Essays on Judgment and Decision Making

Essays on Judgment and Decision Making

A dissertation presented

by

Christopher Umphres

to

The Committee on Higher Degrees in Public Policy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Public Policy

Harvard University

Cambridge, Massachusetts

April 2020

Dissertation Advisors: Jennifer S. Lerner & Julia A. Minson Christopher Umphres

Essays on Judgment and Decision Making

Abstract

Most judgments and choices in life arrive not as disconnected and isolated problems but as the next in a series enmeshed in a social context. The works contained herein, in one way or another, examine the role of context in shaping human judgment and decision making. The first essay reveals the surprising effects of the context of one's own prior judgments of confidence. In seven experiments (N = 5,484), we explore the counterintuitive finding that confidence in one's own judgments decreases over a series of difficult quantitative estimates. Our findings suggest that rather than evaluating confidence in isolation, participants evaluate confidence in reference to their stated confidence on earlier judgments. We theorize that confidence in earlier judgments increases in hindsight due to biased forgetting of disconfirming evidence, resulting in the downward trend we observed. The second essay examines sunk-cost bias in a social context. Viewed through the lens of economic rationality, sunk-cost bias is clearly irrational. In four experiments and a replication (N=2,754 U.S. adults), however, we found that decision makers received rewards precisely when they factored in sunk costs. Across three domains, decision makers who chose to escalate commitment after a prior investment were rated as more competent, warm, and confident than those who did not, a pattern that reversed in the absence of a prior investment. The pattern persisted with real financial stakes, an effect mediated by the perceived competence of the decision maker. The third essay builds upon the second by

examining escalation of commitment to a failing course of action as a signal of trustworthiness. In two experiments (N = 2,198), we found that decision makers who escalate commitment to a failing course of action are trusted 29% more by third-party observers than decision makers who de-escalate. Decision makers who escalate commitment actually are 15% more trustworthy. This signal was surprisingly robust to incentives for strategic signaling. Taken together, these essays contradict a foundational assumption of the rational actor model that history, whether your own recent judgments or the decision process, is irrelevant.

Table of Contents

Acknowledgements

I could hardy begin to thank everyone who helped me to produce what I hope will be a small contribution to this rapidly growing field, but I would be remis not to thank the following. First, there is my wonderful wife, Eleanor. Eleanor contributed in a thousand ways to this journey, including personally securing permission from the Chief of Staff of the Air Force for me to attend Harvard. Thank you.

Next, I owe a debt of gratitude to my fantastic mentors, Julia Minson and Jennifer Lerner. I could not have asked for a better team of advisors. Both were ceaselessly dedicated and selfless in their efforts to support not only me but the entire community of decision science researchers at Harvard. They were generous with their time, their expertise, and their advice, and they complemented each other perfectly. It would not have been possible to complete this work in thirty-two months without you both.

I am grateful to the entire JDM extended family for making HKS such a vibrant, productive, and *fun* place to work these last few years. Specifically, I want to thank Chelsea Zabel, Klara Kabadian, Ke Wang, Molly Moore, Ayse Yemiscigil, Alki Aliopoulou, Hanne Collins, Ariella Kristal, Martha Jeong, Hayley Blunden, Jenn Logg, Mike Yeomans, and David Hagmann. A special thanks to my partners in crime, Major Bradley DeWees and Charlie Dorison.

Finally, a heartfelt thanks to my parents Phil and Nancy Umphres, for instilling in me a thirst for knowledge and what discipline I have.

Essay 1

Confidence in Context:

Perceived Accuracy of Quantitative Estimates Decreases With Repeated Trials[1]

with Julia A. Minson

Although the literature on confidence in judgment is vast, few studies have examined how confidence changes over time. Yet, many highly consequential situations feature repeated judgments. In the course of a day, a financial analyst may evaluate a series of investments; a doctor may diagnose a series of patients; and a recruiter might review a series of resumes. Importantly, each individual judgment commands its own level of confidence, which influences decision-making. For example, a doctor is more likely to prescribe treatment when highly confident in a diagnosis, than when less so.

We document and explore a systematic decline in confidence when individuals make a series of judgments. Because confidence in a particular judgment should be a function of the features of that judgment and the information in one's possession, it should not exhibit systematic time trends. Indeed, lay intuition holds that confidence should increase over time (Study 3).

Many studies show that individuals are overconfident in many domains (Fischhoff et al., 1977; Mannes & Moore, 2013; Prims & Moore, 2017; Soll & Klayman, 2004). Unjustified confidence in one's judgments, sometimes referred to as "overprecision" (Moore & Healy, 2008), is remarkably robust to de-biasing (Moore et al., 2015).

Most research has studied confidence judgments in the aggregate, implicitly assuming that the findings apply equally to individual trials. The small body of work examining confidence over time has found mixed results. Specifically, Sanchez and Dunning (2018) showed that for

[1] Funding from the Harvard Kennedy School Dean's Research Fund partially supported this research.

novel, multi-cue probabilistic learning tasks (diagnosing a zombie disease), participants' confidence grew more rapidly than accuracy, a pattern they termed "the beginner's bubble." Conversely, Pulford and Colman (1997) report a decline in overconfidence throughout a series of difficult trivia questions. They speculate that the decline is due to participants' ability to self-monitor performance.

Our results are consistent with Pulford and Colman. However, after testing several versions of what they may have meant by "self-monitoring," we posit a more concrete explanation. Specifically, we suggest that confidence on subsequent estimates decreases because confidence in prior estimates *increases* in retrospect.

Contextual Confidence and Post-Decisional Processing

Prior research suggests that "confidence *evolves* during the course of the decision process" (Baranski & Petrusic, 1998, p. 942). Broadly, information distortion and motivated reasoning continue post-decision to consolidate preferences and alleviate dissonance (Arkes & Blumer, 1985; Festinger, 1964, p. 31; Festinger & Carlsmith, 1959; Lerner & Tetlock, 1999; Svenson et al., 1994). Specifically, the literature on the "confirmation bias" argues that people selectively seek and process information that supports their hypotheses (Kunda, 1990; Nickerson, 1998). Most relevantly, people demonstrate biased recall, disproportionately retaining arguments that support prior beliefs and forgetting ones that do not (Kunda, 1990).

Research on perceptual tasks confirms that confidence elicited after a decision is partly determined by post-decisional computation processes (Baranski & Petrusic, 1998). However, research examining both perceptual tasks and more cognitively complex assessments diverges over whether the continued collection of evidence is unbiased, leading to improved calibration (Moran et al., 2015; Yu et al., 2015), or biased in favor of the prior choice (Zylberberg et al.,

2012). Whereas prior research has examined the time interval between multiple confidence ratings concerning a single judgment, we examine the time intervals between confidence ratings regarding two different judgments.

Our theory is made up of two components. First, confidence is inflated in retrospect. When participants have to assess their confidence in a prior judgment, they do not go over the entire process of deliberation but rely primarily on the "gist" of that process (Stephen & Pham, 2008). We hypothesize that once a participant has made a judgment, stated their confidence, and moved on to the next item, they have little reason to hold in memory the evidence supporting or contradicting their earlier assessment. If the supporting information is retained more than contradictory information in line with prior research, then when participants recall their prior confidence this "gist" is shifted toward greater certainty.

Second, our theory leverages the influence of reference points and comparisons, common in classic research on sensory perception (Di Lollo, 1964; Helson, 1964; Krantz & Campbell, 1961). Indeed, reference dependence has been employed to explain perceptions of value (Kahneman & Tversky, 1979), and the effects of "anchors" on judgments regarding seemingly objective quantities (Frederick & Mochon, 2012). Here, we suggest that in determining their present confidence, individuals recall their stated confidence on a recent judgment, and compare how they feel about the current judgment to how they feel (currently) about the reference judgment. Because of biased recall, prior judgments appear more certain in retrospect making later judgments appear less certain. This results in the observed decline in reported confidence. We refer to this two-stage mechanism as the "inflation and adjustment" theory.

Overview of Studies

Seven experiments document the decline in reported confidence across a series of estimates. In all studies, participants made estimates regarding randomly ordered stimuli, and then reported confidence. Studies 1 and 2 demonstrate the effect using multiple confidence elicitation methods and incentives for truthful reporting. In Study 3, people mispredict the effect. Study 4 tests whether the effect relies on estimates being topically related (it does not), and demonstrates it across several domains. Study 5 moderates the effect with task difficulty. Studies 6 & 7 test our theory that declining confidence relies on comparing current confidence to that in prior items.

In all studies, we report how we determined our sample size, and all data exclusions, manipulations, and measures. Our pre-registrations (Studies 5-7), materials, data, and code are publicly available via the Open Science Framework (OSF).

Study 1: Confidence Declines Over Trials

Method

In Study 1, participants estimated the weights of nine animals from their photographs and stated their confidence in each estimate. We tested whether confidence decreased over the course of the task.

Participants. We collected 202 participant responses from Amazon Mechanical Turk (MTurk). After removing data from four participants who looked up information online and six participants who responded incorrectly to an attention check, our final sample consisted of 192 participants (45% Female, M_{age} = 38.5).[2] We paid participants $0.50 for their time. Furthermore,

[2] We discovered some participants would begin but not finish a survey and then re-take the survey using private browser mode or a VPN to defeat "ballot box stuffing" filters on Qualtrics. As a result, 91 of the 2,981 (3.1%) participants in Studies 1-5 had previously viewed some portion of the study they ultimately participated in. We recognized this issue and excluded these participants in studies 6 and 7.

a single question was selected at random for each participant. If the participant's estimate was within 10% of the correct answer, the participant received a $0.25 bonus.

Procedure. Participants estimated the weights of nine zoo animals based on photos. The actual weights of the animals ranged from 6.4 to 379 pounds. The photos of the animals were presented in random order. For each weight estimate, participants also reported how confident they were that their estimate fell within 10% of the correct answer. Participants reported confidence on a 5-point scale, anchored at "Not at all" and "Very," and recorded their response by typing a number from 1 to 5 into a text box. Responses regarding each of the nine target animals were collected on separate web pages, enabling us to collect the time spent on each question. After completing all estimates, participants reported how many of the estimates they expected would fall within 10% of the correct answer. Participants then reported demographics, and whether they had looked up any information.

Analytical Approach. Based on the effect sizes observed in earlier studies (see Appendix 1), we predetermined a sample size of 200 participants. The following methods apply to all subsequent studies unless otherwise noted.

Because each participant provided observations for nine different animals, we employed a mixed effects linear model computed using the lmer function in the lme4 package in R (Bates et al., 2015). Our models cluster the standard errors at the level of participant and item by including random intercepts for stimulus (animal in this case), and random slopes and intercepts for confidence across participant.

Furthermore, because we are primarily interested in how an individual's confidence responds to repeated questioning and not individual differences in the use of the confidence scale, we z-scored confidence ratings within participant. Rescaling preserves the trend but

eliminates differences in the degree of sensitivity, as well as allows comparisons between confidence ratings elicited on different scales in later studies.

Error was quantified as the absolute percentage deviation from the true value for each item. Accuracy is a binary variable set to 1 if the estimate is within the specified criteria (10% in the present study), and 0 otherwise.

To analyze response time data, we used a generalized linear mixed effects regression predicting response submission time with a fixed effect for question order and random intercepts for subject. We utilized a gamma distribution and a logarithmic link function to account for the right skewed distribution of submission time. Additionally, we use a log transformation of the question response time to predict confidence, error, and accuracy to further investigate the role of fatigue.

Results

Study 1 tested whether question order systematically affects subjective confidence in weight estimates. When we consider all 1,728 estimates made by participants, the average confidence on a 5-point scale was 2.80, SD = 1.06. The task proved quite difficult with only 9% of weight estimates landing within 10% of the true value. Most importantly, confidence z-scored within participant declines over the course of the task at a rate of 0.03 standard deviations per question (β = -0.03, CI_{95} [-0.05, -0.02], t = -3.82, $p < .001$). The mean confidence declined from 2.94 on the first question to 2.76 for the ninth question (see Figure 1.1).

Prior to examining the influence of response time, we excluded 8% of estimates (145 of 1,728) with response times in excess of 33 seconds (three times the median response time) as outliers probably due to task interruption. The gamma regression predicting submission time using question order confirms that time spent on each estimate declines by about 1 second with

each subsequent question (gamma regression b = -0.06, SE_b = 0.003, t = -19.44, $p < .001$). This could be a sign of participant fatigue or simply reflect growing familiarity with the task. When we examine error (operationalized as absolute deviation of the estimate from the truth as a percent of the true value) excluding 35 estimates with an error of more than 1,000%, we observe a small but significant increase over the course of the nine estimates (b = 3.0%, SE_b = 1.43, t = 2.009, p = .037). The log of time does not significantly predict confidence (even when controlling for question order), accuracy (using logistic regression), or error. We find that confidence and error were not strongly correlated (average within-participant Goodman-Kruskal gamma correlation: -.04, SD = 0.39). As a final robustness check, controlling for the log of response time and percent error does not diminish the effect of question order on z-scored confidence (β = -0.04, CI_{95} [-0.06, -0.02], t = 3.68, $p < .001$).

While a 5-point confidence scale does not permit statements about the appropriateness of the expressed level of confidence, it has the advantage of being more intuitive for participants who might struggle to express their confidence as a probability. Studies 2, and 4-7 implement a probabilistic confidence elicitation allowing us to examine overconfidence.

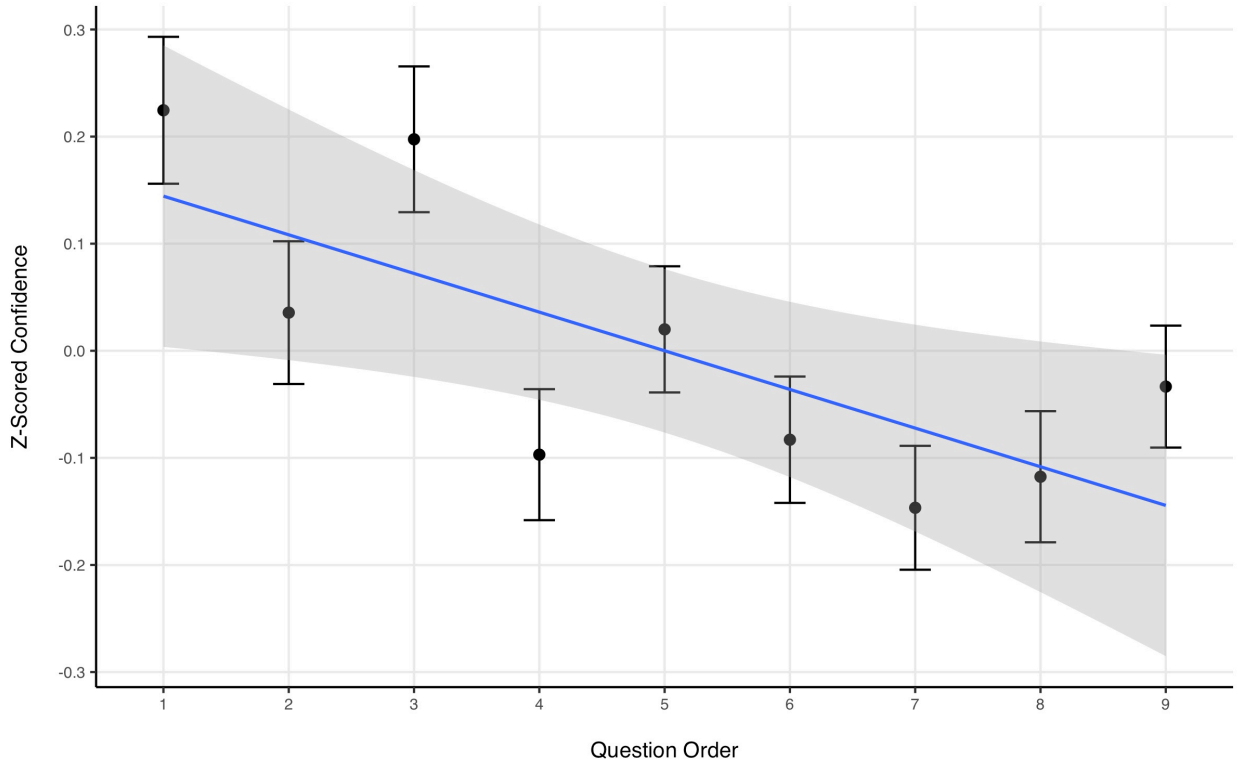

Figure 1.1. *Decline in z-scored confidence across the series of questions in Study 1. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

Study 2: Confidence Expressed as Probability

Method

In Study 2, we elicited confidence ratings using a 5-point Likert scale, or as a subjective probability, or by using the Becker-DeGroot-Marschak (BDM) confidence elicitation method (Becker et al., 1964). The expression of confidence in terms of probability enables us to assess calibration at the population level. The BDM procedure provides financial incentives for truthful reporting of confidence.

Participants. We obtained 606 completed survey responses from MTurk. After removing 29 participants who looked up information online, 11 participants who responded incorrectly to an attention check, and 4 responses with duplicate MTurk identifiers, our final sample consisted

of 562 participants (44% Female, M_{age} = 36.5). We paid participants $0.50 with an opportunity for a $0.25 bonus. Bonus criteria differed by condition.

Procedure. We randomly assigned participants to one of three conditions. In all conditions, the task was to estimate the weight of the same animals used in Study 1 and to indicate confidence in each estimate. The method of confidence elicitation and incentives differed by condition. In the *Likert* condition, participants estimated the weights of nine animals presented in random order and reported their confidence ratings using a 5-point Likert scale, anchored at "Not at all" and "Very." We incentivized accuracy by selecting a question at random and awarding a bonus if that estimate fell within 10% of the truth.

In the *Probability* condition, participants again estimated animal weights, but we elicited confidence ratings by asking "What is the probability that your estimate is within 10% of the actual value?" as a percent probability between zero and 100. The incentives for accuracy remained the same as in the *Likert* condition.

In the *BDM* condition, we altered the *Probability* condition to provide an incentive for subjects to accurately report the probability that their estimate was correct using the Becker-DeGroot-Marschak (Becker et al., 1964) method. We informed participants that they could win an additional $0.25 by either betting that one of their estimates (chosen at random) was correct, or via a lottery. Their confidence rating determined the point at which they would shift from betting on their estimate being correct to betting on the lottery with equal or greater probability of winning, the actual odds of which have not yet been determined. For example, if they stated being 75% confident, they would bet on their estimate unless the lottery odds (not yet known) provided a better than 75% chance of winning. This incentive structure dis-incentivizes over-stating one's confidence. Such a strategy would prevent participants from betting on lotteries that

offer a better chance of winning than their true level of confidence, but a lower chance than their (inflated) stated level of confidence. Similarly, it dis-incentivizes under-stating one's confidence: such a strategy would potentially force participants to bet on lotteries that offer a worse chance of winning than their true level of confidence.

We used a comprehension check to ensure that participants understood the BDM incentive structure. Consistent with the other conditions, 5% of the 220 participants assigned to the BDM condition quit before attempting the comprehension question. 78% correctly answered the question "What should you do if you want to maximize your chance of winning a prize?" on their first attempt (Answer: "Report your best estimate and how likely you believe it is that this estimate is correct"). 10% answered correctly on their second attempt. 7% of participants incorrectly answered the comprehension on their second attempt and were dropped from the survey or quit following their first failed attempt. Losses due to screening or attrition in the instructional phase reduced the rate of completion in the *BDM* condition to 81%, significantly lower than that of the other conditions ($X^2(2)= 12.36$, $p = .002$).

After completing all estimates, participants reported how many of the estimates they expected would fall within the 10% window and were given a chance to provide more information about their estimation process via a free text response. Participants then reported demographic information, and whether they had looked up any information online.

Analytical Approach. A target sample size of 200 participants per cell was predetermined based on observed effect sizes in prior studies. The statistical approach remained the same as in Study 1.

Results

Summary statistics for all three conditions are presented in Table 1.1. In all conditions, we again observed the main effect of decreasing confidence with subsequent questions (Table 1.2, Figure 1.2). While a decline in confidence of 0.5% per question may seem small, over the course of 10 questions this results in a 5% decline in confidence lacking an apparent normative explanation. The interaction between question order and condition was not significant, demonstrating that the decline in confidence is robust to variations in elicitation scale and incentives for accurate reporting of confidence.

Prior to examining the influence of response time, we excluded 6% of estimates (348 of 5,058) with response times in excess of 33 seconds (three times the median response time) as outliers probably due to task interruption. Time spent on each estimate declined with each subsequent question (gamma regression $b = -0.06$, $SE_b = 0.002$, $t = -31.84$, $p < .001$), and response time significantly predicted confidence when controlling for participant and stimulus, increasing 1.5% for each additional log(second) spent ($b = 1.5$, $SE_b = 0.56$, $t = 2.60$, $p = .009$). When we examined error over the course of the nine items, excluding 105 estimates with an error of more than 1,000%, we did not observe any systematic change over the course of the task in any of the three conditions or overall. Collapsing across condition and controlling for the log of response time and percent error does not diminish the effect of question order on z-scored confidence ($\beta = -0.04$, CI_{95} [-0.05, -0.03], $t = -7.94$, $p < .001$).

Mean confidence ratings and accuracy levels for each question by condition are presented in Figure 1.3. Overconfidence can be computed by comparing mean confidence levels expressed on a probability scale to the accuracy achieved by all participants in a given condition for each

question. This measure reveals substantial overprecision (about 45%), though calibration improves slightly with subsequent questions due to the decline in confidence.

In summary, Study 2 demonstrates that our effect is robust to confidence elicitation on probability scales and with financial incentives for accurate reporting of confidence.

Table 1.1. *Summary Statistics for Study 2*

Measure	Likert	Probability	BDM
Mean Confidence	2.54	45.8	55.9
(SD)	(0.88)	(26.0)	(24.8)
% of Estimates within 10% of the truth	9.8%	7.9%	9.4%

Table 1.2. *Regression Results for Study 2*

Condition	b (SE) p	β [95% CI] p
Likert (5-pt scale)	-0.03 (0.01) $p < .001$	-0.05 [-0.06, -0.03] $p < .001$
Probability (100-pt scale)	-0.6 (0.16) $p < .001$	-0.04 [-0.06, -0.02] $p < .001$
BDM (100-pt scale)	-0.4 (0.15) $p = .007$	-0.03 [-0.05, -0.01] $p = .002$

Note. Effects of question order on confidence and z-scored confidence for three elicitation methods in Study 2.

