MEASURING PREFERENCES FOR UNCERTAINTY

Robert Aron Mislavsky

A DISSERTATION

in

Operations, Information and Decisions

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

____

Uri Simonsohn
Professor of Operations, Information and Decisions, Professor of Marketing

Graduate Group Chairperson

____

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee:

Joseph Simmons, Associate Professor of Operations, Information and Decisions

Deborah Small, Laura and John J. Pomerantz Professor of Marketing, Professor of Psychology

DEDICATION

To Mom, Miles, Sheri, and Susan

ACKNOWLEDGMENTS

My five years here have convinced me that the Wharton Decision Processes group is the best place in the world to do decision-making research. As a result, there are too many people to acknowledge in a relatively short space, but I'll do my best.

First and foremost, I would like to thank my advisor, Uri Simonsohn, for his guidance, encouragement, patience, and general thoughtfulness throughout my career.

Berkeley Dietvorst (Chapter 2) and Celia Gaertig (Chapter 3) have been terrific collaborators on this dissertation and even better friends.

I am grateful to the rest of the faculty and students in the group who have been great role models and friends. In particular, I would like to thank Joe Simmons, Deb Small, Katy Milkman, Maurice Schweitzer, Brad Bitterly, and Shalena Srna.

John Sperger, Sara Sermarini, Sargent Shriver, Laura Kuder, Johanna Matt-Navarro, and Catherine O'Donnell provided valuable research assistance, and the Wharton Behavioral Lab and Wharton Risk Center provided important financial support.

I would not have made it here at all without Erich Studer-Ellis, Carey Morewedge, Eyal Pe'er, and Colleen Giblin, who gave me a great introduction to academic life and research.

Finally, none of this would have been possible without my family's love, support, and encouragement. You all are the best.

ABSTRACT

MEASURING PREFERENCES FOR UNCERTAINTY

Robert Mislavsky
Uri Simonsohn

Understanding decision making under uncertainty is crucial for researchers in the social sciences, policymakers, and anyone trying to make sense of another’s (or their own) choices. In this dissertation, my coauthors and I make three contributions to understanding preferences for uncertainty regarding (a) how preferences are measured, (b) how these preferences may (or may not) manifest in a consequential real-world context, and (c) how different types of advice influence opinions about uncertain events. In Chapter 1, we examine methods that researchers use to study preferences for uncertainty. We find that the presence of uncertainty is often confounded with the presence of “weird” transaction features, dramatically overstating the presence of uncertainty aversion in these experiments. In Chapter 2, we show that extreme uncertainty does not exist in the context of corporate experimentation, despite speculation by pundits and researchers. In fact, people judge experiments similarly to how they would judge simple gambles, with the experiment being judged near the “expected value” of the policies it implements. In Chapter 3, we find that the format in which uncertainty is presented impacts how people combine forecasts from multiple sources. Numeric probability forecasts are averaged, while verbal forecasts are combined additively, with people making more extreme judgments as they see additional forecasts.

TABLE OF CONTENTS

LIST OF TABLES

TABLES IN CHAPTER 1

TABLES IN CHAPTER 2

TABLES IN CHAPTER 3

LIST OF FIGURES

FIGURES IN CHAPTER 1

FIGURES IN CHAPTER 2

FIGURES IN CHAPTER 3

INTRODUCTION

Understanding decision making under uncertainty is crucial for researchers in the social sciences, policymakers, and anyone trying to make sense of another's (or their own) choices. In this dissertation, my coauthors and I make three contributions to understanding preferences for uncertainty. First, we examine how preferences for uncertainty are typically measured, finding that many common methods may overstate the presence of uncertainty aversion in experiments. Second, we test preferences for uncertainty in a consequential real-world context, corporate experimentation, where a company's employees and customers are randomly assigned to different outcomes. Finally, we show that when presented with multiple forecasts for uncertain events, people combine these forecasts differently depending on whether they are provided numerically or verbally, resulting in potentially drastic differences in internal judgments of an event's uncertainty. Taken together, our findings have consequences for researchers, organizational decision makers and policymakers, and individuals, which we discuss throughout.

In the first chapter of my dissertation, "When Risk is Weird: Unexplained Transaction Features Lower Valuations," we examine a potential cause of a major behavioral anomaly in risk preference, the uncertainty effect (Gneezy, List, & Wu, 2006). Although prior research on the uncertainty effect finds the introduction of risk causes substantial violations of the internality axiom, where participants value a gamble less than its worst outcome, we find that this is likely caused by the presence of what we call “weird” transaction features. In typical risk preference studies, valuations of risky gambles are typically compared to valuations of certain outcomes. However, risky gambles often include additional features, such as purchasing a lottery ticket or flipping a coin, whereas the certain outcomes do not. We propose that an aversion to weird features, rather than uncertainty itself, drives the extreme risk aversion found in paradigms such as the uncertainty effect. As a result, we believe that these studies typically overstate the presence of risk aversion. In an incentivized experiment under high stakes, participants are essentially risk neutral when comparing gambles with “weird” transaction features to certain outcomes that have the same transaction features.

In the second chapter, “Critical Condition: People Only Object to Corporate Experiments If They Object to a Condition,” we investigate preferences for uncertainty in a consequential real-world context—corporate experimentation. Although experimentation is one of the most effective tools for determining the impact of a given policy, which could allow organizations to test whether certain policies are beneficial before rolling them out more broadly, there is a widespread perception that people dislike corporate experimentation as a general rule (e.g., M. N. Meyer, 2015; M. N. Meyer & Chabris, 2015). If such “experiment aversion” exists, it could severely hamper the ability of researchers to learn about the world and test theories in real-world settings. However, in 5 studies, we show that a general experiment aversion does not exist. Rather, people dislike experiments only when they dislike a specific policy that the experiment implements. Further, people evaluate experiments with an objectionable policy more favorably than the policy itself.

In the final chapter, “60% + 60% = 60%, but Likely + Likely = Very Likely,” we find differences in how people combine probability forecasts from multiple advisors depending on whether those forecasts are given numerically or verbal. Specifically, we find that, consistent with prior research (e.g., Budescu & Yu 2006, 2007), participants average numeric probability forecasts. For example, if two weather forecasters predict that there is a “60% chance” and a “70% chance,” respectively, that it will rain, participants’ own predictions typically lie between 60% and 70%. However, participants combine verbal forecasts more additively. That is, if the forecasters say rain is “probable” and “likely,” participants tend to make predictions that are more extreme than each forecaster individually (e.g., “very likely”).

CHAPTER 1.

WHEN RISK IS WEIRD:

UNEXPLAINED TRANSACTION FEATURES LOWER VALUATIONS

Robert Mislavsky

Uri Simonsohn

ABSTRACT

We define transactions as weird when they include unexplained features, that is, features not implicitly, explicitly, or self-evidently justified, and propose that people are averse to weird transactions. In six experiments, we show that risky options used in previous research paradigms often attained uncertainty via adding an unexplained transaction feature (e.g., purchasing a coin flip or lottery), and behavior that appears to reflect risk aversion could instead reflect an aversion to weird transactions. Specifically, willingness to pay drops just as much when adding risk to a transaction as when adding unexplained features. Holding transaction features constant, adding additional risk does not further reduce willingness to pay. We interpret our work as generalizing ambiguity aversion to riskless choice.

The amount people are willing to pay for a given item is influenced by the context in which the purchase takes place (Ariely, Loewenstein, & Prelec, 2006; Jung, Perfecto, & Nelson, 2016; Lichtenstein & Slovic, 2006). Transactions, the necessary steps to acquire the item, are a part of every purchase context.

In this paper, we identify a transaction attribute which negatively influences willingness to pay: the extent to which it contains features that lack an explanation. These explanations may be (i) implicit, based directly on consumers' past experiences with similar transactions, (ii) explicit, explained by the seller, or (iii) self-evident, based on reasonable inferences from context. For brevity, we refer to transactions that include unexplained features as "weird." When using the term weird, we refer *exclusively* to such a definition—the presence of unexplained features.

To illustrate how the presence of unexplained features may manifest itself in a transaction and how the three aforementioned types of explanations might mitigate their impact on willingness to pay, consider a restaurant that sells lunches by placing them in boxes and then asks people to pay for the right to open the box and take the lunch from the box. Placing the lunch in a box and asking to pay to open it could constitute an unexplained transaction feature and may make customers uncomfortable or suspicious (e.g., is the lunch in the box because the restaurant doesn't want you to see what it really looks like?). An explanation could easily mitigate any such consequences. An *implicit* explanation would be if the "box" was simply a vending machine; customers could draw on their prior experience and the transaction feature is no longer unexplained. Alternatively, the restaurant could provide an *explicit* explanation "These are our new self-service boxes, which we’ve introduced to help you get your food more easily.” The transaction feature is again no longer unexplained thus no longer expected to reduce valuations.

Note that unexplained is not the same as novel. A completely novel transaction feature could come with an explanation. For example, imagine a restaurant that requires customers to draw on a piece of glass with their finger in order to get their lunch. That is an unusual transaction feature. But if the glass is an iPad screen, and the drawing is the customer’s signature, customers facing this transaction feature for the very first time would easily generate a *self-evident* explanation for why the transaction feature is there. It would not be expected to lower valuations.

We conjecture that the presence of unexplained features lowers willingness to pay because they trigger reactions akin to ambiguity aversion (Ellsberg, 1961; Frisch & Baron, 1988; Keren & Gerritsen, 1999) in general and comparative ignorance in particular (Chow & Sarin, 2001; Fox & Tversky, 1995; Fox & Weber, 2002). Relevant but unknown information may make consumers less confident in the decision to make the purchase (Chow & Sarin, 2001; Fox & Tversky, 1995; Fox & Weber, 2002) or perhaps make them feel the seller has more information that she may use to her advantage (Frisch & Baron, 1988, p. 153; Keren & Gerritsen, 1999). The presence of unexplained features creates an imbalance between seller and buyer in terms of what relevant information they have for the transaction. Weirdness aversion, the aversion to transactions with unexplained features, may then constitute the generalization of ambiguity aversion to situations that lack (explicit) uncertainty.

We demonstrate the practical relevance of an aversion to unexplained transaction features by focusing on a research paradigm where researchers unintentionally manipulated the presence of unexplained transaction features and obtained a result, often referred to as the “uncertainty effect” (Gneezy et al., 2006). We find that the uncertainty effect may instead be caused by weirdness, or the presence of unexplained transaction features.

Gneezy et al. (2006) documented that people were willing to pay less for a risky prospect than for its worst possible outcome. For instance, people were willing to pay an average of $26.10 for a $50 Barnes and Noble gift card but only $16.12 for a gamble where participants were guaranteed to win either a $50 or $100 gift card, each with a 50% probability. This general finding has been replicated by many independent research teams (e.g., Andreoni & Sprenger, 2011; Newman & Mochon, 2012; Simonsohn, 2009; Yitong Wang, Feng, & Keller, 2013; Yang, Vosgerau, & Loewenstein, 2013).[1]

These uncertainty effect studies pit valuations of a risky option against valuations of a riskless one. The risky option requires a mechanism that introduces risk, while the riskless option does not. For example, researchers have generated risky prospects by asking participants to buy coin flips, lottery tickets, unlabeled envelopes, and gift cards of unknown value and have compared participants’ valuations of these transactions to that of buying a gift card outright. There is no explicit nor implicit justification to sell gift

[1] Keren and Willemsen (2009) report results where the uncertainty effect is not observed when comparing average valuations. Gideon Keren shared the raw data from that article with us. We analyzed it as in Simonsohn (2009), comparing the entire distributions of responses and found that a substantial share of participants do show the effect. Rydval et al. (2009) provide the only failure to replicate the uncertainty effect that we are aware of. Their favored explanation is that participants in other experiments misunderstood the task and/or payoffs. Yang et al. (2013) find that the uncertainty effect is only observed for willingness to pay and not for willingness to accept measures.

cards of unknown value or to utilize a coin flip to determine their value. Therefore, while these mechanisms do generate risk, they also introduce unexplained features to the transaction.

Uncertainty effect studies, therefore, have included a risky transaction *with* unexplained features and a not risky one *without* unexplained features, perfectly confounding risk with weirdness. In this paper, we report studies that manipulate the presence of unexplained features independently of risk. Our results are consistent with an aversion to unexplained features accounting for somewhere between the preponderance and the totality of the uncertainty effect. After presenting our empirical results, we discuss how unexplained features could be present in other paradigms used to study consumer behavior.

TRANSPARENT REPORTING

Studies 1-5 were run on Amazon's Mechanical Turk (MTurk) and were administered through Qualtrics. Study 6 was incentive compatible and run in a behavioral lab. For all studies we decided sample size before collecting any data. MTurk participants were not allowed to participate in more than one study. We included attention checks for Studies 5A and 5B. Studies 5A, 5B, and 6 were preregistered. For all studies we report all data exclusions (if any), all manipulations, and all measures. Data, analysis code, preregistrations, and survey materials are available at http://osf.io/x8cqm.

STUDIES 1-3: WEIRD, BUT NOT RISKY

Our first three studies are similar, so we present them together. In all three we modified the traditional uncertainty effect paradigm to disentangle the effect of risk from the effect of unexplained features on valuations. For a more fluent reading experience, we refer to transaction that include unexplained features as "weird" and to the presence of such features as "weirdness." The uncertainty effect paradigm pits the valuation of a riskless prospect (e.g., buying a $50 Target gift card) against that of a risky one (e.g., flipping a coin to determine if the gift card is for Target or for Walmart). This paradigm confounds risk and weirdness because the manipulation that introduces risk also introduces unexplained features to the transaction (e.g., flipping a coin). To examine the importance of this confound, we created a third type of transaction, one that was *weird but not risky*. Specifically, this was a transaction that includes the same unexplained features present in the risky transactions (e.g., buying a token redeemable for a gift card) but with a certain outcome (e.g., the value of the gift card is known).

Method

Design. In Study 1 (N = 603; 29.6% female), we randomly assigned participants to one of three conditions asking them indicate their maximum WTP for a transaction. The first two were analogous to traditional uncertainty effect studies:

Condition 1. *Neither weird nor risky:*[2]

We want to know how much you would be willing to pay for two different items, a $50 Walmart gift card and a $50 Target gift card.

If you could buy only the $50 Walmart gift card, what is the most you would pay for it?

If you could buy only the $50 Target gift card, what is the most you would pay for it?

Condition 2. *Weird and risky:*

Imagine that you are standing in front of a table that has a locked box on it. The box has two gift cards inside: a $50 Walmart and a $50 Target gift card.

You can pay to open the box and choose a gift card, which will be yours to keep. The gift cards do not have the names of the stores printed on them, so you will not know which gift card is which.

What is the most you would be willing to pay to open the box?

Uncertainty effect studies compare the valuation of similar pairs of transactions. Any difference in WTP can therefore be caused by the risk difference (having a known vs. unknown outcome) or by the weirdness difference (buying outright vs. paying to open a box). We addressed this confound by adding a *weird but not risky* condition. Participants read the same scenario as those in the *weird and risky* condition, except the gift cards were labeled, so participants knew which card they were getting before choosing. Specifically, it read (differences between Conditions 2 and 3 underlined here but not in original materials):

Condition 3. *Weird but not risky*:

Imagine that you are standing in front of a table that has a locked box on it. The box has two gift cards inside: a $50 Walmart and a $50 Target gift card.

[2] In Study 1, some participants valued Walmart/Target gift cards and others valued Amazon/Barnes & Noble gift cards. Because subsequent studies only included the former, we report results for the latter in footnote 3. We also collected data on self-reported average expenditures in other purchases to use as covariates to increase power, but they were uncorrelated with the dependent variable and therefore not useful. We did not collect these in subsequent studies. See Supplement 2 for covariate results.

> You can pay to open the box and choose a gift card, which will be yours to keep. The gift cards **have** the names of the stores printed on them, so you **will know** which gift card is which.
>
> What is the most you would be willing to pay to open the box?

After running this study, we identified a potential confound. The weird transactions (paying to take one of two gift cards from a box) had two possible outcomes, while the not weird transaction had only one. We believed this difference, rather than weirdness, could explain any observed differences (e.g., because people are averse to explicitly rejecting an outcome). In Study 2 (N = 308; 35.5% female) we reran the two weird conditions and added a new weird condition that had only one possible outcome. Across the three conditions, then, participants paid to open a box and take a card from it. The conditions differed on whether the box contained *one labeled* gift card (new condition), *two labeled* gift cards, or *two unlabeled* gift cards. We did not rerun the *neither risky nor weird* condition.

