American Economic Journal: Microeconomics 2023, 15(2): 271–308
https://doi.org/10.1257/mic.20210031

Reference Dependence and Attribution Bias: Evidence from Real-Effort Experiments[†]

By Benjamin Bushong and Tristan Gagnon-Bartsch*

We document a form of attribution bias wherein people wrongly ascribe sensations of positive or negative surprise to the underlying disutility of a real-effort task. Participants in our experiments learned from experience about two unfamiliar tasks, one more onerous than the other. We manipulated expectations about which task they would face: some participants were assigned their task by chance, while others knew their assignment in advance. Hours later, we elicited willingness to work again on that same task. Participants assigned the less (more) onerous task by chance were more (less) willing to work than those who knew their assignment in advance. (*JEL* C91, D84, D91, M54)

Evidence from the lab and field suggests that our experiences are reference dependent: how we feel about an outcome often depends on both its intrinsic value and how that value compares to expectations (e.g., Kahneman and Tversky 1979; Medvec, Madey, and Gilovich 1995; Card and Dahl 2011; Abeler et al. 2011). But do we properly account for how our past impressions were shaped by our prior expectations? For instance, after a surprisingly good meal at an unassuming restaurant, a diner may not appreciate that his pleasant experience stemmed from both the food and the surprise itself. In neglecting this latter component, he may come to believe the food was better than it really was. This intuitive mistake resembles "attribution bias," wherein state-dependent features of utility are wrongly attributed

* Bushong: Department of Economics, Michigan State University (email: bbushong@msu.edu); Gagnon-Bartsch: Department of Economics, Florida State University (tgagnonbartsch@fsu.edu). Leeat Yariv was coeditor for this article. We thank Ned Augenblick, Katherine Coffman, Stefano DellaVigna, Benjamin Enke, Christine Exley, David Laibson, Muriel Niederle, Devin Pope, Matthew Rabin, Joshua Schwartzstein, Andrei Shleifer, Charles Sprenger, the referees, and seminar audiences at the ASSA annual meeting (2020), Boston College, Boston University School of Management, California Institute of Technology, Cornell University, Harvard University, Harvard Business School, University of Michigan, Michigan State University, Norwegian School of Economics (NHH), the North American ESA Conference, Purdue University, Stanford University, Stanford Institute for Theoretical Economics, University of Tennessee, and Rady School of Management at the University of California, San Diego for comments. We thank Jeffrey Yang for excellent research assistance. We thank Alexander Millner for sharing the annoying audio stimulus used in these experiments. The experiments were approved by the Harvard University-Area IRB (protocol numbers IRB15-0365 and IRB16-0944); the replication study was determined exempt by the Michigan State University HRPP. We gratefully acknowledge financial support from the Eric M. Mindich Research Fund for the Foundations of Human Behavior.

[†] Go to https://doi.org/10.1257/mic.20210031 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

to a stable quality of a person or good.[1] In this paper we conduct two real-effort experiments to determine whether such a bias operates over expectations-based reference dependence. Behavior in our experiments suggests that people misattribute sensations of elation or disappointment to the intrinsic (dis)utility of working on a real-effort task, consistent with attribution bias over reference-dependent utility.

To provide an example of how this form of attribution bias can arise in an environment similar to our experiment, consider a worker completing a series of short-term jobs. Each day, the worker is randomly assigned to one of two tasks—one more desirable than the other. The job she faces each day therefore comes with an element of elation or disappointment. When a "misattributing" worker is randomly assigned to the less desirable task, she misattributes the sensation of disappointment to the intrinsic disutility of that task. In doing so, she becomes too reluctant to work in that role in the future. By contrast, when she is assigned to the desirable task, she wrongly attributes the positive feelings of surprise to the intrinsic enjoyment of the task and becomes too enthusiastic about the role. In both cases the worker forms biased impressions of the task because she neglects the degree to which her experienced utility was shaped by her expectations.

Following the example above, we conducted experiments in which participants learned from experience about one of two real-effort tasks. In Experiment 1 we endowed participants with differing chances of facing either of these previously unexperienced tasks, one clearly more onerous than the other. Immediately after resolving this uncertainty, participants worked on their assigned task. Several hours later, we elicited their willingness to continue working on that task. Comparing those who were assigned to the same task, we find that the ex ante chance of facing each of the two tasks significantly altered participants' subsequent willingness to work, even though this initial uncertainty was long since resolved. In Experiment 2 we manipulated initial expectations within subjects to examine how a participant's willingness to work changed over one week as their expectations changed. As with Experiment 1, we find that a participant's willingness to work was shaped by the elation or disappointment they experienced while forming their initial impressions, suggesting a specific, previously unexplored form of attribution bias. As we discuss below, this expectations-based attribution bias leads to judgments of outcomes that are excessively swayed by deviations from expectations, which has important implications for how firms, policy makers, or employers set or manage those expectations.

We first present a simple theoretical model in Section I (following Gagnon-Bartsch and Bushong 2021) that guides our experimental designs. We then describe Experiment 1 (conducted online) in Section II. Subjects ($N = 866$) listened to audio recordings of book reviews and had to determine whether each review was endorsing or criticizing the book. This simple yet tedious classification task came in two variants. One variant—which we call *noise*—included an annoying sound

[1] Alternative forms of attribution bias are well established in psychology. For example, Dutton and Aron (1974) show that opinions of a newly met person depend on unrelated situational factors—e.g., current state of excitement or fear. Meston and Frohlich (2003) replicate and extend this seminal result to broader settings. More recent evidence in economics (Simonsohn 2007, 2010; Haggag et al. 2019) demonstrates that when assessing the value of a good or service, people incorrectly attribute state-dependent sensations caused, for instance, by weather or thirst to the underlying quality of the good. We further discuss this evidence in the related-literature section.

layered on top of the audio review. The second variant—which we call *no-noise*—had no additional sound added to the audio review.

We endowed participants with different chances of facing either task. In one treatment, participants were assigned to a task from the onset of the experimental instructions (i.e., they faced no uncertainty). In another, participants flipped a coin to determine which task they would face (i.e., they faced a 50 percent chance of either task). In a final treatment, participants were assigned to a task with near certainty (i.e., they faced a 99 percent chance of one task and a 1 percent chance of the other). Put together, this design generates six groups, which result from crossing the three manipulations in expectations described above with the ultimate task a participant faced: $\{\textit{control}, \textit{coin-flip}, \textit{high-probability}\} \times \{\textit{noise}, \textit{no noise}\}$. After reading the instructions (and resolving any uncertainty about task assignment), each participant completed eight rounds of their assigned task. Knowing that they would later be asked about their willingness to continue working on this task, these initial trials gave participants an opportunity to learn their preferences. In a second session, which participants could access only after eight hours elapsed, we elicited their willingness to continue working on their assigned task for additional pay.

We examine how willingness to work (WTW) differed between participants across the three treatments. Our misattribution model predicts that participants who were assigned the noiseless task via the coin flip would form the most optimistic beliefs about that task, since their initial impressions came with the greatest sense of positive surprise. That is, participants in the *coin-flip* + *no noise* group would exhibit higher WTW than those in the *control* + *no noise* and *high-probability* + *no noise* groups, even though all of these people ultimately faced the same task. By contrast, our model predicts that those assigned the noisy task via the coin flip would exhibit lower WTW than those in the *control* + *noise* and the *high-probability* + *noise* groups.

Indeed, we find these effects, as previewed in Figure 1. For example, when the stakes were highest, participants who were assigned the noiseless task via the coin flip were 20 percent more willing to work than those who faced that task with certainty, while those who were assigned the noisy task via the coin flip were 25 percent less willing to work than those who faced that task with certainty. To help clarify the mechanism underlying these results, we address several alternative explanations. We first discuss how classical models and reference-dependent models without misattribution struggle to predict these results. We then highlight how our data and design suggest that short-term mood effects do not drive our results.[2] Perhaps most importantly, we illustrate how our high-probability treatment helps rule out informational explanations stemming from participants in the control and coin-flip groups drawing different inferences from the experimental design itself.[3]

[2] The time gap between participants forming their impressions and our elicitation of WTW helps distinguish misattribution from short-term mood effects (as in, e.g., Saunders 1993; Hirshleifer and Shumway 2003; Edmans, Garcia, and Norli 2007).

[3] The coin-flip and high-probability treatments utilized identical instructions aside from the probability of task assignment; thus, comparing WTW across these groups cleanly reveals the effect of changing this probability. In contrast, participants in the control treatment only knew about the single task they faced.

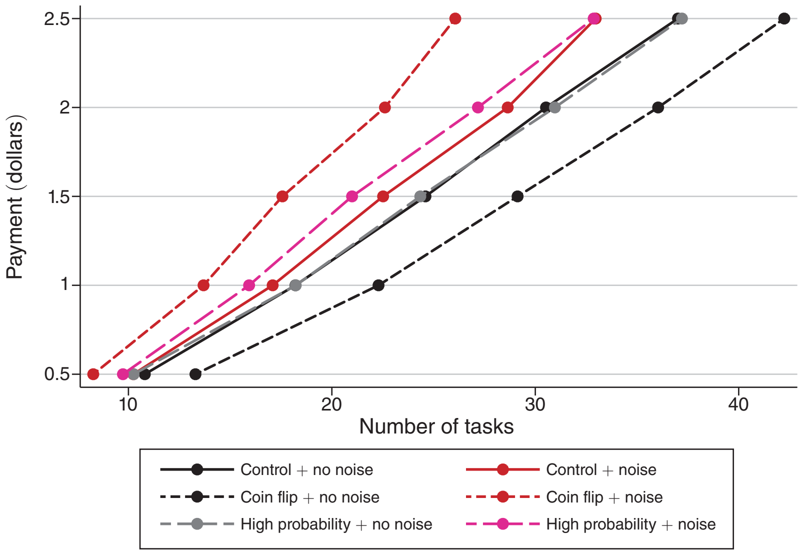


Figure 1. Labor Supply Curves across Treatments

Notes: Each point represents the average WTW for a fixed payment as elicited using the BDM mechanism. Those assigned to their task after facing the most uncertainty (coin-flip groups) demonstrated greater WTW when assigned the noiseless task and less WTW when assigned the noisy task than the control and high-probability groups.

Experiment 2, presented in Section III, adopts a within-subject design $(N = 87)$ in a laboratory setting. We elicited each participant's WTW in two different sessions, separated by one week. In the first session, each participant flipped a coin to determine whether they faced a noiseless or noisy task and then completed five trials of that task. Directly after this learning phase, we elicited the participant's WTW. One week later, the same participants returned, but there was no coin flip: each knew ahead of time that they would again face the same task as before. In that second session, each participant completed five trials of their previously assigned task and then stated their WTW.