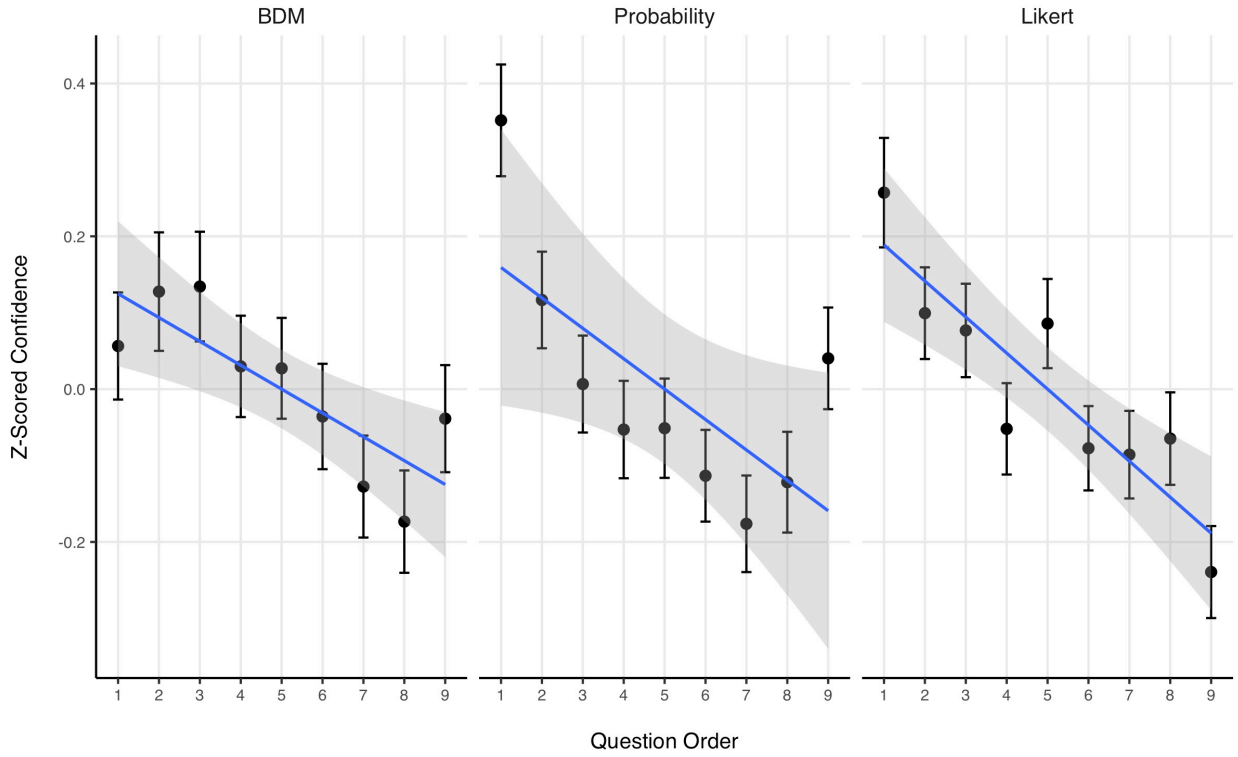

Figure 1.2. *Trends in z-scored confidence across the series of questions for each condition illustrate that the decline in confidence is robust to probability and incentivized elicitation methods. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

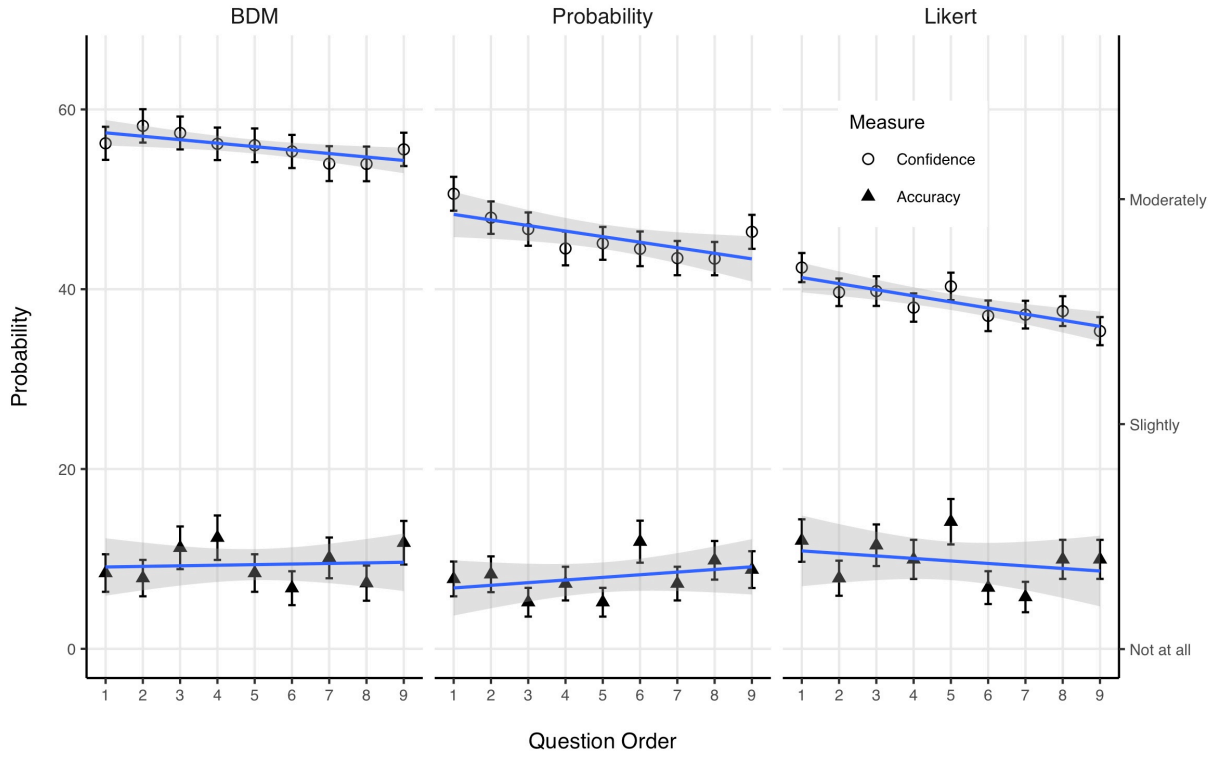

Figure 1.3. *The difference between subjective confidence and aggregate accuracy is a measure of overconfidence. In all conditions, the accuracy was constant while confidence declined, resulting in a slight improvement in calibration over time. The Likert results are rescaled to approximate a probabilistic scale for the purposes of illustration, but a comparison of confidence to accuracy is not possible strictly speaking. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

Study 3: Forecasting Confidence

Method

In Study 3, we examine lay beliefs regarding changes in confidence over a series of estimates. We do this by comparing participants' beliefs about how their confidence would change (if at all) to actual estimates and confidence ratings provided by another sample.

Participants. We collected 403 completed survey responses from MTurk. After removing five participants who looked up information online, five who failed an attention check and two

responses with the same MTurk ID, our final sample consisted of 391 participants (55% Female, M_{age} = 35.9). Incentive and bonus procedures were the same as in Study 1.

Procedure. We randomly assigned participants to one of two conditions. In the *Control* condition, participants estimated the weights of five zoo animals (a subset of the stimuli in Study 1) and reported their confidence for each estimate using a five-point Likert scale. In the *Forecast* condition, participants completed one randomly selected weight estimate and confidence assessment before being asked to imagine completing four more estimates of similar difficulty. Participants reported how confident they expected to be in the hypothetical fifth estimate (using the same five-point scale) and whether they expected their confidence to increase, decrease, or remain the same across the series of estimates. We also asked participants to provide a written justification for their answer. Participants then reported demographic information and were given an opportunity to report cheating before being presented with results for a single randomly selected question for bonus payment purposes.

Analytical Approach. A sample size of 200 participants per cell was predetermined and the statistical approach remained the same as in Study 1.

Results

When we consider all 1,965 estimates made by participants, the average reported confidence on a 5-point scale was M = 2.63, SD = 0.89, and performance remained poor with a success rate of 7%. In the *Forecast* condition, a paired t-test revealed that participants forecasted a significantly higher level of confidence for the hypothetical fifth question than they reported for the estimate they actually provided (Mean difference = 0.14, CI_{95} [0.04, 0.24], t = 2.76, p = .006, Cohen's d = 0.16, CI_{95} [0.04, 0.28]). When asked explicitly how they thought their confidence would respond to a series of questions of similar difficulty, only 12% reported that

they expected confidence to decline. 51% predicted no change and 37% predicted an increase in confidence.

Consistent with prior findings, in the *Control* condition we observe a significant decline in confidence z-scored within participant (β = -0.07, CI_{95} [-0.11, -0.02], t = -3.07, p = .002). Thus, the vast majority of participants did not accurately anticipate the observed decline and many predicted the opposite (interaction b = -0.08, SE_b= 0.019, t = -4.21, p < .001, see Figure 1.4).

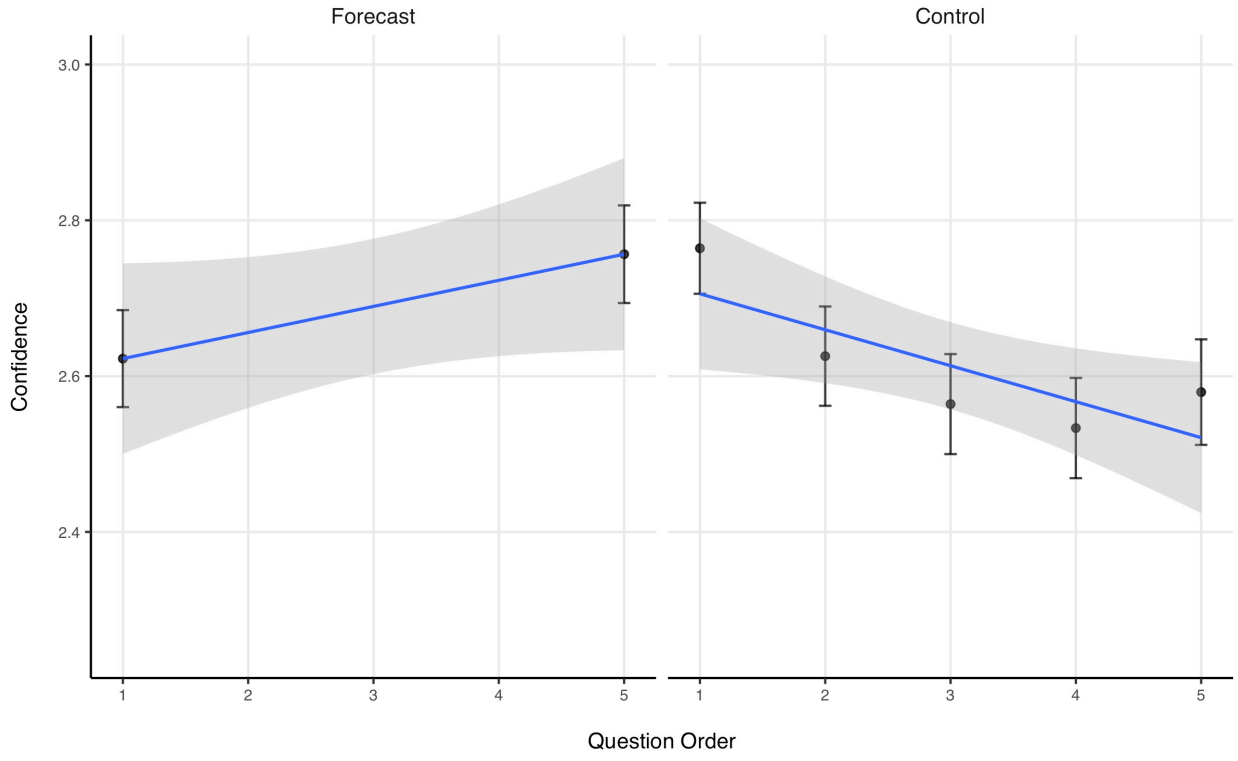

Figure 1.4. *Participants forecasted an increase in confidence for a hypothetical 5th estimate but confidence actually declined. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

Study 4: Multi-Domain Estimates

Method

In Study 4, we manipulate whether the stimuli used to elicit a series of estimates are related (as in prior studies) or drawn from several different topic domains. This manipulation allows us to test two possible explanations for the decline. First, participants might derive their later estimates by adjusting an uncertain amount from a prior (also uncertain) estimate. Participants may recognize that this method would lead estimation error to accumulate from estimate to estimate and therefore reduce their confidence for later estimates. By eliciting estimates from variety of domains (e.g., weights, calories, temperatures) we ensure that later estimates are not being derived from earlier estimates. If the decline in confidence persists, we can conclude that it is unlikely to be due to a feeling of errors adding up from one item to another.

Second, repeated estimation in the same domain may impact confidence by leading individuals to identify previously unrecognized holes in their knowledge (Fernbach et al., 2013; Rozenblit & Keil, 2002). This is one plausible interpretation of what Pulford and Coleman meant by "self-monitoring." Eliciting estimates from unrelated domains should reduce or eliminate the decline in confidence if it is driven by the realization that participants are less knowledgeable than they initially believed.

Participants. We recruited 900 participants from MTurk and obtained 910 complete responses. After removing 10 participants who looked up information online and 19 participants who answered an attention question incorrectly, our final sample consisted of 881 participants (57% Female, M_{age} = 35.7). We paid participants $0.50 for their study participation. Furthermore, we awarded a $0.50 bonus for highly accurate judgments.

Procedure. We randomly assigned participants to one of six conditions. In five of the conditions we asked participants to make four estimates in the same domain, similar to previous studies (number of candies in a jar, weights of animals, calories in foods, weights of people, and temperatures in cities). In the sixth, *Mixed,* condition participants made one estimate from each of these five domains. In all conditions, we randomized the order in which we presented the stimuli.

Specifically, in the *M&M* condition, participants saw photos of four jars of multicolored M&Ms. For each jar, participants estimated the number of a specified color of M&Ms (Brown, Green, Red, and Yellow). The *Animals* condition was the same as Study 1 with just four animals (wild cat, weasel, deer, and monkey). In the *Food* condition, participants estimated the calorie counts of four snack foods (Jelly Beans, raisins, pretzels, sunflower seeds) based on a description of the serving size and weight of each food (ex. Jelly Beans, 35 pieces, 40 grams). The *People* condition requested weight estimates for four people (130-195 lb. range) based on full-length photos. The *Weather* condition required participants to estimate the noon temperatures in four major US cities (New York, San Francisco, Denver and Omaha) a week into the future.

In all conditions, participants reported their confidence that each of their estimates was within 10% of the correct answer on a 5-point Likert scale. At the end of the survey, participants reported demographic information and whether they had looked up any information online during the study. Based on the effect sizes observed in prior studies, we predetermined a sample size of 150 participants per cell.

Results

When we consider all 3,671 forecasts and estimates made by participants, the average confidence on a 5-point scale was $M = 2.53$, $SD = 0.88$. The percentage of estimates that met the

10% accuracy criterion varied from 0.5 to 60 percent per condition (see Table 1.3). As predicted, there was a statistically significant linear decrease in z-scored confidence (and unstandardized confidence) in each of the un-mixed categories (see Table 1.4 and Figure 1.5). Importantly, the *Mixed* condition also showed evidence of declining confidence. The decrease in confidence did not significantly differ between conditions.

When we examine error over the course of the task after excluding 11 estimates with errors more than five standard deviations above the mean error, we did not observe any systematic changes. Collapsing across condition and controlling for percent error does not diminish the effect of question order on z-scored confidence (β = -0.09, CI_{95} [-0.12, -0.07], t = -8.56, $p < .001$).

Table 1.3. *Summary Statistics for Study 4*

Measure	Animals	Weather	Foods	M&Ms	People	Mixed
Mean Confidence (SD)	2.64 (0.78)	2.41 (0.85)	2.52 (0.91)	2.21 (0.86)	2.91 (0.78)	2.48 (0.94)
% of Estimates within 10% of the truth	7.5%	26.9%	16.8%	0.5%	60.3%	18.6%

Table 1.4. *Regression Results for Study 4*

Condition	b	(SE)	p	β	[95% CI]	p
M&Ms	-0.05	(0.015)	< .001	-0.07	[-0.12, -0.03]	.002
Animals	-0.07	(0.019)	< .001	-0.12	[-0.18, -0.06]	< .001
Food	-0.06	(0.021)	.007	-0.08	[-0.14, -0.03]	.005
People	-0.08	(0.018)	< .001	-0.13	[-0.18, -0.08]	< .001
Weather	-0.05	(0.022)	.028	-0.08	[-0.14, -0.02]	.013
Mixed	-0.08	(0.019)	< .001	-0.09	[-0.13, -0.04]	< .001

Note. Effects of question order on confidence and z-scored confidence for five estimation domains and a mixed-domain condition in Study 4.

The persistence of the decline in the mixed condition suggests that it is not driven by participants' feeling that error from earlier estimates is somehow getting accumulated on later estimates. It also makes it less likely that the key to the effect is participants' recognition of the limitation of their knowledge in any one domain. However, it is not clear to what degree participants recognized that the estimates in the mixed condition were truly independent tasks. It is still possible that participants were learning that they are poor at estimation tasks in general, topic notwithstanding.

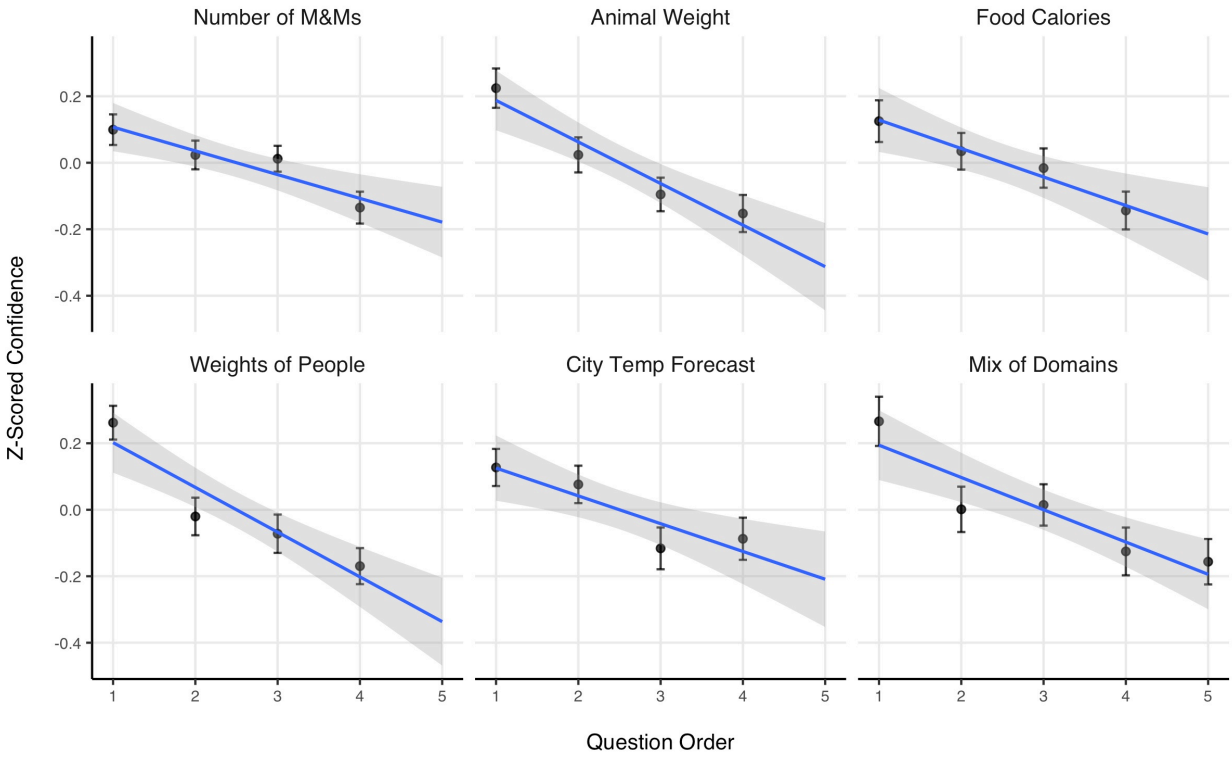

Figure 1.5. *Confidence declined across a variety of estimation domains and when unrelated tasks were mixed. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

Study 5: Easy vs. Hard Tasks

In Study 5 we manipulate question difficulty. On easy estimation items, participants are likely to identify fewer reasons why their estimate might be incorrect. Thus according to our theorizing, when they move on to a subsequent estimate, they will have fewer conflicting cognitions to forget, and experience a lesser or no change in confidence. This suggests that a series of easy estimation questions should show an attenuated decline in confidence. On the other hand, people tend to be overconfident on difficult tasks and underconfident on easy ones (e.g., Moore & Small, 2007). Thus, if learning about the task accounts for the confidence change, gradually learning that the task is easy should lead to increasing confidence.

Method

Participants. We recruited 700 participants from MTurk and collected 692 completed responses after excluding responses with duplicate respondent IDs. In accordance with our pre-registration, we excluded 16 participants who failed an attention check. We excluded an additional 58 participants who missed a later comprehension check. The rate of attrition did not differ between conditions ($X^2(2) < .001$, $p = 1$). Our final sample consisted of 618 participants (56% Female, $M_{age} = 36.0$). Incentive and bonus procedures were the same as in Study 1 except that the bonus criteria varied by condition as described below.

Procedure. We randomly assigned participants to one of two estimate difficulty levels. In both conditions, we showed participants photos of a glass container filled with six different colors of M&M candies. We asked participants to estimate the number of each color of candy in the container. The order of the questions (which color was being asked about) was randomized. In the hard condition, the required tolerance for a "correct" answer was plus or minus 10 candies. In the easy condition, the tolerance was plus or minus 60 candies. Each participant was asked to

make 6 estimates (the number of red, blue, yellow, green, brown and orange M&Ms) in random order and to report their confidence in the accuracy of each estimate.

After submitting each estimate, we showed participants their answer as well as the calculated range for the true value that would allow their answer to count as a "correct" answer. For example: "Your estimate (85) is within 60 of the actual value and your answer will be judged as correct ***IF*** the true value falls between 145 and 25. What is the probability that the true number of Blue M&Ms lies between these two values?" This procedure ensured that participants were aware of the range that would count as "correct" for the purposes of their confidence ratings.

Following the six estimates, we collected demographic information. We also asked participants to estimate how many questions they answered correctly. Finally, we used the experimental software to give them feedback on their total score and a randomly selected result for bonus payment purposes.

Analytical Approach. A minimum post-exclusion sample size of 300 participants per cell was predetermined using power simulations from pilot data. The statistical approach remained the same as in Study 1.

Results

Both confidence and accuracy were higher in the *Easy* condition (M = 69.5%, SD = 22.0%, Accuracy (within 60) = 81.3%) than the *Hard* condition (M = 50.8%, SD = 22.7%, Accuracy (within 10) = 10.5%), $t = 10.39$, $p < .001$, $t = 34.14$, $p < .001$, confirming that our manipulation of task difficulty was successful. Comparing average confidence and accuracy by condition, we find that participants were overconfident in the *Hard* condition (t = 22.71,

$p < .001$) and underconfident in the *Easy* condition ($t = 6.28$, $p < .001$), replicating the well-known easy-hard effect (Lichtenstein et al., 1982; Lichtenstein & Fischhoff, 1977; Moore & Small, 2007).

In the *Hard* condition, we observed the main effect of decreasing z-scored confidence ($\beta = -0.06$, CI_{95} [-0.08, -0.03], $t = 4.18$, $p < .001$). In the *Easy* condition, confidence did not significantly decline ($\beta = -0.02$, CI_{95} [-0.04, 0.01], $t = 1.23$, $p = .221$). As predicted, the interaction between question order and condition was significant ($\beta = -0.04$, CI_{95} [-0.08, -0.002], $t = 2.05$, $p = .041$) (see Figure 1.6). Importantly, this result is distinct from the easy-hard effect because it demonstrates an impact of task difficulty not on calibration overall or on an isolated judgment but on the dynamics of confidence in context.