In Study 3 (N = 403; 36.8% female) we reran all four conditions from Studies 1 and 2 with a different operationalization of risk and weirdness: purchasing a token at an event and redeeming it for a gift card. The four conditions were:

1. *Neither weird nor risky*
   What is the highest amount you would be willing to pay for a $50 [Walmart/Target] gift card?
   (Target and Walmart counterbalanced within-subjects)

2. *Weird but not risky, one option*
   Imagine that you are at an event where there are tokens for sale. These tokens can be redeemed at a cashier for a $50 [Walmart/Target] gift card. What is the highest amount you would be willing to pay for one of these tokens?
   (Target and Walmart counterbalanced within-subjects)

3. *Weird but not risky, two options*

```
Imagine that you are at an event where there are tokens for sale. These tokens can be redeemed at a cashier for your choice of either a $50 Walmart gift card or a $50 Target gift card. What is the highest amount you would be willing to pay for one of these tokens?
```

4. *Weird and risky*

```
Imagine that you are at an event where there are tokens for sale. These tokens can be redeemed at a cashier for either a $50 Walmart gift card or a $50 Target gift card. The cashier will flip the token, and if it lands on heads, you will receive the Walmart gift card. If it lands on tails, you will receive the Target gift card. What is the highest amount you would be willing to pay for one of these tokens?
```

Results

Figure 1 depicts results for Studies 1-3. We identify four main takeaways:

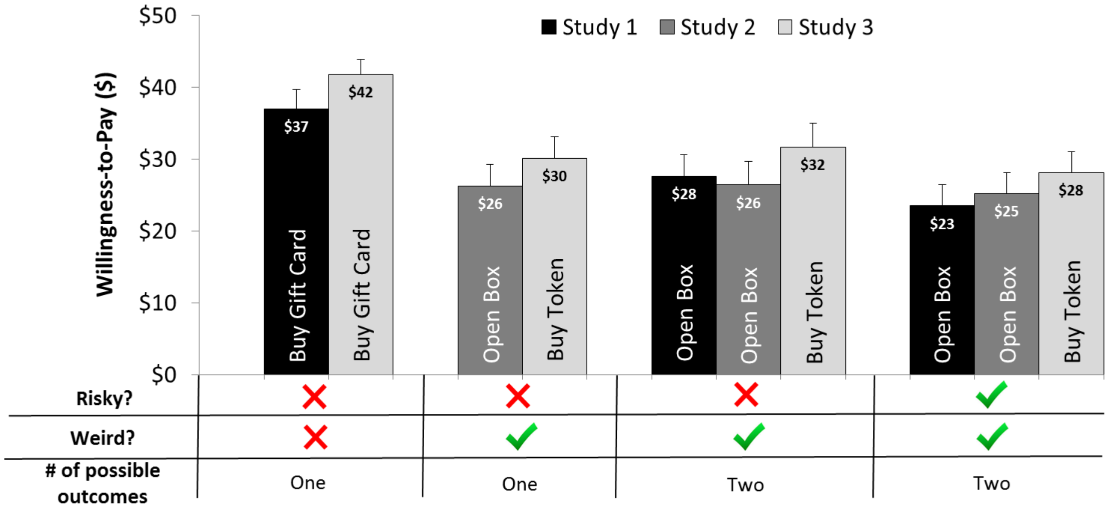
1. In Studies 1 and 3, we replicate the original uncertainty effect (Study 2 does not allow testing it). Participants valued the *weird and risky* prospects (M = $25.80), less than their least-valued *neither weird nor risky* gift card (M = $39.37). The risky option was valued significantly less than its worst outcome in both studies, ts > 6.58, *p*s < .001.
2. Holding weirdness constant, there is no apparent uncertainty effect. Comparing the two weird conditions, *risky* gift cards (M = $25.60 across all studies) were not valued significantly less than the *riskless* gift cards (M = $28.39 across all studies), whether they had one or two options (Study 1: t(199) = 1.92, *p* = .057; Study 3: ts < 1.64, *p*s > .10). Based on point-estimates of the means, the effect of weirdness is two-thirds (Study 1) to three-quarters (Study 3) as large as the uncertainty effect is when weirdness is not accounted

for.[3] We believe some of this residual effect we are attributing to uncertainty is also attributable to weirdness, because it seems likely that, in these scenarios, uncertainty makes the weird scenarios weirder by adding an additional unexplained feature, flipping a token to determine the value of a gift card. We could not estimate this for Study 2, because it did not include a *not weird* condition.

3. Contrary to our initial expectations, these results are not driven by the number of potential options. Valuations for the weird but not risky transactions are similar when they involve one or two possible outcomes, ts < .71, *p*s > .47.
4. Study 3 rules out a potential confound for Studies 1 and 2. In the box scenarios, participants may have believed that they had to make two payments, one to open the box and another to purchase the gift card. Because very few participants paid $0 in the weird scenarios (as would be expected if this were the case; see Supplement 2), we believe this is unlikely, although a reviewer also raised the possibility that participants may have averaged the two payments when reporting their WTP. We obtain very similar results in the token scenario, where this ambiguity is not present, which appears to rule this possibility out.

[3] For the Barnes & Noble and Amazon gift cards in Study 1, the means are $35.91 (neither weird nor risky), $27.77 (weird but not risky), and $22.30 (weird and risky). The total uncertainty effect amounts to $13.61, with weirdness accounting for nearly 60% of the effect.

Figure 1. Average valuations (Studies 1-3) as a function of risk and weirdness



Notes: Hypothetical valuations for $50 gift cards. Risk involves whether it is for Target or Walmart, operationalized via opening a box and selecting one of two *unlabeled* envelopes (Studies 1 & 2), or purchasing a token exchangeable for one of the two gift cards, determined by flipping the token (Study 3). Weird but riskless involves labeled envelopes (Studies 1 & 2), or participants *choosing* what to redeem the token for (Study 3). Transactions with one outcome (bottom row) involve box with 1 gift card (Study 2) or token with predetermined value (Study 3). Error bars represent 95% confidence intervals.

STUDY 4: BIGGER DIFFERENCES IN OUTCOMES

In the first three experiments, the risky prospects involved gift cards with the same face value ($50) for different stores (e.g., Target vs. Walmart). This design, originally used by Newman and Mochon (2012), allowed us to create *weird but not risky* conditions where participants could meaningfully choose between gift cards, whereas choosing between a $50 card and a $100 card is not a meaningful choice. However, minimizing outcome variance may have inflated the importance of the unexplained features. In other words, we may have found risk did not matter much because we created situations without much risk. In this experiment, we created risky prospects with greater outcome variance.

Method

Sample. We recruited 604 participants (39.4% female), each paid $0.25.

Design. Participants were randomly assigned to one of eight conditions in a between-subjects design. Two *not weird* conditions were similar to those in Studies 1 and 3: participants provided their WTP for either a $50 Target gift card or a $100 Target gift card bought outright. The remaining six conditions involved *weird* transactions and conformed to a 2 (*transaction*: box vs. token) x 3 (*value*: $50 vs. $100 vs. risky) design. Participants read either the box or token scenarios from the prior studies, where the outcomes were either a $50 Target gift card for sure, a $100 Target gift card for sure, or a Target gift card that was worth either $50 or $100, each with 50% probability. We did not include a condition where participants could choose either a $50 or $100 gift card because we assumed all participants would choose $100. We decided before data collection began to obtain 120 observations from the *not weird* conditions and 60 from each *weird* condition (since we had two versions of weirdness, 60*2=120).

Results

Beginning with the token conditions, the uncertainty effect was again replicated when not accounting for transaction weirdness. Participants valued the risky token $6.27 less than they did its worst possible outcome purchased outright (M = $37.23 and M = $43.50, respectively), t(179) = 2.70, *p* = .008. Comparing the weird conditions, people paid $5.59 *more* for the risky prospect (*token exchangeable for $50 gift card:* M = $31.64; *risky token:* M = $37.23), t(118) = 1.57, *p* = .12.[4]

The uncertainty effect was also replicated in the box conditions (*$50 Target gift card bought outright*: M = $43.50; *risky box*: M = $25.23), t(180) = 9.13, *p* < .001. The difference between the risky prospect and its least valued outcome was much smaller when comparing the two weird conditions (*$50 gift card in box*: M = $29.44; *risky box with $50 or $100 gift card*: M = $25.23), t(120) = 1.40, *p* = .16. The total uncertainty effect is about $18 ($43.50-$25.23). The effect of weirdness alone is about $14. As argued above, the residual $4 effect could be the result of weirdness if choosing among unlabeled cards seems less justified than taking a labeled card out of a box.

There was also a sizable main effect of weirdness for individual valuations of the $50 and $100 gift cards. Buying a $50 or $100 gift card outright was valued at $43.50 and $86.49, respectively, whereas a $50 or $100 gift card in a box was valued at $29.44 and $51.47, respectively, and a token exchangeable for a $50 or $100 gift card was valued at $31.64 and $65.93, respectively, ts > 5.38, *p*s < .001. We report all pairwise comparisons in Supplement 4. In sum, we obtain results similar to those of Studies 1-3 using risky prospects with greater outcome variance. The data are consistent with unexplained features accounting for somewhere between the preponderance and the totality of the uncertainty effect.

[4] Analyzing the data as in Simonsohn (2009), the lower bound of people paying less for the uncertain item is 3.3% in the *token* conditions and 19.7% in the *box* conditions, neither of which is significantly greater than 0 (*p*s > .09). See Supplement 4.

STUDIES 5A AND 5B: EVALUATING WEIRDNESS OF PRIOR UNCERTAINTY EFFECT STUDIES

In Study 5 we more directly test if prior uncertainty effect studies have unintentionally manipulated weirdness by asking participants to evaluate the weirdness of the underlying transactions in those studies.

One may measure weirdness on absolute or relative scales, although each has its limitations. Absolute scales (e.g., “How weird is this transaction?”) are ambiguous about what a transaction is being compared to, or equivalently, what the values in the scale represent. Relative scales, on the other hand, (e.g., “Which transaction is weirder?”), may create demand effects or change participants’ definitions of weirdness where they think the weirdest transaction is the one that is least like the others (even though it may be the simplest). Since neither approach was obviously superior, we pursued both, and in both cases we explicitly defined weirdness to our participants as involving the presence of unexplained features. Participants judged weirdness on both an absolute scale (Study 5A) and on a relative scale (Study 5B). We obtained consistent results with both methods. Risky transactions in prior uncertainty effect studies are weirder than their riskless counterparts.

STUDY 5A: BETWEEN-SUBJECTS RATINGS OF WEIRDNESS

Method

Sample. We recruited 714 MTurk participants, 600 of whom (53.3% female, M_{age} = 35.3 years) passed an attention check and were able to continue to the rest of the survey, each paid $0.40 (pre-registration: https://aspredicted.org/3mu9d.pdf).

Design. In a between-subjects design, participants evaluated the weirdness of transactions used in prior uncertainty effect studies. Participants began by reading this passage:

> We will show you an example of a purchase that experimenters ask participants to evaluate. We are interested in knowing how "weird" you think the purchase is. By "weird," we mean how much the purchase has unusual and unexplained features.

Participants then read one of eight questions used in prior uncertainty effect studies—two from Gneezy et al. (2006), three from Yang et al. (2013), and three from this paper. Three of these questions were "baseline" questions (i.e., the riskless valuations that were used as control conditions in uncertainty effect studies).[5] We preregistered that we would collapse the ratings for these conditions for analysis. The other five valuations were used in prior studies—Gneezy et al.'s (2006, p. 1301) lottery, Yang et al.'s (2013, p. 737) certain and uncertain coins, our certain and uncertain boxes (Study 4). See the Appendix for the exact text of these stimuli. After reading the question, participants rated its weirdness using the following scale: "How weird is it to buy a gift [card/certificate]

[5] These questions were slightly adapted in order to sound like an actual transaction (e.g., "Imagine you are buying this") rather than an abstract valuation (e.g., "What is the most you are willing to pay for this?").

like this?" (1 = It is not weird at all; 2 = It is a little weird; 3 = It is very weird; 4 = It is extremely weird). If risk and weirdness were confounded in these studies, we would expect that the weird transactions would be rated as weirder than the baseline ones.

Results. Consistent with the notion that prior uncertainty effect studies have confounded risk and weirdness, participants rated all of the weird transactions (1.94 ≤ Ms ≤ 2.68) as weirder than the baseline transaction (M = 1.35), all ts > 4.83, all *p*s < .001. See Figure 2, panel (i). In addition to this pre-registered comparison, we compared the share of participants rating a transaction as "not weird at all." Seventy percent of participants gave this rating to the baseline transaction compared to between 8% and 39% for the weird transactions, Zs > 4.45, *p*s < .001.

STUDY 5B: WITHIN-SUBJECTS RANKINGS OF WEIRDNESS

Method

Sample. We recruited 184 participants, 153 of whom (42.7% female, M_{age} = 35.5 years) passed an attention check and were able to continue to the rest of the survey, each paid $0.40 (pre-registration: https://aspredicted.org/p4hi5.pdf).

Design. All participants were given the same instructions as in Study 5A, but instead of rating them between-subjects, they were shown six transactions (one of the three baseline transactions and all five weird transactions) and asked to *rank* them from weirdest (1) to least weird (6). Ties were not allowed.

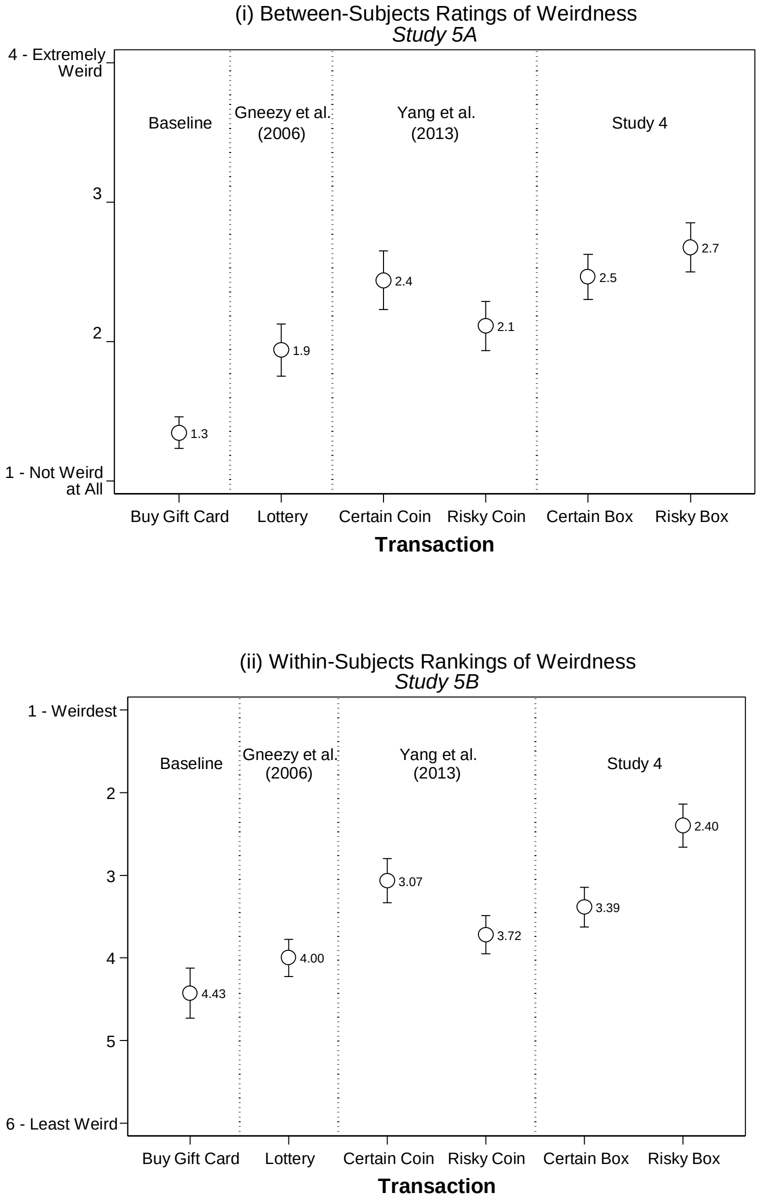
Results

Consistent with Study 5A and more generally with the notion that prior uncertainty effect studies have confounded risk and weirdness, participants ranked purchasing a gift card outright as the least weird (M = 4.43 out of 6) out of all the transactions (between 2.40 for the Risky Box, t(149) = 8.47, *p* < .001, and 4.00 for the GLW Lottery, t(149) = 2.00, *p* = .047). See Figure 2, panel (ii). Here the weirdness difference between the baseline and the original uncertainty effect (Gneezy et al., 2006) seems smaller than in Study 5A. Part of this may be explained by some participants reversing the scale, since 14% of participants ranked the baseline transaction as the *weirdest* (the second most popular answer). Nevertheless, looking at the number of participants who ranked the transaction as least weird, a comparison not included in our pre-registration, we see a more substantial difference. Specifically, while 46% of people ranked the baseline as the least weird, only 15% did for the Gneezy et al. (2006) lottery, Z = 5.97, *p* < .001.

STUDY 6: INCENTIVIZED LAB STUDY

To this point, all of our studies have used hypothetical scenarios. To address the possibility that our findings were driven in part by participants' inattention or lack of motivation, our last study is an incentive-compatible replication (pre-registration: https://aspredicted.org/dq97y.pdf).

Figure 2. Prior uncertainty effect studies are weirder than their baseline comparisons



Notes: Panel (i) shows between-subjects ratings (Study 5A; N=600) of transactions used in prior uncertainty effect studies (see Appendix). The scenarios were described verbatim to participants. The y-axis shows the average response to the question: '*We are interested in knowing how "weird" you think the purchase is . . . By "weird," we mean how much the purchase has unusual and unexplained features.'* Panel (ii) shows within-subjects rankings of weirdness (Study 5B; N=153) of the same scenarios.

