	2/2 (100%)	1/2 (50%)
Contingent incentives communicated during recruiting	1/13 (8%)	12/13 (92%)	7/13 (54%)	12/13 (92%)
$\chi^2(2)$ tests for treatment effects[a]	3.67 (0.160)	3.15 (0.207)	7.80[b] (0.020)	2.79 (0.248)

[a] Pearson's χ^2 test for independence of the frequencies of rejection and the incentive category in a given column of the table.
[b] Significant at the 95% level of confidence.

nearly replicated the original research. From a scientific point of view, this is disturbing. Replication is a cornerstone of the scientific method.

Though the number of resulting data sets is small, we are still able to show a clear effect of incentives on the aggregate pattern of data in preference reversal experiments. Without incentives, choices are inconsistent across the two tasks. However, under truth-revealing incentives, decisions in the choice task and pricing task outcomes become more aligned, revealing more consistency across preference orderings. The pattern of conditional reversal rates accords with what we call noisy maximization: subjects have consistent preferences, but if they do make random errors, they are likely to be corrected as subjects switch between tasks.

With truth-revealing incentives, the overall pattern of behavior is consistent with risk seeking subjects who generally prefer the \$-bet (that is, $(a+b)/(a+b+c+d) \leqslant 50\%$ in all cases and $(a+c)/(a+b+c+d) \leqslant 50\%$ in all but one case). If a risk seeking subject either chooses the \$-bet or prices the \$-bet higher, then conditional reversal rates should be relatively low. In fact, such reversals occur infrequently in the data under truth-revealing incentives. In contrast, if a risk seeking subject chooses the P-bet or prices it higher, it is inconsistent with the subject's underlying preferences and we should see conditional reversal rates that are relatively high, consistent with an error correction interpretation. Again, this holds in the data under truth-revealing incentives. Under such incentives, the reversal rate conditional on choosing the P-bet is always higher than the rate conditional on choosing the \$-bet. Similarly, the reversal rate is higher after pricing the P-bet higher in all but two cases.

Using aggregate level data, we test several existing behavioral models in the preference reversal literature.[39] Our results suggest that the underlying model describing behavior changes when the incentive environment changes. In experiments using purely hypothetical choices, the data strongly contradict noisy maximization models such as expected utility with error. Only models where expressions of preferences or actual preferences depend on the task can explain the data. However, with truth-revealing incentives, an expected-utility-with-error model not only fits the data, it fits the data as well as *any*

[39] All of our results are based on the available aggregate data. Individual data is not available for all data sets we include in the analysis. As a result, any individual analysis we could do on the available data is subject to a data selection criticism. We have taken great pains in the selection of the data sets we use in our analysis to avoid any such selection biases and choose not to undermine this with individual analysis on selected data sets. Nevertheless, we are confident that additional work using individual data will help shed more light on these issues. For example, Butler and Loomes (2007) explore whether imprecision in preference is associated with the range of a gamble's payoffs. Though their model does not accommodate incentive effects, some hybrid of their model and ours may help to understand whether reversal rates are likely to differ systematically with the payoff attributes of a gamble pair.

other model possibly could. Task dependent evaluation of gambles is not needed to explain the data when truth-revealing incentives exist.[40]

The procedures we use for model testing may be useful in other areas of experimental economics. In conducting our analysis, we develop a "best-fit" benchmark and test alternative models of decision making against this best-fit. This method can prove useful in many experimental contexts where data is fit into categories and, as a result, can be described by multinomial distributions.

In addition to our main analysis, we present data from a number of papers where the designs change significantly from the original Lichtenstein and Slovic (1971) design. Nevertheless, the results are largely consistent with the main analysis: incentives or other means of creating a clearer preference across the gambles create more consistent responses.

While compelling, our results are not without qualifications. First, the results indicate a complex relationship between incentives and outcomes.[41] Researchers originally anticipated that incentives would reduce reversal rates. By and large, they do not. But, they do create a pattern of responses consistent with noisy maximization. We are not the first to point out the complex relationship between incentives and outcomes. Our research adds to that documented in Camerer and Hogarth (1999). Here, results are consistent with incentives changing the underlying model of behavior, but not eliminating errors entirely.[42] Our results also highlight the importance of the specific nature of incentives. Incentives for truthful revelation are important.[43] The complexity and importance of the issues argues for more research on the interactions of incentives and behavior.

Second, there are potential confounds in the data. For example, one might argue that the disciplines of the researchers or subject self-selection into incentive treatments may produce the results. While the available data does not allow a complete analysis, it leans against such explanations. Only new research can answer these questions fully. While beyond the scope of the current paper, such research would build on the results of Dohmen and Falk (2006) and Cadsby et al. (2007). Both groups study the payment schemes that subjects select, but not specifically whether subjects select to come to an experiment based on the payment scheme used. The ideal design would be to have treatments in which subjects must participate in a given incentive scheme and another treatment where they choose the incentive scheme. Then one could assess the change in performance resulting from being forced to participate in a less preferred system (i.e. the decrement when there is no self-selection). Assessing risk preferences and abilities prior to the study would make it possible to measure how subjects who self-select into different payment schemes differ from each other. Such a result could begin to shed light on the importance of subject self-selection in interpreting experimental results.