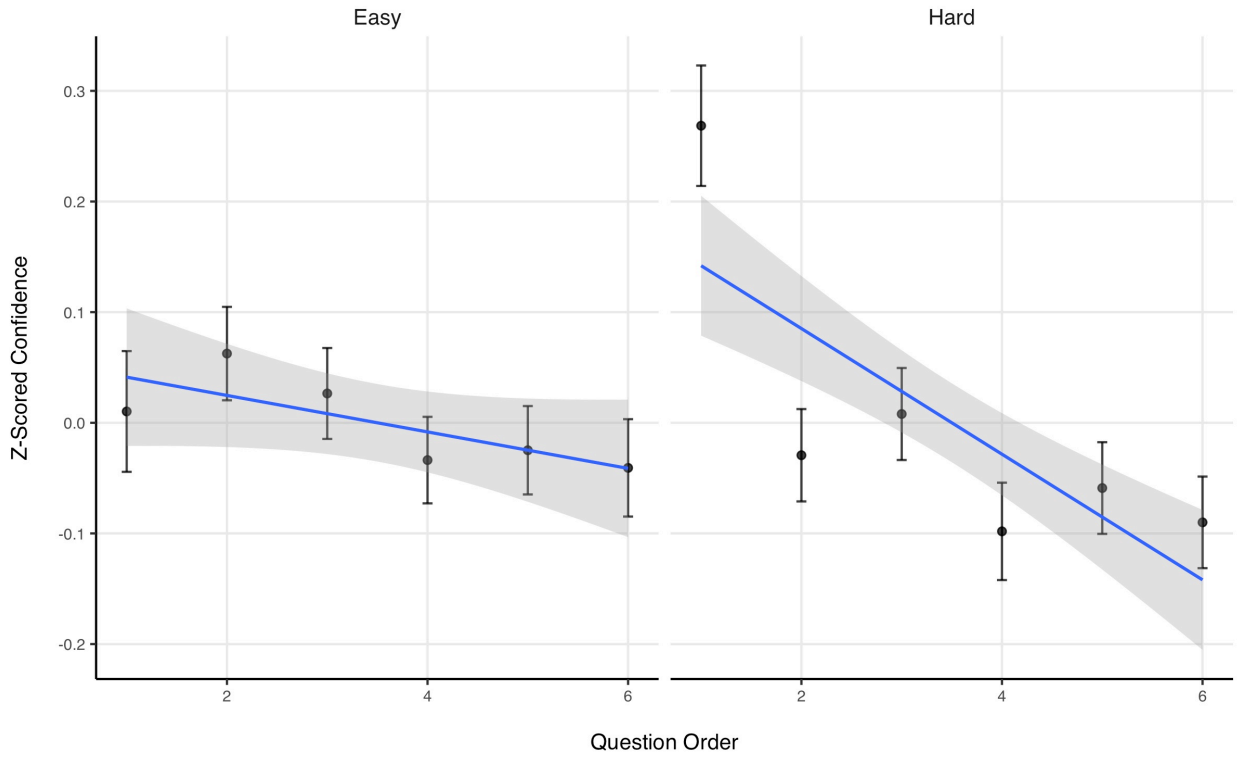

Figure 1.6. *In Study 5, the decline in confidence is significantly attenuated by question difficulty. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

Prior to examining the influence of response time, we excluded 12% of estimates (443 of 3,708) with response times in excess of 45 seconds (three times the median response time) as outliers probably due to task interruption. Time spent on each estimate declined with each subsequent question (gamma regression b = -0.16, SE_b = 0.004, t = -41.8, $p < .001$), and response time marginally predicted confidence when controlling for worker ID and stimulus, increasing .8% for each additional log(second) spent (b = 0.76, SE_b = 0.42, t = 1.82, p = .069).

When we examined error over the course of the six items, excluding 15 estimates with an error of more than 600% (suggesting an estimate for the entire jar rather than one color), we also observe a decrease in percent error over the course of the task in both conditions (b = -1.6, SE_b = 0.26, t = -6.41, $p < .001$). It seems likely that participants were taking advantage of the repeated examination of the same candy jar to improve their estimates. This makes the decline in confidence that we observed in the *Hard* condition particularly intriguing, as it seems that confidence declined while at the same time accuracy improved. However, once again we find that these declines in error and response time do not statistically mediate the effect of question order on confidence. Controlling for the log of response time and percent error does not diminish the effect of question order on z-scored confidence in the *Hard* condition (β = -0.05, CI_{95} [-0.08, -0.02], t = -3.59, $p < .001$). The interaction between question difficulty and question order remains at a similar magnitude though it is only marginally significant with the addition of the control variables (β = -0.04, CI_{95} [-0.08, 0.00], t = -1.91, p = .056).

Study 6: Revisit Earlier Estimates

In Study 6 we explicitly test our inflation and adjustment theory, which proposes that 1) confidence in prior estimates increases in hindsight; and 2) this inflated level of confidence serves as a reference point from which future confidence judgments are adjusted. Prior research

suggests that over time, people become more committed to their judgments, view them as more consistent, more cohesive, and forget or discount contradictory and missing information (Griffin & Tversky, 1992; Koriat, 2012; Koriat et al., 1980; Walters et al., 2016). Following this increase in recalled confidence, the current estimate might appear less certain by comparison. Thus, the confidence decrease we report arises not because latter items are actually more difficult, the estimates less accurate, or the deliberation process more fraught, but because prior estimates seem more coherent in hindsight.

Method

We test this theory by comparing our previous procedure to one in which participants make estimates and confidence ratings, complete a set of "filler" estimates, and then revisit the confidence ratings they offered for earlier items. We are particularly interested in the change in confidence from the end of the filler estimates to the first "revisited" estimate.

If confidence judgments are based on a recollection of the general "gist" of the earlier estimation process and if confidence in previous estimates is inflated due to biased forgetting, then a revisited estimate should receive higher confidence ratings compared to the most recent filler estimate. However, if the participants are monitoring their performance and learning that the task is difficult, then we would observe the opposite pattern. To the extent that self-monitoring has allowed the participant to discern the true difficulty of the task and their own general ineptitude, then revisited estimates should receive lower or comparable ratings to recent estimates.

Participants. We collected 1,008 completed responses from MTurk. After removing 29 participants who looked up information online and 24 participants who failed an attention check

in accordance with our pre-registration, our final sample consisted of 955 participants (51% Female, M_{age} = 36.1). Incentive and bonus procedures were the same as in Study 1.

Procedure. We randomly assigned participants to one of two conditions. In both conditions, participants made estimates of four different animal weights (drawn at random from a set of eight stimuli and presented in random order), and then made 12 unrelated "filler" estimates. As in prior studies, they reported their corresponding confidence ratings after each estimate. Our treatment occurred after the 12 filler estimates. In the *Control* condition, participants then made weight estimates for the four animal stimuli of the original set of eight not sampled in the first four estimates and again reported their level of confidence for each (new) estimate. In the *Repeat* condition, participants were presented with the estimates they had already made for the first four animals (one at a time in a new random order) and asked to re-assess their confidence in those original estimates. All confidence ratings were elicited as the probability that the estimate is within 10% of the actual value. Participants reported demographic information, reported cheating, and were presented with one result for bonus payment purposes.

If participants' confidence ratings reflect how confident they feel on a current estimate relative to the immediately preceding one, we should observe an increase in confidence in the *Repeat* condition between the last of the 12 filler items and the first item of the final set, again featuring animals. We expect no such increase in the *Control* condition. When *Repeat* condition participants encounter an item they had seen previously, the familiar item should feel more certain than the immediately preceding, unfamiliar item. Furthermore, we predicted that the four repeated items would not show a downward trend in confidence ratings because participants would not be actually making the estimates (only rating their confidence), and thus would *not* feel more uncertain and conflicted about each new estimate relative to the previous one.

Analytical Approach. A sample size of 1,000 participants was predetermined in order to provide sufficient statistical power to detect an interaction between question order and experimental condition.

We conducted three pre-registered analyses. First, we tested for an increase in confidence between the last of the 12 filler items and the first item that presented participants with a familiar stimulus (items 16 and 17) in the *Repeat* condition. We expected the magnitude of this change to be larger than the change in confidence on the same items in the *Control* condition. For this analysis we used a linear mixed effects model predicting the dependent variable (confidence z-scored within participant) from question order, condition, and their interaction term, random intercepts for specific stimuli and participants, and random slopes for each participant across question order. We predicted the coefficient of the interaction term to be positive and significant.

Second, we repeated this analysis for the final set of 4 items which featured familiar animals in the *Repeat* condition and new animals in the *Control* condition, with question order mean-centered. We predicted that the *average* level of confidence for these final four items would be higher in the *Repeat* condition than in the *Control* condition.

Finally, we tested the hypothesis that as in prior studies, confidence would decline over the course of the initial questions (i.e. questions 1-4), but that this decline would be attenuated in the repeat condition on the final four questions when participants are restating their confidence regarding the same stimuli. To test this hypothesis, we analyzed a subset of the data within the repeated condition, containing only items 1-4 and 17-20. This allowed us to compare the confidence on the 1^{st}-4^{th} items to the 17^{th}-20^{th} items each participant considers. We centered the question numbers within each set of 4 and assigned a dummy variable for each set (0 = first set,

1 = final set). The significance of the coefficients for set and the interaction term of centered question order and set are the test of our exploratory hypothesis.

Results

When we consider all 19,100 estimates made by participants, the average confidence on a 100-point scale was M = 52.6, SD = 25.1. As in prior studies, we replicated the main effect of decreasing z-scored confidence across the first four estimates in the *Control* condition (β = -0.11, CI_{95} [-0.14, -0.07], t = -5.96, $p < .001$) and the *Repeat* condition (β = -0.10, CI_{95} [-0.13, -0.07], t = 5.79, $p < .001$).

Next, we examine the change in confidence from the last of our 12 filler questions (item 16) to the first item of the second set of animal weight estimates (question 17). In the *Control* condition, we observe a small but significant decline in z-scored confidence (mean difference = -0.17, t = 2.89, p = .004, Cohen's d = -0.19, CI_{95} [-0.33, -0.06]) suggestive of a continuation of the effect as usual. As our theory predicts, the *Repeat* condition shows a significant increase in z-scored confidence from the last new estimate to the first revisited estimate (mean difference = 0.41, t = 6.73, $p < .001$, Cohen's d = 0.44, CI_{95} [0.31, 0.58]). Thus, we observed a significant interaction between item and condition (interaction β = 0.58, CI_{95} [0.42, 0.74], t = 7.21, $p < .001$).

When examining the final four questions, as predicted we observed no significant confidence change in the *Repeat* condition (highlighted in Figure 1.7). We observe a marginally significant increase in the *Control* condition which in the broader context of items 8-20 appears to be noise. In line with our predictions, the mean level of confidence was higher in the *Repeat* condition than in the Control condition for these final four estimates (β = 0.26, CI_{95} [0.19, 0.32],

$t = 7.78$, $p < .001$). Also, as predicted, the significant downward slope in confidence for the first four estimates was significantly attenuated when these same estimates were revisited at the end of the survey (interaction $\beta = 0.10$, CI_{95} [0.05, 0.15], $t = 4.21$, $p < .001$).

The results of Study 6 show a pattern largely consistent with our theory. The fact that confidence in the revisited estimates shows a dramatic increase relative to the immediately preceding estimate, suggests that participants are not recalibrating their evaluation of task difficulty, at least not in a generalizable manner that would cause them to question earlier assessments of their confidence. Overall, the data seem supportive of part 2 of our theory.

Contrary to part 1 of the theory, however, we did not observe an increase in confidence for the revisited estimates compared to the original confidence appraisals, conducted several minutes ago. In fact, a paired t-test of all original and revisited confidence ratings shows a marginally significant decrease (mean difference = 0.85%, $t = 1.70$, $p = .089$). With the benefit of hindsight, we can think of a number of reasons why this might be so. Anchoring on the lower confidences for items 5-16 may decrease reported confidence on the items that immediately follow. It was our intention that the filler questions would make remembering the original stated confidence difficult to remember to prevent participants from simply restating their earlier responses. However, this also prevents drawing a definitive conclusion about whether participants intended to report a higher or lower confidence than that originally reported. Finally, it is of course possible that participants are doing some amount of learning about the difficulty of the task and this mechanism is also at play.

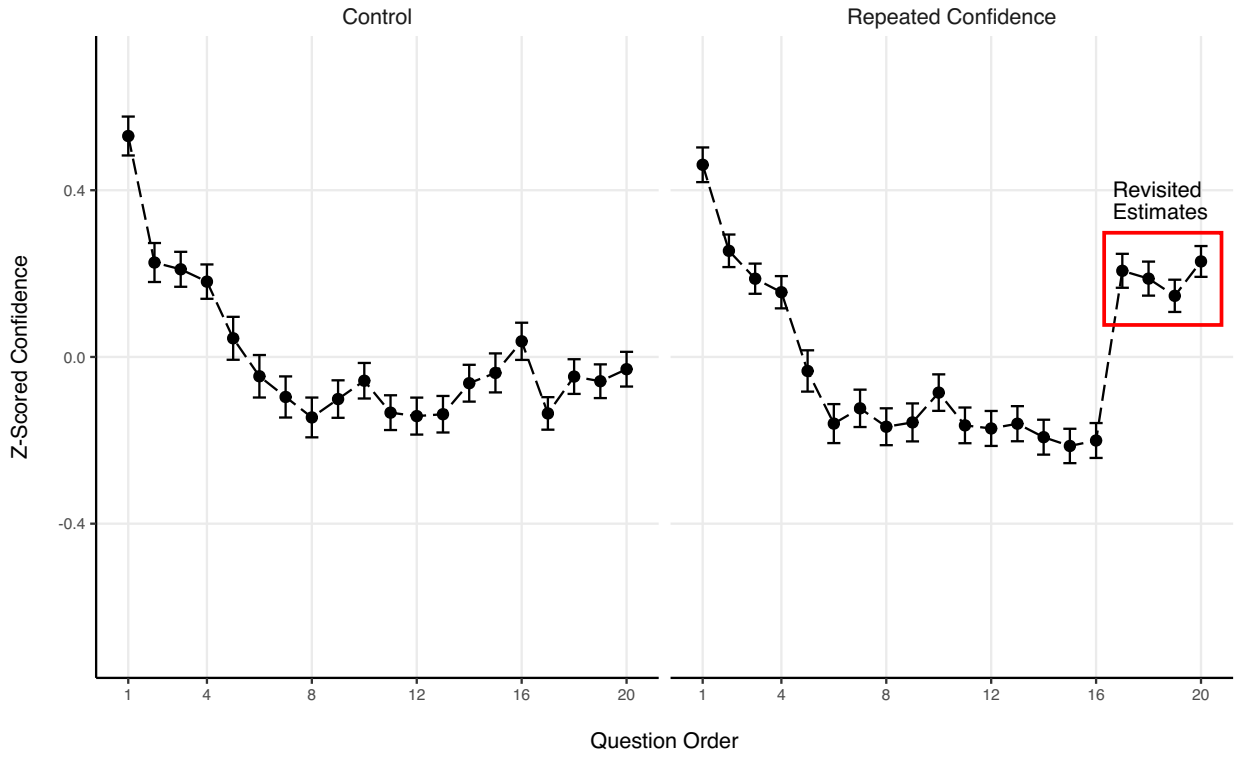

Figure 1.7. *Points 17-20 in the Repeated Confidence condition are a reappraisal of the first four estimates. Error bars indicate standard errors.*

Study 7: Repeated Estimates or Repeated Confidence Ratings

Method

Our final study tests part 2 of our theory by separating the effect of making repeated estimates from the effect of repeatedly reporting one's confidence. If the decline in reported confidence arises from conveying one's degree of confidence *relative* to prior statements of confidence, the effect should require multiple confidence ratings. It should not emerge if a participant made multiple estimates but only offered a confidence rating for one of them. By contrast, if confidence declines because participants are gaining an appreciation for the difficulty of the task, the effect should be observed between-subjects even if any given participant expresses confidence only once during the series of estimates.

Participants. We obtained 1,982 completed responses from MTurk after excluding 72 responses with duplicate respondent IDs. In accordance with our pre-registration, we excluded 80 participants who failed an attention check common to all conditions. After removing 17 participants who looked up information online, our final sample consisted of 1,885 participants (56% Female, M_{age} = 36.1). Incentive and bonus procedures were the same as in Study 1.

Procedure. We randomly assigned participants to one of two conditions. In both conditions, participants estimated the weights of five animals. In the *Control* condition, participants reported their corresponding confidence ratings after each estimate as in prior studies. We randomly selected only one of these confidence ratings from each participant to include in the analysis. Follow-up analysis used bootstrapping to ensure that the results did not rely on which estimate was selected for each participant. In the *Single Confidence* condition, participants estimated the weights of five animals but only reported confidence for one of these five estimates (randomly selected). The result is 2 (condition) X 5 (question # for confidence rating) between-subjects design.

Following the five estimates, participants reported demographic information, and were given an opportunity to report cheating. We then used the experimental software to present each participant with a randomly selected result for bonus payment purposes.

Analytical Approach. The experimental manipulation in this study results in a single confidence measure per participant, requiring a fully between-subjects analysis and precluding the use of confidence z-scored within participant as our main dependent variable. In order to make full use of our available data, we repeated the analysis using 10,000 bootstrapped simulations of the random draw of confidence ratings from the control condition data. We

predetermined a sample size of 2,000 participants (1,000 per condition) in an effort to detect the moderation of our effect.

We expected to observe a decline in confidence over the five estimates in the *Control* condition, but no decline in confidence in the *Single Confidence* condition. We used a linear mixed effects model predicting confidence from question order, condition, and their interaction term, as well as random intercepts for specific stimuli. We predicted the coefficient of the interaction term to be positive and significant.

Results

We replicated our finding of a decline in confidence for the *Control* condition (b = -2.01, CI_{95} [-3.06, -0.96], t = -3.75, $p < .001$). Crucially, there was not a significant decline in the *Single Confidence* condition (Figure 1.8). The interaction between condition and question order was positive and significant, as predicted (b = 1.65, CI_{95} [0.19, 3.11], t = 2.22, p = .026). The supplementary bootstrap analysis confirms the robustness of the decline of confidence in the control condition with 79% of samples showing a significant decline at an average of 1.5% per question. The analysis also shows that while the interaction is consistently positive (98% of estimates > 0, M_b = 1.09) it is significant for only 24% of samples.

These results further support the hypothesis that the decline in confidence over the course of a task is a contextual phenomenon in which prior statements of confidence serve as reference points. When no prior reference points are available as in the *Single Confidence* condition, the decline so consistently observed in the prior studies disappears (b = -0.37, CI_{95} [-1.39, 0.64], t = -0.72, p = .473). The results would be only consistent with a "learning through self-monitoring" explanation if learning required not only experiencing the estimates but also explicitly reporting one's confidence.

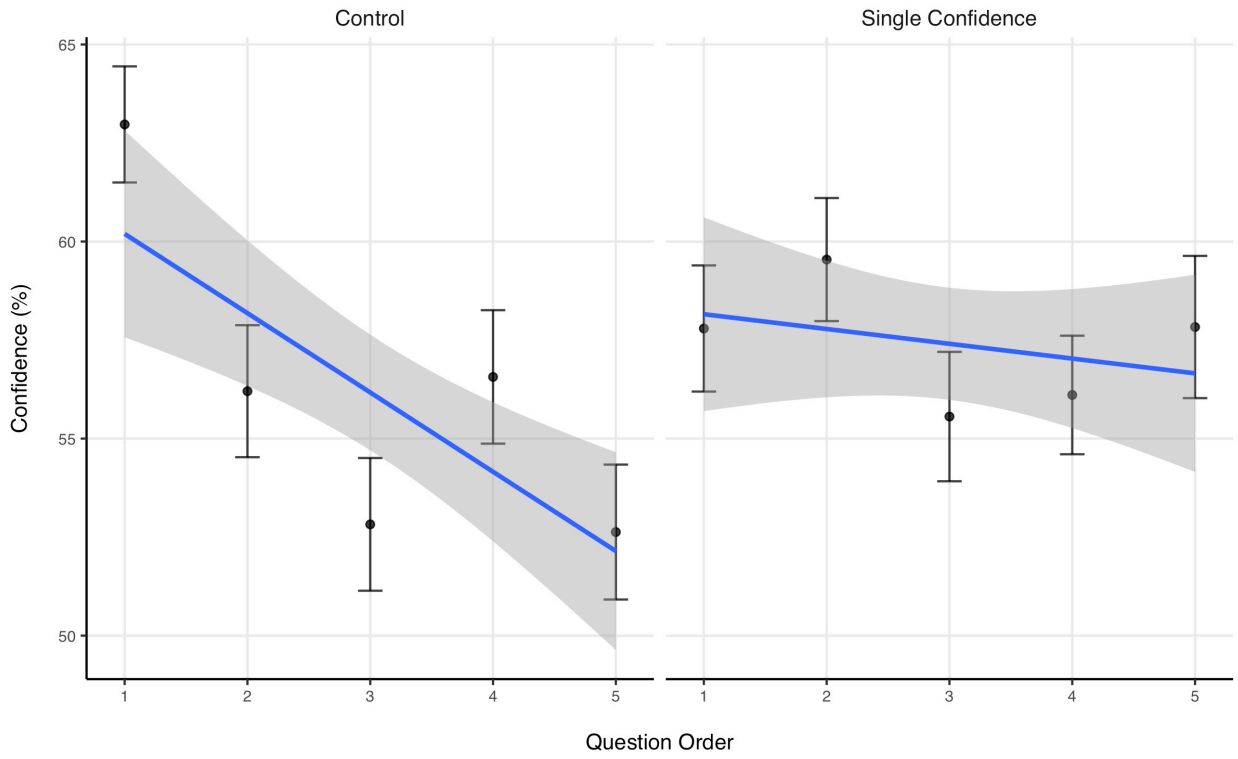

Figure 1.8. *The decline in confidence was eliminated by eliciting confidence for a single randomly selected estimate. Error bars indicate standard errors. The shaded area represents the 95% confidence interval for the regression.*

General Discussion

We document a decline in reported confidence across a set of quantitative estimates that does not appear readily justifiable by normative considerations such as feedback or new information. We observed this decline using a variety of stimuli, several confidence elicitation methods, with financial incentives, and without. The decline does not appear dependent on the factors suggested by existing theories, including changes in accuracy, effort, or subjective knowledge. Study 3 also suggests that the effect violates lay intuition.

An intuitively appealing explanation (Pulford & Colman, 1997) is that the decline reflects the ability of participants to self-monitor their own performance, learning over time that the task is more difficult or their ability more limited than they thought at first. Our results are largely

inconsistent with this explanation. The persistence of the decline when facing a mixed set of stimuli (Study 4), the failure to increase confidence when underconfident (Study 5), the resilience of revisited confidence ratings (Study 6), and the disappearance of the effect when reporting confidence only once (Study 7) all suggest that the decline does not reflect a reappraisal of the overall task difficulty and one's knowledge.

Instead, we suggest an alternative theory in two parts. First, individuals feel more certain about their estimates upon later reflection because contradictory information dissipates from memory more easily than the supporting information. Second, confidence in any given judgment is expressed in reference to confidence in prior judgments. Although the most influential point of comparison is the immediately preceding judgment, it seems likely that the entire set of expressed confidence ratings provides a context for a participants' meta-understanding of the precision of their beliefs. The combined effect is that more recent estimates seem less certain than preceding estimates and are rated lower by comparison not because they accurately assess the task as more difficult but because they misperceive prior tasks as easier in hindsight than in the moment. Thus, we observe movement in the normatively correct direction resulting in improved calibration, but for the wrong reasons. In this case, two wrongs make a right.