Method

Sample. We recruited 219 participants (71.1% female, M_{age} = 20.8 years) at the Wharton Behavioral Lab. This study was part of a larger lab session with several unrelated studies, and all participants were paid $10 for completing the session.

Design. In a three-cell between-subjects design, participants indicated their willingness to pay (WTP) for an item. The three conditions were (i) buying a $50 Amazon gift card (*neither weird nor risky* condition), (ii) paying to open a locked box with a $50 Amazon gift card and taking the card (*weird but not risky* condition), and (iii) paying to open a locked box containing a $50 gift card and a $100 gift card, with values only visible on the inside, and taking a card without knowing its value (*weird and risky* condition).

One in every twenty participants was randomly selected to have their decision count for real and receive a $100 bonus (to fund the purchase). To make the WTP elicitation incentive-compatible, a price was set but not revealed to participants. If participants' WTP was greater than that price, they made the purchase and paid that price. Otherwise, they kept the entire bonus and did not make a purchase. To indicate their WTP, we showed participants a price, starting at $5, and they indicated if they would make the purchase for that amount. If they said yes, we increased the price by $5, and they answered again. This was repeated until they answered "No" or the price reached $100.[6] The highest price participants said "Yes" to is our dependent variable. We

[6] Only one participant (in the *neither weird nor risky* condition) gave a WTP of $100.

purposefully avoided a multiple-price-list and used a multiple-price-*sequence*, concerned that the price list could prompt participants to choose valuations in the middle of the range for the uncertainty condition, attenuating the uncertainty effect (original materials: https://osf.io/x8cqm/).[7]

Results

Without accounting for weirdness, for the presence of unexplained transaction features, participants again acted as if they were extremely risk averse. Willingness to pay for the *weird and risky* transaction (M = $39.24) was similar to that for the *neither weird nor risky* one (M = $38.70), t(143) = .19, *p* = .85, even though the former has an expected value approximately 50% higher than the latter. As in prior uncertainty effect studies, this suggests the presence of direct risk aversion, since neither prospect theory nor expected utility theory can generate such extreme levels of risk aversion. But if defined narrowly, as obtaining a strictly lower mean, this result does not replicate the uncertainty effect.[8] In any case, this comparison confounds risk and weirdness.

Controlling for weirdness, participants appear to show very mild (if any) risk aversion: the risky purchase (M = $39.24) was valued noticeably *above* the not risky one

[7] A reviewer expressed this concern about a multiple-price-sequence that we thought was worth sharing with readers: "[A] price-sequence may not be innocuous, either: The initial, low prices may serve as anchors for subjects' valuations […] which may bias WTPs down. If such anchoring effects were asymmetric, and were more pronounced for risky or weird transactions (because, say, preferences for risky or weird transactions are less stable), then they could make the experimental results difficult to interpret." To respond to this concern we ran a study on MTurk manipulating whether the multiple-price-sequence was increasing or decreasing. The effect of weirdness is significant and of the same magnitude for both. See Supplement 7.

[8] Although we preregistered that we would calculate the proportion of the uncertainty effect explained by weirdness, we could not do this here because we do not directionally replicate the original uncertainty effect.

(M = \$30.47), t(143) = 2.94, *p* = .004. In fact, participants valued the uncertain gift card close to what a risk *neutral* buyer would be expected to value it. In particular, assuming participants would pay twice as much for a \$100 gift card as they would for a \$50 gift card (which is a conservative assumption that does not account for diminishing sensitivity or marginal utility), a risk neutral valuation of the risky gift card is \$45.71 (1.5 * \$30.74), which is not much higher than what we observe (\$39.24), t(143) = 1.77, *p* = .080.[9]

Finally, holding risk constant, we replicate weirdness aversion. The not weird purchase (M = \$38.70) was valued above the weird one (M = \$30.48), t(144) = 3.17, *p* = .002.

GENERAL DISCUSSION

We have documented that the presence of unexplained features lowers willingness to pay (WTP). We manipulated the presence of such features, weirdness, independently of risk and found that the effect of weirdness on WTP is of about the same magnitude as the uncertainty effect, which had previously been attributed to the presence of uncertainty. These results suggest that subtle transaction features can have dramatic effects on WTP—dramatic enough for multiple independent research teams to run successful replications of the original Gneezy et al. (2006) finding, but subtle enough that they did not notice the potential confound when doing so (including one of us; see Simonsohn, 2009).

[9] To perform this t-test we multiplied all valuations in the weird but not risky condition by 1.5, and conducted a standard difference of means t-test comparing this new variable with the observed valuations in the *weird and risky* condition. The comparison, therefore, treats \$45.71 as an estimated magnitude with a standard error (which it is), rather than as a pre-set constant (which it is not). We did not preregister this analysis, because we did not expect this valuation to be so high.

Unexplained features is the key manipulation

We have characterized our key manipulations as increasing weirdness, or introducing unexplained features to transactions. Some of the seven members of our review team proposed alternative interpretations for our manipulations. One reviewer proposed that perhaps we simply manipulated the total number of features (whether weird or not). We do not believe the number of features per se is critical. First, in an experiment included in a prior version of the manuscript, we found that merely adding features did not reduce valuations (see supplement 6). Second, in many empirical studies, valuations are often elicited with procedure that require different numbers of steps (e.g., asking for a price outright vs. going through a multiple-price list), and it has not been previously documented that transactions with more steps lead to lower valuations. Third, there is no obvious psychological process that would seem to justify this prediction. In contrast, we believe that all mechanisms that have been proposed for ambiguity aversion would also predict that *unexplained* features lower valuations.

Another reviewer proposed that perhaps what's special about the features we introduced is not that they are unexplained features, but that they are unusual features that transactions outside the lab would not include. That is to say, people would pay less for opening a box to buy an item, not because they see no reason to have that extra step, but because outside the lab they have never purchased an item by paying to open the box. We do not find this alternative explanation compelling either. First, most transactions in the lab are rather unusual. Take, for example, our baseline condition in incentive-compatible Study 6. Participants completed a multiple-price sequence which was then compared with a pre-set price to determine if they would purchase a $50 gift card held by the experimenter. This is not a transaction they would engage in outside the lab. And yet, their WTP was a rather high $38.70 and comparable to the valuations from prior studies that did not involve the convoluted incentive-compatible mechanism (e.g., $37 in Study 1 here).

Second, we can easily imagine situations where a completely new transaction feature, because it is accompanied by an explanation, would not be expected to lower WTP. Consider again that example from the introduction about a person's first payment by signing on an iPad, or perhaps an American asked to pay in rubles during her first coffeeshop visit in Moscow. In these examples, consumers are facing entirely novel transaction features, but these features have self-evident explanations and would not be predicted to lower WTP.

When risk is not weird

Our studies manipulate unexplained features independently of risk (i.e., we include transactions that are *weird but not risky*), but not risk independently of unexplained features (i.e., we do not include transactions that are *not weird but risky*). The absence of a *not weird but risky* cell in our experiments may pose some problems for the interpretation of our studies. If a *not weird but risky* condition was valued similarly (or lower) than a *weird and risky* scenario, it would imply that unexplained features moderate, rather than account for, the effect of risk in those transactions. Although we think this is unlikely, our data cannot rule this out.

This is a challenge to explore empirically because it requires a situation where risk is an expected feature (e.g., buying stocks), and is therefore not weird. In such situations, however, offering an option with no risk (e.g., a riskless stock) would be weird, since it would involve the presence of a feature that requires an explanation ("why is this stock riskless?"). Yang et al.'s (2013) Experiment 4 provides an example of our concern. They include a condition where participants indicate their WTP for a coin flip that paid a $50 gift certificate if the coin landed on heads *or* tails ("Certain Coin Flip," p. 737). In our Studies 5A and 5B, we asked participants to rate how weird this transaction was, and they rated it as *weirder* than the risky coin flip (i.e., as containing more unexplained features), likely because a coin flip implies risk and removing risk makes the coin flip unnecessary.

Further, even holding all features of a transaction constant, all risk per se may not be equally unexplained. For instance, in most gambling situations, payoffs are inversely proportional to the probability of winning. Therefore, a lottery with a 1% chance of winning $100 and a 99% chance of winning $50 is more typical (i.e., has an implicit explanation) than a gamble with a 99% chance of winning $100 and a 1% chance of winning $50. If this were true, and if unexplained features reduce valuations, people should appear more risk averse for the latter lottery. A closer look at Gneezy et al. (2006, p. 1287) reveals evidence consistent with this conjecture. Participants are risk seeking (i.e., WTP > Expected Value) when there is a 1% chance of winning the larger price and risk averse (i.e., WTP < Expected Value) when there is a 99% chance of winning the larger prize (p. 1287, Table 1). In fact, the median WTP for these two gambles are identical ($37.50) in this study. Of course, this is speculative and there are several potential explanations for these findings that have little to do with the specific transaction features (e.g., probability weighting; McGraw, Shafir, & Todorov, 2010; Rottenstreich & Hsee, 2001).

Attributing the uncertainty effect to unexplained transaction features may reconcile inconsistent findings

The "direct risk aversion" explanation for the uncertainty effect (Gneezy et al., 2006; Simonsohn, 2009) seems at odds with studies that show consumers responding more favorably to risky promotions than to riskless ones. Specifically, Mazar, Shampanier, and Ariely (2016) find that consumers prefer a probabilistic discount to a certain discount of the same expected value (e.g., a 10% chance of getting item for free vs. a certain 10% discount), while Goldsmith and Amir (2010) find that offering a randomly determined prize for making a purchase is nearly as effective as offering the most attractive prize for sure.

If the uncertainty effect were caused by unexplained transaction features, rather than direct risk aversion, at least two explanations arise for the apparent contradiction. First, it may be that consumers can readily identify a reason for a company to offer the type of promotions examined in those studies. They have an explanation, so they are not aversive.[10] Second, in uncertainty effect studies, the focal item (e.g., the gift card participants are purchasing) is uncertain, while in the risky promotion studies, the

[10] A reviewer also suggested that the certain discount may be considered weird in these studies.

"bonus" is uncertain. The focal transaction does not contain an unexplained feature, the bonus does. Perhaps people tolerate (or even prefer) these features in such circumstances.

Another difference is that uncertainty effect studies typically use WTP as their dependent variable, while the risky promotion studies use choice (Mazar et al., 2016) and attractiveness ratings (Goldsmith & Amir, 2010). Perhaps the WTP question implicitly forces a transaction on participants, enhancing the negative suspicions of buyers, but this pressure dissipates in the other tasks. Moon and Nelson (2015) do not replicate the uncertainty effect with a choice task, but Gneezy et al. (2006, p. 1292) do. The role of elicitation mode on the effects of risk and of unexplained features remains an open question, as there are too many differences in these respective designs to meaningfully interpret the differences in results.

Potential transaction feature confounds in other literatures

Much of consumer research involves the comparison of valuations of the same item across different transaction contexts. For example, the endowment effect compares valuations of items being sold against those being purchased, and time preference studies compare the valuations of delayed payments occurring at different points in time (e.g., payments happening today vs. payments happening in the future). Those contextual differences may unintentionally have added unexplained features as well.

For example, it may be the case that giving participants an item and immediately ask them to sell it is an atypical feature, relative to giving them money and offering the opportunity to buy an item. This would depress WTP relative to WTA. Similarly, delaying a payment due today may be perceived as less justified than delaying a payment occurring in the future. This potential confound would lead to more severe discounting of immediate than future delays, typically interpreted as evidence of impatience. In many cases, however, controlling for these differences may be difficult. In our case, for example, we could not find a way to induce risk without adding transaction features, so we added features to the riskless option, this may be the easiest path to control for the weirdness confound in other paradigms as well.

This paper contains a supplement. Table 1 summarizes its contents.

Table 1. Index of supplementary materials (available from http://osf.io/fzjuw)

Section	**Pages**
Supplement 1. Complete age data for Studies 1-4	2
Supplement 2. Additional Analyses for Study 1	3-4
Supplement 3. Within-subject variation in valuation of gift cards in Studies 1-3	5
Supplement 4. Pairwise comparisons across all conditions in Study 4	6
Supplement 5. All means and pairwise comparisons for Studies 5A-B	7
Supplement 6. Study S1 – Isolating and mediating with weirdness	8-10
Supplement 7. Study S2 – Comparing ascending and descending price sequences	11

Appendix. Stimuli used in Studies 5A and 5B

Baseline (randomly selected from the following):

- Imagine that you could buy a $50 gift certificate to Barnes and Noble as part of this study. The gift certificate is good for use within the next two weeks.
- Imagine that you could buy a $50 Target gift card as part of this study.
- We are interested in how much you would pay for a $50 Barnes & Noble gift certificate, which you could buy as part of this study.

Gneezy, List, and Wu (2006, p. 1301) Lottery

Imagine that we offer you a lottery ticket that gives you a 50 percent chance at a $50 gift certificate for Barnes and Noble, and a 50 percent chance at a $100 gift certificate for Barnes and Noble. Whichever gift certificate you win is good for use within the next two weeks.

Yang, Vosgerau, and Loewenstein (2013, p. 737) Certain Coin

We are interested in how much you would be willing to pay for participating in a coin flip. If heads comes up, you will get a $50 gift certificate for Barnes & Noble bookstore. If tails comes up, you will get a $50 gift certificate for Barnes & Noble bookstore.

Yang, Vosgerau and Loewenstein (2013, p. 737) Uncertain Coin

We are interested in how much you would be willing to pay for participating in a coin flip. If heads comes up, you will get a $50 gift certificate for Barnes & Noble bookstore. If tails comes up, you will get a $100 gift certificate for Barnes & Noble bookstore.

Study 4 Certain Box

Imagine that you are standing in front of a table that has a locked box on it. The box has a $50 Target gift card inside. You can pay to open the box and take the gift card, which would be yours to keep.

Study 4 Risky Box

Imagine that you are standing in front of a table that has a locked box on it. The box has two gift cards inside: a $50 Walmart and a $50 Target gift card.

You can pay to open the box and choose a gift card, which will be yours to keep. The gift cards do not have the names of the stores printed on them, so you will not know which gift card is which.

CHAPTER 2.

CRITICAL CONDITION:
PEOPLE ONLY OBJECT TO CORPORATE EXPERIMENTS
IF THEY OBJECT TO A CONDITION

Robert Mislavsky

Berkeley Dietvorst

Uri Simonsohn

ABSTRACT

Why have companies faced a backlash for running experiments? Academics and pundits have argued that it is because the public finds corporate experimentation objectionable. In this paper we investigate “experiment aversion,” finding evidence that, if anything, experiments are rated more highly than the least acceptable policies that they contain. In five studies participants evaluated the acceptability of either corporate policy changes or of experiments testing those policy changes. When all policy changes were deemed acceptable, so was the experiment, even when it involved deception, unequal outcomes, and lack of consent. When a policy change was unacceptable, the experiment that included it was deemed less unacceptable. Experiments are not unpopular, unpopular policies are unpopular.

In June 2014, the Proceedings of the National Academy of Science (PNAS) published an article describing the results of a field experiment where academic authors (Kramer, Guillory, & Hancock, 2014) partnered with Facebook to manipulate content users saw (i.e., "News Feeds"), showing either more positive or more negative emotional content, to measure potential emotional contagion. A month later, the online dating site OkCupid published a blog post titled "We Experiment on Human Beings," which described three experiments they had run on their users (Rudder, 2014). Reaction to the revelation of these experiments was swift and highly negative.

The backlash the Facebook and OkCupid experiments received, described by a Forbes contributor as "one epic freak out" (Muse, 2014), dominated several news cycles despite competing for attention with the 2014 World Cup and major U.S. Supreme Court rulings. Articles describing the negative reaction to the Facebook experiment reached the front page of the Wall Street Journal and were the number one most popular/shared articles on several news outlets, including The Atlantic, The Wall Street Journal, and The BBC.[11] Articles on CNN.com and in the New York Times proclaimed that Facebook treated users like "lab rats" (Goel, 2014; Goldman, 2014). When the OkCupid experiment was revealed, an article in FastCompany declared that the experiment was "way creepier" than Facebook's (Greenfield, 2014). Even legislators got involved, calling for investigations into data collection practices (R. Meyer, 2014; Stampler, 2014). A few months later, Facebook's chief technology officer formally acknowledged that the company was "unprepared" for the reaction elicited by the experiments and admitted that

[11] Internet Archive screenshots from The Atlantic (June 29, 2014), Wall Street Journal (June 30, 2014), and BBC (June 30, 2014) showing lists of most popular articles can be found at https://osf.io/z39aq.

they “should have considered non-experimental ways” to conduct research on the topic (Schroepfer, 2014).

In this paper, we present evidence suggesting that the backlash to these experiments had nothing to do with the experimentation itself. Instead, the backlash was likely driven by the specific policies that these experiments contained (i.e., the individual treatment arms), and reactions would have been at least as negative if these were implemented as standalone policy changes, outside of an experimental context. We conclude that marketing researchers and organizational decision makers should not hesitate to run field experiments using treatment arms that they would also be comfortable implementing as individual policy changes, since experimentation does not make policies *more* objectionable. Similarly, implementing objectionable policies outside of an experiment will not make them more palatable to the public.