Finally, while our research shows a clear change in response patterns when truth-revealing incentives are implemented, it says little about why this occurs. One might simply argue that truth-revealing incentives make the decisions real, resulting in more "rational" choices. However, we believe that the fundamental issues are complex and require a much more thorough understanding. There is some research in the area. Using a process tracing methodology, Schkade and Johnson (1989) document that structural differences in the preference reversal tasks themselves affect the information processing strategies used by subjects. Direct studies of neurological data show that slight changes in context can dramatically alter the apparent behavioral processes that people invoke.[44] In contexts similar to the experiments here, Dickhaut et al. (2002) argue that small changes in environment invoke major changes in the way the brain seems to function in choice studies. It seems to us that additional research should start from the common observation of psychology, neural science and economics that structural differences affect the way stimuli are perceived and processed and how decisions are made.

Appendix A. Papers eliminated by criteria 4 and 5 from Table 1

A.1. Papers eliminated by criterion 4

Our criterion 4 requires that at least one experiment in the paper contains two-prize gambles for real or hypothetical money. The following papers do not meet that criterion. The objects evaluated in each paper are described so that it is easy to see why the objects are not two-prize gambles for real or hypothetical money.

[40] This overall incentive effect is consistent with induced value theory (Smith, 1976) and Camerer and Hogarth's (1999) capital–labor–production theory. While preference reversal is an individual level phenomenon, our results also are consistent with research on multiplayer games that suggests background noise drives observed anomalies (for examples, see Goeree and Holt, 2005, and Palfrey and Prisbrey, 1996, 1997).

[41] As evidence of the complexity of the relationships we have uncovered, we note that several authors who added incentives to preference reversal experiments (including two of the authors of this paper) did not recognize the result reported here in the original reports of their data.

[42] Harrison (1989, 1994) suggests that the level of incentives affects behavior and incentives that are too low may not result in behavior that appears economically rational.

[43] Recently Luce (2000) pointed to Lichtenstein and Slovic's (1971) failure to find an incentive effect in their casino experiments as an example of incentives not mattering in experiments. However, our results suggest that the failure comes from the fact that Lichtenstein and Slovic's (1971) incentives were not necessarily truth-revealing.

[44] For example, consider the Stroop task in which the word red is written in green or red. In cases when the color is matched with the word and the subject is asked to read the word, the reaction time is much less than when the word is not written in its color. Pardo et al. (1990) show that there are not just simple changes in reaction time, there are wide changes in brain behavior overall.

	Paper	Objects evaluated
1.	Bazerman et al. (1992)	Pairs of payoffs consisting of a fixed payoff to self and a fixed payoff to another person, for example $600 for self and $800 for the other person.
2.	Bohm (1994a)	Cars.
3.	Bohm (1994b)	Fixed payoffs received after fixed amounts of time, for example $200 in three months.
4.	Camacho-Cuena et al. (2005)	Multi-outcome gambles or multi-dimensional income distributions each with at least 10 outcomes.
5.	Chapman and Johnson (1995)	Consumer commodities (for example, a one day vacation in Bermuda), health-related items (for example, a treatment that would result in 20/20 vision), or filler items (for example, your house will always be neat and clean).
6.	Colombo et al. (2002)	Skiing holidays, restaurants, movies, cars, apartments, swimming pools.
7.	DeNeufville and Smith (1994)	Construction bidding situations (construction management versus lump sum bid projects).
8.	Ganzach (1996)	Multi-outcome gambles each with 5 equal probability payoffs.
9.	Gonzalez-Vallejo and Moran (2001)	Candidates for computer programmer job, television sets.
10.	Green et al. (1994)	Fixed payoffs received after fixed amounts of time, for example $20 now or $50 in one week.
11.	Hatfield and Seiver (2001)	Restaurants with various health inspection grades.
12.	Hawkins (1994)	Apartments.
13.	Hsee (1996)	Music dictionaries, candidates for computer programmer job, TVs, compact disk changers.
14.	Irwin (1994)	Environmental situations (for example, air quality), consumer goods (for example, bicycles).
15.	Irwin and Davis (1995)	Job candidates (nurses, engineers, sanitation managers).
16.	Irwin et al. (1993)	Air quality improvements, consumer product improvements (for example, a camera with more features).
17.	Kirby and Herrnstein (1995)	Fixed payoffs received after fixed amounts of time (for example $12 in two days or $16 in ten days), consumer goods received after fixed amounts of time (for example, a Sony radio in two days).
18.	List (2002)	Baseball cards.
19.	Maher and Kashima (1997)	Three-outcome gambles with unknown probabilities for two of the outcomes (Ellsberg choices).
20.	Nowlis and Simonson (1997)	Consumer products (for example, Sony TV).
21.	Oliver (2005)	Life expectancy.
22.	Ranyard (1995)	Simple and compound gambles with three or four outcomes.
23.	Schmeltzer et al. (2004)	Four-outcome gambles.
24.	Schmidt and Hey (2004)	Multi-outcome gambles with at least three prizes each.
25.	Seiver and Hatfield (2002)	Restaurants with various health inspection grades.
26.	Selart et al. (1999)	Medical treatment options.
27.	Stalmeier et al. (1997)	Personal health outcomes, for example living with a migraine 4 days a week for 20 years.
28.	Sumner and Nease (2001)	Personal health outcomes (for example, live 20 years with migraines 4 days per month), altruistic tissue donation (for example, friend lives 1 year and 1/1000 chance you die).
29.	Tan et al. (1993)	No objects evaluated; paper is a neurological study of left handedness.
30.	Tornblom (1982)	Distributive injustice situations demonstrated by charts of worker payoff entitlements and outcomes.
31.	Waters and Collins (1984)	New product options in a simulated market setting.
32.	Wicklund (1970)	Consumer items, for example shaving bag and desk lamp.
33.	Wong and Kwong (2005)	HiFi systems, compact disk changers, job candidates for a computer programming job, airline travel.
34.	Zapotocna (1986)	No objects evaluated; paper is a neurological study of hand–eye coordination and mirror images.
35.	Zikmund-Fisher et al. (2004)	Medical care providers, for example, doctors performing laser eye surgery.