Misattribution in this setting predicts a systematic change in WTW across these two sessions as a result of the participant's changing expectations; thus, we examine the difference in a participant's WTW between Session 1—when their task came as a surprise—and Session 2, when that same task was completely expected. We find that participants who faced the noiseless task in the first session were less willing to work in the second week than in the first, while those who faced the noisy task in the first session were more willing to work in the second week than the first. Furthermore, the evidence from Experiment 2 suggests a form of "sequential contrast effect" that is predicted by misattribution but not predicted by alternative explanations for our main findings (such as short-term mood effects or reciprocity toward the experimenter).[4]

[4] Some concerns that might apply to Experiment 1—e.g., effects driven by differences in information or reference points that are slow to adapt—do not apply to Experiment 2, and vice versa. Additionally, while Experiment 1

Jointly, our experimental findings offer a compelling case in favor of a novel form of attribution bias that operates over expectations. In addition to extending the literature on attribution bias, our conceptual framework extends the literature on reference dependence to errors in beliefs. Absent misattribution, reference dependence captures the notion that potential elation or disappointment looms large in preferences. But we provide a mechanism for why past sensations of surprise continue to loom large in both memory and beliefs. We further contribute to the literature on reference dependence by providing evidence in support of expectations-based reference dependence—most clearly documented in our basic treatment effects in Experiment 1.

Our notion of expectations-based attribution bias illuminates the importance of initial expectations on judgements and effort provision. For example, misattribution offers a logic for why the surprisingly high payments in Gneezy and List (2006) increased effort in the short term and why these failed to motivate longer-term changes in behavior after workers' reference points adapted. More broadly, our results offer a caution against raising expectations when agents judge their experiences against these lofty beliefs. For example, while hyping a product may help early sales, such marketing efforts can hurt if early adopters then underestimate the product's quality as a result of contrasting it against a high reference point. This intuition suggests that managers and firms should strategically restrain expectations—a practice commonly observed in marketing, politics, and finance.[5]

Along these lines, our evidence and theoretical framework provide a new lens for understanding existing empirical results. For example, Backus et al. (2021) show that among new shoppers on eBay who lost their first auction, those who were in the lead longer—and hence developed more optimistic expectations—were more likely to quit using the platform. In the domain of policy reform, Adhvaryu, Nyshadham, and Xu (2020) examine a field experiment in which an NGO improved workers' housing conditions in India. The improvements were modest but fell short of what was originally planned. The authors find that workers who knew the original plans ahead of time perceived their conditions as worse than workers who were neither told about nor provided with any improvements at all. Indeed, our framework highlights that falling short of expectations can lead people to form overly pessimistic beliefs and hence prematurely abandon new technologies or reject recently enacted reforms. In this way, we provide an intuition for why informational campaigns or excessive hype can backfire.

Related Literature.—Attribution bias, often referred to in the psychology literature as the "fundamental attribution error" or "correspondence bias" (e.g., Ross 1977; Gilbert and Malone 1995), is the idea that temporary sensations or situational

provides strong evidence of misattribution, Experiment 2 provides a recipe for identifying heterogeneity in the degree of misattribution across subjects (though our current experiment is underpowered for this endeavor).

[5] Political scientists, for example, have argued that discrepancies between a politician's performance and citizens' expectations play a key role in how citizens perceive that politician (see, e.g., Patterson, Boynton, and Hedlund 1969; Kimball and Patterson 1997). Likewise, marketing has emphasized the role of expectations in perceived quality of service (see, e.g., seminal works from Oliver 1977, 1980; and Boulding et al. 1993). Kopalle and Lehmann (2006) and Ho and Zheng (2004) discuss how firms restrain expectations about product quality and delivery times, respectively.

factors are incorrectly attributed to an underlying, stable characteristic of a person or good. Despite its long history in psychology, there are only a handful of studies in economics that examine attribution bias. Simonsohn (2007) demonstrates that college applicants with particularly strong academic qualities were evaluated higher by admissions officers when the weather on that evaluation day was poor, and Simonsohn (2010) shows that incoming freshman were more likely to matriculate at an academically rigorous school when the weather during their visit was cloudy versus sunny. Relatedly, a series of papers show that luck is wrongly attributed to skill or effort for CEOs (Bertrand and Mullainathan 2001) and politicians (Wolfers 2007; Cole, Healy, and Werker 2012). Recent laboratory experiments have replicated this result (Brownback and Kuhn 2019; Erkal, Gangadharan, and Koh 2019).

Most closely related to our study, Haggag et al. (2019) provide clean evidence that people wrongly attribute state-dependent fluctuations in utility to their valuation of a good. Specifically, Haggag et al. show that participants in an experiment value an unfamiliar beverage more if they first drink it while thirsty rather than sated. In a field study, they show that nice weather during a person's visit to a theme park increases the likelihood that they plan to return. While Haggag et al. provide a general framework of attribution bias over state-dependent utility, we apply a similar logic to a distinct state variable: expectations. This form of misattribution generates unique predictions. For example, Haggag et al. do not speak to the notion of expectations management highlighted above; in contrast, this is an immediate implication of our framework. More technically, their framework (contemporaneous with our own) focuses on state-dependent utility without complementaries through which past experiences influence today's consumption utility. Reference dependence naturally introduces these complementarities, since past experiences form the reference point against which today's consumption is evaluated. As a result, misattribution of reference dependence generates dynamic errors in beliefs (discussed in Gagnon-Bartsch and Bushong 2021; see below). This can manifest as sequential contrast effects, wherein a second outcome seems better the worse the first outcome was. We find suggestive evidence for such a contrast effect in Experiment 2 (see Section IIIC).

Given our focus on expectations-based attribution bias, we also connect to a literature that considers how prior expectations can influence impressions through either assimilation or contrast. This research highlights that when outcomes deviate from expectations, a person might either assimilate that experience, interpreting it in favor of their current beliefs (as in, e.g., Rabin and Schrag 1999; Fryer, Harms, and Jackson 2019), or contrast it, interpreting the experience against their expectations (e.g., Oliver 1977, 1980; Boulding et al. 1993). It remains an open empirical question when each force dominates; we find contrast effects are predominant in our environment.

Unlike attribution bias, reference-dependent preferences have been the subject of many papers in economics. Recent papers have demonstrated that reference dependence affects behavior across a wide range of contexts, including labor supply among taxi drivers (Camerer et al. 1997; Crawford and Meng 2011; Thakral and Tô 2021), domestic violence resulting from unexpected football losses (Card and Dahl 2011), and decisions in game shows and sports (Post et al. 2008; Pope and Schweitzer 2011; Allen et al. 2017; Markle et al. 2018). However, what exactly determines the reference point in a given setting remains contested. To discipline their theory, Kőszegi and Rabin (2006) assume that the reference point corresponds to recent expectations. However, laboratory evidence on this has been mixed. Supporting expectations-based reference points are Ericson and Fuster (2011), Abeler et al. (2011), Gill and Prowse (2012), and Karle, Kirchsteiger, and Peitz (2015); against are Wenner (2015), Heffetz and List (2014), and Heffetz (2018). We find that exogenously imposed expectations shape participants' behavior in our experiment and thus provide further support for expectations-based reference dependence.

Our paper also complements recent work by Imas, Sadoff, and Samek (2017) and de Quidt (2018), which suggests that people may fail to fully anticipate their own future loss aversion. Our model is motivated by the related idea that people may fail to retrospectively account for their reference-dependent preferences when learning.

Finally, our evidence grounds our companion theoretical paper, Gagnon-Bartsch and Bushong (2021), which examines the dynamic implications of the basic framework we present here. There we show that with repeated experiences, misattribution leads a decision maker to rely too heavily on recent outcomes when making decisions.[6] We also show that, over the long run, a pessimistic bias emerges and persists as a direct result of loss aversion. Both the short- and long-run dynamics of beliefs suggest that a misattributor is prone to abandon worthwhile prospects (e.g., new technologies) when learning from experience. While these two papers share a common core, the theoretical piece examines how a misattributor's beliefs evolve over time, while this paper documents the bias that is at that core. Thus, the current paper provides a foundation for Gagnon-Bartsch and Bushong (2021) but cannot speak to some of the implications suggested therein.

I. Theoretical Framework

In this section we present a streamlined version of our model of reference dependence with attribution bias (Gagnon-Bartsch and Bushong 2021), which guided our experimental designs. We apply the model to our specific experimental settings in Sections IIB and IIIB to derive testable predictions.

Reference-Dependent Preferences: Following Kőszegi and Rabin (2006; henceforth KR), we assume that the agent's overall utility has two additively separable components. The first component, consumption utility, corresponds to the material payoff traditionally studied in economics, which we denote by $v \in \mathbb{R}$.[7] The second component, gain-loss utility, derives from comparing v to a reference level of utility. We take this reference point to be the agent's prior expectation of her consumption utility (as in Bell 1985), and we consider a simple piecewise-linear specification of gain-loss utility. Specifically, if the agent believes that consumption

[6] Recency biases have been documented in a range of economic contexts, such as stock market participation (Malmendier and Nagel 2011) and hiring decisions (Highhouse and Gallo 1997); see Fudenberg and Levine (2014) for additional references.

[7] We interpret v as if it derives from a classical Bernoulli utility function $u_C : \mathbb{R}_+ \rightarrow \mathbb{R}$ over consumption realizations $x \in \mathbb{R}_+$ such that $v = u_C(x)$, but we work directly with consumption utility v to reduce notational clutter.

utility is distributed according to CDF $\widehat{F}_V$ with a mean value $\widehat{E}[V]$, then gain-loss utility from outcome v is

$$n\left(v \,|\, \widehat{E}[V]\right) = \begin{cases} v - \widehat{E}[V], & \text{if } v \geq \widehat{E}[V]; \\ \lambda\left(v - \widehat{E}[V]\right), & \text{if } v < \widehat{E}[V]; \end{cases} \tag{1}$$

where parameter $\lambda \geq 1$ captures loss aversion. The agent's total utility is then

$$u\left(v \,|\, \widehat{E}[V]\right) = \underbrace{v}_{\text{Consumption utility}} + \underbrace{\eta n\left(v | \widehat{E}[V]\right)}_{\text{Gain-loss utility}}, \tag{2}$$

where $\eta > 0$ is the weight given to sensations of gain and loss relative to absolute outcomes.[8]

Attribution Bias: We now introduce misattribution, which can arise when the agent is learning about the typical consumption utility she derives from a prospect. A misattributing agent uses her experienced utility to infer this consumption utility but neglects the extent to which her total utility was shaped by reference dependence. That is, following an outcome v she correctly recalls how happy she felt, but she underappreciates how sensations of elation or disappointment affected her total utility. We model this form of attribution bias by assuming that the agent infers v using a misspecified model that weights the gain-loss component of her utility by a diminished factor $\hat{\eta} \in [0, \eta)$. Specifically, she infers outcome $\hat{v}$ as if her utility function were $\hat{u}\left(\hat{v} \,|\, \widehat{E}[V]\right) = \hat{v} + \hat{\eta} n\left(\hat{v} \,|\, \widehat{E}[V]\right)$; thus, $\hat{v}$ solves $\hat{u}\left(\hat{v} \,|\, \widehat{E}[V]\right) = u\left(v \,|\, \widehat{E}[V]\right)$. Equations (1) and (2) imply that this misencoded outcome, $\hat{v}$, takes the following form:

$$\hat{v} = \begin{cases} v + \left(\dfrac{\eta - \hat{\eta}}{1 + \hat{\eta}}\right)\left(v - \widehat{E}[V]\right), & \text{if } v \geq \widehat{E}[V]; \\ v + \lambda\left(\dfrac{\eta - \hat{\eta}}{1 + \hat{\eta}\lambda}\right)\left(v - \widehat{E}[V]\right), & \text{if } v < \widehat{E}[V]. \end{cases} \tag{3}$$

Thus, the encoded outcome is biased upward when the true outcome beats expectations and biased downward when it falls short. This bias is proportional to the deviation between the true outcome and expectations. Additionally, a loss is misencoded by a greater extent than an equal-sized gain when the agent suffers loss aversion (i.e., $\lambda > 1$).