FIELD EXPERIMENTS AND MARKETING SCIENCE

Experimentation provides an unrivalled source of actionable intelligence for businesses, governments, and non-profit organizations (Zoumpoulis, Simester, & Evgeniou, 2015), allowing researchers to identify the causal effects that alternative policies have on behavior.[12] Field experiments overcome the lower external validity of stylized lab experiments by taking place in the precise environment where specific policy changes will occur (DellaVigna, 2009). In part because of these advantages, field experimentation has become a popular tool for marketing scholars that is used to test and

[12] We define an experiment as an instance where an organization implements different policies for different groups with the intention of learning how they differently influence a specific outcome.

complement existing theory, as well as develop new insights into buyer behavior on wide-ranging topics. Within marketing, field experiments have been used to explore charitable giving behavior (Sudhir, Roy, & Cherian, 2016), the effect of social influence on the adoption of new technologies (Miller & Mobarak, 2015), strategies for inducing multi-channel buying (Montaguti, Neslin, & Valentini, 2016), and consumer purchasing habits after the end of a promotion (Yanwen Wang, Lewis, Cryder, & Sprigg, 2016).

Given the value of field experimentation, concerns about its acceptability must be taken seriously. Many pundits and scholars have interpreted the backlash to well-known field experiments as evidence that people have a broad and substantial aversion to experimentation. Gino (2015), for instance, proposed that managers are hesitant to run experiments within their own organizations, in part because they believe that customers and employees do not want to be experimented on. Hill (2014) found that companies that do run experiments often resort to using terms like "diagnostic test" or "A/B test" to avoid presumed negative associations with experimentation (see also, Luca, 2014). M. N. Meyer (2015) stated that people view field experiments as "more morally suspicious than an immediate, universal implementation of an untested practice" (p. 278) and titled this preference the "A/B illusion."

If consumers are indeed averse to experimentation, it would constitute an important barrier to evidence-based marketing and future collaborations between academics and organizations. Organizational decision makers may hesitate to run or publicize the results of experiments for fear of negative publicity, and customers may fear engaging with companies that they believe will experiment on them. In this article, we investigate whether or not such an aversion to experimentation exists.

THREE FORMS OF "EXPERIMENT AVERSION"

We define three different forms that experiment aversion could take and preview our ability to empirically distinguish among them in this article:

1. *Absolute* experiment aversion – All experiments are deemed unacceptable, independent of the policies they include.
2. *Relative* experiment aversion – An experiment is less acceptable than the policies it contains, either because experimentation is a negative attribute (i.e., a main effect), or because the underlying policies are deemed less acceptable when they are part of an experiment (i.e., an interaction). This means experiments with acceptable policies could still be considered acceptable in absolute terms, but less acceptable than their underlying policies.
3. *Critical condition* – There is no experiment aversion. The acceptability of an experiment is instead a weighted average of the acceptability of its policies. Most importantly, this implies an experiment is no less acceptable than its least acceptable policy. Thus, an experiment is only viewed negatively if one of its conditions is viewed negatively.

In Studies 1 and 2, we test for absolute experiment aversion and find several instances where experiments are, in fact, rated positively. Thus, we reject absolute experiment aversion. In Studies 3 and 4 we directly pit the acceptability of experiments against the acceptability of their underlying policies, finding that experiments are rated as no less acceptable than their least acceptable policies, consistent with the *critical condition* account of experiment aversion. Experiments, however, were also rated as less acceptable than the simple average acceptability of the underlying policies. This may reflect either moderate relative experiment aversion or negativity bias, where people give more weight to negative attributes than to positive ones (e.g., Folkes & Kamins, 1999; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). In Study 5, we tease these two apart by asking participants to evaluate experiments with two positive policies that are similarly acceptable (thus negativity bias should be absent), and find no evidence of even modest experiment aversion. Therefore, our combined results support the "critical condition" account of experiment evaluation.

TRANSPARENT REPORTING

In all 5 studies, participants read scenarios describing an action that a company could take (either an experiment or a universal policy change) and indicated how acceptable each action is. We ran all studies, except for Study 3b, on Amazon's Mechanical Turk (MTurk) using Qualtrics. Study 3b was a pen-and-paper survey of non-academic university staff.

Study materials, data, analysis code, and supplements for all studies as well as preregistrations for Studies 3b-5 are available at https://osf.io/z39aq. We report studies in the order they were conducted (except for Study 3b, which was added at the request of reviewers and conducted after Study 4) and discuss all additional studies conducted but not reported in the paper in Supplements 5 and 6. For all studies, we determined sample size before beginning data collection.[13] We report all data exclusions, all manipulations, and all measures.

STUDY 1: PEOPLE DO FIND (SOME) EXPERIMENTS ACCEPTABLE

Our first study tests for absolute experiment aversion—people always object to experiments, even if all conditions are unambiguously beneficial. We presented participants with descriptions of corporate experiments that contained unambiguously positive conditions (e.g., giving $5 to employees for visiting the gym) or unambiguously negative conditions (e.g., taking $5 from employees for not visiting the gym). If absolute experiment aversion exists, participants should find all experiments objectionable. If experiments are instead evaluated based on their conditions, participants should only object to experiments that contain unambiguously negative conditions. Throughout these scenarios, we also added various aspects of experimentation that may contribute to experiment aversion, such as deception and lack of consent. If these specific features cause experiment aversion, participants should view these experiments negatively, even if they have only unambiguously positive conditions.

Method

Sample. We recruited 577 participants on MTurk, of which 505 successfully passed the attention check (37.5% female, M_{age} = 34.1 years). Participants were paid $0.75 for completing the study.

[13] In our online studies, we typically obtained sample sizes that slightly exceeded our goals because some participants did not submit a completion code, allowing additional participants to take the survey. Participants, identified by their MTurk ID number, were not able to participate in more than one study. We included an attention check (Oppenheimer, Meyvis, & Davidenko, 2009) in the first question, and only those who answered correctly were able to participate in the studies. All participant responses are included in analyses, regardless of whether or not they completed the entire survey.

Design. Participants were assigned to one of ten experimental conditions. Fifty-three participants were assigned to the *policy change* condition. The remaining participants (N = 452) were assigned to one of nine *experiment* conditions.

Participants in the *policy change* condition read descriptions of nine possible policy changes. These involved *bad*, *good* or *very good* outcomes, in three different contexts. See Table 2. Participants evaluated all nine policies in random order, answering three questions about their acceptability. We average them (Cronbach's α = .96) to construct the "policy acceptability index." These ratings served as a manipulation check for our stimuli in the *experiment* conditions.

Participants in the nine *experiment* conditions read one scenario about a company running an experiment that randomly assigned employees/customers to one of two policy changes from one of the three contexts in Table 2. The condition pairs were *bad/good*, *control/good*, or *good/very good*. For example, the *shipping control/good* scenario read:

> "A shopping company runs an experiment on their shipping system where one group of customers is randomly picked and the company starts upgrading all 'Standard 5-day' shipped packages to 'Priority 3-day' shipping (without changing the cost to the customer). Another group of customers is randomly picked and gets no change in their shipping. The company will then compare customer satisfaction across the two groups."

Participants then answered the same three questions from the *policy change* condition (measures 1-3 in Table 2), but now focusing on the experiment as a whole rather than the underlying policies. They also answered three additional questions designed to more unambiguously evaluate the acceptability of the experiment (rather than willingness to

Table 2. Stimuli and measures for Study 1

Policy changes			
Context	**Bad**	**Good**	**Very good**
1. Shipping	Slower delivery	Faster delivery	Much faster delivery
2. Company gym	$5 penalty for not going	$5 bonus for going	$10 bonus for going
3. Product recommendations	Poorly rated products	Highly rated products	Highest rated overall

Measures of Acceptability

Participants indicated agreement (1=Strongly Disagree; 7=Strongly Agree), with these statements.

Acceptability of policy changes

1. It is okay for the company to do this.
2. If I were [an employee/a customer], I would object to this. (reverse-coded)
3. If I were [an employee/a customer] and was asked, I would agree to this.

Acceptability of experiment

4. It is immoral to run this experiment (reverse-coded)
5. People in this experiment are being treated like guinea pigs (reverse-coded)
6. The company should be not allowed to run this experiment (reverse-coded)

Notes: Participants in the policy change condition rated all nine policy changes. Participants in the experiment conditions rated one of nine experiments created by pairing two policy changes within a context. The pairs consisted of bad/good, control/good or good/very good. Control consists of keeping the status quo (e.g., shipping item as promised). The average of questions 1-3 is the policy acceptability index, the average of questions 4-6 the experiment acceptability index.

participate in it). We average only these additional three questions (α = .86) to construct the "experiment acceptability index." [14]

[14] In hindsight we found questions 1-3 to be ambiguous for interpreting the evaluation of experiments. Therefore, the experiment acceptability index in the main text is based only on questions 4-6. We report results aggregating over all 6 questions in footnote 15.

Participants also answered five comprehension checks to ensure they noticed potentially controversial attributes of the experiments (e.g., "People will be included in this study without agreeing to be included"). No other measures were collected in this condition. Results for measures not reported below are reported in Supplement 1.

Results

Acceptability of policy changes. Validating our choice of stimuli, the overall policy acceptability index for *bad* policy changes (M = 1.91) was below the midpoint (4) and below both the *good* (M = 6.18) and *very good* policy changes (M = 6.11), which were both above the midpoint. All t-tests vs. midpoint are ts > 20.9, *p*s < .001. The *good* and *very good* policies were rated as similarly acceptable, t(312) =.48, *p* = .63, and were close to the highest possible rating (medians of 6.7 and 7 respectively, on a 7-point scale).

Acceptability of experiments. Figure 3 shows the average *experiment acceptability index* for the nine *experiment* conditions. The results are inconsistent with absolute experiment aversion. In particular, when experiments did not include an objectionable condition (*control/good*, M = 5.11; *good/very good*, M = 5.17), they were rated above the midpoint and as more acceptable than when experiments did include an objectionable condition (*bad/good*, M = 3.25). The experiments with objectionable conditions were in turn rated below the midpoint. All t-tests vs. midpoint are ts > 5.9, *p*s < .001. People found experiments to be acceptable when all conditions in the experiment were acceptable and found experiments to be unacceptable when a condition in the experiment was unacceptable.[15]

[15] These results are based on questions 4-6 in Table 2 (see footnote 14). Including all six questions, the results are very similar. Experiments with a bad condition (*bad/good*, M = 3.01) were rated below the

Discussion

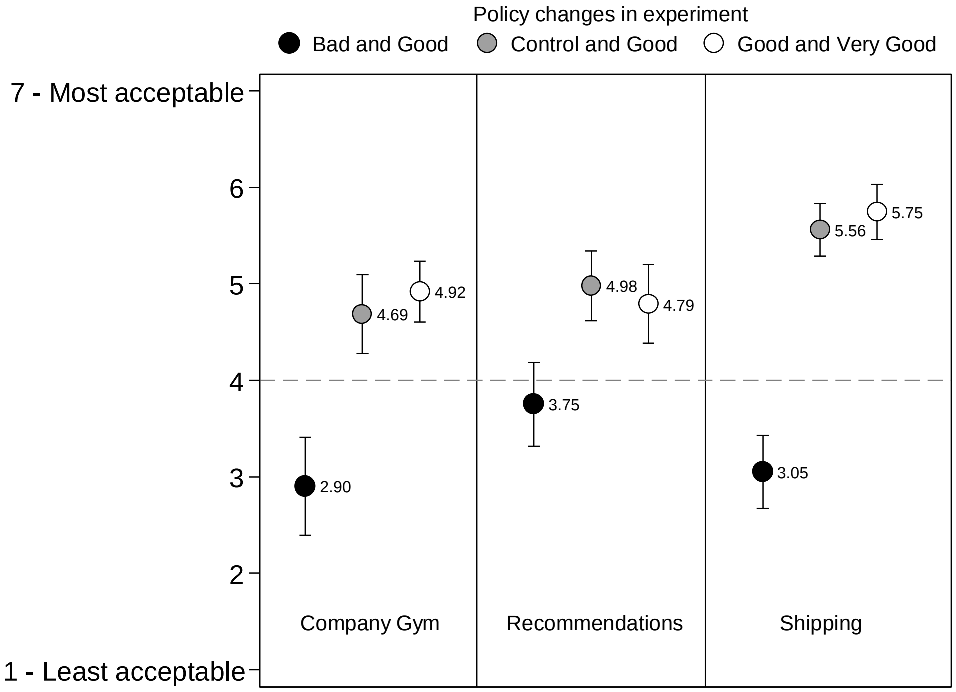
The results from Study 1 are inconsistent with absolute experiment aversion and consistent with a critical condition account of experiment evaluation. Additionally, participants found experiments with deception (e.g., one shipping speed was promised, another was actually delivered), unequal outcomes (e.g., some participants get $5 for attending the gym, others get $10), and lack of consent, to be acceptable, as long as all conditions were themselves acceptable.

However, Study 1 has some important limitations. First, the experiments evaluated as acceptable had unambiguously beneficial outcomes (e.g., free shipping upgrade) and may not generalize to more routine corporate experiments where benefits to participants, if any, are less obvious. Second, we measured agreement with statements rather than absolute measures of acceptability, making it difficult to know whether the experiments are sufficiently acceptable. For example, the *good/very good* experiments were rated M = 5.17 on a 7-point scale where 7 implies strong agreement with the experiment being acceptable. While this is significantly above the midpoint, is it high enough to suggest people would not object to the experiment? Third, participants' ratings in the *policy change* and *experiment* conditions are not directly comparable because: (i) the sets of dependent variables, and their interpretation, are different in the *policy* and *experiment* conditions (see footnotes 14 and 15) and (ii) participants saw all nine policies in the *policy change* condition and only two in the *experiment* conditions. Fourth,

midpoint and below experiments without a bad condition (*control/good*, M = 5.17; *good/very good*, M = 5.30), which were both above the midpoint. All t-tests vs. midpoint are ts > 8.4, *p*s < .001.

participants in this experiment may have higher than average tolerance for experiments because they routinely volunteer for experiments on Amazon Mechanical Turk.

Figure 3. Experiments without bad policies (gray and white circles) are rated positively (Study 1)



Notes: Each participant (N = 452) rated the acceptability of one experiment (out of 9 possible experiments). Markers depict sample averages; error bars represent 95% confidence intervals.

In Studies 2-5 we address all of these issues. We use a wider variety of stimuli (Studies 3a and 3b) and have participants evaluate experiments similar to (controversial) experiments that companies have actually run (Studies 2 and 4). We use questions with less ambiguous endpoints (Studies 2-5) and with neutral and labeled midpoints (Studies 3-5). We have participants in the *policy change* condition rate only the two policy changes that are included in the corresponding *experiment* condition (Studies 3-5) and use the same measures of acceptability across conditions (Studies 2-5). Finally, in Study 3b, we recruited participants who do not routinely volunteer for experiments.

STUDY 2: PREDICTING EXPERIMENT RATINGS FROM CONDITION RATINGS

Kramer et al. (2014) ran an experiment studying emotional contagion through social networks. They manipulated mood by modifying the emotional content of Facebook users' status updates and measured its effect on users' subsequent emotion expression, which upset many users and spurred public outrage (Albergotti, 2014). If, as we have conjectured, people objected to the study because of its polices and not just because it was an experiment, then they should not object to a similar experiment with only acceptable conditions. In Study 2a, we conduct an exploratory search for acceptable and unacceptable mood inductions Facebook could have employed. In Study 2b, we test if the acceptability of the experiment hinges on the acceptability of the mood inductions used.

STUDY 2A: FINDING (UN)ACCEPTABLE MOOD INDUCTIONS

Method

Sample. We recruited 382 participants on MTurk, of which 303 passed the attention check (40.7% female, M_{age} = 30.3 years). Participants were paid $0.30 for completing the study.

Table 3. Overview of study design and contributions

Study 1	• Test of absolute experiment aversion. • People find experiments with unambiguous benefits acceptable.
Studies 2a & 2b	• Extends Study 1 with more realistic stimuli. • Acceptability of conditions predicts acceptability of experiments.
Study 3a	• Direct comparison of experiments with underlying conditions. • Experiments rated at least as acceptable as worst condition is. • Results hold for variety of stimuli.
Study 3b	• Replicates Study 3a results using a sample that does not regularly volunteer for experiments
Study 4	• Best known example of experiment aversion is not an instance of experiment aversion
Study 5	• Experiments with similar and positively-viewed policies are rated identically to the average policy

Design. We generated six interventions, involving positive and negative versions of three possible changes to the site—showing only sad ads, showing only happy ads, showing sad status updates first, showing happy status updates first, showing the least liked status updates first, and showing the most liked status updates first. Each participant evaluated three alternative policies, one for each possible change to the site, randomizing whether participants saw the positive or negative change. We counterbalanced the order of the stimuli.

Measures. Participants answered two questions for each policy change: “Is it okay for a company to do this?” and “Would you object to a company doing this?” These questions were answered on 7-point scales, with endpoints labeled “1. It’s definitely not okay"/ "7. It's definitely okay" and "1. I would definitely object"/ "7. I would definitely not object," respectively. We average the two items (r = 0.69; second question reverse-coded) to construct the "policy acceptability index."