A.2. Papers eliminated by criterion 5

Our criterion 5 requires that an experiment that has satisfied criteria 1 through 4 also be a classic preference reversal task — a single paired-choice task where subjects see all of the parameters of both bets simultaneously and individual pricing tasks where scalars are chosen for the bets' "prices" or certainty equivalents. The following papers do not meet that criterion. The non-compliant preference reversal task(s) is(are) described so that it is easy to see why the paper fails this criterion. Occasionally, a paper consists of several treatments rejected for different reasons. In those cases, separate descriptions are provided for each treatment.

	Paper	Non-classic choice task	Non-classic pricing task
1.	Bohm and Lind (1993)	Iterative choice task. First choice is from a triple (a fixed prize option is included), subsequent choice is between the two remaining options.	Pricing task is multi-player. Subjects participate in a market, buying or selling based on market clearing price.
2.	Bostic et al. (1990)	Each gamble compared to every other gamble. Each gamble pair presented twice. It is not clear how this is handled in the data analysis.	Not a single pricing task. Subjects stated an indifference point and participated in an iterative choice between the gamble and fixed amount.
3.	Cox and Epstein (1989)		No individual pricing task — both gambles in a pair priced simultaneously. Description of gambles (prizes) differ from descriptions in paired choice.
4.	Johnson et al. (1988)	No paired choice task.	
5.	Li (2006) — Experiment 2 (Experiment 1 was eliminated previously)		Gambles not priced, instead individual components of the gambles are priced.
6.	Li (1994)		
	Group 1		No individual pricing task. Both gambles in a pair are priced simultaneously.
	Group 2		No pricing task.
	Group 3	No paired-choice task.	
7.	Loomes (1990)		
	Experiment 1		Subjects experience two valuation methods. One is the standard BDM task, but the other includes iterative choice thus mixing choice and pricing. Presentation format of bets changes between choice (three distinct payoff areas) and pricing tasks (two distinct payoff areas).
	Experiment 2		Iterative pricing task (mixes choice and pricing).
	Experiment 3	No paired-choice task.	
8.	Loomes et al. (1989)		
	Experiment 1		No pricing task.
	Experiment 2		No pricing task.
	Experiment 3	Bets in pair have nearly the same variances, so neither is a "P-bet" or a "$-bet" as in the classic preference reversal experiment.	
9.	Mellers et al. (1992)	No paired choice task.	
10.	Ordonez et al. (1995)		No individual pricing task. Gambles priced simultaneously. Choices and prices on same page and subjects can revise their decisions. (Note: this paper is discussed in the additional evidence section.)
11.	Schkade and Johnson (1989)[a]	Not all attributes of the gamble were displayed at the same time (to facilitate process tracing method). Some subjects might not see all attributes of the gamble before making a choice.	Not all attributes of the gamble were displayed at the same time (to facilitate process tracing method). Some subjects might not see all attributes of the gamble before pricing.
12.	Wedell (1991)		
	Experiment 1	Choices are triads, not pairs.	No pricing task.
	Experiment 2	Choices are triads, not pairs.	No pricing task.
	Experiment 3	Eliminated by Criterion 4: Not monetary bets; objects are cars, restaurants, TV sets.	
13.	Wedell and Bockenholt (1990)		
	Experiment 1		Both bets in a pair priced simultaneously. Four bet pairs to a page. In half the cases, "bets" consist of 10 plays of the bet.
	Experiment 2[b]		Both bets in a pair priced simultaneously. Four bet pairs to a page. In two-thirds of the cases, "bets" consist of either 10 or 100 plays of the bet.

[a] We discuss this paper in the conclusions.
[b] Bets in a pair have intentionally different expected values.