[8] Our predictions do not substantively depend on whether we assume a deterministic reference point (à la Bell and Equation 1, above) or a stochastic reference point (à la KR); we utilize the former for simplicity. Furthermore, unlike KR, we do not impose rational expectations; indeed, a key feature of our framework posits that the agent's (potentially biased) beliefs determine her reference point.

To close the model, we assume the agent uses this misencoded outcome to update her beliefs. The agent takes actions to maximize her expected utility (equation (2)) given these biased beliefs.[9]

To illustrate the model, recall the example from the introduction wherein a worker's daily task is assigned at random: some days she faces a relatively enjoyable task, and other days she faces an onerous one. When the worker is assigned the onerous task, she simultaneously experiences both a bad material outcome and a sensation of disappointment. A misattributor fails to fully account for this disappointment and wrongly attributes this feeling to the underlying disutility of the task. She thus recalls her assigned task as more onerous than it really was. When the worker is assigned the more pleasant task, she simultaneously faces an easier job and a pleasant surprise and recalls the task as even better than it really was.[10]

Our experiments examine a setting with two distinct dimensions of consumption utility—money (m) and effort (e). Thus, when applying the model above to our specific experiments, we will consider a simple extension to two dimensions. Given expectations $\widehat{E}\left[V^k\right]$ along each dimension $k \in \left\{m, e\right\}$, the agent's total utility from realization $v = \left(v^m, v^e\right)$ is

$$u\left(v \,|\, \widehat{E}[V]\right) = \sum_{k \in \{m,e\}} \left[v^k + \eta n\left(v^k \,|\, \widehat{E}\left[V^k\right]\right)\right]. \tag{4}$$

Each misencoded outcome, $\hat{v}^k$, is then defined as in equation (3) dimension by dimension. That is, a misattributor recalls an outcome $\hat{v}^k$ such that $\hat{v}^k + \hat{\eta} n\left(\hat{v}^k \,|\, \widehat{E}\left[V^k\right]\right) = v^k + \eta n\left(v^k \,|\, \widehat{E}\left[V^k\right]\right)$.

II. Experiment 1

In this section we present our between-subject experiment, which we conducted on MTurk. We first describe the experimental design. Next, we provide theoretical predictions of both rational-learning models and our model of misattribution. We then analyze our experimental data, noting throughout how the results are consistent with our notion of misattribution yet inconsistent with rational-learning models with or without reference-dependent preferences. Finally, we present a replication study.

[9] This implies that the agent makes decisions according to the true value of η. While this is our preferred approach, it is worth emphasizing that our predictions are robust to the agent making decisions according to the misspecified parameter value, $\hat{\eta}$. This latter approach may be a reasonable way to incorporate the insights from Imas, Sadoff, and Samek (2017) and de Quidt (2018).

[10] There are at least two plausible interpretations of how and when these biased perceptions are formed: first, the agent improperly encodes each outcome as it happens, which seems most plausible in settings where the determinants of consumption utility are not directly observable (e.g., one's disutility of working on an unfamiliar task or the quality of a meal); second, the agent retrieves a distorted memory of an outcome when attempting to recall its value (e.g., one might remember an unexpectedly high price from a previous transaction as higher than it truly was despite knowing the true price when the transaction took place).

A. *Design*

We recruited approximately 900 participants for a two-session experiment.[11] In the first session—an initial learning phase—participants gained experience with a real-effort task. In the second session we elicited participants' willingness to complete additional work on the same task they previously faced. Participants took an average of 10 and 15 minutes to complete the first and second sessions, respectively. Participants were paid \$4 for successfully completing both sessions and could earn up to \$6.50 depending on chance and their willingness to work.

Each participant worked on one of two tasks. In both tasks participants listened to reviews of books and had to classify whether each review was positive or negative.[12] Figure 2 depicts the interface. Our two tasks differed in a single way: one version used unaltered audio, while the other used audio that was overlaid with an annoying noise. This noise was a composite of a fork scraping against a record and a high-frequency tone. The noise played at approximately 15 decibels lower than the peak levels of the audio in the review when played at moderate volume; it was annoying but did not hinder participants' ability to classify the reviews.

Importantly, participants who faced the annoying noise could not avoid the noise and still successfully complete the task. We also took three additional measures to ensure that participants actually listened to the audio reviews: (1) participants were required to answer at least six out of the eight mandatory classifications correctly during the first session or else they would be removed from the study without pay; (2) response buttons were hidden for the first ten seconds of each review, preventing participants from quickly guessing; and (3) many of the reviews featured revealing details only in the late part of the review. To prohibit participants from reloading the web session in an attempt to be reassigned without the noise, we blocked multiple logins and required unique email authentication to access each session of the experiment.

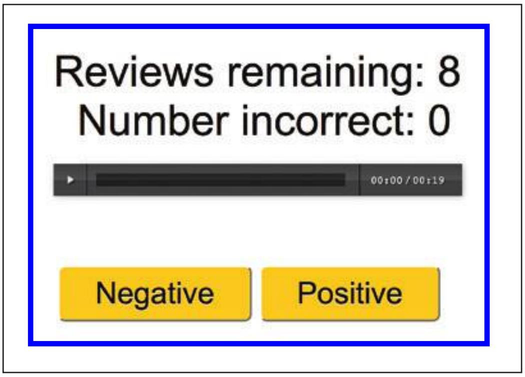
The two sessions of the experiment were conducted as follows.

Session 1 (Initial Learning Phase): Participants were instructed that the purpose of this session was to learn about how much they enjoyed the task, since they would later have an opportunity to complete additional rounds of that task for extra pay.

We ran several treatment arms to investigate how initial expectations altered subsequent evaluations. Participants in the known-assignment group ($N = 292$) were told from the start which task they would face, while participants in the coin-flip ($N = 294$) and high-probability ($N = 300$) groups were initially uncertain. (We call these groups *control*, *coin-flip*, and *high-probability*, respectively.) The former two treatment arms were conducted one month before the high-probability treatment.

[11] Participants were recruited between July and August 2016 and were required to be located within the United States and to have completed at least 100 prior jobs on MTurk with a 95 percent approval rating.

[12] We used digital-voice software to read reviews collected from Amazon.com. Unbeknownst to participants, all reviews were either one-star reviews or five-star reviews, to make the task straightforward (though tedious). Reviews were edited to last approximately 20 seconds, to remove any specific references to authors' names or book titles, and for grammar. See online Appendix H for sample text from the reviews.



Reviews remaining: 8

Number incorrect: 0

Figure 2. Screenshot of the Classification Task from Experiment 1

Notes: Buttons appeared after ten seconds. Participants clicked the appropriate button to classify whether a review was positive (i.e., endorsing the book) or negative.

We now describe how Session 1 differed across treatments. Participants in the control treatment were randomly assigned—unbeknownst to them—to one of two subgroups prior to entering the experiment: *noise* or *no noise*. Participants in the *control* + *noise* group completed classifications with an annoying noise overlaid. Participants in the *control* + *no noise* group completed classifications without the overlaid noise. Participants in each subgroup were not aware of the possibility of facing the alternate task—they were only told about the one they were assigned. Each participant completed eight mandatory trials of their assigned classification task to conclude the first session.[13]

In contrast, participants in the *coin-flip* treatment were told that they faced a one-in-two chance of doing the task without noise and a one-in-two chance of doing the task with noise. They were then given a sample task (without noise) and a short sample of the unpleasant noise (eight seconds in duration; repeatable if desired). This sample and the remaining instructions were designed to provide time for this uncertainty to sink in and form a reference point. After these additional instructions, each participant flipped a digital coin to determine whether they would ultimately face the task with noise or without. Immediately thereafter, each participant then completed the eight mandatory classifications prescribed by the result of their coin flip.

Last, participants in the *high-probability* treatment faced identical instructions to the *coin-flip* treatment, except they were told that they were very likely to face a given task (either *noise* or *no noise*; see online Appendix H for full text with highlighted differences). Half of the participants were assigned to a $p = 0.99$ treatment and the other half were assigned to a $p = 0.01$ treatment, where p corresponds to the probability of facing the task with noise. Each participant drew a random integer z from $[1, 100]$. Participants in the $p = 0.99$ arm were assigned the task without noise if $z = 100$; otherwise, they faced the task with noise. Participants in the

[13] Prior to completing the eight mandatory trials, participants in each control subgroup completed one practice trial (which matched their assigned version of the task) to teach them how to use the interface.

$p = 0.01$ treatment were assigned the task with noise if $z = 100$; otherwise, they faced the task without noise. As in the groups above, each participant completed eight trials of their assigned task immediately after the resolution of this uncertainty.

In each group, the first session concluded after the participant completed the eight mandatory trials of her assigned task. Before exiting Session 1, they were reminded that they would face the same task when they returned for Session 2.

Session 2 (Willingness to Work): We emailed each participant a link to the second session exactly eight hours after they finished the first.[14] Upon logging into the second session, participants were reminded of their prior task assignment (noise or no noise). They were then given the option to complete additional trials (of that same task) for a bonus payment.