Results

Participants found negative changes less acceptable than positive ones and manipulating status updates less acceptable than manipulating ads. From most to least acceptable, they ranked happy ads (M = 5.67), most liked status updates (M = 4.63), happy status updates (M = 4.58), sad ads (M = 3.90), least liked status updates (M = 3.62) and sad status updates (M = 3.08). For Study 2b, we used the highest rated (happy ads) and lowest rated (sad status updates) changes to test our prediction that experiments are only objectionable if they contain objectionable conditions.

STUDY 2B: EXPERIMENTS WITH (UN)ACCEPTABLE MOOD INDUCTIONS

Method

Sample. We recruited 255 participants on MTurk, of which 201 passed the attention check (43.9% female, M_{age} = 34.2 years). Participants were paid $0.30 for completing the study.

Design. Participants were randomly assigned to one of two conditions in a between-subjects design. In both conditions, participants read descriptions of a social networking company that ran an experiment, assigning half of its customers to a control condition and the other half to a treatment condition. The treatment condition in those experiments was either the *happy ads* or *sad status updates* policy described in Study 2a. Participants answered the same two acceptability questions from Study 2a.

Results

The results were consistent with the critical condition account of experiment evaluation and inconsistent with absolute experiment aversion; only the experiment with an objectionable condition was considered objectionable. Participants rated the *happy ads* experiment significantly above the midpoint (M = 4.72), t(98) = 3.47, $p < .001$, and the *sad status updates* experiment below it (M = 2.59), t(99) = 9.30, $p < .001$.

Although Study 2 shows that experiments with acceptable conditions are acceptable in an absolute sense, relative experiment aversion may still exist if experiments are rated as being *less* acceptable than their underlying conditions. In Study 3 we examine this possibility by directly comparing ratings of individual policies to experiments that use these policies as conditions.

STUDY 3A: TESTING FOR RELATIVE EXPERIMENT AVERSION

Method

Sample. We recruited 533 participants on MTurk, of which 423 passed the attention check (43.5% female, M_{age} = 36.0 years). Participants were paid $0.50 for completing the study.

Design. Participants were randomly assigned to one of six conditions, in a 2 (action: *policy change* vs. *experiment*) x 3 (policy combination: *negative/positive* vs. *no change/positive* vs. *negative/no change*) fully between-subjects design.

Participants in the *policy change* conditions were told that a company was deciding between two policies. They were then told to imagine the company chose one of the policies and answered three questions about the acceptability of this action. They then answered the same questions, but imagining that the other policy had been chosen.

Participants in the *experiment* conditions were told that a company was running an experiment that randomly assigned customers to one of two policies (from the same pool of policy pairs as the *policy change* conditions) and answered the same questions as the *policy change* conditions.

Stimulus selection and sampling. To reduce the probability that the results would be driven by idiosyncratic features of the selected stimuli (Wells & Windschitl, 1999), we presented policy changes for seven different contexts (e.g., showing emotionally charged ads, changing a product recommendation system, and changing frequency of issuing coupons). See Supplement 2 for a full list of stimuli.

Measures. Participants in all conditions answered the following three questions containing labeled neutral midpoints:

1. How okay is it for the company to do this?
   (1 = It's really bad; 4 = It's okay; 7 = It's really good)
2. If you were a customer of this company and learned about the company's plans, how would this influence your opinion of the company?
   (1 = I would view the company much more negatively; 4 = […] not view the company any differently; 7 = […] much more positively)
3. If you were a customer of this company and learned about the company's plans, how likely would you be to switch to a different company?
   (1 = […] definitely not switch […]; 4 = […] not change how likely I am to switch [...]; 7 = […] would definitely switch […]; reverse-coded)

Participants in the *policy change* condition answered these questions twice, once for each policy (in counterbalanced order). Participants in the *experiment* condition answered these questions once, evaluating only the experiment. We average these items (α = .86) to construct an "acceptability index."

Results

Evaluating policy changes. Validating our choice of stimuli, the *negative* policies were rated as the least acceptable (M = 2.65), followed by the *no change* (M = 4.61) and *positive* (M = 5.46) policies. The *negative* policies were rated below the midpoint (4), while the *no change* and *positive* policies were rated above the midpoint, all ts > 6.4, *p*s < .001.

Evaluating experiments. Replicating the results from Studies 1 and 2, and again inconsistent with absolute experiment aversion, experiments that only included acceptable policy changes (*no change/positive*) were rated as acceptable (M = 4.38); significantly above midpoint, t(79) = 3.54, *p* < .001. Conversely, experiments with an unacceptable policy (*negative/positive,* M = 3.22; *negative/no change,* M = 3.31) were rated below the midpoint, ts > 4.3, *p*s < .001. Because, in this study, we used a labeled neutral midpoint (see '*Measures'* above), evaluations above/below the midpoint are unambiguously positive/negative.

Because participants may not all have the same opinion of which policy is "worst," we compare participants' average ratings of each experiment in the *experiment* conditions to the average rating of each participant's less preferred policy in the corresponding *policy change* conditions. When comparing average experiment ratings to the average of the lowest rated corresponding policies, participants found experiments to be significantly more acceptable in the *no change/positive,* t(139) = 2.53, *p* = .013, and *negative/positive*, t(139) = 4.23, *p* < .001, conditions, and marginally more acceptable in the *negative/no change* conditions, t(137) = 1.77, *p* = .079. Collapsing across all policy combinations, experiments were rated as significantly more acceptable than the policy that represented their least acceptable condition, t(419)=5.16, *p* < .001.[16] Most importantly, experiments were not rated as less acceptable than their worst conditions (see Figure 4). This suggests that participants rate experiments as some weighted average of its policies.

STUDY 3B: REPLICATION WITH FIELD SURVEY

One concern about the generalizability of our findings may be that our results to this point have relied on a sample (MTurkers) that regularly opts-in to taking experiments and may therefore be less experiment averse than the general public. In this study, following suggestions of the review team, we replicated our findings using a sample of participants from outside an established participant pool.

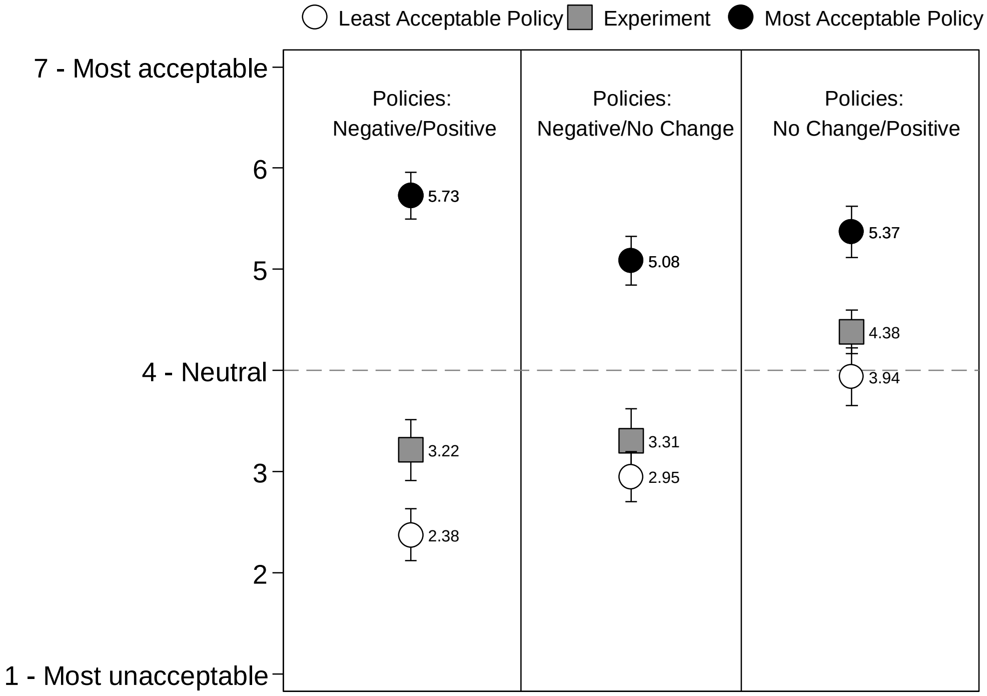
Method

Sample. Three research assistants walked around a university campus, approached non-academic staff members, and asked them if they were willing to take a short, one-page pen-and-paper survey. We specifically instructed the research assistants to approach staff in and around non-academic buildings (e.g., the student union and library) to reduce

[16] These results are consistent when comparing each experiment to the policy change with the lowest average rating (as opposed to the average of each participant's lowest rated policy). Experiments were rated directionally more acceptable than their worst policies in all three cases (significantly so for the negative/positive experiment; t(139) = 3.94, *p* < .001, negative/no change experiment, t(132) =2.06, *p* = .04 and when collapsing across all policy pairs, t(419) = 4.27, *p* < .001).

the likelihood that our participants themselves would be involved in conducting research. It is also important to note that our respondents did not initiate participation in the study (reducing potential selection effects), nor were they compensated for completing the survey (which may have caused them to view academic research and experimentation more favorably). In total, we obtained 247 responses (68.4% female, M_{age} = 33.4 years).

Figure 4. Experiments (gray squares) are no less acceptable than their least acceptable condition (white circles) (Study 3a)



Notes: Each participant (N=423) rated the acceptability of a company choosing one of two policies or running an experiment using those two policies as conditions. The policies involved a negative change, a positive change, or no-change. Circular markers depict means evaluation of each policy, squared markers the evaluations of the experiment that combines them. Error bars represent 95% confidence intervals.

Design. Participants were assigned to one of two conditions (*policy change* vs. *experiment*) in a between-subjects design.

The design of the study was nearly identical to that of Study 3a, with two changes. First, participants only evaluated the *negative/positive* stimuli (i.e., the left-most panel from Figure 4). Second, to make the survey fit on one page, we only included one of the three dependent variables ("How okay is it for the company to do this?") from Study 3a.

Results

Replicating our results from Study 3a, participants rated the experiments (M = 3.54) more favorably than their worst conditions (M = 2.41), t(239) = 7.06, p < .001.[17] These ratings are similar to MTurker ratings of identical stimuli in Study 3 (Experiments: M = 3.35; Worst Conditions: M = 2.26).[18]

Discussion

The results from Studies 3a and 3b are inconsistent with absolute experiment aversion, where people find all experimentation objectionable. Additionally, these results are inconsistent with a version of relative experiment aversion that is large enough to make an experiment less acceptable than its "worst" condition. In our next study, we apply the paradigm from Study 3 to directly examine the potential role of experiment aversion in the backlash to Kramer et al. (2014)'s Facebook experiment. Specifically, we assess whether the backlash may actually be attributed to the policies people were assigned to rather than experimentation per se.

[17] This analysis was done using a regression with fixed effects for each stimulus. We preregistered that we would also conduct a simple t-test collapsing across stimuli. The results are consistent, t(245) = 6.24, p < .001.

[18] These numbers are not the same as those in Study 3a (and in the left panel of Figure 4) because in Study 3a we used a composite of three measures. Here, we compare only results for the question ("Is it okay for the company to do this?") that we used in both studies.

STUDY 4: WAS FACEBOOK BACKLASH REALLY EXPERIMENT AVERSION?

As in Study 2, we investigated perceptions of an experiment based on Kramer et al. (2014). Unlike in Study 2, we used only stimuli that represented the specific conditions used in that experiment, rather than modifying certain aspects to find an "acceptable" version. We also used the same bipolar scales as Study 3, with labeled neutral midpoints, to evaluate policy changes and experiments.

Method

Sample. We recruited 748 participants on MTurk, of which 608 passed the attention check (41.3% female, M_{age} = 32.2 years). Participants were paid $0.30 for completing the study.

Design. The overall design of Study 4 was nearly identical to that of Study 3, but used different stimuli. Participants were randomly assigned to one of six conditions in a 2 (action: *policy change* vs. *experiment*) x 3 (policy combination: *sad/happy* vs. *no change/happy* vs. *sad/no change*) fully between-subjects design.

Participants in the *policy change* condition read that Facebook was considering making two policy changes (randomly selected from: sorting status updates to prioritize *happy* ones, to prioritize *sad* ones, or making *no change*). They then read that Facebook chose to implement one of the two policies. Participants in the *experiment* condition read that Facebook was considering running an experiment where they would randomly assign customers to two of the policy changes described above.

Measures. Participants answered the same acceptability questions from Study 3. However, because Facebook does not have an obvious competitor, we did not ask if participants would switch to a different company.[19] We average these two variables (r = .80) to construct the "acceptability index." Participants then indicated whether or not they had previously heard of Facebook taking similar actions in the past. This was collected to account for participants that may have been influenced by media coverage of the Facebook study.[20]

Results

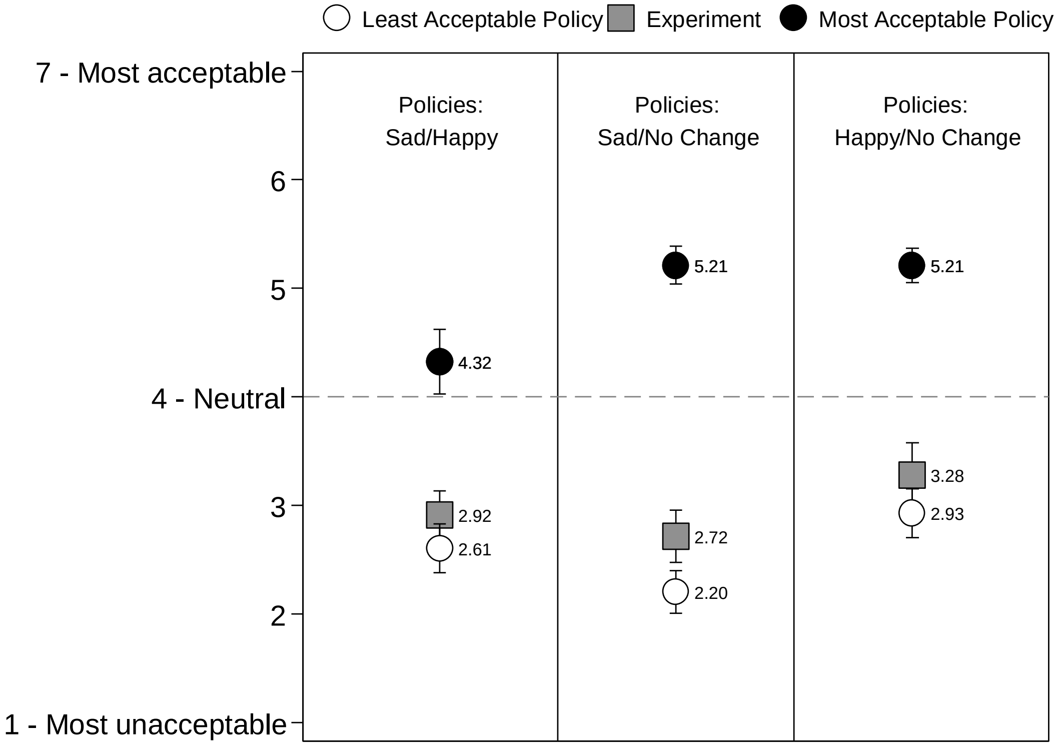
Figure 5 shows the main results from Study 4. All three experiments (grey squares), even the experiment with ostensibly "good" conditions (i.e., *happy/no change*), were rated significantly below the acceptability midpoint (ts > 4.8, *ps* < .001). At first glance, this could be consistent with absolute or relative experiment aversion. However, this conclusion is not supported once we take into account the fact that the underlying policies are unacceptable even outside of an experimental context. The lowest rated condition in each experiment was rated no higher than a 2.93 on a 7 point scale; significantly below midpoint, ts > 9.4, *p*s < .001.[21]

[19] We exploratorily asked if participants would be inclined to cancel their Facebook membership; see preregistration file.

[20] Most participants said that they had not heard of Facebook doing something similar (70.3% in the *experiment* condition and 82.2% in the *policy change* condition). Those with prior knowledge in the *experiment* condition rated Facebook's actions slightly more negatively (M = 2.77) than those with no prior knowledge (M = 3.06), t(301) = 1.75, *p* = .08. There was no difference between ratings in the *policy change* condition (*p* = .81). Therefore, we report results from all participants in our analysis.

[21] The only specific policy that was rated above the midpoint was making no change (M = 5.10). Both sad status updates (M = 2.48), and happy status updates (M = 3.67) are viewed as unacceptable (all pairwise ts > 8.0, all ts vs. midpoint > 3.0).

Figure 5. Facebook experiments (gray squares) are no less acceptable than their least acceptable condition (white circles) (Study 4)



Notes: Each participant (N = 601) rated the acceptability of Facebook changing how status updates are sorted or of running an experiment randomly assigning users to one of those changes. Circular markers depict mean evaluations of the least and most acceptable change in the pair, squared markers mean evaluations of an experiment randomly assigning users to them. For example, the first panel shows that people evaluating sorting status updates by sad/happy rated the worst of these with M=2.70, the highest with M = 4.22, and an experiment with M = 2.92. Error bars represent 95% confidence intervals.