Appendix B. Categorization of preference reversal experiments by incentive types

B.1. No-incentives experiments

"No-incentives" experiments are those where rewards are independent of subjects' decisions. Falling in this category are: Lichtenstein and Slovic (1971) experiments 1 and 2, Goldstein and Einhorn (1987) experiment 2a, and Grether and Plott (1979) experiment 1a. The design for these experiments is simple. Each subject arrives and participates in a number of paired choice tasks and a number of pricing tasks. They are paid for participation, independent of their actions. Each experiment yields one data set for our analysis. We label the data sets L&S1, L&S2, G&E2a, and G&P1a, respectively.

B.2. Indeterminate-incentives experiments

Because of design choices in these experiments, one cannot unequivocally assert that incentives exist for an expected utility maximizing subject to truthfully report preferences. The reward structures may give utility maximizing subjects some preference over gambles, but the reward structures could lead to systematic misreporting of preferences for some utility functions. These experiments include Lichtenstein and Slovic (1971) experiment 3, two Lichtenstein and Slovic (1973) experiments conducted in Las Vegas, four experiments from Pommerehne et al. (1982) and the "warm up" exercises from Mowen and Gentry (1980).

Lichtenstein and Slovic (1971) experiment 3 yields one data set, which we label L&S3. They use the following design:

Subject chooses between a pair of gambles (6 pairs)	→	Subject states a selling price for a gamble (12 gambles)	→	Subject rewarded for each choice made

Payoffs were stated in points with points converted to dollars using a "conversion curve." The curve guaranteed a minimum dollar payoff of \$0.80 (even with negative points) and a maximum payoff of \$8.00. Before making any decisions, subjects were informed of the reward process and minimum and maximum amounts to win, but not the actual conversion curve.

All decisions were played at the end of the experiment. For each paired choice task, the subject played the gamble indicated as preferred and received points according to the outcome. If the decision involved a pricing decision, then a "counter offer" was randomly selected from a bet-specific distribution. These distributions were not disclosed at the beginning of the experiment. If the "counteroffer" was greater than the price that the subject stated, then the subject received points in the amount of the counteroffer. If the counteroffer was less than the number the subject stated, then the subject played the bet and received points according to its outcome.

Without prior specification of the conversion curve, subjects cannot determine which bet is the expected utility maximizing choice in the paired choice task, or what price response is the expected utility maximizing response in the pricing task. Even if subjects assume a linear increasing conversion curve, wealth effects could result in economically rational reversals of preference. Such rational reversals violate the assumption of a stable underlying preference in the two-error-rate analysis. Because the tasks are taken in a particular order, wealth effects could easily make subject choices appear task dependent, when in fact these differences are due to wealth effects.

Lichtenstein and Slovic (1973) report two studies conducted at the Four Queens Hotel and Casino in Las Vegas. Subjects in the experiments purchased chips with their own money. They chose to play with either \$0.05, \$0.10, or \$0.25 chips for the duration of the session. Subjects made decisions about 10 pairs of positive expected value bets and 10 pairs of negative expected value bets. We group the data into two data sets based on this distinction. L&SLV+ denotes the set of positive expected value bets and L&SLV− denotes the set of negative expected value bets. The experiment proceeded as follows:

Stage 1 (20 Paired Choices)				Stage 2 (40 Selling Price Decisions)		
Subject chooses one bet from each of two pairs	→	Chosen bets are played	→	Subject reports selling price for a bet	→	Counteroffer generated and bet played according to the counteroffer

Subjects could drop out at any time. Data is reported for 53 completed sessions (from 44 different subjects).

In these experiments, the conversion of points to dollars is known prior to making any choices. However, subjects play bets based on their decisions immediately following each decision. This could result in wealth effects and economically rational reversals. Such economically rational reversals violate the two-error-rate model's assumption of a constant underlying preference. Thus, the model again makes no clear prediction. Again, the sequencing of tasks can make wealth effects appear as task dependent preferences.

The Pommerehne et al. (1982) experiments give four data sets, which we label PSZ1.1, PSZ1.2, PSZ2.1, and PSZ2.2. In these experiments, subjects are rewarded a pro rata share of a fixed reward. Their reward depends on their own decisions

as well as decisions of others in the experiment. There is no dominant strategy in such a reward scheme, so subjects have no clear incentive to truthfully report their preferences. For example, consider a two-person experiment where each person makes one paired choice decision and then based on the outcomes of their decisions, each person receives a pro rata share of $10.00. Let the paired choice be between two gambles, (0.70, $10; 0.30, $1) and (0.30, $17.33; 0.70, $3), where the four-tuple is read as (the probability of winning the high prize, the point amount of the high prize; the probability of winning the low prize, the point amount of the low prize). A risk averse player with $U(x) = x^{0.5}$ does not have a dominant strategy: he prefers bet 1 if the other player chooses bet 1 and prefers bet 2 if the other player chooses bet 2. Thus, the payoff scheme adds a significant amount of complexity to the reward structure. The level of complexity added depends on the task. (In particular, the pricing task comparison involves comparing two gambles to two prices when the values of the stated prices are uncertain.) Thus, because the uncertainty faced in the two comparisons differs, utility-maximizing subjects could easily make inconsistent choices in a manner that appears as though they have task dependent preferences.