Our work complements and extends earlier work on the dynamics of confidence. Although our findings appear at odds with the "beginner's bubble," Sanchez and Dunning suggest task difficulty and familiarity are likely to be potential boundary conditions of their effect (2018). Our task used stimuli that participants are likely to have encountered many times before. In effect, many participants began the experiment at the peak of the bubble.

If experience is enough to reduce overconfidence, how is it that the bias persists in so many domains among adults? Study 7 shows that the effect depends on repeated confidence

assessments, something that most individuals outside of the laboratory might rarely explicitly engage in. Yet, research should consider important exceptions of individuals who professionally make repeated judgments (diagnosticians, financial analysts, security analysts, etc.), and examine our phenomenon in those contexts. Future research inserting delays between estimates and exploring various levels of difficulty and expertise is also needed to determine the durability of the observed decline. In Supplemental Study 2 (Appendix 1) we test the effect of feedback (which on a difficult task, is predominantly negative). We find that accurate feedback speeds the confidence decline.

Future work should further explore the behavioral consequences of our phenomenon. The BDM measure employed in Study 2 suggests that individuals are willing to act on their confidence ratings, since they chose to make real bets based on their confidence. However, the speculative, probabilistic, and abstract nature of this incentive structure make it a less than ideal behavioral measure.

Whether the observed decline in reported confidence leads to behavioral implications or merely reflects a subjective and temporary assessment, knowledge of the phenomenon is important for contexts in which a person's estimates are inputs into others' decision-making. We rely on such interpersonal expressions of confidence in many areas such as medicine, criminal trials, sports betting and intelligence analysis. The confidence levels that accompany judgments are generally considered as truly reflective of the rater's beliefs. Our studies indicate that this assumption is somewhat erroneous.

Identifying circumstances in which confidence declines could prove useful in designing environments that calibrate decision makers for important tasks. Engaging in a series of "warm up" estimates may provide a means of reducing overconfidence in concert with other proven

methods such as providing feedback, increasing the salience of unknown information, and consciously considering alternative viewpoints. Consumers of others' judgments should be aware of how the context of these judgments influences reports. Most importantly, the contextual nature of confidence revealed in these experiments illustrates that expressed confidence is a dynamic process, and ought to be studied as such.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929–945. https://doi.org/10.1037/0096-1523.24.3.929

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Becker, G. M., Degroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232. https://doi.org/10.1002/bs.3830090304

Di Lollo, V. (1964). Contrast effects in the judgment of lifted weights. *Journal of Experimental Psychology*, *68*(4), 383–387. https://doi.org/10.1037/h0042094

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political Extremism Is Supported by an Illusion of Understanding. *Psychological Science*, *24*(6), 939–946. https://doi.org/10.1177/0956797612464058

Festinger, L. (1964). *Conflict, decision, and dissonance*. Stanford U. Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(4), 552–564. https://doi.org/10.1037/0096-1523.3.4.552

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*(1), 124–133. https://doi.org/10.1037/a0024006

Griffin, D., & Tversky, A. (1992). The weighting of evidence and the determinants of confidence. *Cognitive Psychology*, 411–435.

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. New York.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. JSTOR. https://doi.org/10.2307/1914185

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Koriat, A., Lichtenstein, S., Fischhoff, B., & Bourne, L. E. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118.

Krantz, D. L., & Campbell, D. T. (1961). Separating perceptual and linguistic effects of context shifts upon absolute judgments. *Journal of Experimental Psychology*, *62*(1), 35–42. https://doi.org/10.1037/h0040386

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychology Bulletin*, *125*(2), 255–275.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. https://doi.org/10.1016/0030-5073(77)90001-0

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art in 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge University Press.

Mannes, A. E., & Moore, D. A. (2013). A Behavioral Demonstration of Overconfidence in Judgment. *Psychological Science*, *24*(7), 1190–1197. https://doi.org/10.1177/0956797612470700

Moore, D. A., & Healy, P. J. (2008). The Trouble with Overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moore, D. A., & Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. *Journal of Personality and Social Psychology*, *92*(6), 972–989. https://doi.org/10.1037/0022-3514.92.6.972

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in Judgment. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (pp. 182–209). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333.ch6

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 46.

Prims, J. P., & Moore, D. (2017). Overconfidence over the lifespan. *Judgment and Decision Making*, *12*(1), 29.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, *23*, 125–133.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28. https://doi.org/10.1037/pspa0000102

Soll, J. B., & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30*(2), 299–314.

Stephen, A. T., & Pham, M. T. (2008). On Feelings as a Heuristic for Making Offers in Ultimatum Negotiations. *Psychological Science*, *19*(10), 1051–1058. https://doi.org/10.1111/j.1467-9280.2008.02198.x

Svenson, O., Rayo, A. O., Andersen, M., Sandberg, A., & Svahlin, I. (1994). Post-decision consolidation, as a function of the instructions to the decision maker and of the decision problem. *Acta Psychologica*, *87*(2), 181–197. https://doi.org/10.1016/0001-6918(94)90050-7

Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2016). Known Unknowns: A Critical Determinant of Confidence and Calibration. *Management Science*, *63*(12), 4298–4307. https://doi.org/10.1287/mnsc.2016.2580

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General*, *144*(2), 489–510. https://doi.org/10.1037/xge0000062

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*. https://doi.org/10.3389/fnint.2012.00079

Essay 2

The Benefit of Bias:

Decision Makers Who Exhibit Sunk-Cost Bias
Receive Social and Economic Rewards for Doing So[3]

with Charles A. Dorison, Bradley R. DeWees, and Jennifer S. Lerner

Prescriptive decision theory holds that decision makers should choose options when future benefits exceed future costs (Edwards, 1954; Friedman, 1953). Yet, a large, interdisciplinary literature (e.g., Arkes & Blumer, 1985; Thaler, 1980) makes clear that decision makers are influenced not only by future costs and benefits, but also by sunk costs (i.e., prior investments of time, money, or effort that can no longer be recovered). Economic and decision science literatures predominantly characterize this "sunk-cost bias" as irrational and maladaptive, demonstrating that it contributes to sub-optimal resource allocations (e.g., Baron, 1990) and escalations of commitment to failing courses of action (Sleesman et al., 2012).

An alternative view, prevalent among practitioners, is that sunk costs are politically, and potentially even economically, relevant to future costs and benefits (Brest & Krieger, 2010, p. 436). One hears the telltale refrain "we've got too much invested to quit now" in settings ranging from infrastructure projects (Nagouney, 2018) to military campaigns (Schwartz, 2006). Tetlock (2000) argued that escalating commitment due to sunk costs could be acknowledged as a "private cognitive vice" and yet still be justified by political necessity. To date, however, we know of no systematic experiments examining the potential social and material benefits of affirming sunk costs. The present studies attempt to fill that gap.

[3] This paper is based upon work supported by: the National Science Foundation under Grant No. (1559511), the National Institute of Health under Grant No. (1R01CA224545-01A1), the Harvard Program on Negotiation, and the Harvard Mind Brain Behavior Initiative.

A Benefit of Bias?

On the one hand, individuals who consider only future costs and benefits (i.e., ignore sunk costs) might receive reputational rewards for adhering to prescriptive decision theory by writing off sunk costs. We term this possibility the *reward for rationality hypothesis*. In some situations, those evaluating decision makers are less prone to bias than the decision makers themselves, allowing evaluators to recognize biased behavior in others. For example, John, Jeong, Gino, and Huang (2019) found that evaluators judge individuals who do not change their minds in the face of incontrovertible evidence as lacking good judgment. To the extent that sunk-cost bias reflects an unwillingness to change one's mind in the face of evidence, evaluators might perceive individuals who pay attention to sunk costs more negatively than individuals who do not.

On the other hand, some researchers theorize that factoring in sunk costs may signal competence and/or commitment, leading evaluators to perceive decision makers who do so more positively (Kanodia et al., 1989; McAfee et al., 2010; Tetlock, 2000). We term this possibility the *benefit of bias hypothesis*. Evidence from other judgment and choice environments suggests that social rewards for such signaling behaviors are possible even when evaluators disagree with decision makers' choices. For example, leaders who held fast to their moral commitments were perceived as less hypocritical, more effective, and more worthy of support as compared to leaders who changed their moral commitments, even when the participant disagreed with the moral stance of the leader (Kreps et al., 2017). In addition, agreement between evaluators and decision makers might strengthen the benefit of bias. Recent work finds that people consider sunk costs in their own judgments even when they have not personally incurred the cost in

question (Olivola, 2018), suggesting that in situations in which a majority of decision makers favor escalation, a large number of evaluators would likely agree.

Overview of Studies

The present studies test the two competing hypotheses: the *reward for rationality hypothesis* and the *benefit of bias hypothesis*. Study 1 provides an initial test of these hypotheses using social perceptions in a business domain. Studies 2-3 test generalizability to consumer and political decisions. Finally, Study 4 examines how these processes play out when evaluators' judgments affect financial consequences.

Study 1: Are There Social Benefits of Escalating Commitment to a Sunk Cost?

Study 1 provides an initial test of the *reward for rationality* and *benefit of bias* hypotheses. We predicted, contrary to prescriptive decision theory, that sunk-cost bias would yield social benefits. Specifically, we predicted that (1) decision makers who invested after incurring a sunk cost would be perceived more positively than would decision makers who pivoted from sunk costs while (2) decision makers who invested in something with identical prospects but without sunk costs would be perceived more negatively than decision makers who declined to invest. Also, given that gender stereotypes can influence social perceptions of decision makers, we systematically varied the stated sex (male/female) of the decision maker.

Method

Open Practices and Ethical Conduct. In all studies, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures. All data and materials are publicly available via the Open Science Framework (OSF). The design and analysis plans for these studies were preregistered at AsPredicted; copies can be found on OSF. All studies received Institutional Review Board approval.

Participants. We recruited 400 respondents from the United States using Amazon's Mechanical Turk (MTurk). We advertised the study as a "decision making survey." We pre-determined our target sample size to ensure 80% power to detect moderate simple effects (Cohen's $d = 0.25$) in each sunk-cost condition. Attrition was below 2% and did not vary by condition.

Procedure. All participants learned that they would play the role of an "evaluator" who would read a scenario about a decision maker and make judgments about the decision maker based on the decision maker's choices. We randomly assigned participants to one of eight evaluator conditions in a 2 (Decision Maker's Gender: male versus female) x 2 (Presence/Absence of Sunk Cost: sunk cost versus no sunk cost) x 2 (Decision Maker's Choice: invest versus not-invest) fully between-subjects design.

Participants read a commonly-used scenario from the sunk-cost literature about a "radar-blank plane" (see Arkes & Blumer, 1985; Olivola, 2018), adapted to read in the third person. Specifically, they were told that the CEO of an airline company, whom we named either John (Male condition) or Jessica (Female condition) Morrison, was deciding whether to invest $1 million to develop a radar-blank plane. A competing company had just begun marketing a faster and more economical radar-blank plane. We told participants either that the CEO's company had already invested $10 million and was 90% complete with the project (sunk cost condition) or that no money had been spent (no sunk cost condition). The CEO decided either to invest the $1 million (invest condition) or not to invest any money in the radar blank plane project (not-invest condition). In order to ensure participants actually read and considered the scenario, they were asked to answer a set of three comprehension questions regarding the scenario.

Dependent Variables. Three social perception scales served as our dependent variables: perceived decision maker warmth, competence, and confidence. Warmth and competence are two fundamental dimensions of person perception (Fiske et al., 2007, 2002), and confidence specifically is a key component of competence shown to influence consequential decisions such as hiring (John et al., 2019). For the respective measures of warmth and competence, we drew on work by Fiske and colleagues (Fiske et al., 2002). Specifically, participants evaluated the target decision maker on a 5-point scale (1 = not at all, 5 = extremely) on such adjectives as "good natured" and "intelligent." For the measure of confidence, we drew on work by Fast and colleagues (Fast, Sivanathan, Mayer, & Galinsky, 2012; see also John et al., 2019). Participants evaluated the target decision maker on a 7-point scale (1 = strongly disagree, 7 = strongly agree) on such items as: "is very sure about what she knows" and "seems confident about her beliefs." Because each of the three scales achieved high reliability (all alphas > .81), we averaged scale items to compute scores for a given scale.

Exploratory Variables and Demographics. Following the primary dependent variables, and consistent with our preregistration, we also collected exploratory measures of perceived trustworthiness, consistency, wastefulness, and rationality. Results for these measures largely mirror the results for the three focal dependent variables. Full details are available in our online materials (OSF).

After completing their social evaluations, participants reported what they would have chosen to do if they themselves were in the situation described. Finally, participants reported demographic information (age, gender, education), and whether they had previously been taught the principle of sunk cost or encountered a scenario similar to the ones described in the study.

Results

Preliminary Analyses. In keeping with our preregistration and before conducting any inferential analyses, we cleaned the data by excluding: incomplete survey responses, duplicate responses from the same person, and responses that failed to pass one simple comprehension question common to all conditions. Unfortunately, accuracy rates for two other comprehension checks differed by condition. Following guidance for avoiding selective attrition (Zhou & Fishbach, 2016), and consistent with our preregistration, we did not exclude for failure to correctly answer these two questions. The final sample for inferential analyses thus consisted of N = 390 participants (45% female; age: M = 35.9, SD = 11.4, Median = 33). Importantly, applying the preregistered rules for sample cleaning did not substantively alter the results reported (see OSF page for details).

Inferential Analyses. If the *reward for rationality hypothesis* correctly characterizes social perceptions in this context, then evaluators will view decision makers who chose to invest unfavorably (compared to decision makers who do not invest) when there is not a sunk cost *and* when there is a sunk cost. Similarly, they will evaluate decision makers who chose *not* to invest favorably when there is not a sunk cost and when there is a sunk cost. If, on the other hand, the *benefit of bias hypothesis* holds, evaluators' views will depend on the presence of prior investment. Specifically, evaluators will view decision makers who chose to invest unfavorably when there is *not* a sunk cost, but favorably when there *is* a sunk cost (i.e., there will be an interaction between Decision Maker's Choice and Presence/Absence of Sunk Cost). Additionally, given that men and women frequently differ in the amount of competence ascribed to them (Foschi, 1992; Ridgeway, 2001), a negative evaluation for a decision maker who pivots away from a sunk cost might be more extreme if the decision maker were a female.

Analytic Plan. We examined the data using OLS regression and planned 2x2 contrasts. Because none of the two-way interactions between investment and sunk cost were qualified by gender, we collapsed across gender. See Appendix 2 for detailed analysis of the gender condition. Results for all two-way interactions in Study 1 appear in the top row of Figure 2.1. Means and confidence intervals appear in Table 2.1, simple contrasts in Table 2.2, and interactions in Table 2.3.

Competence. Consistent with the *benefit of bias hypothesis*, there was a significant interaction between the presence of a sunk cost and the decision maker's choice to invest on perceptions of competence (b = -0.97, CI_{95} [-1.27, -0.68], β = -0.53, CI_{95} [-0.70, -0.37], t = 6.50, $p < .001$). In the absence of a sunk cost, participants rated decision makers who chose to invest (M = 3.63, CI_{95} [3.47, 3.78]) as less competent than decision makers who chose not to invest (M = 3.98, CI_{95} [3.86, 4.11]), $t(178.79) = 3.52$, $p < .001$, d = 0.51, CI_{95} [0.22, 0.80]. As predicted by the *benefit of bias hypothesis*, this pattern was reversed when a significant sunk cost had been incurred prior to the investment decision. In the presence of a sunk cost, participants rated decision makers who chose to invest (M = 4.08, CI_{95} [3.95, 4.21]) as more competent that decision makers who chose not to invest (M = 3.46, CI_{95} [3.28, 3.65]), $t(168.98) = 5.49$, $p < .001$, d = 0.80, CI_{95} [0.50, 1.09].

Confidence. Also consistent with the *benefit of bias hypothesis*, there was also a significant interaction between the presence of a sunk cost and choice to invest on perceptions of confidence (b = -1.45, CI_{95} [-1.92, -0.97], β = -0.49, CI_{95} [-0.65, -0.33], t = 5.98, $p < .001$). In the absence of sunk cost, participants rated decision makers who chose to invest (M = 5.38, CI_{95} [5.15, 5.61]) as less confident than decision makers who chose not to invest (M = 5.75, CI_{95} [5.58, 5.93]), $t(176.99) = 2.54$, p = .012, d = 0.37, CI_{95} [0.08, 0.65]. This pattern was reversed

when a significant sunk cost had been incurred prior to the investment decision. In the presence of a sunk cost, participants rated decision makers who chose to invest (M = 5.75, CI_{95} [5.56, 5.95]) as more confident than decision makers who chose not to invest (M = 4.68, CI_{95} [4.34, 5.02]), $t(147.32) = 5.43$, $p < .001$, $d = 0.80$, CI_{95} [0.50, 1.09].

Warmth. Finally, and again consistent with the *benefit of bias hypothesis*, there was a significant interaction between the presence of a sunk cost and choice to invest on perceptions of warmth ($b = -0.47$, CI_{95} [-0.73, -0.21], $\beta = -0.30$, CI_{95} [-0.47, -0.14] $t = 3.57$, $p < .001$,). In the absence of sunk cost, participants rated decision makers who chose to invest ($M = 3.52$, CI_{95} [3.39, 3.65]) as less warm than decision makers who chose not to invest ($M = 3.62$, CI_{95} [3.50, 3.74]), $t(189.27) = 1.10$, $p = .273$, $d = 0.16$, CI_{95} [-0.44, 0.13]. When sunk costs were present, however, participants rated decision makers who chose to invest ($M = 3.66$, CI_{95} [3.53, 3.79]) as more warm than decision makers who chose not to invest ($M = 3.28$, CI_{95} [3.14, 3.43]), $t(186.69) = 3.79$, $p < .001$, $d = 0.54$, CI_{95} [0.26, 0.83].

Discussion

Study 1 provided support for the *benefit of bias hypothesis:* While decision makers who persisted in the presence of a sunk cost were perceived more positively than decision makers who pivoted from a sunk cost, decision makers who made an investment with identical prospects but *without* sunk costs were perceived more negatively than decision makers who passed on the investment. This was true for perceptions of competence, confidence, and warmth. Moreover, gender of the decision maker did not qualify any of our key findings. Thus, Study 1 provided evidence for generality across multiple social perceptions and across gender of the decision maker.

Studies 2 and 3: Will the Benefit of Bias Generalize Across Decision Domains?

In order to test generalizability to different decision-making domains, Study 2 used a personal decision domain and Study 3 used a political decision. We again tested the *benefit of bias hypothesis*. Because these studies follow a similar procedure, we describe them in parallel.

Method

Studies 2-3 were identical to Study 1 except for three key differences. First, we varied the setting of the decision. We used a consumer preference domain in Study 2 as a potentially more conservative test than Study 1, reasoning that evaluators may be more reticent to judge decision makers for decisions contingent on personal preferences. We used a political domain in Study 3 as another kind of conservative test, reasoning that evaluators may be less likely to reward the honoring of sunk costs when common (taxpayer) resources are potentially wasted.

Second, because we found no moderating effects of gender of the decision maker in Study 1, we dropped this factor in Studies 2-3. Third, given that the exploratory social judgments from Study 1 showed an identical pattern to the confirmatory variables, we did not collect the exploratory social judgments from Study 1 in these studies (although we did add one new exploratory item in Study 3, described below). Lastly, in order to gain more inferential insight into the judgment patterns, we included a free text response to elicit participants' justifications for their own preferred course of action.

Participants. Recruitment was increased to 440 participants for each study to ensure adequate power post exclusions but otherwise remained the same as in Study 1. The final sample for inferential analyses was N = 415 participants (47% female; age: $M = 37.2$, $SD = 11.0$,

Median = 35) for Study 2 and N = 400 (55% female; age: *M* = 39.2, *SD* = 12.3, Median = 36) for Study 3. For each study, attrition was below 4% and did not vary by condition. As in Study 1, our results remain consistent regardless of exclusion criteria (see OSF page for details).

Study 2. Participants read a commonly-used Hotel-TV movie scenario (Frisch, 1993; Olivola, 2018) adapted to the third person. Specifically, participants were told that Casey, sick in bed while on vacation, was deciding whether to continue watching a movie on the TV in his hotel room or to change the channel. After five minutes, Casey was bored with the film. Participants were told either that Casey had already paid $19.95 to watch the movie (Sunk Cost "SC" condition) or that the movie was free (No Sunk Cost "NSC" condition). Casey either decided to finish the movie (Invest condition) or change the channel (Not Invest condition). We then collected the same three social judgments as Study 1 (competence, warmth, and confidence).

Study 3. Study 3 adapted the structure of the radar-blank plane scenario (Arkes & Blumer, 1985) to a political rather than a commercial context. It was loosely based on the real-life high-speed rail project in California (Nagouney, 2018). Specifically, we informed participants that the governor of a large U.S. state was deciding whether to invest $4 billion on a major infrastructure project when a new study revealed that technological innovations drastically reduced demand for the project and created more economical alternatives. We told participants either that the state had already invested $36 billion and was 90% complete with the project (Sunk Cost "SC" condition) or that no money had been spent (No Sunk Cost "NSC" condition). The governor either decided to continue the project (Invest condition) or to discontinue the project (Not Invest condition). Immediately following the social perception scales (competence, warmth, and confidence), we also asked whether the governor's decision made the participant

more or less likely to vote for him. This variable showed an identical pattern to the social perception results; full details are available in our online materials.

Results

Building on the results from Study 1, we predicted a social benefit of honoring sunk costs in both personal and political decision domains. We predicted that decision makers would be evaluated more positively when they invested (vs. did not invest) in the presence of a sunk cost, but would be evaluated more negatively when they invested (vs. did not invest) in the absence of a sunk cost, despite the fact that the two scenarios represented equivalent choices according to rational economic theory.

Analytic Plan. We tested this hypothesis using a 2 (prior investment: yes, no) x 2 (future investment: yes, no) interaction. We again examined the data using OLS regression and planned 2x2 contrasts by prior investment condition. In both studies, all three dependent variable scales achieved a high level of reliability (all alphas > .81).

Interactions. We observed a significant interaction for all three variables in both studies (six interactions; all *p*s $\leq$ .004). Further, we observed the predicted cross-over interaction in 5 of 6 possible cases. Specifically, participants rewarded decision makers for escalating commitment in the presence of a sunk cost by rating them as more competent, more confident, and warmer. In all but one case, the participants rewarded the opposite decision—despite identical prospects—when there was not a prior investment. The perception of warmth in Study 2 (the Hotel-TV scenario) was the sole exception, with participants rating those who chose to finish the movie as directionally warmer even when the movie was free, though this difference was not statistically significant. Results for all interactions are again depicted in Figure 2.1, this time in rows 2 and 3. Means and confidence intervals are listed in Table 2.1. Simple contrasts appear in Table 2.2 and

interactions appear in Table 2.3. The contrasts consistently reveal that individuals who invest in the presence of sunk costs are perceived more positively than individuals who do not invest in the presence of sunk costs. The opposite pattern occurs in the absence of sunk costs.