As was the case in Study 3, when we directly compare the acceptability of experiments to the acceptability of their treatments' in the corresponding *policy change* conditions, we see that experimentation does *not* decrease the acceptability of the company's actions relative to some weighted average of its policy ratings. Indeed, experiments were again rated as at least marginally more acceptable than their worst conditions when considering each experiment individually, ts > 1.88, *p*s < .061, and significantly more acceptable when collapsing across all three experiments, t(599) = 3.94, *p* < .001.[22]

Discussion

Again, if there is relative experiment aversion, it is not large enough to push the experiment's ratings below the ratings of its policies. Thus, it is probable that participants were not reacting negatively to experimentation per se but to each experiment's underlying policies. Although the reaction to the Kramer et al. (2014) Facebook experiment is held up as evidence of a public distaste for corporate experiments, in Study 4 we find that Facebook probably did not face backlash because they ran an experiment, but because they implemented unacceptable policies. This suggests the public's reaction would have been even worse had Facebook modified how status updates are sorted for all (rather than for a random subset) of its users.

STUDY 5: RELATIVE EXPERIMENT AVERSION VS. CRITICAL CONDITION

Studies 3 and 4 demonstrate that relative experiment aversion, if it exists, may not be strong enough to drive ratings of an experiment below some weighted average of its policies. However, we cannot conclusively reject the existence of *some* relative experiment aversion. Even though the experiments were not rated worse than the least preferred policy, they were still rated below the equally-weighted average of its policies.

[22] As indicated in our pre-registration, we ran a regression estimating ratings using fixed effects for each policy pair and an indicator for whether the participant rated a policy or an experiment. The coefficient for experiments was positive (b = .39; *p* < .001), indicating that experiments were rated more highly than policies when controlling for which policies participants saw.

This could be consistent with the critical condition account of experiment aversion if participants are taking a weighted average of their ratings of the two policies and giving more weight to the worse rated policy, as they might if they exhibit negativity bias (Skowronski & Carlston, 1989). However, this finding could also be consistent with the existence of moderate relative experiment aversion. For example, participants may be averaging their opinions of the policies and then applying some fixed “experiment penalty.” Alternatively, participants’ ratings of policies could be lower when those policies are part of an experiment. We ran Study 5 to more directly tease apart these two explanations by creating an experiment where both policies would be deemed equally acceptable. If there is relative experiment aversion, an experiment over both policies would be rated as lower than either, which would not happen if people evaluate experiments based on their critical conditions. We view this design as one which maximizes the ability to detect relative experiment aversion.

Method

Sample. We recruited 502 participants on MTurk, of which 406 passed the attention check (46.4% female, M_{age} = 35.0 years). Participants were paid $0.40 for completing the study.

Design. Participants were randomly assigned to one of two between-subjects conditions (*policy change* vs. *experiment*). We pretested the acceptability of 30 policies (see Supplement 4) and chose two that had nearly identical means (Ms = 5.54 and 5.59 out of 7) and distributions of responses (SDs = 1.40 and 1.32). The general design of Study 5 was similar to that of Studies 3 and 4. Participants read that a ride-sharing company (e.g., Uber, Lyft) was considering implementing two discounts (either a flat 10% discount or a $1 credit for every $10 spent) and either chose one of the two (*policy change* condition) or ran an experiment where they randomly assigned customers to receive one of the two discounts (*experiment* condition).

In both conditions, participants answered the following question: “How okay is it for the company to do this?” (1 = It’s really bad; 4 = It’s okay; 7 = It’s really good).

Results

Participants rated both discounts (10% discount: M = 5.84; $1 credit for every $10 spent: M = 4.85) significantly above the midpoint, ts > 9.14, *p*s < .001, indicating that they viewed both discounts positively.[23] Participants rated the experiment that assigned participants to one of two discounts (M = 5.32) nearly identically to the *average* discount (M = 5.34), t(399) = .21, *p* = .83, and well above the least preferred discount (M = 4.61), t(399) = 5.24, *p* < .001. Participants in this study do not show even small levels of experiment aversion.[24]

GENERAL DISCUSSION

Taken together, the results of our studies are inconsistent with both absolute and relative experiment aversion, while consistent with the critical condition account of

[23] We should point out that the mean ratings of the individual discounts diverged more in Study 5 (Ms = 4.85 and 5.85) than they did in the pilot (Ms = 5.54 and 5.59). We believe that this is because evaluating only two discounts (compared to 10 in the pilot), made those discounts seem less similar.

[24] The 95% confidence interval for the difference between the acceptability of the experiment and the average policy is (-.21, +.26), thus we reject experiment aversion that is larger than .26 on our 7 point scale. With a pooled standard deviation of 1.19, we can reject experiment aversion having a Cohen’s d > .22.

experiment evaluation. In particular, experiments that include only acceptable policies are deemed acceptable, and whether they include acceptable or unacceptable policies, experiments are deemed to be at least as acceptable as their least acceptable policy.

These results are good news for companies that want to learn from experiments. Companies should not be more hesitant to run an experiment that includes a certain policy than to implement that policy outright. A practical takeaway for organizations interested in running experiments is to first determine if their planned policy changes are objectionable (e.g., through a survey) and then run an experiment to determine which acceptable policy best achieves their desired objective.

Limitations

We have identified two key limitations with our studies. The first limitation is that our samples consist primarily of people who volunteered to complete our studies, possibly excluding individuals who most strongly oppose evidence gathering in general or experiments in particular. We are optimistic this is not a consequential limitation for two main reasons. First, our respondents did negatively evaluate experiments that included negative policies, indicating that they do not have universally positive opinions of experiments, and that they do discriminate between acceptable and unacceptable practices. Second, Study 3b surveyed a sample of non-academic university staff, who do not regularly participate in experiments. Their responses were indistinguishable from those of our online samples. It is nevertheless obviously impossible to obtain data on the attitudes of people who are unwilling to participate in an experiment.

The second limitation is that it is difficult to specify the threshold of acceptability that an action must reach to prevent a backlash. For example, a small group of motivated people (e.g., activists or media personalities) could be vocal enough to cause backlash against an experiment that most people find acceptable. At the same time, this concern applies to any action an organization can take and not solely experiments. Comparing the most extreme ratings across policy and experiment evaluations in our studies suggests experiments are not more polarizing than are policies. In Study 3a, for example, 12.5% of participants gave the negative policy the lowest possible rating and 7.6% of participants gave the experiment the lowest possible rating, a pattern that holds in all studies where this comparison is possible.[25]

This also speaks to a larger issue of how different people may view different policy changes—what some may consider fine, others may find completely unacceptable. For this reason, we compared experiments to each participant's *least preferred* policy, rather than the average of each specific policy. Additionally, it is important to examine distributions of responses (rather than means) to determine if a certain policy, although it may have a high mean, may be especially divisive (i.e., having a high variance). We encourage researchers and practitioners to pretest the acceptability of policies using surveys and measures like those we used in Studies 3 through 5.

[25] In Study 3b, 35.5% gave the lowest possible rating to the worst policy, compared to 9.8% for the experiment. In Study 4, these values are 20.7% and 12.9%, respectively, and in Study 5, they are 2.5% and 1.0%, respectively.

Experiment aversion is an interaction

Finally, there are many factors that could influence how acceptable experiments are. For example, much research has examined how people view the ethics of corporate practices that can be included in experiments, such collecting sensitive data (e.g., Awad & Krishnan, 2006; Culnan & Armstrong, 1999; Miyazaki, 2008), changing pricing practices (e.g., Bolton, Warlop, & Alba, 2003; Campbell, 1999; Haws & Bearden, 2006), or introducing new marketing strategies (e.g., Smith & Cooper-Martin, 1997).

Using the more specific context of our motivating example, it may be that Facebook's experiment was more objectionable because it involved emotions (or specifically *negative* emotions).[26] Our review team, in particular, proposed that perhaps people view experiments as less acceptable if they are *in* an experiment compared to if they simply heard about it or that it is less acceptable to tell customers about experiments after the fact than before they are run. We report two studies that test these two hypotheses in the supplement (Studies S4 and S5). We find that people prefer to hear about experiments before (rather than after) they are run, and that people rate hypothetical experiments that they were in similarly to those they merely heard about.

However, asking "Do these factors impact the acceptability of experiments?" will not teach us about experiment aversion, because these factors can be present in corporate actions within and but also outside of an experiment. A company can take an action and only later tell customers about it. A company can also take an action and some non-participating observer then evaluate it. The critical question for the purposes of this

[26] See Supplements 5-6 for descriptions of studies that test these questions.

paper, then, is “Do these factors impact the acceptability of experiments *more than they impact the acceptability of underlying policies*?” That is, is there an *interaction* between these factors and whether or not they are part of an experiment? In Studies S4 and S5, we find none of these hypothesized interactions (Study S4: t(794)=.99, *p*=.32; Study S5: t(793)=.56, *p*=.58). For example, in Study S4 we find that the negative effect of learning about an experiment after it is conducted (versus before it is conducted) is not larger than the negative effect of learning about a policy change after it is conducted (versus before it is conducted). We would expect the same to be true for other potential factors that could influence opinion of experiments and universal policy changes. Experiments are not unpopular, unpopular policies are unpopular.

This paper contains a supplement. Table 4 summarizes its contents.

Table 4. Index of supplementary materials (available from https://osf.io/z39aq)

Section	**Pages**
Supplement 1. Additional Study 1 analysis	2-3
Supplement 2. Full list of Study 3 stimuli	4
Supplement 3. Additional analyses for Study 4 included in pre-registration	5-7
Supplement 4. Study 5 pilot results	8-9
Supplement 5. Overview of studies not included in main manuscript	10-11
Supplement 6. More details on studies not included in main manuscript	12-18

CHAPTER 3.

60% + 60% = 60%, BUT LIKELY + LIKELY = VERY LIKELY

Robert Mislavsky

Celia Gaertig

ABSTRACT

How do we combine others' probability forecasts? Prior research has shown that when advisors provide *numeric* forecasts, people typically average them together. If two advisors think an event has a 60% chance of occurring, we will also believe it has a 60% chance (more or less). However, what happens if two advisors say that an event is "likely" or "probable"? In four studies, we find that people combine verbal forecasts additively, making their forecasts more extreme than each advisor individually. If two advisors say something is "likely," people then believe that it is "very likely."

Imagine that you are heading out the door and wondering if you should bring an umbrella. You check a weather app, which says there is a 30% chance of rain, and just to be sure, you turn on the TV where the weatherperson says that there is a 50% chance of rain. Given these two forecasts, do you bring an umbrella? Situations like these are common in daily life, from the mundane, such as bringing an umbrella, to the serious, such as getting a second opinion about a medical diagnosis. So how do we combine these forecasts to make our own judgments? Well, it depends.

In the above example, the forecasts were numeric. We know from prior research that we generally combine numeric probability forecasts by averaging them (Biswas, Zhao, & Lehmann, 2011; Budescu & Yu, 2006, 2007; Wallsten, Budescu, & Tsao, 1997). If one advisor says there is a "30% chance" and another says there is a "50% chance," our own forecasts will typically be somewhere between 30% and 50%. However, we generally don't use numeric probabilities in daily speech. Instead of saying there is a "60% chance," we use verbal probabilities, saying that an event will "probably" happen, or that it is "likely" (Erev & Cohen, 1990; Zimmer, 1983). Despite this, there has been no study of how we combine verbal probability forecasts, for example, how our own beliefs update when two people tell us that something is "likely." In the four studies that follow, we find that people tend to combine verbal probability forecasts *additively*. Beliefs about an event's likelihood move closer to certainty when another person says an event is likely and closer to impossibility when another says an event is unlikely, *regardless of prior beliefs*.

DIFFERENCES BETWEEN NUMERIC AND VERBAL FORECASTS

The differences between how we combine verbal versus numeric forecasts may be traced to several documented differences between how numeric and verbal probabilities are interpreted more generally. First, numeric probabilities are precise, while verbal probabilities are vague (Beyth-Marom, 1982; Lichtenstein & Newman, 1967; Zimmer, 1983). While a "60% chance" has a precise mathematical meaning, in a seminal study, Lichtenstein and Newman (1967) found that "likely" was interpreted to mean anything from 25% to 99%. Second, the subjective interpretations of numeric probabilities are more context-dependent than verbal probabilities, which are processed more intuitively (Bilgin & Brenner, 2013; Teigen, 2001; Teigen & Brun, 1995, 1999, 2000; Windschitl & Weber, 1999; Windschitl & Wells, 1996). It is easier to evaluate verbal probabilities as a positive or negative sign than it is for numeric probabilities. For example, a candidate that is "likely" to win an election should feel confident, but a candidate with a "30% chance" might feel confident if there are 10 other candidates, but not if there are two.

Recognizing these differences, organizations that provide subjective probability forecasts have tried to standardize the interpretation of verbal probabilities in their reports. The Intergovernmental Panel on Climate Change (IPCC)[27] defines "likely" as "greater than 66%," and the United States Director of National Intelligence (DNI)[28] defines it as "between 55% and 80%," although research suggests that these guidelines

[27] https://www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf
[28] https://fas.org/irp/dni/icd/icd-203.pdf

are mostly ineffective (Budescu, Broomell, & Por, 2009; Budescu, Por, & Broomell, 2012; Budescu, Por, Broomell, & Smithson, 2014).

However, even if these guidelines worked perfectly, they assume that verbal and numeric probabilities differ only in how they are initially interpreted. We show that this is not the case. They also differ in how they are aggregated, which can lead to drastically different judgments from relatively similar inputs. For example, imagine a group of military officials deciding to launch a risky operation. If the collected experts all agree that the operation has a 60% chance of success, the ultimate decision-maker should also think that the operation has a 60% chance of success. However, if the experts all agree that success is "somewhat likely," we show that the decision-maker might think that success is nearly certain.

TRANSPARENT REPORTING

We report four studies in this manuscript and include six more in the supplement. All studies were run on Amazon's Mechanical Turk (MTurk) and were administered through Qualtrics. Studies 2-4 were preregistered. For all studies we report all data exclusions, all manipulations, and all measures. Preregistered exploratory measures are mentioned in footnotes and discussed Supplement 1, along with preregistered secondary analyses (e.g., robustness checks). Sample size for each study was determined before data collection, and participants in all studies were excluded from participating in any related studies run within one month. We analyze all answers participants provided, regardless of whether they completed the survey. Supplementary materials, including data, analysis code, preregistrations, and survey materials, are available at http://osf.io/atruq.

STUDIES 1-3: PREDICTING FUTURE EVENTS

Our first three studies have a relatively common design. Participants are asked to predict the likelihood of an event. To help make their forecasts, they are shown forecasts from one or two advisors. Forecasts are given either verbally (e.g., "Rather Likely") or numerically (e.g., "60%"), which we refer to as forecast *formats*. Participants then make their own forecasts on scales using the same format that the advisors used.

Analyses

To test the combination strategies that participants use, we look exclusively at the proportion of participants that make *extreme* forecasts (i.e., forecasts that are closer to impossibility or certainty than any individual advisor). For example, if the two advisors in the study say that an event has a 60% and 65% chance of occurring, an extreme forecast is anything that is 66% or higher. We predict that as the number of advisors increases, more participants will make extreme forecasts when using verbal probabilities than when using numeric probabilities. That is, we predict a positive interaction between *format* and the *number* of advisors on the likelihood of making an extreme forecast.

We use this strategy because it is the most diagnostic test of whether participants use an additive strategy, compared to, say, testing for mean differences. For example, means can increase if participants move from far below the advisors' average to slightly below the advisors' average. This could be consistent with both an averaging or additive combination strategy. On the other hand, a participant who moves from a non-extreme forecast to an extreme forecast could not possibly be using an averaging strategy.

Finally, unless specified otherwise, analyses for Studies 1-3 are conducted using probit regressions, including indicator variables for condition (and their interactions where appropriate), fixed effects for each stimulus (e.g., stock), and clustering standard errors by participant. Full regression tables can be found in Supplement 2.

STUDY 1: LIKELY + LIKELY = VERY LIKELY

Method

Sample. We recruited 205 participants (35.0% female, M_{age} = 33.7 years), each paid $0.30.

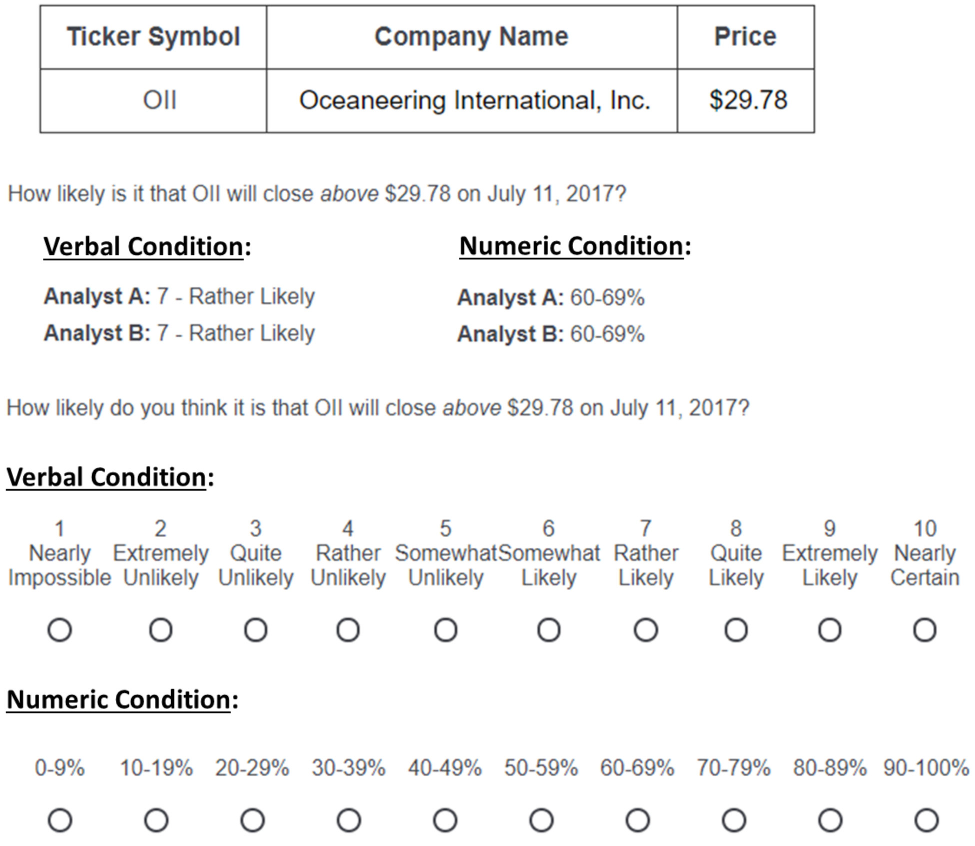
Design. Participants were randomly assigned to one of two between-subjects conditions. All participants saw information about a stock[29] (ticker symbol, company name, and most recent closing price) and predicted how likely it was that the stock's price would be higher in one year. Before making their own forecasts, participants also saw forecasts from two (fictional) advisors.