In Mowen and Gentry (1980), the subjects state prices for both bets in a pair. The outcome is determined only by the higher priced bet in the pair. A random offer was generated from a distribution around the bet's expected value. If the random offer was higher than the subject's price, the subject sold the bet at the subject's stated price. This breaks the incentive compatibility of the Becker et al. (1964) procedure, which relies on a dominant revelation strategy based on the price being that of the offer, not the subject's own stated price. Effectively, this turns the Becker et al. (1964) procedure into the analog of a first-price, instead of a second-price, auction. In addition, the subjects played the bets immediately, resulting in wealth effects.

B.3. Truth-revealing-incentives experiments

We report on twelve experiments in this category: three from Grether and Plott (1979), two from Reilly (1982), two from Selten et al. (1999), four from Berg et al. (1985), the Reilly replication in Chu and Chu (1990), two from Selten et al. (1999) and one of the data sets in Chai (2005).

Grether and Plott (1979) experiment 1b (data set denoted G&P1b) has the following sequence:

Subject chooses between a pair of gambles (3 pairs)	→	Subject states a selling price for a gamble (12 gambles)	→	Subject chooses between a pair of gambles (3 pairs)	→	One decision is randomly chosen and subject plays the gamble

Thus each subject makes 18 decisions, and each decision has a 1/18 chance of being selected and played at the end of the experiment.

If a choice task is selected at the end of the experiment, then the subject plays the gamble indicated as preferred. If a pricing task decision is selected, a random number is generated from a (commonly known) uniform distribution on the interval [0, 999]. If that number is greater than the subject's stated selling price (in cents), the subject receives the random amount. Otherwise, the subject plays the gamble. This procedure is referred to as the Becker et al. (1964) procedure.

The subjects here have incentives for truthful revelation. If they select the less preferred gamble in a paired choice task and it is selected, they lose the difference in expected utilities between the two choices. Similarly, in the pricing tasks, the Becker et al. (1964) procedure makes it a dominant strategy to reveal true values.

Grether and Plott (1979) experiments 2a and 2b test whether strategic considerations in the pricing task are responsible for preference reversal. Beyond the 18 decisions described above, each subject is also required to state a dollar equivalent for each of the 12 gambles. This task is structurally identical to the pricing task — the sole difference is in wording (no reference is made to selling price). We separate their data into two data sets based on this design feature, denoting the selling price data set G&P2SP and dollar equivalent data set G&P2DE. At the end of the experiment, one of the 30 decisions is chosen, and the subject is rewarded based on the decision chosen. Paired choice and pricing tasks are rewarded as described above. The dollar equivalent task is rewarded using the same procedure as the pricing task.

Reilly (1982) gives two data sets (Stage 1, groups 1 and 2, denoted R1.1 and R1.2) intended to replicate Grether and Plott (1979). The only difference between Group 1 and Group 2 is that in Group 2 subjects receive information about the expected value of each bet and the meaning of expected value. However, the reward mechanism used in these two experiments differed slightly from that used in Grether and Plott. Rather than choosing one of the 18 decisions at the end of the experiment and rewarding based on that decision, Reilly chooses one of the six paired choices at random and then the subject is rewarded using the pricing decision associated with the preferred bet in the paired choice. In this reward scheme, subjects always have an incentive to truthfully report both prices and preferences.

Chu and Chu (1990) include replications of the Reilly experiment that are run separately on subjects recruited from psychology classes and subjects recruited from economics classes (data sets C&CPs and C&CEc). Although the subject pools differ from that of Reilly, the procedures and reward mechanisms used are identical to those of Reilly.

Berg et al. (1985) replicate and extend the Grether and Plott (1979) design. They present four experiments, each in two stages. In a two-by-two design, they vary the pricing procedure and whether the experimenter exploits preference reversal

driven arbitrage opportunities. Based on our selection criteria, we include only the no-arbitrage experiments in our data set. Each experiment consists of two parts: an initial set of 18 decisions (6 paired choices and 12 pricing decisions following the same format as Grether and Plott) and a within-subjects replication consisting of 18 decisions using a different set of bets. We denote these data sets by experiment and part (e.g., BDO1.1 and BDO1.2).

Selten et al. (1999) contrast monetary incentives alone to risk-neutral preferences induced by the Berg et al. (1986) procedure. They run two sessions with direct monetary payoffs that include the Becker et al. (1964) procedure, one where subjects were able to get statistical measures on the gambles by request and one where they could not. We denote these data sets by SSA1 and SSA2.

Finally, we include a data set found in Chai (2005). Chai has subjects choose between pairs of gambles, then elicits rankings through four other methods. One of these, Chai's "Minimum Selling Price" treatment, retains all of the important features of the Becker et al. (1964) procedure. We use the direct comparison choices and rankings according to this procedure as a data set and denote it by C.MSP. We are forced to choose one of the four ranking methods for analysis because there is only one choice for each bet pair and, hence, the reversal data for other rankings are not independent data sets. We chose this ranking in particular because it most closely mirrors the Becker et al. (1964) procedure use in the other papers included in our analysis.

References

Ballinger, T.P., Wilcox, N.T., 1997. Decisions, error and heterogeneity. Econ. J. 107 (443), 1090–1105.