We elicited participants' willingness to continue working in exchange for five different payment values. We utilized the Becker-Degroot-Marshak (BDM) mechanism to incentivize their responses. The mechanism operated as follows: for each possible bonus payment $m \in \{\$0.50, \$1.00, \$1.50, \$2.00, \$2.50\}$, we asked participants the maximum number of tasks they would complete in order to receive $\$m$. They responded by using a slider to select an integer $e \in [0, 100]$, which we call "willingness to work" or WTW. We then (uniformly) drew a random integer $y \in [0, 100]$. If $y \leq e$, then the participant completed e additional tasks and received $\$m$. If $y > e$, then the participant completed no additional tasks and earned no bonus pay. We utilized simple instructions and two sample questions to illustrate the mechanism to participants. After eliciting participants' WTW, we employed the mechanism and participants completed additional tasks as required.

Session 2 was identical across all treatment groups, conditional on the assigned task.

B. *Theoretical Predictions*

In this section we apply a model of reference dependence and attribution bias to our experimental setting and derive our key theoretical prediction: fixing her assigned task, a misattributing participant's willingness to work (WTW) is increasing in the ex ante probability of being assigned the noisy task. In contrast, the WTW of an agent who does not suffer misattribution is independent of this probability. While this analysis motivates our empirical strategy, the eager reader may skip to the experimental results (Section IIC).

Theoretical Setup: Following our experimental design, there are two periods. In the first period ($t = 1$), participant i is randomly assigned to one of two tasks $a \in \{h, l\}$, where h is the noisy task and l is the noiseless one. Let probability $p_i \in \{0, 0.01, 0.5, 0.99, 1\}$ denote the participant's ex ante belief that she will be assigned to task $a = h$. Participant i completes eight trials of her assigned task a in

[14] Fourteen subjects emailed the authors stating that they had not received an invitation to the second session after more than eight hours (despite our use of an automated notification system). All were sent an additional invitation and completed the second session.

period 1 and is informed that she will face this same task with certainty in period 2. In the second period ($t = 2$), the participant chooses the maximum number of trials of task a she is willing to complete in exchange for a monetary payment $m > 0$.

We consider a participant who is uncertain about the cost function associated with her assigned task and who updates her perception of this function based on her work experience. Along the effort dimension, we assume participant i's consumption utility from completing $e_{i,t} \geq 0$ rounds of task a in period t is

$$v^e_{i,t} = -\left[\theta_i(a) + \epsilon_{i,t}\right] c(e_{i,t}), \tag{5}$$

where $c(\cdot)$ is an increasing function with $c(0) = 0$, $\theta_i(a)$ is a cost parameter that depends on $a \in \{h, l\}$, and $\epsilon_{i,t}$ are i.i.d. mean-zero random cost shocks that are independent of $\theta_i(a)$. The structure of $v^e_{i,t}$ (equation (5)) is known to the participant, but she is initially uncertain about the cost parameter, $\theta_i(a)$. Let $\hat{\theta}_{i,0}(a)$ denote the participant's expected value of $\theta_i(a)$ under her prior. We assume the participant (rightfully) has priors such that $\hat{\theta}_{i,0}(h) > \hat{\theta}_{i,0}(l) > 0$—i.e., the noisy task seems more onerous than the noiseless one—and these priors are independent of her treatment group—i.e., each $\hat{\theta}_{i,0}(a)$ is independent of p_i.

Belief Updating: We consider a participant who cannot separately observe $\theta_i(a)$ and $\epsilon_{i,1}$ and who thus uses her experienced utility in period 1 as a signal to update her beliefs about $\theta_i(a)$. Importantly, when the participant has reference-dependent preferences, this experienced utility depends on her initial expectations. In this case her experienced utility in period 1 follows equation (4):

$$u_{i,1} = v^e_{i,1} + \eta n\left(v^e_{i,1} \,|\, \widehat{E}_{i,0}\left[V^e_{i,1}\right]\right),^{15} \tag{6}$$

where $\widehat{E}_{i,0}\left[V^e_{i,1}\right]$ represents her expected consumption utility in period 1. More specifically, because she is assigned task $a = h$ with probability p_i, the participant's expected consumption value on the effort dimension entering period 1 is $\widehat{E}_{i,0}\left[V^e_{i,1}\right] = -\left[p_i \hat{\theta}_{i,0}(h) + (1 - p_i)\hat{\theta}_{i,0}(l)\right] c(8)$.

As described in Section I (equation (3)), a participant uses $u_{i,1}$ to infer her consumption utility from effort and subsequently updates her belief about $\theta_i(a)$. Misattribution implies that when the task is less burdensome than expected (i.e., $v^e_{i,1} > \widehat{E}_{i,0}\left[V^e_{i,1}\right]$), the participant encodes $\hat{v}^e_{i,1} > v^e_{i,1}$. If instead the task is worse than expected, then she encodes $\hat{v}^e_{i,1} < v^e_{i,1}$.[16] The participant then uses Bayes' Rule as if the realized value of $V^e_{i,1}$ was $\hat{v}^e_{i,1}$ to form her updated expectation of $\theta_i(a)$, denoted by $\hat{\theta}_{i,1}(a)$. For tractability, we assume that $\theta_i(a)$ and $\epsilon_{i,t}$ are normally

[15] Since there is no payment in period 1, the participant only experiences utility along the effort dimension.

[16] An agent who fully appreciates the extent to which her utility depends on expectations (i.e., $\hat{\eta} = \eta$) encodes the correct value, $\hat{v}^e_{i,1} = v^e_{i,1}$.

distributed, which implies that her updated belief has a simple negative linear relationship with $\hat{v}^e_{i,1}$.[17]

Effort Choice: We now show how biased learning about $\theta_i(a)$ in period 1 distorts her WTW in period 2 once she fully expects to face her previously assigned task. To illustrate this most cleanly, we will examine the effort level $e^*_i(a\,|\,p_i)$ such that the agent is indifferent between completing $e^*_i(a\,|\,p_i)$ rounds of the task for m dollars and not working at all. We call this value the participant's "maximal WTW." In the subsection that follows, we discuss how we attempt to elicit this value.

When the agent has reference-dependent preferences, indifference between completing $e^*_i(a\,|\,p_i)$ tasks for m dollars and not working at all implies that $e^*_i(a\,|\,p_i)$ solves

$$(7)\qquad \widehat{E}_{i,1}\big[u_{i,2}\,|\,e_{i,2}\big] \;=\; \widehat{E}_{i,1}\big[V^e_{i,2}\big] + \eta\widehat{E}_{i,1}\Big[n\Big(V^e_{i,2}\,|\,\widehat{E}_{i,1}\big[V^e_{i,2}\big]\Big)\Big] + m \;=\; 0.$$

This equation demonstrates that uncertainty over $\theta_i(a)$ and, hence, the disutility of effort—captured in the subjective expectation $\widehat{E}_{i,1}\big[V^e_{i,2}\big]$—induces gain-loss utility. This makes solving for the maximal WTW somewhat more complicated than the standard case absent reference dependence. Building on equation (7), we show in Online Appendix C that $e^*_i(a\,|\,p_i)$ in fact solves

$$(8)\qquad h\Big(\hat{\theta}_{i,1}(a)\Big)c\big(e^*_i\big) \;=\; m,$$

where $h(\,\cdot\,)$ is an increasing function of $\hat{\theta}_{i,1}(a)$. This function depends on the participant's preference parameters (η, λ) and her subjective distribution of $V_{i,2}$. However, absent misattribution, it is independent of p_i. Intuitively, an agent who does not suffer misattribution properly accounts for how p_i influenced her experienced utility in period 1. Thus, p_i does not distort her inferred value of $v^e_{i,1}$ nor her beliefs about $\theta_i(a)$. Therefore, p_i does not influence $e^*_i(a\,|\,p_i)$.[18]

[17] If the agent believes that $\theta_i(a) \sim N\big(\hat{\theta}_{i,0}(a), \rho^2\big)$ and $\epsilon_{i,t} \sim N\big(0, \sigma^2\big)$, then

$$\hat{\theta}_{i,1}(a) \;=\; -\alpha\left[\frac{\hat{v}^e_{i,1}}{c(8)}\right] + (1-\alpha)\hat{\theta}_{i,0}(a) \text{ where } \alpha \equiv \frac{\rho^2}{\rho^2+\sigma^2}.$$

Our basic predictions for both experiments hold under weaker assumptions that are likely met even if the participant does not precisely follow Bayes' rule. Namely, our results extend so long as the following properties hold. First, the agent's updating is monotonic: $\hat{v} < \hat{v}'$ implies $\hat{\theta}_{i,1}(a\,|\,\hat{v}) > \hat{\theta}_{i,1}(a\,|\,\hat{v}')$. Second, beliefs update in the direction of the signal: $\hat{v} > \widehat{E}_{i,0}\big[V^e_{i,1}\big]$ implies that $\hat{\theta}_{i,1}(a\,|\,\hat{v}) < \hat{\theta}_{i,0}(a)$, and $\hat{v} < \widehat{E}_{i,0}\big[V^e_{i,1}\big]$ implies that $\hat{\theta}_{i,1}(a\,|\,\hat{v}) > \hat{\theta}_{i,0}(a)$. Both assumptions are implied by Bayesian updating for a range of distributional assumptions, including the case where $\theta_i(a)$ and $\epsilon_{i,t}(a)$ are independent and normally distributed. See Chambers and Healy (2012) for general sufficient conditions for Bayesian updating in the direction of the signal.

[18] This conclusion immediately extends to the case where the agent does not have reference-dependent preferences. In this case, $\eta = 0$ and $h(\,\cdot\,)$ from equation (8) reduces to the identity function.

OBSERVATION 1: *Let $e^*(a|p)$ denote the maximal WTW averaged over participants who faced task a and held prior beliefs that there was a chance p of facing the noisy task. Absent misattribution, $e^*(a|p) = e^*(a|p')$ for all p,p'.*

We now describe $e_i^*(a|p_i)$ under misattribution. As in the case above, $e_i^*(a|p_i)$ solves equation (8); however, the misattributor makes this choice based on her biased assessment of $\theta_i(a)$. In particular, she encodes an overly optimistic value $\hat{\theta}_{i,1}(a)$ whenever the task she faces beats expectations, and she encodes an overly pessimistic value whenever her realized task falls short. Thus, fixing the task she faces, raising initial expectations tends to generate a more pessimistic view of the underlying task, and lowering expectations tend to a rosier view. We therefore predict that for each $a \in \{h,l\}$, $e^*(a|p)$ is increasing in p.