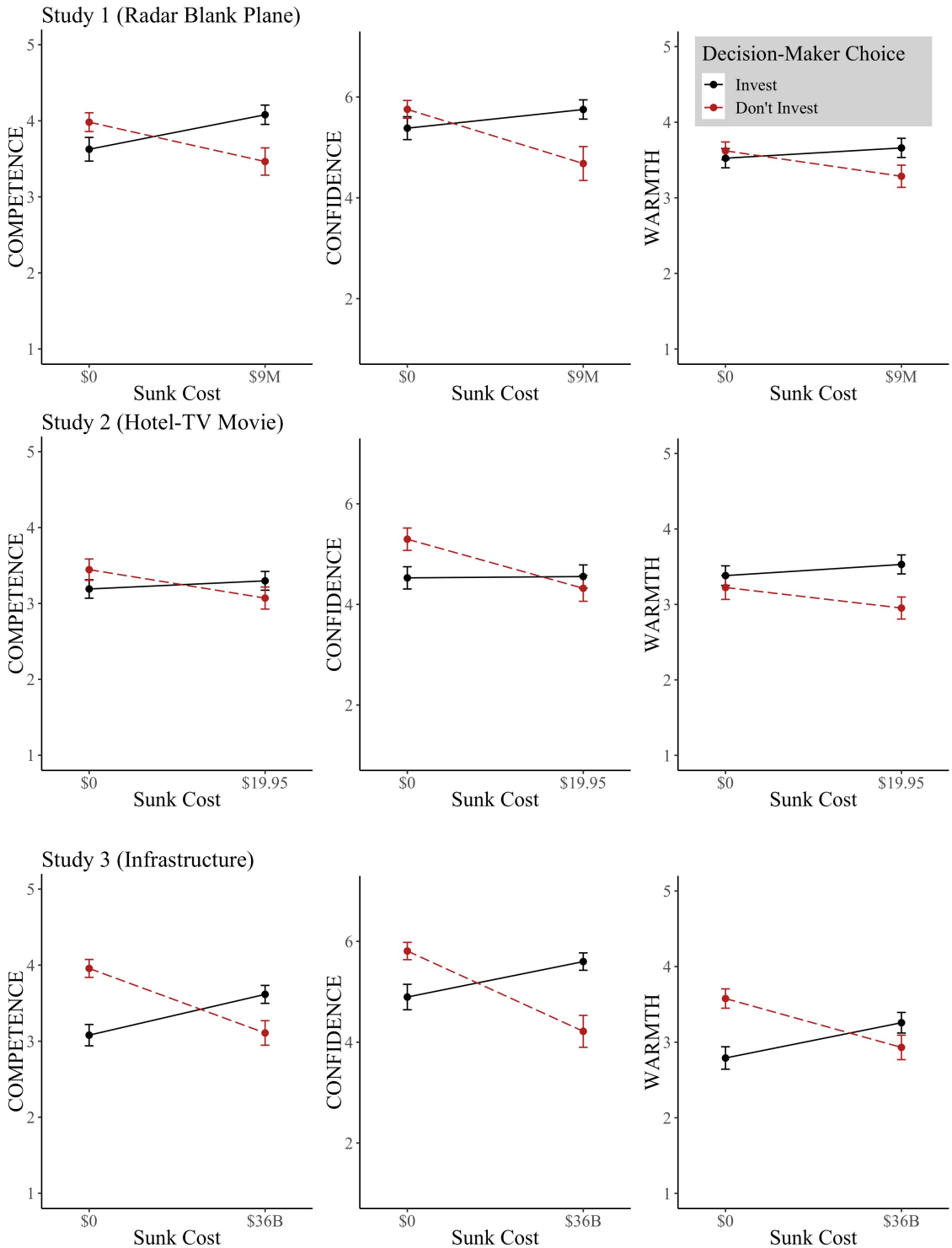

Figure 2.1. *Results from Studies 1 through 3: Perceived competence, confidence, and warmth of a decision maker as a function of sunk cost and future investment. Decision makers were evaluated more positively for investing (vs. not investing) when costs had been sunk, but evaluated more negatively for investing (vs. not investing) in the absence of a sunk cost. Error bars represent 95% CI.*

Table 2.1. *Studies 1-3 Social Perceptions by Condition*

Experiment	No Sunk Cost condition Decision-maker choice		Sunk Cost condition Decision-maker choice	
	Invest	Don't Invest	Invest	Don't Invest
S1: Radar-blank plane	*n* = 93	*n* = 101	*n* = 103	*n* = 93
Competence	3.63 [3.47, 3.78]	3.98 [3.86, 4.11]	4.08 [3.95, 4.21]	3.46 [3.28, 3.65]
Confidence	5.38 [5.15, 5.61]	5.75 [5.58, 5.93]	5.75 [5.56, 5.95]	4.68 [4.34, 5.02]
Warmth	3.52 [3.39, 3.65]	3.62 [3.50, 3.74]	3.66 [3.53, 3.79]	3.28 [3.14, 3.43]
S2: Hotel-TV movie	*n* = 108	*n* = 97	*n* = 100	*n* = 110
Competence	3.19 [3.07, 3.31]	3.45 [3.30, 3.59]	3.30 [3.17, 3.42]	3.07 [2.92, 3.21]
Confidence	4.53 [4.30, 4.75]	5.30 [5.07, 5.52]	4.56 [4.32, 4.79]	4.32 [4.06, 4.58]
Warmth	3.38 [3.25, 3.51]	3.22 [3.38, 3.22]	3.53 [3.40, 3.66]	2.95 [2.80, 3.10]
S3: Infrastructure	*n* = 98	*n* = 107	*n* = 101	*n* = 94
Competence	3.08 [2.94, 3.22]	3.96 [3.84, 4.08]	3.62 [3.50, 3.74]	3.11 [2.94, 3.27]
Confidence	4.90 [4.64, 5.15]	5.81 [5.64, 5.98]	5.60 [5.42, 5.77]	4.22 [3.89, 4.54]
Warmth	2.79 [2.64, 2.94]	3.58 [3.45, 3.71]	3.26 [3.12, 3.39]	2.93 [2.77, 3.09]

Note. Studies 1-3: Mean ratings of perceptions of competence (5-point scale 1-5), confidence (7-point scale 1-5), and warmth (5-point scale 1-5) grouped by sunk cost condition and investment choice. 95% confidence intervals for the mean estimates are in brackets.

Table 2.2. *Studies 1-3 Contrasting Social Perceptions of Investors and Non-Investors*

Experiment & Contrast	Mean difference	*t*	*p*	*d*
S1: Radar-blank plane				
Competence				
No Sunk Cost; Invest vs. Not Invest	-0.36 [-0.56, -0.16]	-3.52	<.001	-0.51 [-0.80, -0.22]
Sunk Cost; Invest vs Not Invest	0.62 [0.39, 0.84]	5.49	<.001	0.80 [0.50, 1.09]
Confidence				
No Sunk Cost; Invest vs. Not Invest	-0.37 [-0.66, -0.08]	-2.54	.012	-0.37 [-0.65, -0.08]
Sunk Cost; Invest vs Not Invest	1.07 [0.68, 1.46]	5.44	<.001	0.80 [0.50, 1.09]
Warmth				
No Sunk Cost; Invest vs. Not Invest	-0.10 [-0.27, 0.08]	-1.10	.273	-0.15 [-0.44, 0.13]
Sunk Cost; Invest vs Not Invest	0.38 [0.18, 0.57]	3.79	<.001	0.54 [0.26, 0.83]
S2: Hotel-TV movie				
Competence				
No Sunk Cost; Invest vs. Not Invest	-0.25 [-0.44, -0.07]	-2.69	.008	-0.38 [-0.66, -0.10]
Sunk Cost; Invest vs Not Invest	0.23 [0.03, 0.42]	2.33	.020	0.32 [0.05, 0.59]
Confidence				
No Sunk Cost; Invest vs. Not Invest	-0.77 [-0.95, -0.38]	-4.80	<.001	-0.67 [-0.95, -0.39]
Sunk Cost; Invest vs Not Invest	0.23 [-0.11, 0.58]	1.33	.186	0.18 [-0.09, 0.46]
Warmth				
No Sunk Cost; Invest vs. Not Invest	0.16 [-0.05, 0.36]	1.51	.132	0.21 [-0.06, 0.49]
Sunk Cost; Invest vs Not Invest	0.58 [0.38, 0.77]	5.87	<.001	0.80 [0.52, 1.09]
S3: Infrastructure				
Competence				
No Sunk Cost; Invest vs. Not Invest	-0.88 [-1.06, -0.69]	-9.37	<.001	-1.32 [-1.62, -1.01]
Sunk Cost; Invest vs Not Invest	0.51 [0.31, 0.71]	4.97	<.001	0.72 [0.43, 1.01]
Confidence				
No Sunk Cost; Invest vs. Not Invest	-0.91 [-1.22, -0.61]	-5.86	<.001	-0.83 [-1.12, -0.54]
Sunk Cost; Invest vs Not Invest	1.38 [1.02, 1.75]	7.51	<.001	1.10 [0.79, 1.40]
Warmth				
No Sunk Cost; Invest vs. Not Invest	-0.79 [-0.98, -0.59]	-7.87	<.001	-1.11 [-1.40, -0.81]
Sunk Cost; Invest vs Not Invest	0.33 [0.11, 0.54]	3.03	.003	0.44 [0.15, 0.72]

Note. Studies 1-3: Simple contrasts of social perceptions show that individuals who invest in the presence of sunk costs are consistently perceived more positively than individuals who do not invest in the presence of sunk costs. The opposite pattern occurs in the absence of sunk costs. 95% confidence intervals are displayed in brackets.

Table 2.3. *Studies 1-3 Interactions Between Sunk Cost and Investment Conditions*

Experiment & Contrast	Estimate of interaction	*t*	*p*	β
S1: Radar-blank plane				
Competence	-0.97	-6.50	<.001	-0.53
	[-1.27, -0.68]			[-0.70, -0.37]
Confidence	-1.45	-5.98	<.001	-0.49
	[-1.92, -0.97]			[-0.65, -0.33]
Warmth	-0.47	-3.57	<.001	-0.30
	[-0.73, -0.21]			[-0.47, -0.14]
S2: Hotel-TV movie				
Competence	-0.48	-3.54	<.001	-0.30
	[-0.75, -0.21]			[-0.47, -0.13]
Confidence	-1.00	-4.18	<.001	-0.35
	[-1.47, -0.53]			[-0.51, -0.18]
Warmth	-0.42	-2.93	.004	-0.24
	[-0.70, -0.14]			[-0.41, -0.08]
S3: Infrastructure				
Competence	-1.38	-10.09	<.001	-0.76
	[-1.65, -1.11]			[-0.90, -0.61]
Confidence	-2.30	-9.72	<.001	-0.73
	[-2.76, -1.83]			[-0.88, -0.58]
Warmth	-1.11	-7.61	<.001	-0.60
	[-1.40, -0.83]			[-0.75, -0.44]

Note. Studies 1-3: Regressing social perceptions on sunk cost, future investment, and their interaction term reveals that the effect of decision maker choice on evaluator perceptions of competence, confidence, and warmth significantly depends on the presence or absence of a sunk cost. 95% confidence intervals are displayed in brackets.

Discussion

Studies 2-3 provided consistent evidence that participants held more positive views of individuals who persisted after incurring a sunk cost but more negative views of those who pursued a new investment with the same outlook. We found this pattern of results across two new settings (consumer, political) and across three different social perceptions (competence, confidence, warmth).

Although Studies 1-3 provide compelling evidence for the overall pattern of results in the domain of social perceptions, their respective implications may be limited by the lack of real-world consequences. We thus designed Study 4 to overcome this limitation.

Study 4: Will the Benefit of Bias Persist With Real Financial Consequences?

Studies 1-3 demonstrated that in the presence of sunk cost social perceptions favoring escalation of commitment can conflict with economic norms favoring pivoting to new opportunities. In Study 4, we test this effect when social judgments carry real financial consequences. Further, we examine the influence of warmth and competence perceptions on these financial rewards. (We discontinued assessments of confidence because confidence did not differ meaningfully from the broader concept of competence in Studies 1-3.)

Drawing on prior work suggesting that decision makers might rationally escalate commitment (i.e., invest in the presence of sunk costs) in order to preserve a reputation for competence (Kanodia et al., 1989; McAfee et al., 2010), we predicted that perceptions of competence, but not warmth, would underpin financial rewards in a dictator game framed as a reward for competent decision making (Bardsley, 2008).

Finally, in Study 4, the person being evaluated was assigned to their role as the decision maker only after the initial investment had already been made. Because they were responsible only for the decision to escalate or not, this design allowed us to isolate the second decision as the sole basis for evaluations.

Method

Participants. Recruiting, base payment, and advertising were identical to prior studies. We recruited 1,000 participants based on effect size estimates from prior studies. The final sample for inferential analyses consisted of N = 925 participants (55% female; age: M = 36.7, *SD* = 11.8, Median = 34). Attrition was below 5% and did not vary by condition.

Procedure. The procedure closely resembled prior studies with the addition of a financial consequence outcome variable. Participants again observed a choice by a decision maker in a

situation where we varied the presence or absence of sunk cost as well as the decision to invest or not (radar-blank plane, Arkes & Blumer, 1985, slightly revised for clarity). While we told participants that their partner for the game was another MTurk worker, in reality there was not another player; their partner's choice was randomly assigned according to condition. A suspicion check was added at the conclusion of the survey. This pairing of partners was necessary to obtain our behavioral measure but also to addresses a potential confound from the first three studies. In the sunk cost condition, evaluators in Studies 1-3 evaluated a decision maker responsible for both the initial investment and the decision to escalate while only the escalation decision was made in the no sunk cost condition. In Study 4, the person being evaluated was assigned to their role as the decision maker only after the initial investment had already been made. Because they were responsible only for the decision to escalate or not, it allowed us to isolate the second decision as the sole basis for evaluations.

Participants made social judgments of warmth and competence about their partner before making a choice with real financial consequences. Participants received a bonus allocation of $0.20 to allocate between themselves and their partner in a dictator game (e.g., Forsythe, Horowitz, Savin, & Sefton, 1994). The allocation was framed as a performance reward for the partner based on their performance in the role of CEO in the radar-blank plane scenario. Participants received the portion of the bonus that was not allocated to the (fictitious) partner only if they correctly answered all comprehension questions.

Exploratory Variables. As in prior studies, we collected a variety of exploratory variables. Participants reported what they would have chosen to do if they were in the situation described. We also collected perceptions of the probability that investing would be financially successful (that the profit from sales of the plane would exceed the cost of finishing

development). Finally, participants reported demographic information (age, gender, education), and whether they had previously been taught the principle of sunk cost.

Results

Based on the *benefit of bias hypothesis*, we predicted that factoring in sunk costs would lead to social benefits for decision makers, even when such benefits required real financial costs from evaluators. We thus predicted an interaction between sunk cost (none or $9 million) and future investment (invest or don't invest) on financial reward. Specifically, we predicted that, having incurred substantial sunk costs, future investment would lead to increased financial reward; however, without financial sunk costs, we predicted that future investment would lead to decreased financial reward (both in comparison to lack of future investment). We predicted that perceptions of competence, but not warmth, would mediate the relationship between future investment and financial reward.

Analytic Plan. We coded the amount of money sent as the proportion of available funds transferred [0,1]. Based on pilot data, we had concerns about violating the normality assumptions for parametric tests and OLS regression. Many participants chose to allocate either all or none of their endowment to their partner resulting in a "U" shaped distribution for the dependent variable. We pre-registered and conducted an analysis approximating the data using a beta distribution rather than a normal distribution in an effort to better fit the data; we also used non-parametric tests where appropriate. Simulations based on our data, however, confirm that a standard t-test remains a valid test of our hypothesis. Because the results are consistent regardless of the analysis used, we report the results of the more familiar tests here; full details of simulations and alternative analyses are available in Appendix 2.

Interaction. Reinforcing the results of Studies 1-3, the present results with real financial incentives mirrored the results observed for social perceptions. The effect of investing on financial rewards significantly depended on the presence or absence of sunk costs, b = -0.18, CI_{95} [-.25, -.10], t = -4.58, p<.001, β = -0.27, CI_{95} [-0.38, -0.15] (Figure 2.2). In the presence of sunk costs, participants financially rewarded decision makers who persisted. Specifically, participants awarded 26% more money to decision makers who persisted (M =0.34, CI_{95}[0.30, 0.38]) compared to decision makers who pivoted (M = 0.27, CI_{95}[0.23, 0.31]), t(443.8) = -2.59, p = .010, d = -0.24, CI_{95}[-0.43, -0.06]. In the absence of sunk cost, this pattern reversed: participants awarded 30% more money to decision makers who passed on the opportunity to invest (M = .36, CI_{95}[0.32, 0.40]) compared to decision makers who plunged ahead despite warning signs (M = .25, CI_{95}[0.22, 0.29]), t(426.88) = 3.83, p < .001, d = 0.36, CI_{95}[0.18, 0.55].

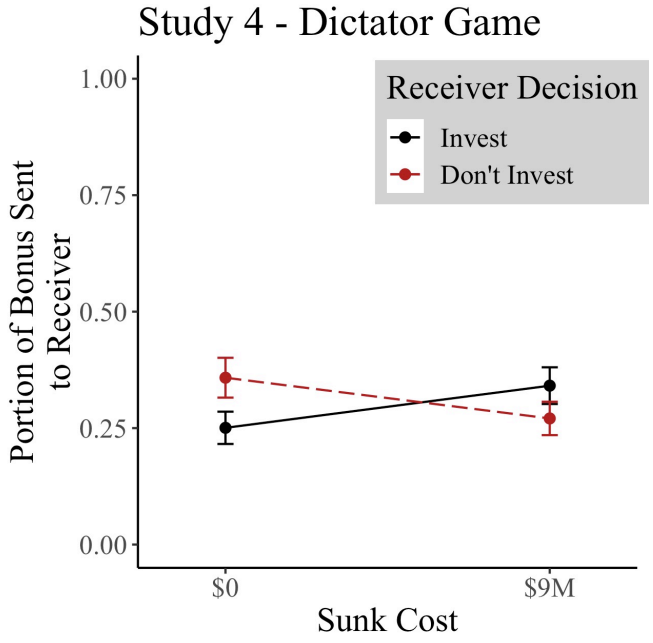

Figure 2.2. *In the presence of sunk costs, participants awarded 26% more money to decision makers who persisted. In the absence of sunk cost, this pattern reversed: participants awarded 30% more money to decision makers who passed on the opportunity to invest. Error bars represent 95% CI.*

Replication of Key Result. Given that we were predominantly interested in the reactions of evaluators in the problematic sunk cost condition, we ran a direct replication of the sunk cost condition from Study 4 in order to further examine the simple effect of investment on financial reward. Using a slightly larger sample for the two sunk cost cells of interest (N=622 post pre-registered exclusions), we found a similar effect: participants awarded 20% more money to decision makers who persisted (M =0.35, CI_{95}[0.32, 0.39]) compared to decision makers who pivoted (M = 0.29, CI_{95}[0.26, 0.32]), t(615.46) = -2.46, p = .014, d = -0.20, CI_{95}[-0.36, -0.04].

Mediation Analysis: Warmth vs. Competence. We now turn our attention to the role of perceptions of warmth and competence in mediating this interaction. We predicted that perceptions of competence, but not perceptions of warmth, would mediate the effect of future investment on financial rewards. We fit three structural equation models to test this hypothesis.

Model 1: Mediation in the Presence of Sunk Cost. We first examined a parallel mediation model in which there was prior investment (Figure 2.3, top panel). When there was prior investment, decision makers who invested were perceived as more competent (b = 0.52, CI_{95} [0.383, 0.650], z = 7.64, p < .001) and warmer (b = 0.22, CI_{95} [0.081, 0.368], z = 3.07, p = .002) than individuals who did not invest. While perceptions of competence were associated with financial rewards (b = 0.092, CI_{95} [0.055, 0.126], z = 5.007, p < .001) perceptions of warmth were not (b = .00, CI_{95} [-0.041, 0.040], z = 0.145, p = .884). Taken together, we observed a significant indirect effect of investment on financial rewards through increased perceptions of competence (b = .048, CI_{95} [0.026, 0.070], z = 4.179, p < .001), but not through increased perceptions of warmth (b = -0.001, CI_{95} [-0.010, 0.010], z = 0.136, p = .892).

Model 2: Mediation in the Absence of Sunk Cost. We next examined a parallel mediation model in which there was no prior investment (Figure 2.3, bottom panel). The results

provided an opposite, albeit weaker, pattern of results than those reported above. When there was no prior investment, decision makers who invested were perceived as marginally less competent (b = -0.11, CI_{95} [-0.23, 0.02], z = 1.692, p = .091) but equally warm (b = .06, CI_{95} [-.077, 0.193], z = 0.831, p = .406) compared to individuals who did not invest. While perceptions of competence were again associated with financial rewards (b = .08, CI_{95} [0.039, 0.122], z = 3.707, p < .001), perceptions of warmth were again unassociated with financial rewards (b = -.01, CI_{95} [-0.055, 0.043], z = 0.223, p = .823). In this model, neither the indirect path through competence (b = -.01, CI_{95} [-0.021, 0.001], z = 1.497, p = .134) nor warmth (b = 0.00, CI_{95} [-0.005, 0.004], z = 0.138, p = .89) reached statistical significance.

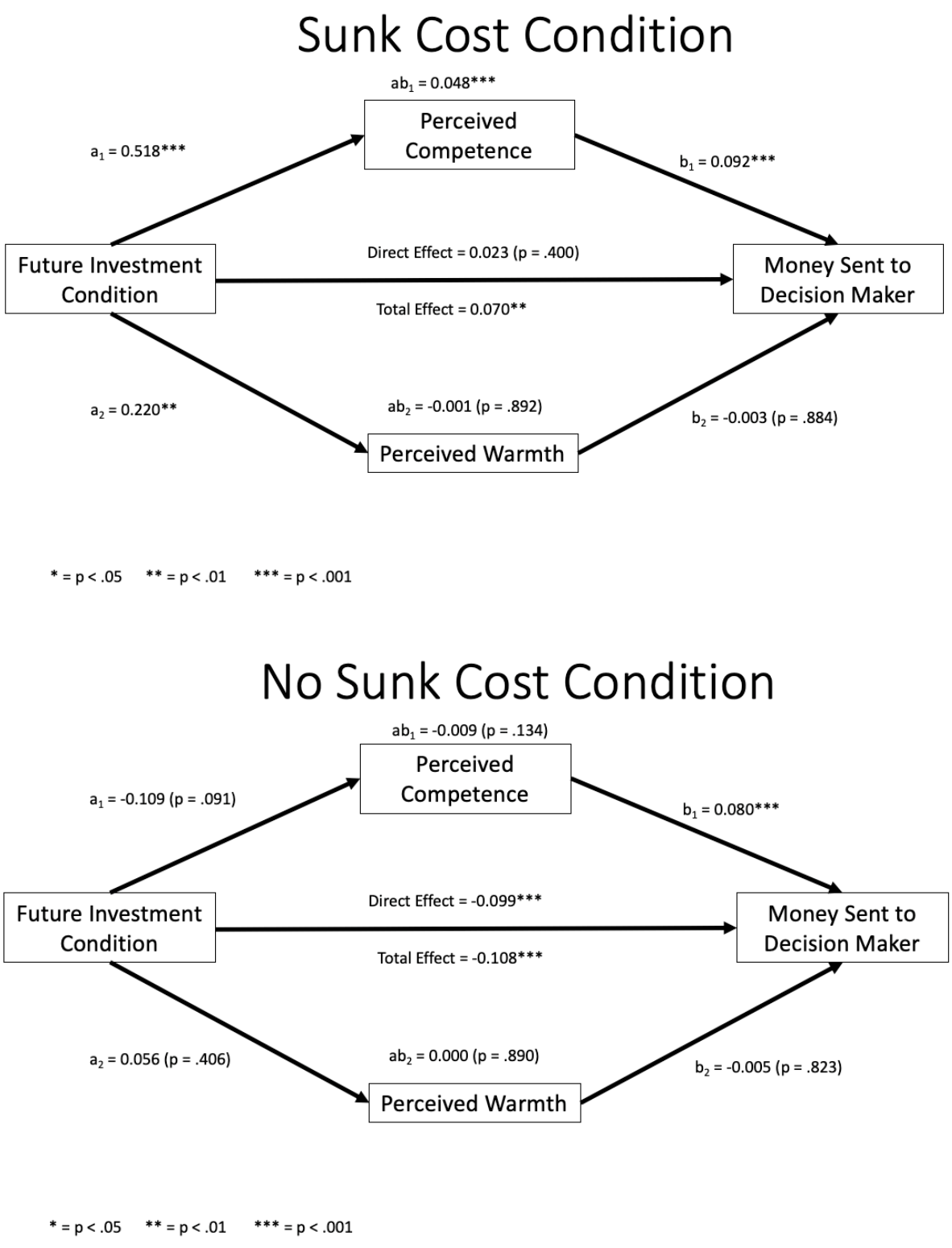

Figure 2.3. *Parallel Mediation Results from Study 4*

Model 3: Moderated Mediation. Finally, we fit a moderated mediation model to test the hypothesis that the effect of future investment on financial rewards through competence depends on the presence of prior investment (Figure 2.4). As in the previous models, future investment served as the independent variable, perceived competence served as a mediating variable, and financial reward served as the dependent variable. In this model, we entered prior investment condition as a moderating variable. We dropped perceptions of warmth because it did not explain significant variance in either prior model.