In the *numeric* condition, the advisors' forecasts were "60-69%," and participants made their forecasts on a 10-point *numeric* probability scale (1 = "0-9%"; 10 = "90-100%"). In the *verbal* condition, the two advisors' forecasts were "7 – Rather Likely,"

[29] For Studies 1 and 2, stocks were randomly selected from a list of 10. Our analyses in both studies include fixed effects for each stock. See Supplement 3 for full list of stimuli used.

and participants made their forecasts on a 10-point *verbal* probability scale (1 = “1 – Nearly Impossible”; 10 = “10 – Nearly Certain”; adapted from Windschitl & Weber, 1999). In both conditions, the advisors’ advice corresponded to the 7th point on their respective scales, keeping the extremity of advisor forecasts constant across condition. See Figure 6 for example stimuli and response scales.

Figure 6. Sample Study 1 stimuli and response scale



Ticker Symbol	Company Name	Price
OII	Oceaneering International, Inc.	$29.78

How likely is it that OII will close *above* $29.78 on July 11, 2017?

Verbal Condition:

Analyst A: 7 - Rather Likely
Analyst B: 7 - Rather Likely

Numeric Condition:

Analyst A: 60-69%
Analyst B: 60-69%

How likely do you think it is that OII will close *above* $29.78 on July 11, 2017?

Verbal Condition:

1 Nearly Impossible	2 Extremely Unlikely	3 Quite Unlikely	4 Rather Unlikely	5 Somewhat Unlikely	6 Somewhat Likely	7 Rather Likely	8 Quite Likely	9 Extremely Likely	10 Nearly Certain
○	○	○	○	○	○	○	○	○	○

Numeric Condition:

0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
○	○	○	○	○	○	○	○	○	○

Results

We classify participant forecasts as "extreme" if they are closer to certainty than *both* advisors' forecasts (i.e., answers from 8 to 10 on the response scale). More participants in the *verbal* condition made extreme forecasts (29.4%) than in the *numeric* condition (10.9%), Z = 3.29, *p* = .001.

This result suggests that participants are more likely to use an additive strategy when combining verbal forecasts than they are when combining numeric forecasts. However, this study has a key limitation: we do not know what participants' forecasts would have been if they had only seen one advisor forecasts. It may be that participants always make more extreme forecasts when using verbal probabilities. Therefore, we need to compare the amount of extreme forecasts participants make when only seeing one advisor forecast and see how that proportion changes as they see additional advisor forecasts. We do this in Study 2.

STUDY 2: SEQUENTIAL EVALUATION (AND UNLIKELY + UNLIKELY = VERY UNLIKELY)

Method

Sample. We recruited 854 participants, of which 806 passed an attention check (39.0% female, M_{age} = 33.4 years). Participants who completed the survey were paid $0.35.

Design. The general design of Study 2 was extremely similar to that of Study 1, with two changes. As in Study 1, all participants saw information about a stock and estimated how likely it was that the stock's price would be higher in one year on either a verbal or numeric scale. Participants again saw forecasts from two fictional advisors, given either numerically or verbally.

Unlike in Study 1, advisor forecasts were shown one at a time (i.e., manipulated within-subjects). Participants saw a forecasts from the first advisor, made their own forecast, saw a forecast from a second advisor, and could revise their first forecast. Further, we tested if our results held for forecasts below even chance by randomizing advisors' forecasts to be the 7th point on the response scale (i.e., "60-69%" or "Rather Likely") or the 4th point on the scale (i.e., "30-39%" or "Rather Unlikely").

In summary, participants were assigned to one of four between-subjects conditions in a 2 (format: numeric vs. verbal) x 2 (direction: above vs. below midpoint). Number of advisors was manipulated within subjects for all participants.[30]

Results

Again, we classify participant forecasts as "extreme" if they are closer to certainty than each advisor's forecast (8 to 10 in the *above midpoint* conditions; 1 to 3 in the *below midpoint* conditions). We preregistered that we would analyze the *above* and *below*

[30] We also asked participants two exploratory questions about their perceptions of advisor consensus. Participants perceived more consensus in the *verbal* condition (*p*s < .003), but this did not mediate our effect. See Supplement 1.

midpoint conditions together, but for ease of interpretation, we discuss them separately here and include results from the combined analyses in footnote 31.

In the *above midpoint* condition, participants' forecasts became more extreme when they saw a second advisor. Specifically, 18.3% of participants made an extreme forecast after seeing the first advisor in the *verbal* condition, which increased to 29.7% after seeing the second advisor, Z = 4.13, *p* < .001. In contrast, the proportion of extreme forecasts in the *numeric* condition directionally decreased as participants saw more advisors (11.4% to 9.0%), Z = 1.24, *p* = .21. The interaction between format and number of advisors is significant, Z = 3.62, *p* < .001.

This pattern also held in the *below midpoint* condition. As they saw the second advisor's forecast, the number of participants making extreme forecasts increased in the verbal condition (13.1% to 23.1%, Z = 3.28, *p* = .001) but decreased in the numeric condition (18.3% to 13.4%, Z = 2.46, *p* = .014). The interaction between format and number of advisors is significant, Z = 4.14, *p* < .001.[31] See Figure 7.

STUDY 3: REAL EVENTS AND ADVICE

In Studies 1 and 2, we tested how participants used forecasts from two fictional advisors that gave identical forecasts. In Study 3, participants make forecasts for real events using real advice (and as a result, had natural variation between advisors).

[31] Combining the *above* and *below* midpoint conditions into one regression, the interaction between format and number of advisors is significant, Z = 3.99, *p* < .001. There is no significant 3-way interaction between format, number of advisors, and above/below midpoint, Z = -.41, *p* = .68, indicating that the effect is approximately the same size for forecasts better or worse than even chance.

Method

Sample. We recruited 626 participants (42.7% female, M_{age} = 35.6 years), each paid $0.50.

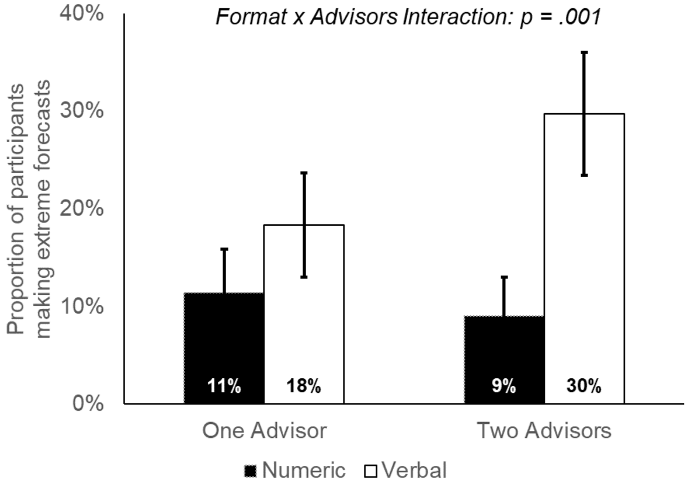
Design. Participants were assigned to one of four between-subjects conditions in a 2 (format: numeric vs. verbal) x 2 (number of advisors: 1 vs. 2) design. All participants were shown information about ten Major League Baseball games (randomly selected from the 15 games played that day) and asked to predict how likely it was that the favorite would win each game. For each game, participants either saw one or two forecasts, randomly selected from Fivethirtyeight.com, Fangraphs.com, or VegasInsider.com.[32]

In the *numeric* condition, advisor forecasts were given as percentages (e.g., "55%"), and in the *verbal* condition, they were given as a number with a verbal label (e.g., "55 – Somewhat Likely"). The number was added to the verbal condition to keep the extremity of the advice consistent across conditions. Because advisor forecasts could take any value between 0 and 100, participants answered on a 0 to 100 slider scale with numeric or verbal labels depending on condition.

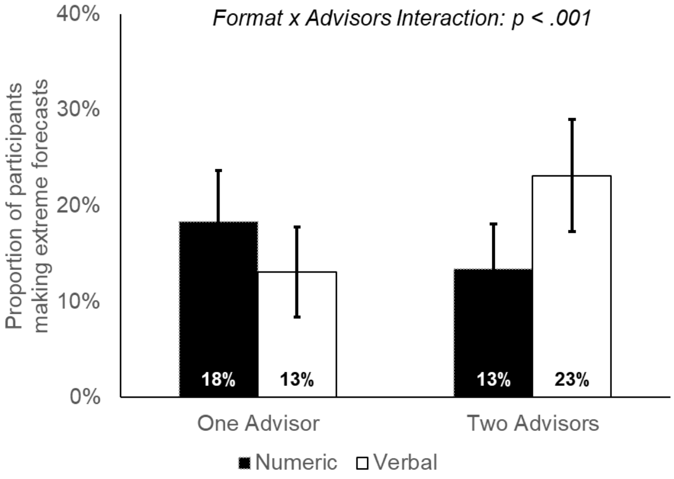
[32] We also collected measures of participants' baseball knowledge, favorite team, motivation, and trust in the advisor websites. See Supplement 1.

Figure 7. Participants' forecasts become more extreme when they see additional verbal forecasts, but not when they see equivalent numeric forecasts

Above Midpoint Condition



Below Midpoint Condition



Notes: Advisor forecasts in the *above midpoint* condition correspond to the seventh point (out of 10) on the response scale ("60-69%" in the *numeric* condition and "Rather Likely" in the *verbal* condition). Advisor forecasts in the *below midpoint* condition correspond to the fourth point on the response scale ("30-39%" in the *numeric* condition and "Rather Unlikely" in the *verbal* condition. Extreme forecasts are those that are above the seventh point in the *above midpoint* condition and below the fourth point in the *below midpoint* condition. Error bars represent 95% confidence intervals.

Results

Although participants made forecasts for all games played that day, we preregistered that we would only analyze forecasts for games where both advisors agreed on a winner, since those are the only games where we can meaningfully distinguish extreme forecasts from average ones.

Again, we classify participant forecasts as "extreme" if they are closer to certainty than each advisor forecast for each game.[33] When participants saw only one advisor forecasts, there were no differences between the proportion of extreme forecasts in the *numeric* (50.0%) and *verbal* (55.5%) conditions, $Z = 1.39$, $p = .17$. However, participants that saw two advisor forecasts made many more extreme forecasts in the *verbal* condition (46.6%) than in the *numeric* condition (29.8%), $Z = 5.11$¸ $p < .001$.[34] The interaction between format and number of advisors is significant, $Z = 2.93$, $p = .003$.[35]

STUDY 4: DECISIONS BASED ON FORECASTS

In Studies 1 to 3, we find a common pattern. When participants see multiple *verbal* probability forecasts from advisors, they are much more likely to combine them

[33] We preregistered that we would classify "extreme" as above the *average* advisor's forecast and that the classification in the main text would be a secondary analysis. However, the analysis reported here is a more conservative test and consistent with our definitions from Studies 1 and 2. The results using the original definition are nearly identical, and we report them in Supplement 1.

[34] We should note that, unlike in Study 2, the number of extreme forecasts *decreased* in both conditions when participants saw two advisor forecasts. We believe that this is due to the granularity of the scale in this study. That is, it is more difficult to give an exactly average forecast in this study (when it is 1 out of 101 points) than in Study 2 (when it is 1 out of 10 points). Indeed, in this study, only 13.0% of forecasts were exactly "average" when participants only saw one advisor forecast, compared to 44.2% in Study 2.

[35] We preregistered that we would include participant motivation, baseball knowledge, and average advisor forecast as control variables. Without these controls, the interaction is significant, $Z = 2.56$, $p = .01$.

additively than they are when they see multiple *numeric* probability forecasts. However, it may be that this is caused by differences in how participants use the respective response scales rather than, as we hypothesize, becoming more certain of an event's outcome. We test this in Study 4, where participants make a decision that should be informed by their beliefs about the event's likelihood.

Method

Sample. We recruited 809 participants (44.8% female, M_{age} = 35.3 years), each paid $0.40.

Design. All participants were randomly assigned to read one of two scenarios about making a purchase that involved uncertainty. In one scenario, participants read that they were buying a plane ticket, where the price could change in the future. In the other scenario, participants read that they were buying a cell phone, and a new model could be released shortly. See Supplement 4 for full stimuli.

In both scenarios, participants were told that they checked a forecasting website (e.g., Kayak.com in the plane ticket scenario). The website recommended waiting to make the purchase, giving either a *verbal* or *numeric* forecast that the price of the plane ticket would drop or that a new model would be released. Participants then indicated whether they would make the purchase on a 7-point scale (1 = Definitely buy; 7 = Definitely wait).

After indicating their purchase intent, participants were told that they checked a second website, which gave the same forecast as the first. Finally, participants again indicated whether they would make the purchase on the same 7-point scale.[36] For example, participants in the *plane ticket* condition saw the following:

> You want to buy a plane ticket for a vacation you are taking. You found a ticket that fits your budget, but know prices can drop if you wait (although they can also go up or the flight could sell out). You're willing to wait up to two weeks. You check a price prediction website, which says the following:
>
> "It is [somewhat/rather/55%/65%][37] likely that prices will drop in the next two weeks."
>
> Would you buy the ticket or wait to see if the price goes down?
>
> [page break]
>
> You decide to get a second opinion and check a different site that also makes price predictions. The second site says:
>
> "We think that it is [somewhat/rather/55%/65%] likely that prices will decrease within the next two weeks."
>
> Would you buy the ticket or wait to see if the price goes down?

Results

We preregistered that we would analyze the data collapsed across both scenarios, comparing the proportion of participants who became *more likely to wait* (i.e., more strongly agreed with the advice) when they saw the second website's forecast. Over a third of participants (33.8%) in the *verbal* condition became more likely to wait,

[36] We also asked participants four exploratory questions measuring the extent to which the two sites used different information, the extent to which the second site provided new information, the usefulness of the second site, and which forecast the participant weighed more when making their decision. See Supplement 1 and General Discussion.

[37] We included multiple probability levels for the sake of stimulus sampling, but the second website always made the same forecast as the first site. We also counterbalanced the order of the precise wording of the advice. We preregistered that we would collapse results across probability levels.

compared to 20.5% in the *numeric* condition (20.5%), Z = 4.23, *p* < .001, indicating that participants updated their beliefs more when seeing an additional verbal forecasts. Considering participants' untransformed responses, we find that although they were more willing to follow the websites' verbal advice overall (5.32 vs. 4.92), t(806) = 2.98, *p* = .003, they increased their answers more after getting a second opinion in the verbal condition than in they did after getting a second opinion in the numeric condition (i.e., there is a positive interaction between format and number of websites), t(806) = 2.28, *p* = .02.

GENERAL DISCUSSION

In four studies we find that people use different strategies to combine verbal compared to numeric probability forecasts. Specifically, when combining verbal forecasts, participants use an *additive* strategy, where their own forecasts move closer to certainty or impossibility as they see new advisor forecasts. Conversely, when combining numeric forecasts, participants' forecasts move closer to advisor's *average* forecast. These differences, if unaccounted for, could have substantial effects on how we understand judgments made from aggregating others' forecasts and how we should present multiple forecasts to others.

Our research raises two primary questions. First, are people acting more optimally or less optimally when they use an additive strategy, compared to an averaging strategy? Second, why are participants using an additive strategy to combine verbal forecasts? We discuss these below.

Is this optimal?

If given two verbal probability forecasts or two numeric probability forecasts, how *should* we combine them? The unsatisfying answer is that it depends. When there are few advisors using similar information to make their decision, averaging forecasts is typically most effective, since reduces the impact of each advisor's idiosyncratic error (Ashton & Ashton, 1985; Wallsten, Budescu, Erev, & Diederich, 1997; Wallsten & Diederich, 2001). However, when the number of advisors is sufficiently large, it is often best to use a more additive approach, since individual forecasts are often too conservative, particularly for hard to predict events (Ariely et al., 2000; Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Wallsten, Budescu, Erev, et al., 1997). Additionally, if the advisors are using *different* information, then it may be optimal to use an additive strategy regardless of the number of advisors (Baron et al., 2014; Wallsten & Diederich, 2001). In these cases, the decision-maker has more information than any individual advisor and therefore has "a right to much higher confidence" (Baron et al., 2014, p. 134).[38]

Why does this happen?