OBSERVATION 2: *Let $e^*(a|p)$ denote the maximal WTW averaged over participants who faced task a and held prior beliefs that there was a chance p of facing the noisy task. Suppose each participant's prior beliefs over $\theta(a)$ are independent of treatment with $\hat{\theta}_{i,0}(l) < \hat{\theta}_{i,0}(h)$. Under misattribution, $e^*(a|p)$ is increasing in p.*

The two observations together highlight our empirical strategy. Recall that in Session 2 of our experiment, the participant announced how many additional tasks she was willing to do for a bonus payment of m dollars. Our main interest is whether and how this WTW depended on the likelihood that the participant was assigned to the noisy task, p_i. For a given task, we compare the WTW of participants across the different assignment probabilities. As highlighted in Observation 2, misattribution predicts that, conditional on the assigned task, WTW will be increasing in the ex ante likelihood of facing the onerous task.

Discussion of Assumptions.—We now discuss some of the assumptions underlying the results above. First, we discuss utilizing the BDM mechanism to measure maximal WTW when agents have reference-dependent preferences. Second, we clarify the extent to which our results rely on participants holding well-calibrated priors. Finally, we discuss our assumption that priors are independent of treatment, which was the motivation behind our *high-probability* treatment.

The BDM Mechanism and Agents with Reference-Dependent Preferences: In equation (8) we demonstrated that when agents have expectations-based, reference-dependent preferences, uncertainty about the effort dimension complicates matters, since this uncertainty influences the reference point. Given that the BDM mechanism itself creates even more uncertainty (over how much a participant might eventually work), it is not immediate that it is a useful tool to measure the maximal WTW of reference-dependent agents. In online Appendix D we allow for the possibility that participants incorporated the uncertainty induced by the BDM into their reference points along the effort and money dimensions, and we solve for the optimal response. There, we show that under these conditions the BDM mechanism does not generically reveal $e_i^*(a|p_i)$. Critically, however, we show that the key predictions highlighted above remain: under misattribution, the participant's optimal response is increasing in p_i (i.e., her initial chance of facing the noisy task); absent misattribution, her optimal response is independent of p_i.[19]

Robustness to Poorly Calibrated Priors: The observations above do not require subjects to have held well-calibrated priors about the tasks (i.e., about $\theta_i(a)$). If prior beliefs are biased on average, our observations continue to hold so long as participants' priors are independent of the treatment. In this case, learning absent misattribution leads to the same posterior beliefs regardless of the treatment (fixing the task a participant faced), since the treatment does not influence the interpretation of signals or priors. Even with misattribution, the prediction from Observation 2 still holds so long as priors are "reasonable"—that is, participants believe the noisy task is more onerous than the noiseless one. Given that participants in the *coin-flip* and *high-probability* treatments sampled each task during the instructions, such priors seem likely.

Priors That Are Independent of Treatment Group Assignment: Observations 1 and 2 rely on independence between a participant's priors about the tasks and the likelihood she is assigned the noisy task. However, participants in the *control* group were exposed to only a single task during the instructions; thus, it is plausible that they held initial beliefs about a given task that systematically differed from those in the *coin-flip* treatment who were exposed to both tasks. For instance, the existence of both an easy and hard version of the task might have led a participant in the *coin-flip* group to infer that the noisy task was particularly onerous, while an analogous participant in the *control* group was only aware of the noisy task and might have expected it to resemble a typical MTurk task. Our *high-probability* treatment was designed to address this concern. Participants in the *high-probability* treatment were exposed to both tasks exactly as in the *coin-flip* treatment. This mitigates concerns about differential inference. In this sense, we use the *high-probability* group (where participants were very likely to face task a) as a cleaner alternative to the associated *control* group (where participants were certain to face task a). In both groups participants strongly expected to face task a, but in the *high-probability* version they were perfectly aware of the alternative task.

Adjustment of the Reference Point over Time: The observations above leverage a particular assumption about participants' reference points: we assumed that participants anticipated their task assignment by the onset of the second session. While this assumption generates crisp distinctions between effort under misattribution and rational learning (with or without reference-dependent preferences), reference points that adapt very slowly can muddy these distinctions. In particular, if participants had sluggish reference points (i.e., expectations still depended on the lottery hours later) and held reference-dependent utility over effort but not money, then

[19] By focusing Observations 1 and 2 on $e_i^*(a \mid p_i)$, we further demonstrate that our main predictions hold when participants respond to the BDM in an intuitive way, consistent with the wording of our survey (which asked participants to truthfully report the maximum number of tasks they were willing to do for each payment level). This also corresponds to a form of "narrow bracketing," which is commonly assumed in the literature on eliciting risk preferences (see, e.g., Bernheim and Sprenger 2020). Finally, our focus on $e_i^*(a \mid p_i)$ highlights that our predictions do not stem from some interaction between the BDM mechanism and reference dependence.

reference dependence without misattribution may predict effort patterns similar to those predicted by our model of misattribution. While this particular constellation of assumptions is perhaps plausible, it is inconsistent with existing evidence demonstrating reference-dependent preferences over money.

To alleviate this issue, our design utilizes a relatively long gap between sessions to provide time for reference points to adapt by Session 2. Furthermore, we discuss below how the observed treatment effect supports fast reference point adjustment (Section IIC). If participants did not (at least partially) incorporate the coin flip into their expectations, we would expect no differences across treatments; our data suggest otherwise.

C. *Results*

To test the theoretical predictions above, we first take a simple nonparametric approach to demonstrate that willingness to work (WTW) in Session 2 depends significantly on participants' initial expectations regarding their task assignment. We then estimate a reduced-form version of our model that utilizes our multiple observations to control for potential curvature in the effort-cost function and individual-specific characteristics. Both approaches demonstrate that behavior is consistent with participants wrongly learning the underlying difficulty of their assigned task as a function of their priors.

Summary of the Data: Our experimental design generates six subgroups: treatment (i.e., whether participants faced certain assignment, coin-flip assignment, or high-probability assignment) crossed by eventual task assignment (i.e., noise or no noise). For each subgroup, Table 1 shows the demographic characteristics of participants who successfully completed the first session (886 participants in total) and the proportion of those who returned for the second session.[20] Note that variability in subgroup sizes resulted from random treatment assignment. Also, while there are some differences in attrition rates across groups (e.g., between *coin-flip* + *noise* and *high-probability* + *noise*), we discuss below how this pattern is unlikely to drive our results.

We implemented some data-cleaning procedures to form our primary dataset. We removed participants who either did not answer all five elicitations of WTW (three participants) or stated a WTW equal to either the maximum or minimum amount for every payment level, which prevented us from estimating their responsiveness to payment (eleven participants).[21] Additionally, we omit participants who did not return for the second session—and whose WTW we therefore did not

[20] There is a significant age difference between the first two treatments and the *high-probability* treatment. The first two treatments were run approximately one month prior to the latter and the *high-probability* treatment was launched at a slightly later time of day. We suspect time-of-day effects account for the age difference between groups. Our regression analyses control for demographics.

[21] This first restriction was the result of coding that should have forced all participants to answer all questions but did not function properly on some obsolete browsers. We believe that nonresponsiveness to incentives (the second restriction above) likely resulted from confusion, inattention, or wrongly attempting to manipulate the BDM mechanism. Note that a participant who is supposedly willing to complete 100 tasks for $0.50 is revealing that they command an extremely low hourly wage rate.

Table 1—Demographics and Summary Statistics, Experiment 1

	Control		Coin flip		High prob.	
Variable	noise = 0	noise = 1	noise = 0	noise = 1	noise = 0	noise = 1
Age	38.24	39.71	39.36	39.63	33.61	33.29
	(12.04)	(12.30)	(11.45)	(11.96)	(9.777)	(9.352)
1(Male)	0.468	0.464	0.428	0.387	0.529	0.487
	(0.501)	(0.500)	(0.496)	(0.489)	(0.501)	(0.501)
Income	2.712	2.582	2.901	2.613	2.357	2.462
	(1.009)	(1.092)	(1.066)	(1.103)	(1.011)	(1.069)
1(Return)	0.921	0.882	0.862	0.944	1	0.931
	(0.271)	(0.323)	(0.346)	(0.231)	(0)	(0.254)
Observations	139	153	152	142	140	160

Notes: Standard deviations are in parentheses. Income is coded as a discrete variable that takes a value from one to five corresponding to the following income brackets: (1) less than \$15,000, (2) \$15,000–\$29,999, (3) \$30,000–\$59,999, (4) \$60,000–\$99,999, (5) \$100,000 or more.

measure—though we present their demographics where applicable. With this set of restrictions, we are left with a sample of 803 participants.[22]

Nonparametric Analysis.—Our main hypothesis is that participants' WTW on a given task is increasing in their initial likelihood of facing the bad task. We first compare the average WTW in the *control* and *coin-flip* treatments, averaging over both individuals and the five payment levels about which we elicited WTW. This is presented in columns 1–4 of Table 2. This comparison provides a simple assessment of how uncertainty over task assignment in the initial learning session affected subsequent behavior. Relative to the control group, participants who faced the noiseless task were more willing to work when their initial impressions were formed after the resolution of the coin flip ($p = 0.039$ for difference; statistical results obtained via Wald tests with standard errors clustered at individual level unless otherwise noted). In contrast, participants who faced the noisy task were less willing to work (relative to control) when their initial impressions were formed after the resolution of the coin flip ($p = 0.025$ for difference).[23]

While Table 2 provides a rough sense of the treatment effect, Figure 1 (presented in the introduction) further disaggregates WTW by payment level. Figure 1 shows the average WTW at each of the five payment levels $\{\$0.50, \$1.00, \$1.50, \$2.00, \$2.50\}$ for each group (crossing treatment with task assignment).

These baseline results reveal economically meaningful magnitudes. For instance, consider a hypothetical firm seeking workers to complete 25 of our classification tasks. Workers who faced no uncertainty when forming their initial impressions required (on average) \$1.70 and \$1.50 to complete 25 noisy and noiseless tasks, respectively. This difference is significantly exaggerated when workers experience sensations

[22] In this main sample there were very few mistakes in the classification task: only two participants were removed from the study for inaccurate responses. Since this occurred before they returned for the second session, we do not consider this a data-cleaning step.

[23] In online Appendix Figure B1, we present smoothed CDFs of the aggregate WTW for the *control* and *coin-flip* treatments and present some statistical tests validating their differences.

Table 2—Baseline Results, Experiment 1

Variable	Control		Coin flip		High prob.	
	noise = 0	noise = 1	noise = 0	noise = 1	noise = 0	noise = 1
Willingness to Work (WTW)	24.23 (1.358)	22.29 (1.574)	28.60 (1.622)	17.64 (1.362)	24.20 (1.295)	21.42 (1.276)
Observations	615	665	645	665	690	735

Notes: Willingness to work is averaged over five payment levels. Standard errors (in parentheses) are clustered at the individual level. Differences between columns 1–3, 3–5, 2–4, and 4–6 are all significant at $p < 0.05$.

of surprise when forming initial impressions: workers whose initial impressions were confounded by sensations of disappointment or elation required \$2.30 and \$1.20 to complete 25 noisy and noiseless tasks, respectively. Thus, required payments increased by 35 percent for the noisy task and decreased by 20 percent for the no-noise one. Furthermore, the payment premium for the noisy task—the additional payment required to incentivize the noisy task over the noiseless one—increased from \$0.20 to \$1.10. Across all payment levels, those who formed initial impressions of the noiseless task when it came as a positive surprise were more willing to work than those who faced the same task with certainty. In contrast, we find that a negative surprise had the opposite effect for the noisy task.