The model provides evidence that the indirect effect of future investment on financial rewards depended on the presence of prior investments or sunk costs (a_3b_1 = .058, CI_{95} [0.036, 0.085], z = 4.709, p<.001). Specifically, the results demonstrate that in all cases, perceived competence positively predicts money sent to the decision maker. However, the effect of future investment on perceived competence depends on the presence of prior investment. On the one hand, when there is no prior investment, future investment serves as a negative signal of competence. On the other hand, when there is prior investment, future investment serves as a positive signal of competence.

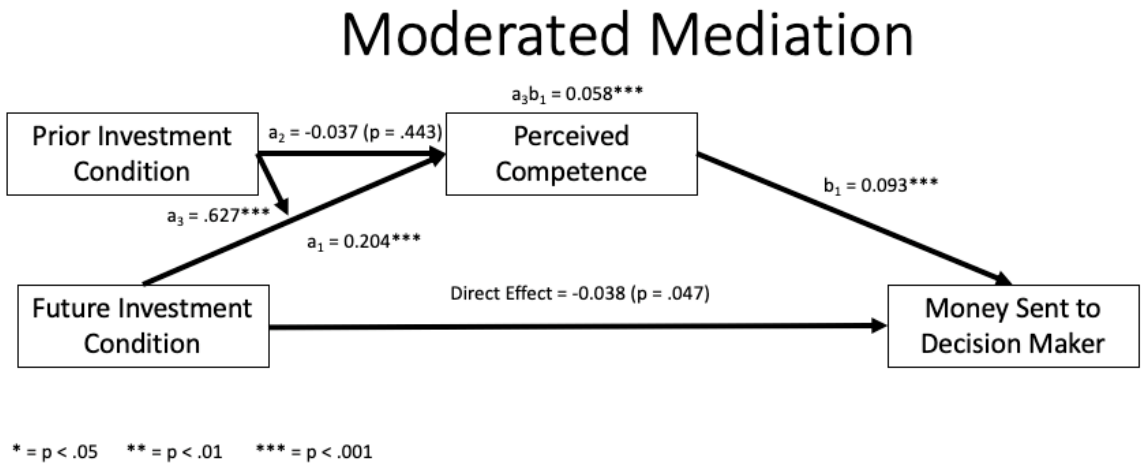

Figure 2.4. *Moderated Mediation Model for Study 4*

Discussion

Prescriptive decision theory suggests that perceptions of decision makers' competence—and subsequent financial rewards—should depend solely on the future costs and benefits associated with a decision option. Study 4, however, shows that when decision makers face a losing course of action, perceptions of competence depend on the presence of a sunk cost. In the presence of sunk costs, the decision to invest signals competence and leads to financial rewards for decision makers; the opposite is true in the absence of sunk costs. Warmth is evidently not a relevant factor for financial compensation in this context. Thus, the financial benefit of the sunk-cost bias we observed arises from participants' belief that—contrary to economic principles—escalation in response to sunk cost reflects competence.

General Discussion

We tested two competing hypotheses regarding social judgments of decision makers facing sunk costs—the *reward for rationality hypothesis* and the *benefit of bias hypothesis*. Four experiments supported the *benefit of bias hypothesis*: Decision makers who chose to invest more funds after a prior investment were rated more favorably than decision makers who did not, a pattern that reversed in the absence of a prior investment. Studies 1-3 revealed that, in business, consumer, and political domains, individuals who escalated in response to sunk costs were perceived as warmer, more competent, and more confident than individuals who pivoted from sunk costs. In the absence of sunk costs, de-escalation was favored by evaluators despite identical prospects. Study 4 revealed that perceptions of competence – but not warmth – mediate the effect of escalation on real financial rewards (or lack thereof) from evaluators.

That decision makers seek not only to maximize economic value but also to generate socially/politically acceptable choices is not a new idea (for reviews, Lerner & Tetlock, 1999;

Tetlock, 2002). What remained unknown was whether flouting maxims for rational economic decision making could be tangibly beneficial. The present evidence fills this gap: Decision makers received larger benefits precisely when they factored sunk costs into their decisions than when they ignored sunk costs. Thus, the present results add empirical content to the hypothesis (Tetlock, 1992) that ostensibly biased tendencies to escalate commitment become pragmatic when viewed through a functionalist lens that portrays decision makers as intuitive politicians rather than as intuitive economists. Through this lens, people seek more to protect their social identities in the eyes of constituencies than to efficiently maximize monetary return on investment (cf. Grossmann et al., 2020; Tenney et al., 2019; Jordan et al., 2016).

The present evidence also helps to explain why sunk-cost bias is so prevalent across decision contexts: Most high-impact decisions take place in social/organizational settings that make social/political incentives salient. Relatedly, the evidence explains why politicians are more prone to escalate commitment in response to sunk costs than the average person (Sheffer et al., 2017): Politicians who win elections are good at recognizing social incentives.

This game-theoretical dynamic, where an actor's payoffs depend in part on the actions and preferences of others, can, however, have suboptimal outcomes. Like a tragedy of the commons, although the incentives for any given individual may be to escalate, society as a collective is better off when policymakers objectively weigh future investment options based on future net expected value (Mankiw, 2020, p. 261). Reducing sunk-cost bias thus remains a worthwhile endeavor. Prior studies have focused on debiasing individuals, providing evidence that teaching decision makers about normative economic models can improve decision making (Larrick et al., 1990). While these attempts have been at least moderately successful, the present

results suggest that the success of existing debiasing efforts may be attributable to the creation of a localized culture of economically rational decision making.

In our studies, we found that participants who reported having previously learned about sunk cost were no less likely to prefer escalation (X^2 (1, N = 925) = 1.325, p = .250). Thus, knowing one should ignore sunk costs is of limited value without a supportive audience who allows decision makers to do so. Indeed, the most successful strategies for reducing escalation (for discussion, Simonson & Staw, 1992) may succeed precisely because they reduce the *social* costs of de-escalation (see Thompson, 2017). The success of such approaches (i.e., changing the social incentives) derives from the fact that human decision makers are fundamentally social beings. Building on classical sociology—Mead, Cooley, Durkheim—as well as on contemporary social-functionalism (e.g., Fiske, 1992; Pettigrew, 2018), the present evidence underscores the principle that thinking processes operate in service of social interaction and social order.

These results provide the first empirical evidence, to our knowledge, that honoring sunk costs can confer social/reputational and financial benefits. Inasmuch as the present work used real financial incentives, examined multiple decision contexts, and drew from an adult population distributed across the United States, the results may generalize across many contexts. While we found no significant effects of age, gender, education, or familiarity with sunk cost, future research should examine the effects of local norms concerning sunk costs and evidence-based decision making, and should test replication in other countries.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

Baron, J. (1990). Harmful Heuristics and the Improvement of Thinking. *Developmental Perspectives on Teaching and Learning Thinking Skills*, *21*, 28–47. https://doi.org/10.1159/000418979

Brest, P., & Krieger, L. H. (2010). *Problem solving, decision making, and professional judgment: A guide for lawyers and policymakers* (1st ed). Oxford University Press.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417. https://doi.org/10.1037/h0053870

Fast, N. J., Sivanathan, N., Mayer, N. D., & Galinsky, A. D. (2012). Power and overconfident decision-making. *Organizational Behavior and Human Decision Processes*, *117*(2), 249–260. https://doi.org/10.1016/j.obhdp.2011.11.009

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from Daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, *63*(6), 877–889. https://doi.org/10.1037/0022-3514.63.6.877

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037//0022-3514.82.6.878

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, *6*(3), 347–369. https://doi.org/10.1006/game.1994.1021

Foschi, M. (1992). Gender and Double Standards for Competence. In C. L. Ridgeway (Ed.), *Gender, Interaction, and Inequality* (pp. 181–207). Springer. https://doi.org/10.1007/978-1-4757-2199-7_8

Friedman, M. (1953). *Essays in Positive Economics*. University of Chicago Press.

Frisch, D. (1993). Reasons for Framing Effects. *Organizational Behavior and Human Decision Processes*, *54*(3), 399–429. https://doi.org/10.1006/obhd.1993.1017

Grossmann, I., Eibach, R. P., Koyama, J., & Sahi, Q. B. (2020). Folk standards of sound judgment: Rationality Versus Reasonableness. *Science Advances*, *6*(2), eaaz0289. https://doi.org/10.1126/sciadv.aaz0289

John, L. K., Jeong, M., Gino, F., & Huang, L. (2019). The self-presentational consequences of upholding one's stance in spite of the evidence. *Organizational Behavior and Human Decision Processes*, *154*, 1–14. https://doi.org/10.1016/j.obhdp.2019.07.001

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113

Kanodia, C., Bushman, R., & Dickhaut, J. (1989). Escalation Errors and the Sunk Cost Effect: An Explanation Based on Reputation and Information Asymmetries. *Journal of Accounting Research*, *27*(1), 59–77. https://doi.org/10.2307/2491207

Kreps, T. A., Laurin, K., & Merritt, A. C. (2017). Hypocritical flip-flop, or courageous evolution? When leaders change their moral minds. *Journal of Personality and Social Psychology*, *113*(5), 730–752. https://doi.org/10.1037/pspi0000103

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the Use of Cost-Benefit Reasoning in Everyday Life. *Psychological Science*, *1*(6), 362–370. https://doi.org/10.1111/j.1467-9280.1990.tb00243.x

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychology Bulletin*, *125*(2), 255–275.

Mankiw, N. G. (2020). *Principles of Economics* (9th ed.). Boston, Mass.: Cengage Learning.

McAfee, R. P., Mialon, H. M., & Mialon, S. H. (2010). Do Sunk Costs Matter? *Economic Inquiry*, *48*(2), 323–336. https://doi.org/10.1111/j.1465-7295.2008.00184.x

Nagouney, A. (2018, July 30). *A $100 Billion Train: The Future of California or a Boondoggle?* New York Times. https://www.nytimes.com/2018/07/30/us/california-high-speed-rail.html

Olivola, C. Y. (2018). The Interpersonal Sunk-Cost Effect. *Psychological Science*, *29*(7), 1072–1083. https://doi.org/10.1177/0956797617752641

Pettigrew, T. F. (2018). The Emergence of Contextual Social Psychology. *Personality and Social Psychology Bulletin*, *44*(7), 963–971. https://doi.org/10.1177/0146167218756033

Ridgeway, C. L. (2001). Gender, Status, and Leadership. *Journal of Social Issues*, *57*(4), 637–655. https://doi.org/10.1111/0022-4537.00233

Schwartz, B. (2006, September 17). The "sunk-cost fallacy." *Los Angeles Times*. https://www.latimes.com/archives/la-xpm-2006-sep-17-oe-schwartz17-story.html

Sheffer, L., Loewen, P. J., Soroka, S., Walgrave, S., & Sheafer, T. (2017). Nonrepresentative Representatives: An Experimental Study of the Decision Making of Elected Politicians. *American Political Science Review*, *112*(2), 302–321. https://doi.org/10.1017/S0003055417000569

Simonson, I., & Staw, B. M. (1992). Deescalation strategies: A comparison of techniques for reducing commitment to losing courses of action. *Journal of Applied Psychology*, *77*(4), 419–426. https://doi.org/10.1037/0021-9010.77.4.419

Sleesman, D., Conlon, D. E., McNamara, G., & Miles, J. (2012). Cleaning Up the Big Muddy: A Meta-Analytic Review of the Determinants of Escalation of Commitment. *The Academy of Management Journal*, *55*(3), 541–562. https://doi.org/10.5465/amj.2010.0696

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, *116*(3), 396–415. https://doi.org/10.1037/pspi0000150

Tetlock, P. E. (1992). The Impact of Accountability on Judgment and Choice: Toward A Social Contingency Model. *Advances in Experimental Social Psychology*, *25*, 331–376. https://doi.org/10.1016/S0065-2601(08)60287-7

Tetlock, P. E. (2000). Cognitive Biases and Organizational Correctives: Do Both Disease and Cure Depend on the Politics of the Beholder? *Administrative Science Quarterly*, *45*(2), 293–326. https://doi.org/10.2307/2667073

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471. https://doi.org/10.1037/0033-295X.109.3.451

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60. https://doi.org/10.1016/0167-2681(80)90051-7

Thompson, D. (2017, November). *Inside X, Google's Moonshot Factory*. The Atlantic. https://www.theatlantic.com/magazine/archive/2017/11/x-google-moonshot-factory/540648/

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

Essay 3

Trust Me, I'm Irrational:

Escalation of Commitment Is a Reliable Signal of Trustworthiness[4]

with Charles A. Dorison and Jennifer S. Lerner

Escalation of commitment—the tendency to persist in a losing course of action despite negative prospects (Brockner, 1992; Staw, 1976)—is ubiquitous. People finish books they don't like, continue business ventures that are fatally flawed, and continue wars that are unwinnable. They do so despite the widely-held prescription that weighing options based solely on future net expected value achieves better material outcomes (Mankiw, 2020, p. 261).

If escalation is costly, why does it persist? Prior research identified a variety of cognitive factors that drive this robust phenomenon (for review, see Sleesman et al., 2012). In addition, conceptual papers have theorized a role for social factors (Staw, 1981; Tetlock, 2000). To date, however, relatively little work has examined the role of social/structural factors experimentally.[5]

In the present work, we test whether the social incentives typically excluded from economic models of rationality might compensate for the direct material costs of escalation. We hypothesize that escalation serves as a signal of the decision maker's trustworthiness. Because trust is essential for effective personal and professional relationships (Arrow, 1974; Dirks & Ferrin, 2002; Kramer, 1999), the benefits of signaling trustworthiness may partially offset the costs of escalation, contributing to the prevalence of this behavior. This hypothesis gives rise to a set of inter-related questions.

---

[4] This paper is based upon work supported by: The National Science Foundation under Grant No. (1559511), the National Institute of Health under Grant No. (1R01CA224545-01A1), the Harvard Program on Negotiation, and the Harvard Mind Brain Behavior Initiative.

[5] Meta-analysis by Sleesman and colleagues (2012, pg. 557) concluded that "Researchers have emphasized project and psychological determinants at the expense of social and structural factors."

(1) Do third-party observers trust decision makers who escalate commitment to a failing course of action more than they trust decision makers who de-escalate? Trusting behavior is influenced by the perceived ability and motivations of the trusted party (Mayer et al., 1995). Escalation contradicts prescriptive economic models, potentially lowering perceptions of ability. On the other hand, escalation may be driven in part by non-economic considerations which could signal benevolent motives, increasing trust where motives are more important than ability. We hypothesize that decision makers who escalate commitment will be trusted more in a simple trust game than decision makers who do not.

(2a) Does escalation of commitment truly signal trustworthiness and, (2b) if so, is this signal robust to incentives to deceptively signal trustworthiness? Regarding (2a), escalation is multiply determined and reflects a more integrative set of preferences than simple utility maximization (for review, see Sleesman et al., 2012). Specifically, both escalation and trustworthy behavior imply subordination of material consumption to some other interests. We therefore predict escalation will correlate with trustworthy behavior.

Regarding (2b), the traditional rational actor model (Edwards, 1954) and decades of behavioral decision research predict that decision makers are sensitive to changes in social and financial incentives (Ashraf & Bandiera, 2018; Gneezy et al., 2011; Lerner & Tetlock, 1999; but see Shafir & LeBoef, 2002). Thus, if decision makers can intuit the potential benefits of signaling trustworthiness through escalation, strategic decision makers should be more likely to escalate when the cost of signaling is lower and/or the benefits are larger.

The Current Research

We address the preceding questions in two pre-registered experiments with real financial stakes (N = 2,198 U.S. adults). Study 1 tests whether escalation of commitment is perceived as a

signal of trustworthiness. Study 2 tests whether escalation truly signals trustworthiness and, if so, whether this signal is robust to incentives to strategically signal trustworthiness.

Open Science Statement

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in all studies (Simmons et al., 2012). All data, materials, and preregistrations are publicly available via the Open Science Framework (OSF).

Study 1: Do Observers Trust Decision Makers Who Escalate Commitment to a Failing Course of Action More Than They Trust Decision Makers Who De-Escalate?

Overview and Method

We employed a two-stage experimental design with two players: Observer and Actor. In the first stage, the Actor made a choice to escalate or de-escalate commitment to a failing course of action. The participant played the role of Observer. We randomly paired participants with Escalators (i.e., Actors who escalated commitment) or De-Escalators (i.e., Actors who de-escalated commitment). In the second stage, the Observer decided how much money to allocate to the Actor in a Trust Game (TG), described below. We predicted that, on average, Observers would trust Escalators more than De-Escalators.

Participants. We recruited 660 respondents from the United States using Amazon's Mechanical Turk (MTurk). We pre-determined our target sample size to ensure 80% power to detect a Cohen's d = .20 assuming a 10% loss due to exclusions (see Appendix 3). The final sample for analyses consisted of N = 602 (50% female; age: M = 37.3, SD = 11.7, Median = 35).

Procedure. We adapted procedures from Jordan and colleagues (Jordan, Hoffman, Nowak, et al., 2016), substituting an escalation choice as the signal event. We began by explaining the rules of the Trust Game to be played in Stage 2 (Berg et al., 1995). In the game,

Observers received a bonus allocation and had the opportunity to transfer any amount they chose (including none) to the Actor, another MTurk worker. We (the experimenters) tripled the amount transferred. The Actor then decided what percentage of the tripled amount, if any, they wanted to transfer back to the Observer. We asked two comprehension questions to ensure participants understood their incentives.

In Stage 1, participants read a commonly-used scenario (e.g., Arkes & Blumer, 1985; Olivola, 2018) in which the Actor assumed the role of a CEO who after completing 90% of product development learned that a competitor has launched a superior product. The Actor must decide whether to spend the money to finish developing the inferior product (i.e., escalate commitment) or cut their losses (i.e., de-escalate commitment). Following two more comprehension questions, we informed participants of the Actor's choice to escalate or de-escalate. Unbeknownst to the participant, this choice was randomly assigned by condition.

In Stage 2, Observers indicated how many cents they would like to send to the Actor in the TG, our primary dependent measure. To make the design incentive compatible, Observers received 50% of the tripled amount sent if they correctly answered all comprehension questions. Finally, participants responded to several exploratory and demographic questions (see Appendix 3 for details).

Results and Discussion

Consistent with our main hypothesis, escalation increased trusting behavior. Observers entrusted 29% more of their endowment money to Escalators (M = .60, SD = .39) compared to De-Escalators (M = .47, SD = .41), $t(599.2) = 4.18$, $p < .001$, $d = 0.34$, CI_{95} [.18, .50] (see Figure 3.1A). These results support the hypothesis that escalation is perceived as a signal of trustworthiness.

We also examined a pre-registered exploratory hypothesis that the effect of escalation on trust would depend on the observer's own preference. Specifically, we predicted that, while observers who themselves preferred escalation would trust Escalators more, observers who preferred de-escalation would show an attenuation of this effect. Research on other signals of trustworthiness, such as deontological moral judgments (Everett et al., 2016) and preference for equality over efficiency (Dorison et al., 2020), have shown a similar pattern. Results supported the hypothesis (interaction $b = 0.23$, $t = 3.27$, $p = .001$, $\beta = 0.27$, CI_{95} [0.11, 0.43]). Observers who would have escalated trusted Escalators more ($M = .66$, $SD = .36$) than they trusted De-Escalators ($M = .47$, $SD = .41$), $t(379.9) = 5.07$, $p < .001$, $d = 0.51$, CI_{95} [0.30, 0.71]. However, Observers who would have de-escalated did not show this difference ($M_{escalated} = .44$, $SD = .40$ vs. $M_{de\text{-}escalated} = .47$, $SD = .41$), $t(145.8) = -0.48$, $p = .630$, $d = -0.06$, CI_{95} [-0.33, 0.20] (Figure 3.1B). Taken together, the results of Study 1 make clear that escalation is perceived as a signal of trustworthiness by those observers who would have escalated. Appendix 3 contains preregistered beta regression results and robustness checks which support the main findings.

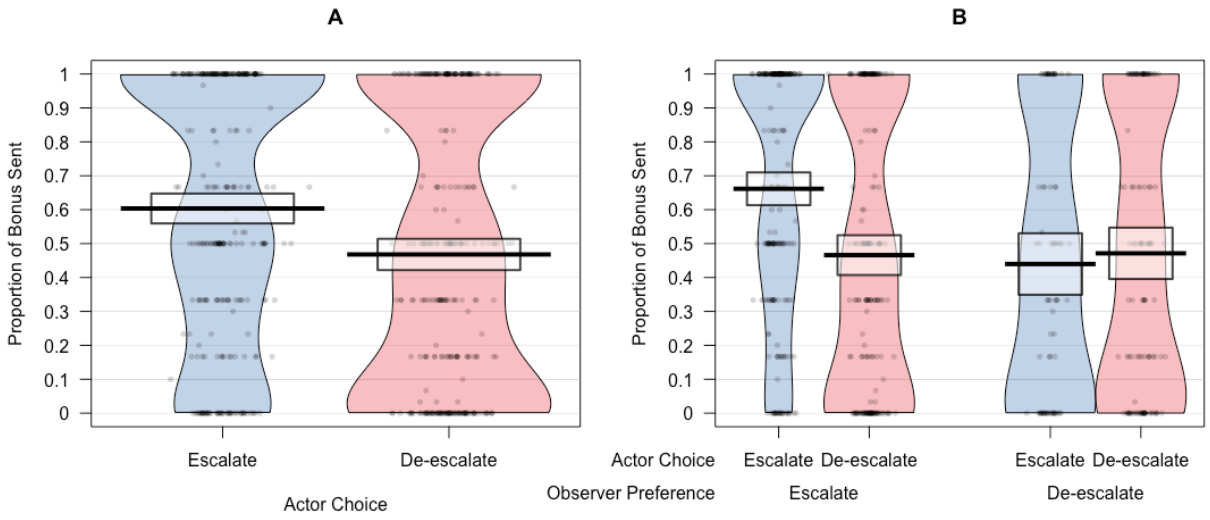

Figure 3.1. *Results from Study 1: Observers sent 29% more money in a Trust Game to Actors who escalated commitment following a sunk cost.* **(A).** *Observers who say they themselves would escalate trusted Actors who did so more than Actors who de-escalated. Observers who say they themselves would de-escalate did not modify their allocation as a function of the Actor's choice.* **(B).** *Violin plots display the distributions, and bars indicate the means. Rectangles show 95% confidence intervals.*

Study 2: Does Escalation Actually Signal Trustworthiness? If So, Is the Signal Robust to Incentives to Deceptively Signal Trustworthiness?