Bazerman, M.H., Loewenstein, G.F., White, S.B., 1992. Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. Administr. Sci. Quart. 37 (2), 220–240.

Becker, G., DeGroot, M., Marschak, J., 1964. Measuring utility by a single-response sequential method. Behav. Sci. 9 (3), 226–232.

Berg, J.E., Dickhaut, J.W., O'Brien, J.R., 1985. Preference reversal and arbitrage. In: Smith, V. (Ed.), Research in Experimental Economics, vol. 3. JA Press, Greenwich, pp. 31–72.

Berg, J.E., Dickhaut, J.W., Rietz, T.A., 2003. Preference reversals and induced risk preferences: Evidence for noisy maximization. J. Risk Uncertainty 27 (2), 139–170.

Berg, J.E., Daley, L.A., Dickhaut, J.W., O'Brien, J.R., 1986. Controlling preferences for lotteries on units of experimental exchange. Quart. J. Econ. 101 (2), 281–306.

Bohm, P., 1994a. Behaviour under uncertainty without preference reversal: A field experiment. Empirical Econ. 19 (2), 185–200.

Bohm, P., 1994b. Time preference and preference reversal among experienced subjects: The effects of real payments. Econ. J. 104 (427), 1370–1378.

Bohm, P., Lind, H., 1993. Preference reversal, real-world lotteries, and lottery-interested subjects. J. Econ. Behav. Organ. 22 (3), 327–348.

Bostic, R., Herrnstein, R.J., Luce, R.D., 1990. The effect on the preference-reversal phenomenon of using choice indifferences. J. Econ. Behav. Organ. 13 (2), 193–212.

Butler, D.J., Loomes, G.C., 2007. Imprecision as an account of the preference reversal phenomenon. Amer. Econ. Rev. 97 (1), 277–297.

Cadsby, C.B., Song, F., Tapon, F., 2007. Sorting and incentive effects of pay for performance: An experimental investigation. Academy Manage. J. 50 (2), 387–405.

Camacho-Cuena, E., Seidl, C., Morone, A., 2005. Comparing preference reversal for general lotteries and income distributions. J. Econ. Psych. 26 (5), 682–710.

Camerer, C.F., 1989. An experimental test of several generalized utility theories. J. Risk Uncertainty 2 (1), 61–104.

Camerer, C.F., Hogarth, R.M., 1999. The effects of financial incentives in experiments: A review and capital–labor–production framework. J. Risk Uncertainty 19 (1–3), 7–42.

Casey, J.T., 1991. Reversal of the preference reversal phenomenon. Organizat. Behav. Human Dec. Proces. 48 (2), 224–251.

Casey, J.T., 1994. Buyers pricing behavior for risky alternatives: Encoding processes and preference reversals. Manage. Sci. 40 (6), 730–749.

Chai, X., 2005. Cognitive preference reversal or market price reversal? Kyklos 58 (2), 177–194.

Chapman, G.B., Johnson, E.J., 1995. Preference reversals in monetary and life expectancy evaluations. Organizat. Behav. Human Dec. Proces. 62 (3), 300–317.

Chu, Y.P., Chu, R.L., 1990. The subsidence of preference reversals in simplified and market-like experimental settings: A note. Amer. Econ. Rev. 80 (4), 902–911.

Colombo, L., Nicotra, E., Marino, B., 2002. Preference reversal in decision making: The attraction effect in choice and rejection. Swiss J. Psych. 61 (1), 21–33.

Cox, J.C., Epstein, S., 1989. Preference reversals without the independence axiom. Amer. Econ. Rev. 79 (3), 408–426.

Cox, J.C., Grether, D.M., 1996. The preference reversal phenomenon: Response mode, markets and incentives. Econ. Theory 7 (3), 381–405.

Cubitt, R.P., Munro, A., Starmer, C., 2004. Testing explanations of preference reversal. Econ. J. 114 (497), 709–726.

DeNeufville, R., Smith, J.T., 1994. Improving contractors bids using preference reversal phenomenon. J. Construct. Engin. Manage. ASCE 120 (4), 706–719.

Dickhaut, J., McCabe, K., Nagode, J.C., Rustichini, A., Smith, K., Pardo, J.V., 2002. The impact of the certainty context on the process of choice. Proc. Natl. Acad. Sci. 100 (6), 3536–3541.

Dohmen, T., Falk, A., 2006. Performance pay and multi-dimensional sorting: Productivity, preferences and gender. IZA Bonn and University of Bonn Discussion Paper No. 2001.

Ganzach, Y., 1996. Preference reversals in equal-probability gambles: A case for anchoring and adjustment. J. Behavioral Decision Making 9 (2), 95–109.

Goeree, J.K., Holt, C.A., 2005. An explanation of anomalous behavior in models of political participation. Amer. Polit. Sci. Rev. 99 (2), 201–213.

Goldstein, W.M., Einhorn, H.J., 1987. Expression theory and the preference reversal phenomenon. Psychological Rev. 94 (2), 236–254.

Gonzalez-Vallejo, C., Moran, E., 2001. The evaluability hypothesis revisited: Joint and separate evaluation preference reversal as a function of attribute importance. Organizat. Behav. Human Dec. Proces. 86 (2), 216–233.