We now present the results of our *high-probability* treatment. Recall that this was designed to mitigate concerns that the differences between the *control* and *coin-flip* treatments in fact reflect differences in information rather than misattribution. We first note that participants in the *high-probability* treatment exhibited a lower WTW on the noisy task than on the noiseless task (aggregating across all payment levels; $p = 0.064$). This validates that participants perceived a difference in the onerousness of the two tasks.

Comparing across treatments, columns 3–6 of Table 2 further demonstrate that participants' WTW depended on the expectations they held prior to the initial learning session. Participants who were assigned the noiseless task based on the coin flip were, on average, significantly more willing to work than those who strongly expected the noiseless task ($p = 0.034$ for difference). In contrast, participants who were assigned the noisy task based on the coin flip were significantly less willing to work than those who strongly expected the noisy task ($p = 0.047$ for difference).

The results above may slightly understate the impact of misattribution. Given the (albeit small) uncertainty over task assignment present in the *high-probability* groups, our model predicts that participants in those groups will demonstrate greater differences in WTW across the two tasks than those in the *control* groups. These differences should theoretically be small, as they stem from the difference between 1 and 0 percent. However, probability weighting—people's tendency to overweight small probabilities (e.g., Kahneman and Tversky 1979; Prelec 1999; Gonzalez and Wu 1999)—implies that behavior in the *high-probability* treatment may deviate from the corresponding *control* by more than a 1 percent chance would suggest. If this 1 percent looms much larger than the objective probability, participants may treat *high-probability* as closer to *coin-flip* than is merited by the objective probabilities, leading us to understate the effect of misattribution.

We now discuss other potential explanations for these baseline results: attrition, mood effects, and reference points that fail to adjust between sessions. We also present results from a replication experiment designed to mitigate concerns stemming from the fact that the *high-probability* treatment was run after the other two. We delay discussion about reciprocity toward the experiment as a potential explanation until after presenting our results from Experiment 2.

Differential Attrition across Treatments: The summary statistics presented in Table 1 suggest that differential attrition—that is, failing to return to the second session—cannot explain our treatment effects. As that table demonstrates, there is not a consistent pattern of attrition between treatments and whether participants were assigned the noisy task. In Table B4 in online Appendix B, we demonstrate that the observables we collect (e.g., task assignment and demographics) do not predict attrition.[24] Overall, attrition was quite low. Accordingly, we believe attrition-based explanations are unlikely to explain the observed effects.

Mood Effects: Our experiment was designed to combat the concern that short-term, transient moods induced by resolving uncertainty (e.g., anger) might explain our effects. Specifically, the time gap between participants forming their impressions and our elicitation of WTW was designed to mitigate such short-term mood effects. In order to explain our effects, the coin flip must continue to influence a participant's mood hours later, when they return for the second session.

Examining the heterogeneity in this time gap between sessions allows us to further speak to this point. In Tables B1 and B2 (in online Appendix B), we reproduce Table 2 but divide the sample in two: those who returned after the mandatory eight-hour gap between sessions but before the median return time ($\approx$ 11.5 hours), and those who returned after the median return time. The average time gap between sessions for this latter group was nearly 24 hours, and their second-session log-in times suggest that these participants slept between sessions. Nevertheless, we find qualitatively similar results across these two groups, though our statistical power is diminished. This suggests that if such mood effects were to drive our results, they would need to be rather persistent. Our results from Experiment 2 challenge this explanation; we return to this discussion in Section IIIC.

Reference Point Adjustment: As noted in the theoretical discussion, if reference points failed to adjust between the first and second sessions, then participants' choices may demonstrate the basic pattern we observe. Our data cannot refute this possibility. However, the fact that we observe a marked difference in WTW across our treatments is evidence that reference points adjusted rather quickly. Participants only sat with the treatment-induced uncertainty in task assignment for a few moments before it was resolved. Thus, the significant treatment effect we observe

[24] An alternative type of attrition is possible given the MTurk setting: some participants may have exited the survey when assigned to the noisy task without ever completing Session 1. We reviewed all partially completed surveys and found that only nine participants closed the survey prematurely after the task assignment was revealed. Of those partial completions, six were assigned to the noiseless task and three were assigned to the noisy task.

suggests that their reference points adjusted to incorporate this uncertainty within that short period of time: if reference points did not adjust quickly, we would expect no treatment effect. Although we cannot rule out that reference points subsequently failed to adjust before the second session, we note they seem to move easily and quickly in the first session.[25]

Parametric Analysis.—Motivated by our simple nonparametric results, we now consider a more structured, regression-based approach. Although this imposes some strong assumptions, doing so allows us to account for the fact that effort costs in our experiment may be nonlinear and to better utilize the multiple observations we obtain from each participant. Thus, we provide better estimates of the aggregate effort-supply curves illustrated in Figures 1 and 3 with the appropriate confidence intervals and, in effect, address the lack of error bars in those figures. Finally, we provide a back-of-the-envelope calculation for the effect size of misattribution and show that our evidence suggests people fully neglect how reference dependence shaped their initial impressions.

Following the learning model in Section IIB, we estimate participants' revealed perception of the underlying cost parameters for each task, $\theta(a)$, conditional on their treatment. For participant *i* who expected to face the noisy task with probability $p \in \{0, 0.01, 0.5, 0.99, 1\}$ and is ultimately assigned task a, let $\hat{\theta}_{i,1}(a|p)$ denote her expectation of $\theta_i(a)$ following Session 1. We estimate the average value of this expectation, denoted $\hat{\theta}_1(a|p)$, among participants in each subgroup.

In order to estimate these parameters, we assume $c(e) = (e+\omega)^{\gamma}$, where ω is a Stone-Geary background parameter.[26] Thus, participant *i* chooses e_i^* such that $\hat{\theta}_{i,1}(a|p)(e_i^*+\omega)^{\gamma} = m$.[27] Rearranging, setting $\omega = 0$, and taking logs yields

$$\log(e_i^*) = \frac{\log(m)}{\gamma} - \frac{\log\left[\hat{\theta}_{i,1}(a|p)\right]}{\gamma}. \tag{9}$$

[25] This accords with the small body of evidence on reference point adjustment. Song (2016) shows that reference points incorporate new information over the course of approximately ten minutes. Likewise, Smith (2019) and Buffat and Senn (2015) provide evidence of relatively quick reference point changes in laboratory settings with small stakes. Using field evidence from taxi drivers, Thakral and Tô (2020) note that "earnings in the first four hours [of a driver's shift] have little or no effect on the decision of whether to end a shift at 8.5 hours." Taken together, we share Song's (2016) interpretation of the broader literature: for small stakes, reference points seem to adjust within minutes.

[26] This functional form has been utilized in similar real-effort experiments (e.g., Augenblick, Niederle, and Sprenger 2015). For the analysis presented below, we take $\omega = 0$. In Table B3 we show that our qualitative results are robust to this assumption: over a wide range of ω, we estimate significant differences in parameters across our treatments.

[27] This effort choice follows from equation (8), which predicts that a participant chooses e_i^* such that $h(\hat{\theta}_{i,1}(a))c(e_i^*) = m$. Thus, our estimates of $\hat{\theta}_{i,1}(a|p)$ are technically estimates of $h(\hat{\theta}_{i,1}(a|p))$. We drop the h notation going forward to simplify exposition. In online Appendix C we present a closed-form solution for $h(\cdot)$ under specific distributional assumptions (e.g., normal priors and noise; see equation C.7). This yields a linear structure, which we implicitly utilize to interpret our results: differences in average estimates of $h(\hat{\theta}_{i,1}(a|p))$ across treatments are directly proportional to differences in average expectations across treatments.

Assuming an additive error structure, equation (9) suggests the following regression model:

$$(10) \qquad \log(e_i) \; = \; \beta_0 \log(m) + \sum_{j=1}^{6} \beta_j \big[\mathbb{D}_i(treatment) \times \mathbb{I}_i(noise)\big] + \delta_i,$$

where $\mathbb{D}_i(treatment)$ is a dummy variable for each treatment (*control*, *coin-flip*, or *high-probability*) and $\mathbb{I}_i(noise)$ is an indicator variable designating whether the person ultimately faced the task with noise. Variation in payouts, m, delivers identification of the curvature parameter, γ, and variation in treatment crossed with task assignment delivers identification of $\hat{\theta}_1(a|p)$. Thus, mapping equation (9) onto our econometric specification, we find the parameters of interest are $\gamma = 1/\beta_1$ and $\hat{\theta}_1(a|p) = \exp(-\beta_j/\beta_0)$. For example, in order to estimate $\hat{\theta}_1(h|p = 1)$, the average belief of participants in the *control* + *noise* subgroup, we combine the coefficient on $\mathbb{D}_i(control)\mathbb{I}_i(noise)$ with the coefficient on $\log(m)$ as prescribed above. We estimate this model using two-limit Tobit regressions with random effects at the individual level, where standard errors are computed using the delta method. This estimation technique is appropriate given that observed WTW is censored at a minimum value of 0 tasks and a maximum value of 100 and we have five observations for each person.

Table 3, column 1 presents the estimates of the baseline specification in equation (10). We find support for our model of misattribution: perceived effort cost is increasing in the probability of facing the bad task. For ease of interpretation, the rows of Table 3 (beyond the first) are ordered to match the predictions of our model. These rows demonstrate that when participants formed their initial impressions immediately after an unfavorable coin flip, they acted as if they formed more pessimistic views of the underlying task than those who faced near-certain task assignment ($\hat{\theta}_1(h|0.5) - \hat{\theta}_1(h|0.99) = 0.0142$; $\chi^2(1) = 4.22$, $p = 0.040$) or faced no uncertainty prior to task assignment ($\hat{\theta}_1(h|0.5) - \hat{\theta}_1(h|1) = 0.0149$; $\chi^2(1) = 4.27$, $p = 0.039$). Conversely, when participants formed their initial impressions after a favorable coin flip, they acted as if they formed more optimistic views of the underlying task (i.e., of $\theta(l)$) than those who faced near-certain task assignment ($\hat{\theta}_1(l|0.5) - \hat{\theta}_1(l|0.01) = -0.0087$; $\chi^2(1) = 4.06$, $p = 0.044$) or faced no uncertainty prior to task assignment ($\hat{\theta}_1(l|0.5) - \hat{\theta}_1(l|0) = -0.0064$; $\chi^2(1) = 2.49, p = 0.115$).