Overview and Method

Study 2 sought to examine whether Escalators were actually more trustworthy than De-escalators. We predicted that, on average, Escalators would return more money in a TG, validating escalation as a signal of trustworthiness. Further, we tested whether participants would be more likely to escalate commitment when their choice was public rather than private. Given the opportunity, would some participants escalate deceptively only to exploit other's trust or would escalation persist as a genuine signal of trustworthiness?

Participants. We recruited 2,787 respondents from MTurk.[6] The final sample after pre-registered exclusions for analyses consisted of N = 1,589 (54% female; age: $M = 36.9$, SD = 12.1, Median = 34). All reported results are qualitatively consistent regardless of exclusion (see Appendix 3).

Procedure. We employed a two-stage design similar to Study 1, except participants played the role of Actor and we manipulated the observability of their choice in Stage 1. In the unobserved condition, Actors read that the Observer would be unable to see their choice and that their choice could not possibly affect how much the Observer would send in the TG. In the observed condition, Actors read that the Observer would be able to see their choice and that this could affect how much the Observer chose to send. After making their choice to invest or not, and without knowing how much had been sent to them by the Observer, Actors decided what percentage of their tripled amount they would return. We used the data from the final sample of Observers in Study 1 to simulate outcomes of the game. Participants received the resulting bonus only if they correctly answered all comprehension questions. Finally, we collected some exploratory and demographic questions. Most important, the second wave of collection included a question in which respondents rated their own trait level reasonableness and rationality (see Appendix 3).

[6] We initially collected 803 respondents and did not detect a significant effect of observation on escalation, N = 465, $p = .894$, $OR = 0.974$, CI_{95} [0.655, 1.448]. Due to significantly higher than expected exclusion rates (42%) resulting from following our pre-registered exclusion plan, however, the estimate obtained was imprecise. We therefore collected an additional 1,985 respondents and report results from the combined sample to ensure a more precise estimate. Appendix 3 contains analyses for the individual samples, which are also qualitatively consistent with the combined results reported here.

Results and Discussion

Results supported the hypothesis that escalation would signal trustworthiness (Figure 3.2A): Escalators returned a 15% greater portion of their bonus (M = .35, SD = .23) compared De-Escalators (M = .31, SD = .25), $t(695.4) = 3.30$, p = .001, d = 0.19, CI_{95} [0.08, 0.31]. Escalators returned more money than De-Escalators regardless of exclusions or method of analysis (see Appendix 3).

Study 2 also considered whether escalation would be used to strategically signal trustworthiness when making the choice in public. Observation had no discernable impact on the probability of investing (b = 0.053, SE = 0.11, z = 0.462, p = .644, OR = 1.05, CI_{95} [0.844, 1.317]). Additionally, there was no effect of observation on the size of the relationship between escalation and trustworthiness (interaction p = .67, Figure 3.2B). Thus, escalation was a reliable signal of trustworthiness when observed. We consider possible explanations for this in the general discussion.

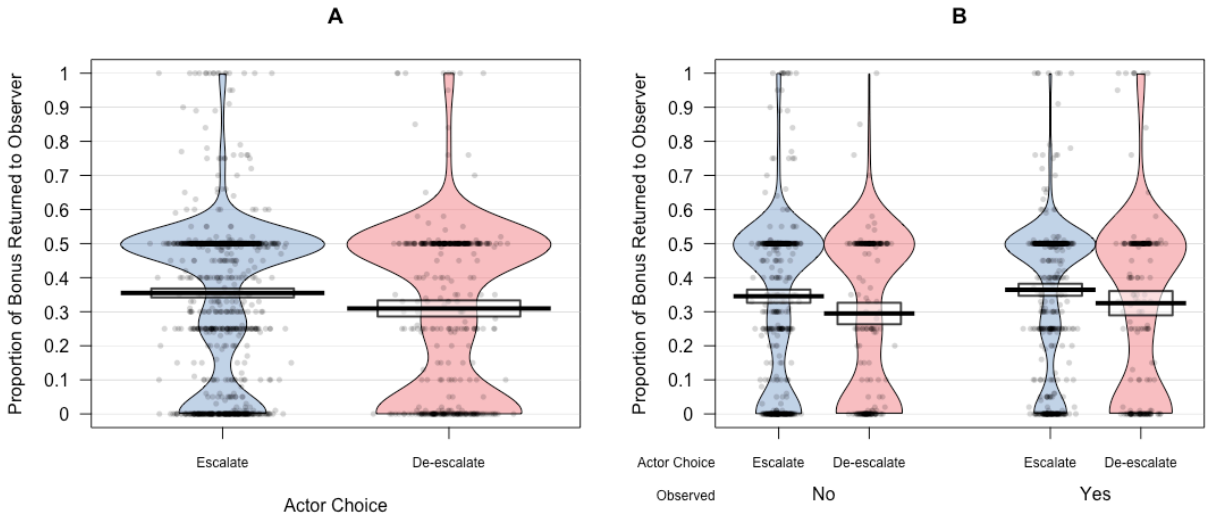

Figure 3.2. *Results from Study 2: Actors who chose to escalate commitment returned 15% more money in a Trust Game.* **(A).** *This relationship was robust to incentives to signal strategically when observed.* **(B).** *Violin plots display the distributions, and bars indicate the means. Rectangles show 95% confidence intervals.*

A final set of exploratory analyses sought to examine whether individual differences in reasonableness (versus rationality) could add explanatory value to the findings. Lay people view *reasonableness* as skill in balancing material self-interest with social, affective, and moral concerns. In contrast, *rationality* is abstract, instrumental and preference maximizing in an individual focused way (Grossmann et al., 2020).

We examined self-ratings of reasonableness and rationality in the second round of data collection by conducting two sets of regression analysis. A simultaneous logistic regression predicting escalation from standardized rationality and reasonableness ratings found that while reasonableness was associated with higher probability of escalation (b = 0.22, SE = 0.09, z = 2.54, p = .011, OR = 1.25, CI_{95} [1.05, 1.49]), rationality was associated with lower probability of escalation (b = -0.22, SE = 0.09, z = -2.46, p = .014, OR = 0.80, CI_{95} [0.67, 0.95]). A second simultaneous regression predicting standardized proportion of bonus money returned (i.e., trustworthiness) from escalation controlling for standardized reasonableness and rationality found that only reasonableness was significantly associated with more trustworthiness ($\beta_{reasonable}$= 0.13, CI_{95} [0.05, 0.20], SE = 0.04, t = 3.40, p < .001; $\beta_{rational}$ = -0.03, CI_{95} [-0.10, 0.05], SE = 0.04, t = -0.68, p = .498; $b_{escalate}$ = 0.10, CI_{95} [-0.04, 0.23], SE = 0.07, t = 1.42, p = .155). These results thus suggest that individual differences in reasonableness (versus rationality) partially explain the observed relationship between escalation and trustworthiness.

General Discussion

The present research examined the social causes and consequences of escalating commitment to failing courses of action. Across two pre-registered economic game experiments with real financial stakes, we found that escalation of commitment serves as a reliable and robust signal of trustworthiness. Observers entrusted 29% more money to partners who escalated

commitment. As predicted, this effect was driven by observers who would have escalated themselves, suggesting that escalation is perceived as an indicator of trustworthiness only by those with more integrative standards for decision making. Consistent with Observers expectations, Escalators were 15% more trustworthy than De-Escalators and this relationship was robust to observation.

The present work makes at least two key contributions. First, it demonstrates that economically irrational choices can be rewarded in some social situations. Escalation engendered trust in the decision maker, partially offsetting the potential costs of this behavior (for related work, see Everett et al., 2016; Grossmann et al., 2020; Tenney et al., 2019; Jordan, Hoffman, Nowak, et al., 2016; Jordan, Hoffman, Bloom, et al., 2016). As a complement to the more traditional approach to educating decision makers about standards for rationality, these findings suggest that a social/structural focus would be fruitful.

Second, the findings reveal that individual differences in reasonableness and rationality help explain and predict escalation behavior. Observers seemed to interpret de-escalation as an indicator that actors may be *too* rational and responded with distrust. By placing more value on cooperation, emotion, and others' welfare than rational self-interest models predict, "reasonable" actors enabled cooperative gains. Consistent with this interpretation, we found in Study 2 that those who described themselves as being more reasonable were more likely to escalate commitment and to be more trustworthy.

Limitations and Future Directions

Although any null effect must be interpreted with caution, we consider several explanations for why participants in Study 2 were not sensitive to observation. First, following social interactionist schools of thought (Goffman, 1983; Mead, 1930), decision makers may have

already internalized how others would evaluate them, signaling even when no one was explicitly observing them. Consistent with this idea, Jordan and Rand (2020) provide evidence that individuals employ a reputation heuristic, engaging in signaling behavior even when no one is watching. Second, it may be that Escalators did not recognize the opportunity to signal trustworthiness. Consistent with this idea, Observers in Study 1 who favored escalation did not differentiate between Escalators and De-Escalators. Future research is needed to tease apart these possibilities.

Future research should also examine cultural variation in the consequences of escalation. While our participants included a diverse set of adults across the United States, it could be the case that certain cultures (e.g., tight versus loose cultures; Gelfand et al., 2011) or sectors (e.g., Silicon Valley versus the Pentagon) value de-escalation to a greater extent than the population in the current studies.

Conclusion

Most choices in life arrive not as disconnected and isolated problems but as the next in a series of choices enmeshed in a social context. When this context causes material and reputational incentives to conflict, our choices provide insight into our values. De-escalation of commitment to failing courses of action maximizes return on investments of time and money, but observers accurately interpret this choice as a signal of untrustworthiness. Whether those who avoid escalation of commitment and other irrational biases do so because they are unaware of the social penalty this will incur or because they are indifferent to social consequences is an important area for future research.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*(1), 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

Arrow, K. J. (1974). *The limits of organization*. W. W. Norton & Company.

Ashraf, N., & Bandiera, O. (2018). Social incentives in organizations. *Annual Review of Economics*, *10*(1), 439–463. https://doi.org/10.1146/annurev-economics-063016-104324

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Brockner, J. (1992). The escalation of commitment to a failing course of action: Toward theoretical progress. *The Academy of Management Review*, *17*(1), 39–61. https://doi.org/10.2307/258647

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology*, *87*(4), 611–628. https://doi.org/10.1037/0021-9010.87.4.611

Dorison, C. A., DeWees, B., Rahwan, Z., Robichaud, C., & Lerner, J. S. (2020). *Inefficient (but seemingly fair) resource allocations are used to signal trustworthiness* [Working Paper].

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417. https://doi.org/10.1037/h0053870

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772–787. https://doi.org/10.1037/xge0000165

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Ferrer, M., Fischlmayr, I. C., … Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, *25*(4), 191–210. https://doi.org/10.1257/jep.25.4.191

Goffman, E. (1983). The interaction order: American sociological association, 1982 presidential address. *American Sociological Review*, *48*(1), 1–17. JSTOR. https://doi.org/10.2307/2095141

Grossmann, I., Eibach, R. P., Koyama, J., & Sahi, Q. B. (2020). Folk standards of sound judgment: Rationality versus reasonableness. *Science Advances*, *6*(2), eaaz0289. https://doi.org/10.1126/sciadv.aaz0289

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. https://doi.org/10.1073/pnas.1601280113

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569–598. https://doi.org/10.1146/annurev.psych.50.1.569

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychology Bulletin*, *125*(2), 255–275.

Mankiw, N. G. (2020). *Principles of economics* (9th ed.). Cengage Learning.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mead, G. H. (1930). Cooley's contribution to American social thought. *American Journal of Sociology*, *35*(5), 693–706. https://doi.org/10.1086/215189

Olivola, C. Y. (2018). The interpersonal sunk-cost effect. *Psychological Science*, *29*(7), 1072–1083. https://doi.org/10.1177/0956797617752641

Shafir, E., & LeBoef, R., A. (2002). Rationality. *Annual Review of Psychology*, *53*(1), 491–517.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*.

Sleesman, D., Conlon, D. E., McNamara, G., & Miles, J. (2012). Cleaning up the big muddy: A meta-analytic review of the determinants of escalation of commitment. *The Academy of Management Journal*, *55*(3), 541–562. https://doi.org/10.5465/amj.2010.0696

Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance*, *16*(1), 27–44. https://doi.org/10.1016/0030-5073(76)90005-2

Staw, B. M. (1981). The Escalation of Commitment to a Course of Action. *The Academy of Management Review*, *6*(4), 577–587. https://doi.org/10.2307/257636

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, *116*(3), 396–415. https://doi.org/10.1037/pspi0000150

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the politics of the beholder? *Administrative Science Quarterly*, *45*(2), 293–326. https://doi.org/10.2307/2667073

Appendix 1 – *Essay 1*

Supplemental Study 1: Forecasting Temperature

Method

In this supplementary study, participants made four predictions of temperature in major U.S. cities one week into the future and stated their confidence in each estimate. We tested whether confidence decreased for later questions.

Participants. We collected data from 303 participants from Amazon Mechanical Turk (MTurk). After removing data from two participants who looked up information online contrary to instructions and four who failed an attention check, our final sample consisted of 297 participants (56% Female, Mean age = 35.7).[7] We paid participants $0.50 for their time and awarded a $0.50 bonus for highly accurate forecasts.[8]

Procedure. Following an attention check and a question about their level of expertise about weather, participants forecasted temperatures in four large U.S. cities (New York, NY; San Francisco, CA; Denver, CO; and Omaha, NE) at noon seven days after study launch. The order of the cities was counterbalanced. After making each estimate, participants reported how confident they were that their estimate fell within an interval of three degrees Fahrenheit higher or lower than the correct answer. They reported their confidence on a 5-point scale, anchored at "Not at all" and "Very." All estimates and confidence ratings were collected on a single web page. Participants then reported demographic information (gender, age, and education level were

[7] Results are not substantially altered by the inclusion of four participants who responded incorrectly to the attention check.

[8] We averaged the correct answers for each item and awarded a bonus if the average estimate made by a participant fell within a 3 degrees Fahrenheit of the average of the correct answers.

collected in all studies), whether they have lived in any of the cities in question, their baseline level of knowledge about weather, and whether they had looked up any information online.

Analytical Approach. We determined our sample size based on pilot data, which showed our effect with 100 participants. To be extremely conservative, we tripled that sample size. The following methods apply to all subsequent studies unless otherwise noted.

Because each of our participants provided observations for four different cities, we employed a mixed effects linear model computed using the lmer function in the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). Our models cluster the standard errors at the level of participant and city by including random intercepts for stimulus (city in this case), and random slopes and intercepts for confidence across participant.

Furthermore, because we are primarily interested in how an individual's confidence responds to repeated questioning and not individual differences in the use of the confidence scales, we z-scored confidence ratings within participant. Rescaling preserves the trend but eliminates differences in the degree of sensitivity, as well as allows comparisons between confidence ratings elicited on different scales in later studies.

Error was quantified as the absolute percentage deviation from the true value for each item. Accuracy is a binary variable set to 1 if the estimate is within the specified criteria (3 degrees in the present study, 10% for most later studies), and 0 otherwise.

Results

Supplemental Study 1 tested whether question order systematically affects subjective confidence in forecasts of temperature. On average, participants were not particularly confident in their temperature forecasts. When we consider all 1,188 forecasts made by participants the

average confidence on the 5-point scale was M = 2.30, SD = 0.94. The task proved quite difficult with only 14% of forecasts being within three degrees of the correct answer.

Most importantly, confidence declined over the course of the task (b = -0.049, CI_{95} [-0.076, -0.021], t = -3.48, p < .001). Examining confidence z-scored within participant, we obtain a similar result (β = -0.055, CI_{95} [-0.093, -0.016], t= -2.80, p = .005). The mean confidence for the first questions was 2.34 which declined to 2.20 for the fourth question. 24% of participants reported a lower confidence on question four than on question one while 14% reported a higher confidence. The remaining 62% reported the same degree of confidence.

We consider the possibility that the decline in confidence over repeated estimates is due to participant fatigue and reduced effort. We reanalyze the trend in confidence controlling for item-level error in order to establish whether the change in error from item to item statistically mediates the decline in confidence. Controlling for error, z-scored confidence declines even more steeply than in the simple model (β = -0.056, CI_{95} [-0.094, -0.018], t= -2.86, p = .004). Furthermore, within participant Goodman-Kruskal gamma correlation (Goodman & Kruskal, 1954) between confidence and error across the four cities varied from 1.0 to -1.0 but averaged just -.02. This low resolution (Koriat, 2012; Lichtenstein, Fischhoff, & Phillips, 1982) is inconsistent with an explanation featuring a reduction in effort/estimate quality.

Supplemental Study 2: Feedback

In Supplemental Study 2 we compare our baseline condition with one that provides immediate and truthful accuracy feedback.

Method

Participants. We collected 597 completed survey responses from MTurk.[9] After removing seven participants who looked up information online and 32 participants who incorrectly answered an attention check, our final sample consisted of 558 participants (54% Female, Mean age = 34.9). Incentive and bonus procedures were the same as in Study 1.

Procedure. We randomly assigned participants to one of two conditions. In the *Control* condition, participants completed the same animal weight task from Study 1. In the *Feedback* condition, the task was the same as in the *Control* condition, but immediately following each estimate and confidence submission, we provided participants with a statement reminding them of the estimate they had just made as well as the correct answer. For example: "You answered 435 lbs. The correct answer is 330.7 lbs." All confidence ratings were elicited on the five-point Likert scale described in earlier studies. At the end, participants reported demographic information and were given an opportunity to report cheating.

Analytical Approach. A sample size of 300 participants per cell was predetermined in order to provide sufficient statistical power for detecting a difference between conditions. The statistical approach remained the same as in Study 1.

Results

When we consider all 5,022 estimates made by participants, the average confidence on a 5-point scale was M = 2.50, CI_{95} [2.47, 2.52]. Approximately 9% of estimates achieved the required accuracy (within 10%), and this rate did not differ between conditions, z(588) = 0.08,

[9] Three participants reported incorrect validation keys preventing the collection of the full 600 participants prior to closing collection.

p = .937. Without feedback, we again observed the main effect of decreasing z-scored confidence with subsequent questions (β = -0.085, CI_{95} [-0.103, -0.0067], t = -9.32, p < .001).[10] The effect was also present in the *Feedback* condition (β = -0.123, CI_{95} [-0.143, -0.103], t = -12.23, p < .001).[11] Indeed, confidence actually declined more rapidly with feedback than without it (Figure A1.1). This interaction between question order and feedback condition was significant (β = 0.038, CI_{95} [0.011, 0.065], t = 2.80, p = .005) and robust to controlling for error and or the interaction between error and condition.[12]

When we examine error over the course of the nine items that participants estimated,[13] we find that in the *Feedback* condition, error decreased by 3% with each question (b = -3.0, CI_{95} [-5.0, -1.0], t = 2.98, p = .003), possibly demonstrating some learning. We did not observe any systematic change in error over the course of the task in the *Control* condition. Analysis of error allowing interactions between condition and question order, confirmed the interaction between question order and condition was significant (b = 3.4, CI_{95} [0.7, 6.2], t = 2.44, p = .015). As in prior studies, the correlation between confidence and error overall remained near zero (mean Goodman-Kruskal gamma correlation = .02).

In sum, this study suggests that feedback expedited the confidence decline. Ironically, the slight reduction in error suggests that the answers to prior questions are being used to improve later answers.

[10] Unstandardized result (b = -0.04, SE_b= 0.005, t = 8.5, p < .001).

[11] Unstandardized result (b = -0.07, SE_b= 0.006, t = 10.7, p < .001).

[12] Unstandardized result (b = 0.02, SE_b= .008, t = 2.7, p < .01).

[13] We excluded 23 estimates with an error of more than 1,000% (10x) as likely data entry errors or unreasonably poor guesses leaving 5,281 observations.

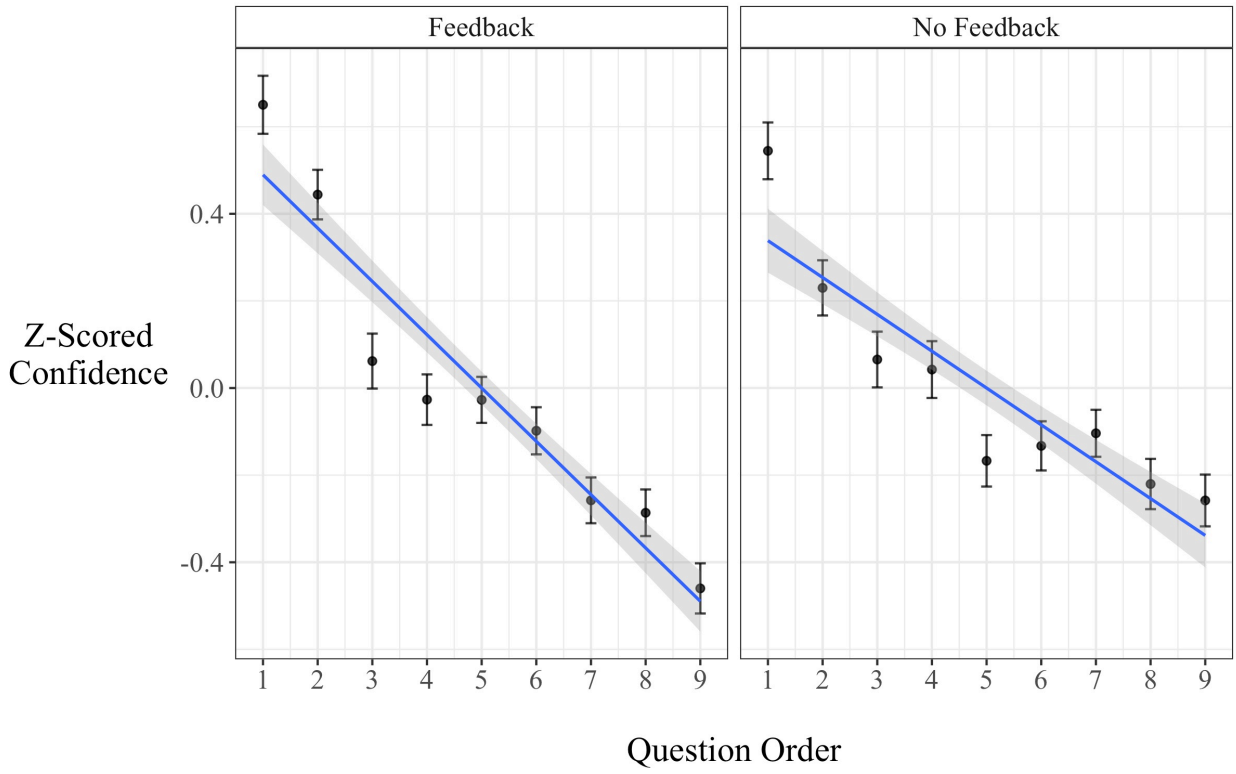

Figure A1.1. *Confidence declined more rapidly when feedback was provided on a difficult task. Error bars indicate standard errors.*

Appendix 2 – *Essay 2*

Supplemental Methods

Participants received $0.50 for their study participation – an amount typical for the duration of this work on the Amazon Mechanical Turk (MTurk) platform.