In the studies presented in this paper and several studies reported in the supplement, we tested several potential mechanisms that may be causing the effect. Although we do not find strong evidence for any of these mechanisms in our studies, they

[38] In Study 3, we found that participants in the *verbal, 2 advisor* condition were more accurate than those in the *numeric, 2 advisor* condition (measured by their average Brier score), but this difference was small and we hesitate to generalize it to other contexts. See Supplement 1.

are worth discussing, and our results do not necessarily mean that they are not present. They may simply be difficult to capture using self-report measures or that they may all contribute a small amount to the strategies that people use to combine forecasts.

Given the discussion in the previous section, participants could just be doing what they believe is optimal. Because numeric probabilities are more precise (Beyth-Marom, 1982; Lichtenstein & Newman, 1967; Zimmer, 1983), if two advisors give identical (or nearly identical) forecasts, it may imply that the advisors used the same information to make their forecasts. On the other hand, if two advisors give identical verbal forecasts, there may be a greater chance that they used different information. Therefore, an additive strategy would more optimal when combining verbal forecasts than when combining numeric forecasts. Because verbal probabilities are considered more intuitive (Windschitl & Wells, 1996), this would be consistent with the idea that people make less accurate judgments when working with percentages but are better at working with more intuitive probability formats, such as frequencies (Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000). Indeed Biswas et al. (2011) find that participants use a more additive strategy to combine frequencies, although they suggest that this occurs when frequencies are more *difficult* to combine.

If this is the case, we do not find evidence that this is a major contributor to our effect. In Study 4, we asked participants if the websites used the same or different information to make their decisions and to what extent the second website provided new information. Participants thought that the second site used more different information and provided more new information in the verbal condition (*p*s < .001). However, these measures mediated approximately 10% of the effect individually and 14% of the effect together. In three additional studies reported in Supplement 5, which all used a similar design to Study 2, we also asked participants if the second advisor had information that the first advisor did not, how much information the second advisor is providing to the participant, whether the advisors are using the same or different information to make their decisions, and how much intuition (vs. deliberation) participants used. There were no differences in participant responses between the numeric and verbal conditions (all *p*s > .09).

The second major candidate mechanism we considered is that because verbal probabilities have an inherent "direction," they are intuitively converted to positive or negative signals and added together.[39] In Supplement 6, we report four studies with the same general design as Study 2. Unlike in Study 2, however, we include an additional *numeric* condition, where we explicitly tell participants to interpret any advisor forecast above 50% as a positive sign. In three of the four studies, we find that participants are directionally more likely to use an additive strategy in this condition compared to the regular numeric condition, but none of these effects reach significance and are practically zero in the two most highly powered replications (with sample sizes of 400 and 800 per condition, respectively).

Finally, in several studies, we included questions on participant confidence, advisor consensus, advisor thoughtfulness, the strength of the advisors' opinions, whether

[39] Yates and Carlson (1986) refer to this as "signed summation."

forecast accuracy relied more on luck or knowledge (i.e., is the uncertainty aleatory or epistemic?), whether the advisor was making a subjective or objective judgment, and participants' trust in the advisors. For most of these questions, there were no differences between responses in the verbal and numeric conditions, and where they did, none of those differences mediated a meaningful proportion of our effect. See Supplement 8.

In summary, we find that individuals use distinct strategies to combine probability forecasts from different sources, where their judgments become more confident as they see more verbal forecasts and converge to the average as they see more numeric forecasts. We also tentatively rule out some potential mechanisms. Future research should delve deeper into possible causes of these differences and test how these strategies might affect decision on a larger scale. Individuals, and particularly organizational decision-makers, should take note of these results and consider their consequences when receiving and presenting probability forecasts from multiple sources.

This paper contains a supplement. Table 5 summarizes its contents.

Table 5. Index of supplementary materials (available from http://osf.io/atruq)

Section
Supplement 1. Additional preregistered measures and analyses
Supplement 2. Full regression tables for Studies 1-4
Supplement 3. List of stimuli used for Studies 1-3
Supplement 4. Full scenario text for Study 4
Supplement 5. Studies testing whether advisors use same vs. different information
Supplement 6. Studies testing effect of adding direction to numeric forecasts
Supplement 7. Studies testing use of intuition vs. deliberation
Supplement 8. Studies testing additional mechanisms

References

Albergotti, R. (2014, June 30). Furor erupts over Facebook's experiment on users. *Wall Street Journal*. Retrieved from http://www.wsj.com/articles/furor-erupts-over-facebook-experiment-on-users-1404085840

Andreoni, J., & Sprenger, C. (2011). *Uncertainty equivalents: Testing the limits of the independence axiom*. Retrieved from

Ariely, D., Loewenstein, G., & Prelec, D. (2006). Tom Sawyer and the construction of value. *Journal of Economic Behavior & Organization, 60*(1), 1-10. doi:http://dx.doi.org/10.1016/j.jebo.2004.10.003

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., . . . Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*(2), 130.

Ashton, A. H., & Ashton, R. H. (1985). Aggregating Subjective Forecasts: Some Empirical Results. *Management Science, 31*(12), 1499-1508. doi:10.1287/mnsc.31.12.1499

Awad, N. F., & Krishnan, M. S. (2006). The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, 13-28.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis, 11*(2), 133-145. doi:10.1287/deca.2014.0293

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1*(3), 257-269. doi:doi:10.1002/for.3980010305

Bilgin, B., & Brenner, L. (2013). Context affects the interpretation of low but not high numerical probabilities: A hypothesis testing account of subjective probability. *Organizational Behavior and Human Decision Processes, 121*(1), 118-128. doi:http://dx.doi.org/10.1016/j.obhdp.2013.01.004

Biswas, D., Zhao, G., & Lehmann, D. R. (2011). The Impact of Sequential Data on Consumer Confidence in Relative Judgments. *Journal of Consumer Research, 37*(5), 874-887. doi:10.1086/656061

Bolton, L. E., Warlop, L., & Alba, J. W. (2003). Consumer perceptions of price (un) fairness. *Journal of Consumer Research, 29*(4), 474-491.

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change. *Psychological Science, 20*(3), 299-308. doi:10.1111/j.1467-9280.2009.02284.x

Budescu, D. V., Por, H.-H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change, 113*(2), 181-200. doi:10.1007/s10584-011-0330-3

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Clim. Change, 4*(6), 508-512. doi:10.1038/nclimate2194 http://www.nature.com/nclimate/journal/v4/n6/abs/nclimate2194.html#supplementary-information

Budescu, D. V., & Yu, H.-T. (2006). To Bayes or Not to Bayes? A Comparison of Two Classes of Models of Information Aggregation. *Decision Analysis, 3*(3), 145-162. doi:10.1287/deca.1060.0074

Budescu, D. V., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making, 20*(2), 153-177. doi:10.1002/bdm.547

Campbell, M. C. (1999). Perceptions of Price Unfairness: Antecedents and Consequences. *Journal of Marketing Research, 36*(2), 187-199. doi:10.2307/3152092

Chow, C. C., & Sarin, R. K. (2001). Comparative Ignorance and the Ellsberg Paradox. *Journal of Risk and Uncertainty, 22*(2), 129-139. doi:10.1023/a:1011157509006

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization science, 10*(1), 104-115.

DellaVigna, S. (2009). Psychology and Economics: Evidence from the Field. *Journal of Economic Literature, 47*(2), 315-372. doi:10.1257/jel.47.2.315

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics, 75*(4), 643-669. doi:10.2307/1884324

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45*(1), 1-18. doi:http://dx.doi.org/10.1016/0749-5978(90)90002-Q

Folkes, V. S., & Kamins, M. A. (1999). Effects of information about firms' ethical and unethical actions on consumers' attitudes. *Journal of Consumer Psychology, 8*(3), 243-259.

Fox, C. R., & Tversky, A. (1995). Ambiguity Aversion and Comparative Ignorance. *The Quarterly Journal of Economics, 110*(3), 585-603. doi:10.2307/2946693

Fox, C. R., & Weber, M. (2002). Ambiguity Aversion, Comparative Ignorance, and Decision Context. *Organizational Behavior and Human Decision Processes, 88*(1), 476-498. doi:http://dx.doi.org/10.1006/obhd.2001.2990

Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making, 1*(3), 149-157. doi:10.1002/bdm.3960010303

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684-704. doi:10.1037/0033-295X.102.4.684

Gino, F. (2015, August 20). Companies Like Amazon Need to Run More Tests on Workplace Practices. *Harvard Business Review*.

Gneezy, U., List, J., & Wu, G. (2006). The Uncertainty Effect: When a Risky Prospect Is Valued Less Than Its Worst Possible Outcome. *The Quarterly Journal of Economics, 121*(4), 1283-1309.

Goel, V. (2014, June 29). Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry. *New York Times*. Retrieved from http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html?_r=0

Goldman, D. (2014). Facebook treats you like a lab rat. *CNN Money*. Retrieved from http://money.cnn.com/2014/06/30/technology/social/facebook-experiment/

Goldsmith, K., & Amir, O. (2010). Can Uncertainty Improve Promotions? *Journal of Marketing Research, 47*(6), 1070-1077. doi:doi:10.1509/jmkr.47.6.1070

Greenfield, R. (2014, July 28). OkCupid's Human Experiments Are Way Creepier Than Facebook's. *FastCompany*.

Haws, K. L., & Bearden, W. O. (2006). Dynamic pricing and consumer fairness perceptions. *Journal of Consumer Research, 33*(3), 304-311.

Hill, K. (2014, July 28). OkCupid Lied To Users About Their Compatibility As An Experiment. *Forbes*.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating Statistical Information. *Science, 290*(5500), 2261-2262. doi:10.1126/science.290.5500.2261

Jung, M. H., Perfecto, H., & Nelson, L. D. (2016). Anchoring in Payment: Evaluating a Judgmental Heuristic in Field Experimental Settings. *Journal of Marketing Research, 53*(3), 354-368. doi:10.1509/jmr.14.0238

Keren, G., & Gerritsen, L. E. M. (1999). On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica, 103*(1–2), 149-172. doi:http://dx.doi.org/10.1016/S0001-6918(99)00034-7

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790. doi:10.1073/pnas.1320040111

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9*(10), 563-564. doi:10.3758/bf03327890

Lichtenstein, S., & Slovic, P. (2006). *The Construction of Preference*: Cambridge University Press.

Luca, M. (2014). Were OkCupid's and Facebook's experiments unethical? *Harvard Business Review*. Retrieved from https://hbr.org/2014/07/were-okcupids-and-facebooks-experiments-unethical/

Mazar, N., Shampanier, K., & Ariely, D. (2016). When Retailing and Las Vegas Meet: Probabilistic Free Price Promotions. *Management Science, 0*(0), null. doi:doi:10.1287/mnsc.2015.2328

McGraw, A. P., Shafir, E., & Todorov, A. (2010). Valuing Money and Things: Why a $20 Item Can Be Worth More and Less Than $20. *Management Science, 56*(5), 816-830. doi:doi:10.1287/mnsc.1100.1147

Meyer, M. N. (2015). Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colorado Technology Law Journal, 13*(2), 60.

Meyer, M. N., & Chabris, C. F. (2015, June 19). Please, Corporations, Experiment on Us. *New York Times*. Retrieved from https://www.nytimes.com/2015/06/21/opinion/sunday/please-corporations-experiment-on-us.html

Meyer, R. (2014, June 28). Everything We Know About Facebook's Secret Mood Manipulation Experiment. *The Atlantic*.

Miller, G., & Mobarak, A. M. (2015). Learning About New Technologies Through Social Networks: Experimental Evidence on Nontraditional Stoves in Bangladesh. *Marketing Science, 34*(4), 480-499. doi:doi:10.1287/mksc.2014.0845

Miyazaki, A. D. (2008). Online privacy and the disclosure of cookie use: Effects on consumer trust and anticipated patronage. *Journal of Public Policy & Marketing, 27*(1), 19-33.

Montaguti, E., Neslin, S. A., & Valentini, S. (2016). Can Marketing Campaigns Induce Multichannel Buying and More Profitable Customers? A Field Experiment. *Marketing Science, 35*(2), 201-217. doi:doi:10.1287/mksc.2015.0923

Moon, A., & Nelson, L. D. (2015). *The Uncertain Value of Uncertainty: When Consumers are Unwilling to Pay for What They Like*. SSRN. Retrieved from https://ssrn.com/abstract=2676699

Muse, T. (2014). The Facebook Experiment: What It Means For You. Retrieved from http://www.forbes.com/sites/dailymuse/2014/08/04/the-facebook-experiment-what-it-means-for-you/#412a95bd1cbc

Newman, G., & Mochon, D. (2012). Why are lotteries valued less? Multiple tests of a direct risk-aversion mechanism. *Judgment and Decision Making*.

Rottenstreich, Y., & Hsee, C. K. (2001). Money, Kisses, and Electric Shocks: On the Affective Psychology of Risk. *Psychological Science, 12*(3), 185-190. doi:10.1111/1467-9280.00334

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review, 5*(4), 296-320. doi:10.1207/s15327957pspr0504_2

Rudder, C. (2014). We experiment on human beings! Retrieved from http://blog.okcupid.com/index.php/we-experiment-on-human-beings/

Schroepfer, M. (2014). Research at Facebook. Retrieved from https://newsroom.fb.com/news/2014/10/research-at-facebook/

Simonsohn, U. (2009). Direct Risk Aversion: Evidence From Risky Prospects Valued Below Their Worst Outcome. *Psychological Science, 20*(6), 686-692. doi:10.1111/j.1467-9280.2009.02349.x

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131.

Smith, N. C., & Cooper-Martin, E. (1997). Ethics and target marketing: The role of product harm and consumer vulnerability. *The Journal of Marketing*, 1-20.

Stampler, L. (2014, July 28). Facebook Isn't the Only Website Running Experiments on Human Beings. *Time*.

Sudhir, K., Roy, S., & Cherian, M. (2016). Do Sympathy Biases Induce Charitable Giving? The Effects of Advertising Content. *Marketing Science, 0*(0), null. doi:doi:10.1287/mksc.2016.0989

Teigen, K. H. (2001). When Equal Chances = Good Chances: Verbal Probabilities and the Equiprobability Effect. *Organizational Behavior and Human Decision Processes*, *85*(1), 77-108. doi:http://dx.doi.org/10.1006/obhd.2000.2933

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica*, *88*(3), 233-258. doi:http://dx.doi.org/10.1016/0001-6918(93)E0071-9

Teigen, K. H., & Brun, W. (1999). The Directionality of Verbal Probability Expressions: Effects on Decisions, Predictions, and Probabilistic Reasoning. *Organizational Behavior and Human Decision Processes, 80*(2), 155-190. doi:http://dx.doi.org/10.1006/obhd.1999.2857

Teigen, K. H., & Brun, W. (2000). Ambiguous probabilities: when does p=0.3 reflect a possibility, and when does it express a doubt? *Journal of Behavioral Decision Making, 13*(3), 345-362. doi:doi:10.1002/1099-0771(200007/09)13:3<345::AID-BDM358>3.0.CO;2-U

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and Combining Subjective Probability Estimates. *Journal of Behavioral Decision Making, 10*(3), 243-268. doi:doi:10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Wallsten, T. S., Budescu, D. V., & Tsao, C. J. (1997). Combining linguistic probabilities. *Psychologische Beitrage*.

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *41*(1), 1-18. doi:https://doi.org/10.1016/S0165-4896(00)00053-6

Wang, Y., Feng, T., & Keller, L. R. (2013). A further exploration of the uncertainty effect. *Journal of Risk and Uncertainty, 47*(3), 291-310.

Wang, Y., Lewis, M., Cryder, C., & Sprigg, J. (2016). Enduring Effects of Goal Achievement and Failure Within Customer Loyalty Programs: A Large-Scale Field Experiment. *Marketing Science, 35*(4), 565-575. doi:doi:10.1287/mksc.2015.0966

Wells, G. L., & Windschitl, P. D. (1999). Stimulus Sampling and Social Psychological Experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115-1125. doi:10.1177/01461672992512005

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of "likely" depends on the context, but "70%" is 70%—right? The influence of associative processes on perceived certainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1514-1533. doi:10.1037/0278-7393.25.6.1514

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343-364. doi:10.1037/1076-898X.2.4.343

Yang, Y., Vosgerau, J., & Loewenstein, G. (2013). Framing influences willingness to pay but not willingness to accept. *Journal of Marketing Research, 50*(6), 725-738.

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including “signed summation”. *Organizational Behavior* *and Human Decision Processes, 37*(2), 230-253. doi:http://dx.doi.org/10.1016/0749-5978(86)90053-1

Zimmer, A. C. (1983). Verbal Vs. Numerical Processing of Subjective Probabilities. In W. S. Roland (Ed.), *Advances in Psychology* (Vol. Volume 16, pp. 159-182): North-Holland.

Zoumpoulis, S., Simester, D., & Evgeniou, T. (2015, November 12). Run Field Experiments to Make Sense of Your Big Data. *Harvard Business Review*.