It is worth noting that we estimate $\gamma = 1.197(0.017)$; thus, we can reject a linear cost function despite the linear appearance of the aggregate data in Figure 3.[28] For robustness, column 2 of Table 3 controls for demographics (age, gender, and income) and for the time spent completing the first session, which we view as a coarse proxy for subjective task difficulty. Finally, column 3 drops participants whose WTW did

[28] As a robustness check, we estimated a model like that of column (1) but introduced a more flexible cost function that allowed γ to depend on whether the person faced the noise or no-noise task. This did not change the qualitative results. Moreover, in that analysis we fail to reject the null hypothesis $H_0 \colon \gamma(h) = \gamma(l); \chi^2(1) = 0.19; p = 0.66$.

Table 3—Parametric Analysis, Experiment 1

	Dependent variable: $\log(e_i)$ Estimated with Tobit regression		
	(1)	(2)	(3)
Cost curvature parameter, γ	1.197 (0.017)	1.197 (0.017)	1.157 (0.015)
$\hat{\theta}_1(\text{noise} \mid p = 0.5)$	0.068 (0.007)	0.063 (0.013)	0.066 (0.013)
$\hat{\theta}_1(\text{noise} \mid p = 0.99)$	0.050 (0.005)	0.050 (0.009)	0.053 (0.009)
$\hat{\theta}_1(\text{noise} \mid p = 1)$	0.053 (0.005)	0.049 (0.008)	0.050 (0.009)
$\hat{\theta}_1(\text{no noise} \mid p = 0)$	0.041 (0.004)	0.038 (0.007)	0.040 (0.007)
$\hat{\theta}_1(\text{no noise} \mid p = 0.01)$	0.043 (0.004)	0.041 (0.007)	0.043 (0.007)
$\hat{\theta}_1(\text{no noise} \mid p = 0.5)$	0.035 (0.003)	0.032 (0.006)	0.035 (0.007)
$H_0 : \hat{\theta}_1(\text{noise} \mid p = 0.5) = \hat{\theta}_1(\text{noise} \mid p = 0.99)$	$\chi^2(1) = 4.22$ $p = 0.040$	$\chi^2(1) = 2.87$ $p = 0.090$	$\chi^2(1) = 2.66$ $p = 0.103$
$H_0 : \hat{\theta}_1(\text{no noise} \mid p = 0.5) = \hat{\theta}_1(\text{no noise} \mid p = 0.01)$	$\chi^2(1) = 4.06$ $p = 0.044$	$\chi^2(1) = 4.82$ $p = 0.028$	$\chi^2(1) = 3.37$ $p = 0.066$
Observations	4,015	4,015	3,470
Clusters	803	803	693
Demographics and Session 1 length controls	No	Yes	No
Restricted to monotonic sample	No	No	Yes

Notes: Recall that p in the left column refers to the ex ante probability of completing the task with noise. Standard errors are clustered at the individual level and recovered via delta method. Eighteen observations are left censored and 43 are right censored in the main sample; 11 are left censored and 43 are right censored in the monotonic sample.

not weakly increase across the five payment levels. This drops a significant portion of the sample, but the point estimates of our effect remain similar.[29]

Finally, following our theoretical framework, we provide a back-of-the-envelope calculation of the parameters of interest. From equation (3), notice that $\left(\frac{\eta - \hat{\eta}}{1 + \hat{\eta}}\right) \equiv \kappa^G$ and $\left(\frac{\eta - \hat{\eta}}{1 + \hat{\eta}\lambda}\right) \equiv \kappa^L$ capture the extent to which misattribution distorts the encoded values of gains and losses, respectively; absent misattribution, $\kappa^G = \kappa^L = 0$. Using simple arithmetic on the results in Table 3, we find a large and asymmetric effect of misattribution: $\kappa^G = 1.1$ and $\kappa^L = 2.6$.[30] Although this calculation requires additional assumptions, it suggests that the aggregate results in Table 2 may be masking significant loss aversion, which the structural analysis in Table 3 helps us recover.

[29] A total of 111 responses were nonmonotonic. Although we observe a seemingly high number of such responses, we believe that our elicitation method (slider) was conducive to small mistakes.

[30] See online Appendix G for details on the theoretical derivation of these results and the underlying assumptions.

D. *Replication*

As noted above, our three treatments were not fully randomized: the *high-probability* treatment was run about a month after the other two. In order to allay any concerns about this driving the differences between the *coin-flip* and *high-probability* groups and to provide additional evidence overall, we ran an exact replication of Experiment 1 in May 2021 with full randomization across treatment arms. We recruited participants using the same platform and same recruitment strategy as before, with 903 total participants ($N = 796$ in our analysis sample, which followed the same exclusion criteria as before). We present the full results from this replication in Appendix A; here, we briefly overview our findings.

Critically, we find a significant effect of initial expectations on WTW when participants faced the noiseless task. Workers were significantly more willing to work when assigned by coin flip versus the high-probability assignment ($p = 0.0425$ for difference). We do not find a statistically significant difference between the *coin-flip* and *high-probability* groups when facing the noisy task, but the result is directionally consistent with our initial results ($p = 0.1352$ for difference).

Our inability to detect a significant difference between *coin-flip* + *noise* and *high-probability* + *noise* may stem from the following: across the board, we observe a marked decrease in WTW as compared to the original Experiment 1 (approximately 4.6 tasks). Since a compression of WTW toward the bottom of the response scale diminishes our statistical ability to detect differences, we ran an additional analysis to help account for this: we pooled the results across the replication and the main study and included a fixed effect for the replication.[31] We then compared the average WTW across groups. In doing so, we find significant differences between the *coin-flip* and *high-probability* groups regardless of task assignment ($p = 0.0173$ for difference when facing noisy task and $p = 0.0036$ for difference when facing the noiseless task; see Table A3 in Appendix A for details). Moreover, despite the caveats above, the magnitude of the effect we observe in the replication—a distortion of WTW around 20 percent—is similar to that we observed in Experiment 1.

III. Experiment 2

In this section we present our within-subject experiment, which was conducted at the Harvard Decision Science Lab. We first describe the design, in which we elicit WTW twice over the span of a week. This design allowed us to firmly set participants' expectations before they entered the second session. We then extend our theoretical setup from Experiment 1 to derive predictions for this new setting. Finally, we analyze the experimental data. Experiment 2 demonstrates a similar effect to that of Experiment 1 but additionally suggests that misattribution dynamically distorts beliefs across the two sessions.

[31] Note that we see similar but slightly diminished variability in WTW in the replication relative to the original sample (see Table 7 and Table 3, respectively).

A. *Design*

We recruited participants $(N = 87)$ from the Harvard student body for a two-session experiment, with sessions separated by a week. A total of 9 groups completed 18 sessions over the course of 1 month. Participants were paid \$7 for successfully completing each of two sessions in addition to any earnings from their choices. To minimize attrition, we paid participants only after they completed both sessions.

Before specifying the details of Experiment 2, we first provide a broad overview of how the design differs from Experiment 1. In the first session each participant was assigned via coin flip to work on one of two tasks. Each participant then returned one week later to work on that same task in a second session. To ensure that participants did not perceive any uncertainty when entering the second session, we instructed them ahead of time that their coin flip in the first session would apply to both sessions, and we sent them an email reminder of their coin-flip outcome approximately two days before their second session. Thus, participants faced uncertainty over their task assignment in the first session, but not in the second. Critically, we measured participants' WTW in both sessions of Experiment 2, and the change in WTW across sessions allows us to identify misattribution.

During both sessions, participants worked on a real-effort task similar to that of Augenblick, Niederle, and Sprenger (2015) and Augenblick and Rabin (2019): "transcribing" handwritten Greek and Russian letters.[32] Each trial of the task consisted of a string of 35 handwritten characters; participants "transcribed" each character by clicking the matching letter from a foreign alphabet. (See Figure 3 for a screenshot.) Each session had the same structure: participants first completed an initial learning phase, which consisted of five mandatory trials, and then we elicited their WTW on additional trials for a bonus payment.

As in the coin-flip condition of Experiment 1, we presented each participant with two variants of the task—a noisy version and a noiseless one. In both variants, participants wore headphones while completing transcriptions. In the noisy version, the annoying noise from Experiment 1 played through the headphones (calibrated to roughly 70–75 decibels) during the transcriptions. In the noiseless version, no sound played through the headphones.

Session 1 (Coin Flip and WTW): Upon entering the experiment, all participants were told that they faced a one-in-two chance of being assigned the noisy task versus the noiseless one. Participants then read the initial instructions, which included an interactive sample of the transcription task and an eight-second sample of the annoying noise (repeatable if desired). Next, participants flipped a coin to determine their assigned task. In order to make this uncertainty salient—and to enhance the sensation of surprise or disappointment—each participant flipped a US quarter to determine their assignment. Immediately after the coin flip, participants completed

[32] Although our task mimics that of Augenblick, Niederle, and Sprenger (2015), we used different visual stimuli, which ended up being easier to transcribe. Participants in our study needed 40 seconds on average to complete one trial, while participants in the first week of Augenblick, Niederle, and Sprenger's study needed 54 seconds on average.

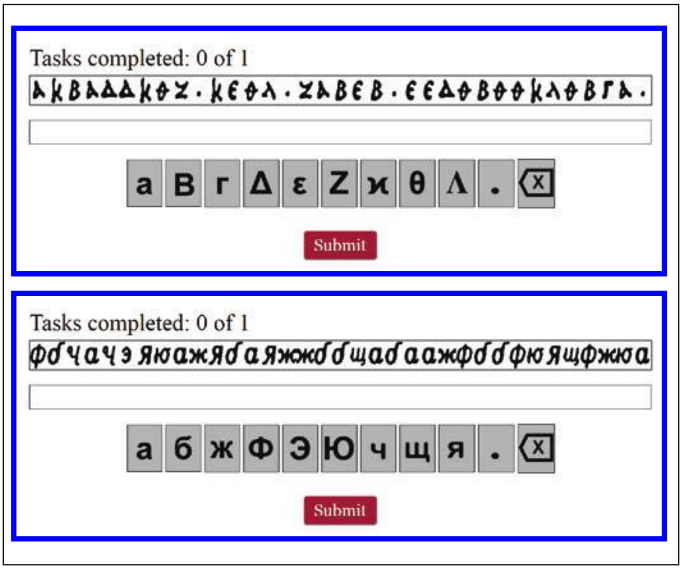


Figure 3. Screenshot of the Transcription Task from Experiment 2

Notes: Participants clicked the gray button that matched the handwritten letter to "transcribe" the text. Participants were required to achieve 80 percent accuracy to advance to the next transcription. Each participant randomly faced one of the two depicted alphabets (Greek or Cyrillic) during their first session and faced the other during their second session.

five mandatory trials of their assigned task. 44 participants were ultimately assigned the noiseless task, while 43 faced the noisy one.