We adapted the confidence scale from a first-person self-report to a third person assessment by replacing relevant pronouns and changing "feel" to "seems."

Study 1: Supplementary Results

Gender. We found little evidence that gender interacted with sunk cost decision making to influence social perceptions. There was a larger spread between the ratings of competence for investors and non-investors in the absence of sunk cost for women than for men, which resulted in a marginally significant 3-way interaction between prior investment, future investment, and decision maker gender (b = 0.53, CI_{95} [-0.06, 1.13], β = 0.24, CI_{95} [-0.03, 0.50], t = 1.76, p = .079). Female decision makers were perceived as more confident (M = 5.56, CI_{95} [5.40 5.73]) than men (M = 5.25, CI_{95} [5.07, 5.44]), d = 0.24, CI_{95} [0.05, 0.45], t(384.55) = 2.44, p = .015. Female decision makers were also perceived as marginally warmer (M = 3.59, CI_{95} [3.49, 3.69]) than men (M = 3.47, CI_{95} [3.38, 3.55]), d = .19, CI_{95} [-0.01, 0.39], t(384.04) = 1.87, p = .062.

Simulations and Robustness Checks for Study 4

The result of our pre-registered analysis using a beta regression with log link function is consistent with the OLS regression reported in the article text. Specifically, we find that the effect of investing on financial rewards significantly depended on the presence or absence of sunk costs (b = -0.41, CI_{95} [-.63, -.19], z = -3.63, p<.001).

We use a bootstrap simulation to illustrate that the parametric tests used in the paper retain the desired ability to discriminate random from causal relationships. By sampling

randomly sampling from investment condition and pairing with actual dictator game behavior observed, we eliminate any possible causal relationship between investment condition and sending behavior. If the test remains valid, we should expect to see a significant result ($p<.05$) approximately 5% of the time due to random sampling variation. This is exactly what we observe when we run the above procedure 10,000 times using the data from the sunk cost condition. 4.7% of the random samples yield a significant t test. The distribution of p values is nearly uniform which further supports the validity of the test despite the non-normal distribution of the dependent variable (Figure A2.1).

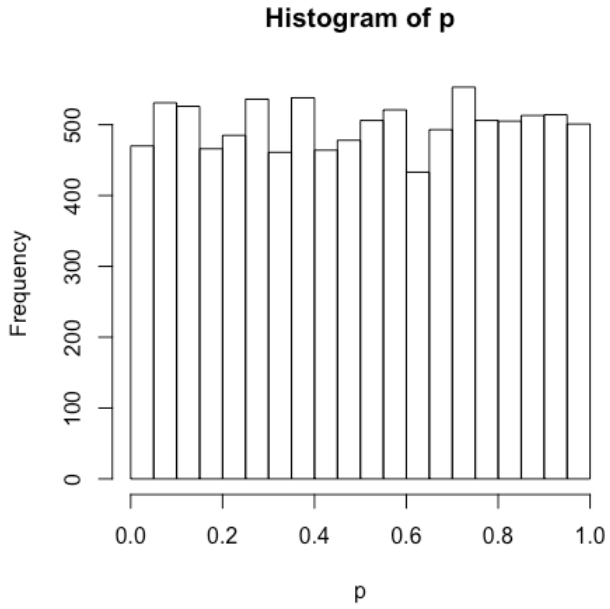

Figure A2.1. *A histogram of the p values resulting from 10,000 simulations of a null effect illustrates that while our distribution is non-normal, a parametric t test retains a 5% false positive rate.*

Appendix 3 – *Essay 3*

General Methods

All studies received Institutional Review Board approval. For practical reasons, the data were collected for Observers and Actors in separate surveys (Study 1 and 2 respectively) requiring the use of simulation to determine outcomes for compensation.

Study 1: Supplemental Methods

Participants. Participants received $0.50 for their study participation – an amount typical for the duration of this work on the Amazon Mechanical Turk (MTurk) platform. We advertised the study as a "decision making survey." The starting Trust Game endowment was $0.30. Attrition was below 3% and did not vary by condition.

Exclusions. In keeping with our preregistration and before conducting any inferential analyses, we cleaned the data by excluding: incomplete survey responses, duplicate responses from the same person, and responses that failed to pass comprehension question common to all conditions. We excluded only participants who missed multiple comprehension questions due to the relatively complicated instructions for the trust game and inclusion of a four comprehension checks. Answer rates did not differ by condition and all 58 respondents who missed two or more comprehension checks were excluded from our analysis. Our results are robust to various exclusion criteria. See Table A3.1 for results of alternative analyses at various levels of exclusion.

Radar-Blank Plane Scenario. "*As the president of an airline company, the Actor decided to allocate $10 million of the company's research budget to a project. The purpose was to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. When the project is 90% completed ($9 million already spent), another firm*

begins marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane the Actor's company is building. The question the Actor now faces is: should (s)he invest the last 10% of the research funds to finish the radar-blank plane?"

Measures. Trust at the behavioral level is not the same as it is at the cognitive level (Dunning et al., 2014). While factors other than cognitive trust influence trust behavior, differences in behavior are most logically attributable to changes in the perceived trustworthiness of the partner and use "trust" to refer to both cognitive and behavioral trust. Anonymity and one-round play eliminate confounds which might explain investment absent trust such as retaliation.

Exploratory Variables and Demographics. Participants reported what they would have chosen to do if they were in the sunk cost situation described above. We also collected their estimates of the probability that investing would be financially successful (the profit from sales of the plane would exceed the cost of finishing development). Participants reported demographic information (age, gender, education), and whether they had previously been taught the principle of sunk cost. At the end of the study, we provided feedback for the attention checks and debriefed participants regarding the use of deception. Eligible participants received their bonus results which were computed by assuming a 50% return from the simulated Actor.

Study 1: Supplemental Results

Inferential Analyses. We coded the amount of money sent as the proportion of available funds transferred [0,1]. Based on pilot data, we had concerns about violating the normality assumptions for parametric tests and OLS regression. Many participants chose to send either all or none of their endowment to their partner resulting in a "U" shaped distribution for the amount sent and the portion returned. We pre-registered and conducted an analysis approximating the

data using a beta distribution rather than a normal distribution in an effort to better fit the data as well as non-parametric tests where appropriate (see Table A3.1). Simulations based on our data, however, confirm that a standard t-test remains a valid test of our hypothesis (see Validation of Methodology below). Because the results are consistent regardless of the analysis used, we report the more results of the more familiar tests in the main text.

The result of our pre-registered analysis using a beta regression with log link function is consistent with the OLS regression reported in the main text. Specifically, beta regression predicts that Observers would entrust 18% more money to decision makers who persisted (58.6% versus 49.5%), $z = 3.27$, $p = .001$. The interaction between participant preference and Actor choice is also robust to regression methodology and remains significant using beta regression, interaction $b = 0.30$, $z = 2.52$, $p = .012$.

Validation of Methodology. We use a bootstrap simulation to illustrate that the parametric tests used in the paper retain the desired ability to discriminate random from causal relationships. By sampling randomly sampling from escalation condition and pairing with actual trust game behavior observed, we eliminate any possible causal relationship between escalation condition and sending behavior. If the test remains valid, we should expect to see a significant result ($p < .05$) approximately 5% of the time due to random sampling variation. This is exactly what we observe when we run the above procedure 10,000 times. 4.8% of the random samples yield a significant t test (Figure A3.1). The distribution of p values is nearly uniform which further supports the validity of the test despite the non-normal distribution of the dependent variable.

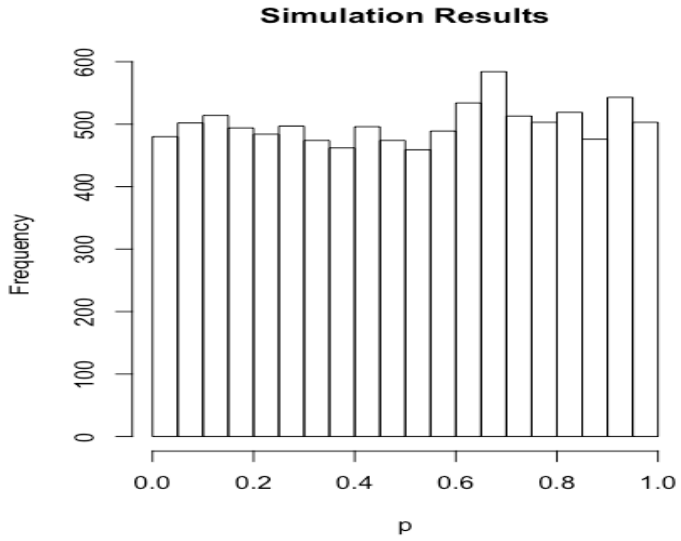

Figure A3.1. *A histogram of the p values resulting from 10,000 simulations of a null effect illustrates that while our distribution is non-normal, a parametric t test retains a 5% false positive rate.*

Table A3.1. *Results of Study 1 With Various Exclusion Criteria*

	Least restrictive ---------- most restrictive				
Exclusion criteria	No Exclusions	Missed All 4	Missed 3+	Missed 2+ (reported)	Missed Any
N =	660	658	649	602	416
Mean portion sent; non-investors	.47	.47	.47	.47	.47
Mean portion sent; investors	.61	.61	.61	.60	.61
t-test	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
Wilcoxon rank test	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
Cohen's d	0.34 [0.19, 0.50]	0.34 [0.19, 0.50]	0.34 [0.19, 0.50]	0.34 [0.18, 0.50]	0.35 [0.15, 0.54]

Study 2: Supplemental Methods

Participants. Participants received $0.30 for their study participation – an amount typical for the duration of this work on the MTurk platform. We advertised the study as a "decision making survey." We pre-determined our target sample size to ensure 80% power to detect an OR

of 1.54 or a Cohen's d of 0.2. However, due to greater than anticipated exclusion (see below), our sample for analysis fell far short of this power. We conducted a second round of collection targeting an additional 2,000 respondents to meet or exceed this power post pre-registered exclusions.

Exclusions. In keeping with our preregistration and before conducting any inferential analyses, we cleaned the data by excluding: incomplete survey responses, responses flagged by Qualtrics as possible spam responses due to suspicious locations or repeated IP addresses, duplicate responses from the same person. We dropped one of the comprehension questions for Stage 1 leaving a total of three comprehension questions. We therefore preregistered a more stringent criteria than in Study 1, excluding participants for any incorrect response to a comprehension question. Only 465 out of 803 participants met what turned out to be a fairly strict exclusion criterion. Only 71% of respondents correctly answered "Imagine that the Observer is deciding how much to send to you. Which decision will result in the highest overall bonus for the Observer?" Correct answer: "It depends on how much you decide to return to the Observer." Only 4% of these remaining participants, however, missed the manipulation check (compared to 9% in the full sample) indicating that the quality of data in this remaining sample is quite good. Our results from this initial round of collection are robust to various exclusion criteria (see Table A3.2). We note that despite the substantially smaller sample size, the most stringent exclusion results in the lowest *p* values and largest estimated effect size due to reduced noise in the sample.

Exploratory Variables and Demographics. A manipulation check followed the Trust Game, asking whether their choice in the investment scenario was observable to the Observer.

We collected age, gender, and education data as well as familiarity with the principle of sunk cost. Actors who reported familiarity were asked for a brief definition of sunk cost.

In the second round of responses, we added two additional exploratory measures. We asked participants to rate the trait descriptors "Reasonable" and "Rational" on a scale from 1 (not at all like me) to 5 (just like me).

At the end of the study, we provided feedback for the attention checks and debriefed participants regarding the use of deception. Eligible participants received their bonus results. We randomly sampled an amount sent from the distribution of Observers in the relevant investment condition from Study 1 to simulate the effects of the participants choice as much as possible. In the unobserved condition, we sampled from the entire distribution of sending behavior collapsed across condition.

Study 2: Supplemental Results

Inferential Analyses. Given that trustworthiness, similar to trust, was not normally distributed, we conducted two sets of robustness checks. First, we again conduct a bootstrapping simulation to confirm a 5% false positive rate. Second, we repeated the parametric analysis reported in the paper using a Wilcoxon-Mann-Whitney test and a beta regression model predicting the trustworthiness behavior as a function of escalation decision.

The result of our pre-registered analysis/exclusions using a beta regression with log link function is consistent with the parametric tests reported in the main text. Specifically, using the combined sample, beta regression predicts that Actors who escalate would return 20% more money to Observers than Actors who de-escalate (31.6% versus 26.4%), $z = 3.72$, $p < .001$.

Appendix 3 Tables 2-4 report the results of analysis on the first round, second round and combined samples at varying levels of exclusion.

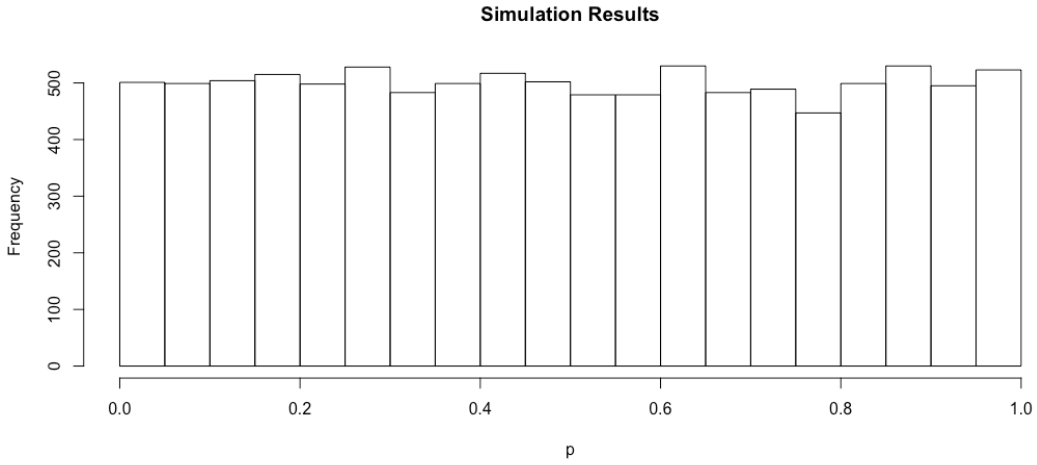

Figure A3.2. *We repeat the same procedure used in Study 1 to confirm that despite the non-normal distribution of responses, a t-test remains a valid test of the significance of any effect that might be present. 10,000 simulations in which conditions have been randomly re-assigned to destroy any causal relationship that might be present results in an essentially uniform distribution of p-values and a false positive rate of 5%.*

Table A3.2. *Results of Study 2 – Round 1 With Various Exclusion Criteria*

	Least restrictive ---------- most restrictive				
Exclusion criteria	No Exclusions	Missed All 3	Missed 2+	Missed Any	Missed Manipulation Check
N =	803	778	703	465	727
p(invest); unobserved	.7157	.7121	.7057	.7016	.7006
p(invest); observed	.7040	.7044	.7134	.6960	.7105
Logistic regression	OR = 0.94 p = .714	OR = 0.96 p = .813	OR = 1.04 p = .811	OR = 0.97 p = .894	OR = 1.05 p = .770
Mean portion returned; De-escalate	.354	.344	.309	.267	.333
Mean portion returned; Escalate	.413	.401	.378	.351	.392
t-test	p = .006	p = .007	p = .001	p < .001	p = .005
Wilcoxon rank test	p = .011	p = .011	p = .002	p < .001	p = .012
Cohen's d	-0.22 [-0.37, -0.06]	-0.22 [-0.37, -0.06]	-0.28 [-0.44, -0.11]	-0.36 [-0.56, -0.16]	-0.23 [-0.39, -0.07]

Table A3.3. *Results of Study 2 – Round 2 With Various Exclusion Criteria*

	Least restrictive ---------- most restrictive				
Exclusion criteria	No Exclusions	Missed All 3	Missed 2+	Missed Any	Missed Manipulation Check
N =	1985	1940	1741	1125	1774
p(invest); unobserved	.7202	.7223	.7188	.7411	.7282
p(invest); observed	.7658	.7662	.7583	.7558	.7630
Logistic regression	OR = 1.27 p = .020	OR = 1.26 p = .027	OR = 1.23 p = .061	OR = 1.08 p = .57	OR = 1.20 p = .093
Mean portion returned; De-escalate	.371	.363	.345	.331	.347
Mean portion returned; Escalate	.412	.404	.381	.356	.395
t-test	p = .003	p = .003	p = .009	p = .133	p < .001
Wilcoxon rank test	p = .017	p = .014	p = .037	p = .229	p = .004
Cohen's d	-0.16 [-0.26, -0.05]	-0.16 [-0.26, -0.06]	-0.15 [-0.25, -0.04]	-0.11 [-0.24, 0.03]	-0.19 [-0.30, -0.08]

Table A3.4. *Results of Study 2 – Combined Samples With Various Exclusion Criteria*

	Least restrictive ---------- most restrictive				
Exclusion criteria	No Exclusions	Missed All 3	Missed 2+	Missed Any (reported)	Missed Manipulation Check
N =	2787	2717	2443	1589	2500
p(invest); unobserved	.7189	.7194	.7150	.7293	.7201
p(invest); observed	.7486	.7491	.7461	.7396	.7485
Logistic regression	OR = 1.16 [0.98, 1.38] p = .077	OR = 1.16 [0.98, 1.38] p = .080	OR = 1.17 [0.98, 1.40] p = .083	OR = 1.05 [0.84, 1.32] p = .644	OR = 1.16 [0.97, 1.38] p = .109
Mean portion returned; De-escalate	.365	.357	.333	.310	.342
Mean portion returned; Escalate	.412	.403	.380	.355	.394
t-test	$p < .001$	$p < .001$	$p < .001$	$p = .001$	$p < .001$
Wilcoxon rank test	$p < .001$	$p < .001$	$p < .001$	$p = .002$	$p < .001$
Cohen's d	-0.18 [-0.26, -0.09]	-0.18 [-0.26, -0.09]	-0.19 [-0.28, -0.10]	-0.19 [-0.30, -0.08]	-0.20 [-0.29, -0.12]

Simulated Outcomes

By combining data from Study 1 and Study 2 we are able to simulate outcomes for both the Actor and Observer side of the trust game using escalation as a signal of trustworthiness. The combination of signaling effects for Observers and effects of type for Actors had an interesting result.

Observers. Observers may have been better off on average keeping their original 30 cent endowment when paired with a partner who de-escalates. Observers averaged a slight gain when paired with escalating Actors (p = .023).

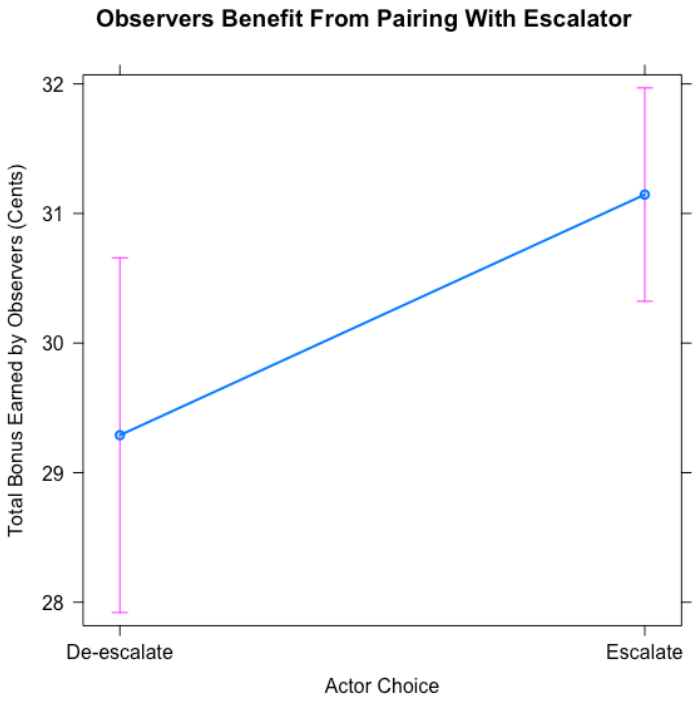

Figure A3.3. *Observers Benefit from Pairing with Escalating Actors*

Actors. From the Actors' perspective, we find a significant interaction between observation condition and escalation decision on the amount Actor's earn ($p < .001$). De-escalators earn the most when unobserved but the least when observed (because senders know not to trust these types if they can identify them, depriving them of the opportunity to take advantage of trusting behavior as they do in the unobserved condition). Escalators' benefit from signaling in the observed condition.

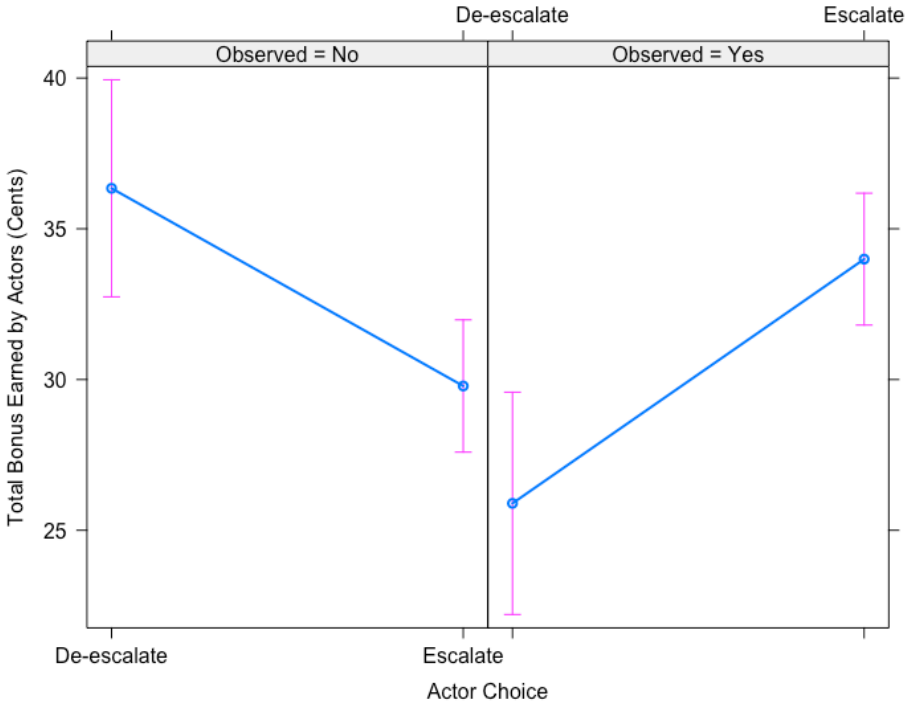

Figure A3.4. *De-escalators Benefit from Privacy*