Green, L., Fristoe, N., Myerson, J., 1994. Temporal discounting and preference reversals in choice between delayed outcomes. Psychonomic Bull. Rev. 1 (3), 383–389.

Grether, D.M., Plott, C.R., 1979. Economic theory of choice and the preference reversal phenomenon. Amer. Econ. Rev. 69 (4), 623–638.

Harless, D.W., Camerer, C.F., 1994. The predictive utility of generalized expected utility theories. Econometrica 62 (6), 1251–1289.

Harrison, G.W., 1989. Theory and misbehavior of first price auctions. Amer. Econ. Rev. 79 (4), 749–762.

Harrison, G.W., 1994. Expected utility theory and the experimentalists. Empirical Economics 19 (2), 223–253.

Hatfield, T.H., Seiver, O.H., 2001. Preference reversals in grading systems for retail food facilities. J. Environ. Health 63 (8), 19–25.

Hawkins, S.A., 1994. Information-processing strategies in riskless preference reversals: The prominence effect. Organizat. Behav. Human Dec. Proces. 59 (1), 1–26.

Hey, J.D., Orme, C., 1994. Investigating generalizations of expected utility theory using experimental data. Econometrica 62 (6), 1291–1326.

Hogarth, R.M., 1980. Judgement and Choice: The Psychology of Decision. John Wiley & Sons, New York.

Hsee, C.K., 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. Organizat. Behav. Human Dec. Proces. 67 (3), 247–257.
Irwin, J.R., 1994. Buying selling price preference reversals: Preference for environmental changes in buying versus selling modes. Organizat. Behav. Human Dec. Proces. 60 (3), 431–457.
Irwin, J.R., Davis, J.H., 1995. Choice matching preference reversals in groups: Consensus processes and justification-based reasoning. Organizat. Behav. Human Dec. Proces. 64 (3), 325–339.
Irwin, J.R., Slovic, P., Lichtenstein, S., McClelland, G.H., 1993. Preference reversals and the measurement of environmental values. J. Risk Uncertainty 6 (1), 5–18.
Johnson, E.J., Payne, J.W., Bettman, J.R., 1988. Information displays and preference reversals. Organizat. Behav. Human Dec. Proces. 42 (1), 1–21.
Judge, G.G., Hill, R.C., Griffiths, W., Lütkepohl, H., Lee, T.C., 1982. Introduction to the Theory and Practice of Econometrics. John Wiley & Sons, Inc., New York.
Kirby, K.N., Herrnstein, R.J., 1995. Preference reversals due to myopic discounting of delayed reward. Psychological Sci. 6 (2), 83–89.
Li, S., 1994. Is there a problem with preference reversals? Psychological Reports 74 (2), 675–679.
Li, S., 2006. Preference reversal: A new look at an old problem. Psychological Record 56 (3), 411–428.
Lichtenstein, S., Slovic, P., 1971. Reversals of preference between bids and choices in gambling decisions. J. Experiment. Psych. 89 (1), 46–55.
Lichtenstein, S., Slovic, P., 1973. Response-induced reversals of preference in gambling: An extended replication in Las Vegas. J. Experiment. Psych. 101 (1), 16–20.
Lichtenstein, S., Slovic, P., 2006. The construction of preference: An overview. In: Lichtenstein, S., Slovic, P. (Eds.), The Construction of Preference. Cambridge University Press, Cambridge.
List, J.A., 2002. Preference reversals of a different kind: The 'more is less' phenomenon. Amer. Econ. Rev. 92 (5), 1636–1643.
Loomes, G., 1990. Preference reversal: Explanations, evidence and implications. Ann. Operations Res. 23 (1), 65–90.
Loomes, G., Starmer, C., Sugden, R., 1989. Preference reversal: Information-processing effect or rational non-transitive choice? Econ. J. 99 (395), 140–151.
Luce, R.D., 2000. Utility of Gains and Losses, Measurement-Theoretical and Experimental Approaches. Lawrence Erlbaum Associates, London.
MacDonald, D.N., Huth, W.L., 1989. Individual valuation, market valuation and the preference reversal phenomenon. J. Behavioral Econ. 18 (2), 99–114.
Maher, P., Kashima, Y., 1997. Preference reversal in Ellsberg problems. Philos. Stud. 88 (2), 187–207.
Mellers, B.A., Ordonez, L.D., Birnbaum, M.H., 1992. A change-of-process theory for contextual effects and preference reversals in risky decision-making. Organizat. Behav. Human Dec. Proces. 52 (3), 331–369.
Mowen, J.C., Gentry, J.W., 1980. Investigation of the preference-reversal phenomenon in a new product introduction task. J. Appl. Psychology 65 (6), 715–722.
Nowlis, S.M., Simonson, I., 1997. Attribute-task compatibility as a determinant of consumer preference reversals. J. Marketing Res. 34 (2), 205–218.
Oliver, A., 2005. Further evidence of preference reversals: Choice, valuation and ranking over distributions of life. J. Health Econ. 25 (5), 803–820.
Ordonez, L.D., Mellers, B.A., Chang, S.I., Roberts, J., 1995. Are preference reversals reduced when made explicit? J. Behavioral Decision Making 8 (4), 265–277.
Palfrey, T.R., Prisbrey, J.E., 1996. Altruism, reputation and noise in linear public goods experiments. J. Public Econ. 61 (3), 409–427.
Palfrey, T.R., Prisbrey, J.E., 1997. Anomalous behavior in public goods experiments: How much and why? Amer. Econ. Rev. 87 (5), 829–846.
Pardo, J.V., Pardo, P.J., Janer, K.W., Raichle, M.E., 1990. The cingulate cortex mediates processing selection in the stroop attentional conflict paradigm. Proc. Natl. Acad. Sci. 87 (1), 256–259.
Pommerehne, W.W., Schneider, F., Zweifel, P., 1982. Economic theory of choice and the preference reversal phenomenon: A reexamination. Amer. Econ. Rev. 72 (3), 569–574.
Ranyard, R., 1995. Reversals of preference between compound and simple risks: The role of editing heuristics. J. Risk Uncertainty 11 (2), 159–175.
Reilly, R.J., 1982. Preference reversal: Further evidence and some suggested modifications in experimental design. Amer. Econ. Rev. 72 (3), 557–584.
Schkade, D.A., Johnson, E.J., 1989. Cognitive processes in preference reversals. Organizat. Behav. Human Dec. Proces. 44 (2), 203–231.
Schmeltzer, C., Caverni, J.P., Warglien, M., 2004. How does preference reversal appear and disappear? Effects of the evaluation mode. J. Behavioral Decision Making 17 (5), 395–408.
Schmidt, U., Hey, J.D., 2004. Are preference reversals errors? An experimental investigation. J. Risk Uncertainty 29 (3), 207–218.
Seidl, C., 2002. Preference reversal. J. Econ. Surveys 16 (5), 621–655.
Seiver, O.H., Hatfield, T.H., 2002. Grading systems for retail food facilities: Preference reversals of environmental health professionals. J. Environmental Health 64 (10), 8–13.
Selart, M., Boe, O., Garling, T., 1999. Reasoning about outcome probabilities and values in preference reversals. Thinking Reasoning 5 (2), 175–188.
Selten, R., Sadrieh, A., Abbink, K., 1999. Money does not induce risk neutral behavior, but binary lotteries do even worse. Theory Dec. 46 (3), 211–249.
Smith, V.L., 1976. Experimental economics: Induced value theory. Amer. Econ. Rev. 66 (2), 274–279.
Stalmeier, P.F.M., Wakker, P.P., Bezembinder, T.G.G., 1997. Preference reversals: Violations of unidimensional procedure invariance. J. Exper. Psychol.: Human Perception Performance 23 (4), 1196–1205.
Starmer, C., Sugden, R., 1989. Probability and juxtaposition effects: An experimental investigation of the common ratio effect. J. Risk Uncertainty 2 (2), 159–178.
Sumner, W., Nease, R.F., 2001. Choice-matching preference reversals in health outcome assessments. Medical Decision Making 21 (3), 208–218.
Tan, U., Komsuoglu, S., Akgun, A., 1993. Inverse relationship between the size of pattern-reversal visual-evoked potentials from the left brain and the degree of left-hand preference in left-handed normal subjects: Importance of the left brain. Int. J. Neurosci. 72 (38719), 79–87.
Tornblom, K.Y., 1982. Reversal in preference responses to 2 types of injustice situations: A methodological contribution to equity theory. Human Relations 35 (11), 991–1013.
Tversky, A., Thaler, R.H., 1990. Anomalies: Preference reversals. J. Econ. Perspect. 4 (2), 201–211.
Tversky, A., Slovic, P., Kahneman, D., 1990. The causes of preference reversal. Amer. Econ. Rev. 80 (1), 204–217.
Waters, L.K., Collins, M., 1984. Effect of pricing conditions on preference reversals by business students and managers. J. Appl. Psych. 69 (2), 346–348.
Wedell, D.H., 1991. Distinguishing among models of contextually induced preference reversals. J. Exper. Psych.: Learning Memory Cognition 17 (4), 767–778.
Wedell, D.H., Bockenholt, U., 1990. Moderation of preference reversals in the long-run. J. Exper. Psych.: Human Perception Performance 16 (2), 429–438.
Wicklund, R.A., 1970. Prechoice preference reversal as a result of threat to decision freedom. J. Person. Social Psych. 14 (1), 8–17.
Wong, K.F.E., Kwong, J.Y.Y., 2005. Comparing two tiny giants or two huge dwarfs? Preference reversals owing to number size framing. Organizat. Behav. Human Dec. Proces. 98 (1), 54–65.
Zapotocna, O., 1986. Lateral-preference aspects of reversal tendency in mirror-image discrimination. Studia Psychologica 28 (1), 17–34.
Zikmund-Fisher, B.J., Fagerlin, A., Ubel, P.A., 2004. Is 28% good or bad? Evaluability and preference reversals in health care decisions. Medical Decision Making 24 (2), 142–148.