After completing the initial learning phase in Session 1, subjects were given the option to complete additional trials for a bonus payment. We asked each participant how many additional tasks they were willing to complete for each of five payments: $\{\$4, \$8, \$12, \$16, \$20\}$. Participants responded by using a slider to select any integer $e \in \{0, \ldots, 100\}$, and we used the BDM mechanism to elicit these responses.

Session 2 (Second Elicitation of WTW): Upon returning to the second session of the experiment, each participant first completed five mandatory trials of the same task variant they faced in Session 1 (i.e., noisy or noiseless). After the five mandatory trials, we elicited participants' WTW on additional trials of that task. The experiment concluded after participants completed any additional trials. Subjects were paid only upon completion of both sessions.

Finally, we note that participants faced different alphabets across the two sessions. Half of them faced a Greek alphabet during the first session and Cyrillic during the second, while the other half faced them in the opposite order. We introduced this minor variation in the task so that participants could plausibly form different perceptions of the task across sessions and hence update their WTW. This was intended to help reduce anchoring or consistency effects: since participants faced a somewhat different task in the second session, they may have been less likely to answer exactly the same as they did during the first session. It also provided subjects with a potential cover story for changing their responses across sessions.

B. *Theoretical Predictions*

Building from the same theoretical setup from Experiment 1 (Section IIB), we now sketch how our predictions extend to this within-subject design.

As in our theoretical discussion of Experiment 1, we assume that a participant's WTW was shaped by her experience with the task. Unlike in Experiment 1, however, a participant in this setting had two separate experiences with the task (across the two sessions), and we elicited her WTW after both. Since we incentivized WTW, some participants were randomly assigned to complete additional tasks in the first session as a result of their WTW and chance (inherent in the BDM mechanism). We focus our theoretical analysis here on participants who did not complete additional tasks in the first session; we relax this focus in our empirical analysis.

As in Experiment 1, suppose that consumption utility from each initial learning phase $t \in \{1,2\}$, in which the participant completes five trials of her assigned task $a \in \{h,l\}$, is given by $v^e_{i,t} = [\theta_i(a) + \epsilon_{i,t}]c(5)$. The participant uses this as a signal to infer the value of $\theta_i(a)$, and we (indirectly) observe her beliefs over $\theta_i(a)$ through her stated WTW. We assume that participants, on average, hold unbiased expectations about the difficulty of the tasks.

Given this assumption, rational learning without reference-dependent preferences immediately predicts that participants' WTW will remain constant (on average) across the two periods. In contrast, learning with reference dependence but without misattribution can lead participants to systematically change their WTW across periods. As we show in detail in online Appendix E, reference dependence absent misattribution creates an incentive for those facing the noisy task to decrease effort over time and for those facing the noiseless task to increase it.[33]

Misattribution introduces an opposing effect that leads those assigned the noisy task to increase effort between Sessions 1 and 2 and those assigned the noiseless task to decrease effort. For a participant who faces the noisy task, her first signal incorporates a sense of disappointment: in Session 1, she anticipates a 50 percent chance of facing the better task. But her second signal comes with less disappointment: in Session 2, she expects the worse task. Put differently, the participant's first experience falls short of expectations by a greater amount than the second and is thus remembered as worse. Thus, on average, a participant assigned the noisy task $(a = h)$ will encode $\hat{v}^e_{i,1} < \hat{v}^e_{i,2}$. In contrast, a participant assigned to the noiseless $(a = l)$ task will (on average) encode values such that $\hat{v}^e_{i,1} > \hat{v}^e_{i,2}$, since the first

[33] These incentives arise if a participant is loss averse and her reference point at the time of her first decision is still based on the expectations she held prior to learning the outcome of the coin flip. If either of these conditions is not met, then reference dependence absent misattribution has no effect on behavior, resulting in effort choices that are, on average, constant across the two periods. If, instead, both of these conditions hold, then reference dependence can generate systematic changes in effort across periods when the participant forms forward-looking strategies aimed at mitigating losses. By planning to exert similar effort (in terms of cost) in the first period regardless of the outcome of the coin flip, a participant can avoid feeling large sensations of disappointment no matter which task she is assigned. If the participant's expectations then adapt to her assigned task by the second period, she no longer has incentive to equalize effort across contingencies. Thus, relative to the first period, she will increase her WTW if she were assigned the less costly (noiseless) task and decrease it if she were assigned the more costly (noisy) task. See online Appendix E for further details.

signal incorporates a sense of elation from the coin flip, but the second signal comes with less (if any) such elation. Thus, participants assigned to the noisy task will, on average, update such that $\hat{\theta}_{i,1}(h) > \hat{\theta}_{i,2}(h)$, while those assigned the noiseless task will update such that $\hat{\theta}_{i,1}(l) < \hat{\theta}_{i,2}(l)$.

Since misattribution acts in opposition to reference dependence absent misattribution, the behavioral implications of these beliefs depend on which force is stronger. If misattribution is relatively strong, then the distorted beliefs described above will be reflected in effort choices. This is the main prediction we empirically test: WTW of participants assigned the noisy task will increase across sessions, while WTW of those assigned the noiseless task will decrease.

Finally, since Experiment 2 involves two elicitations of WTW, it allows us to potentially observe a dynamic contrast effect predicted by misattribution. To illustrate, consider a participant who is assigned to the noiseless task. Since her stated WTW in Session 1 is based on her overly optimistic perception of the underlying effort cost, it is biased upward relative to the case without misattribution. This follows from the theoretical discussion of Experiment 1. In Experiment 2, however, the participant has a second experience with her assigned task in the learning phase of Session 2, and this experience tends to come with an unpleasant surprise: since her prior expectations (stemming from her Session 1 experience) are inflated, her second experience—now devoid of the positive surprise from the coin flip—will not live up to those unrealistic expectations. This typically bad experience pushes her estimated cost upward, reducing her WTW in the second session. If this contrast effect between the first and second rounds is sufficiently strong, then the participant's revealed WTW will decrease over the two sessions. Similar logic implies that a misattributor assigned to the noisy task will increase her effort across sessions: her second experience with the task will typically surpass her overly pessimistic expectations formed in the first session, and this positive surprise will increase her WTW. We discuss some (suggestive) evidence for such a contrast effect in the results that follow.

Discussion of Assumptions: Our theoretical discussion above relies on unbiased priors in aggregate. If participants' priors were systematically biased in a specific direction—namely, they significantly overestimate the disutility of the task with noise and underestimate the disutility of the task without noise—then changes in WTW across sessions may result from rational learning. We believe our assumption of reasonably well-calibrated priors is justified from the experimental design: participants were exposed to both versions of the task before commencing work.

Additionally, our predictions assume that reference points (at least partially) adapted to the assigned task before Session 2. This seems warranted given that participants knew about their task assignment a week in advance and there was no added uncertainty in the second. Furthermore, participants were reminded by email midway through the week. Before beginning Session 2, all participants were required to verbally state which task they had faced in Session 1, and all participants did so successfully. This suggests that the assignment was salient and memorable.

C. *Results*

Our primary analysis considers participants who completed both sessions. Thus, our data come from 71 participants. For completeness, we present an analysis of participant attrition in Table B6.

We first present nonparametric analyses demonstrating that WTW systematically changes over time depending on the resolution of the coin flip in Session 1. We then estimate the parameters of a reduced-form model similar to Experiment 1 but utilizing the within-subject nature of this design. Although our theoretical discussion above focused on participants who did not complete additional tasks in the first session, we show that our results hold whether or not this assumption is maintained. We conclude by discussing reciprocity toward the experimenter and mood effects, both of which might plausibly explain the results in Experiment 1. We argue that these effects are constrained by our experimental designs and provide further evidence that favors misattribution as the underlying mechanism.

Nonparametric Analysis.—Sessions 1 and 2 of this experiment mirrored the *coin-flip* and *control* treatments from Experiment 1, respectively. The difference across sessions stemmed from an uncertain task assignment in the first session changing to a fully anticipated assignment in the second. Following the analysis of Experiment 1, Table 4 presents participants' average WTW—averaged over the five payment levels—in each of these two sessions.

We present aggregate results in columns 1–4 of Table 4; however, these obscure important within-subject variation. Examining within-subject changes in WTW work, we find significant differences across Sessions 1 and 2 (see columns 5–6 of Table 4). Consistent with our theoretical predictions, participants who faced the noiseless task tended to decrease their WTW across sessions while those assigned the noisy task tended to increase it. When assigned the noiseless task, participants were (on average) willing to complete 7.5 more tasks in Session 1 than in Session 2 ($p = 0.004$). In contrast, when assigned the noisy task, participants were (on average) willing to complete 4.3 fewer tasks in Session 1 than in Session 2 ($p = 0.014$). Figure 4 depicts this result by plotting the density of $e_{i,1} - e_{i,2}$ for each task averaged over the five payment levels.[34]

To provide an intuition for the magnitude of this effect, we consider a hypothetical firm paying workers to complete 25 transcriptions (as we did in discussing Experiment 1). To incent the average participant to complete 25 noiseless transcriptions, the firm would have to pay $7.75 right after the worker formed her initial impression (i.e., just after the positive outcome of the coin flip); this would increase to $11 once her assessment of the task is no longer confounded with a sense of elation. In contrast, a firm would have to pay $12 to incent the average participant to do 25 noisy transcriptions right after she formed her initial impression (i.e., just after the negative outcome of the coin flip); this would decrease to $10.50 once her

[34] We present these densities here using kernel smoothing (Epanechnikov kernel); in online Appendix B we show the raw data in Figure B2 and unsmoothed histograms in Figure B3.

Table 4—Baseline Results, Experiment 2

	Session 1		Session 2		$(e_{i,1} - e_{i,2})$	
Variable	noise = 0	noise = 1	noise = 0	noise = 1	noise = 0	noise = 1
Willingness to work (WTW)	31.391 (3.680)	25.927 (3.526)	25.967 (3.006)	26.405 (3.575)	7.472 (2.397)	−4.254 (1.653)
Observations	215	220	180	185	360	370

Notes: Standard errors (in parentheses) are clustered at the individual level. The difference between columns 1–3 is significant at $p = 0.026$, columns 2–4 at $p = 0.865$, and columns 5–6 at $p < 0.001$. Columns 5–6 both significantly differ from zero at $p = 0.004$ and $p = 0.014$, respectively.

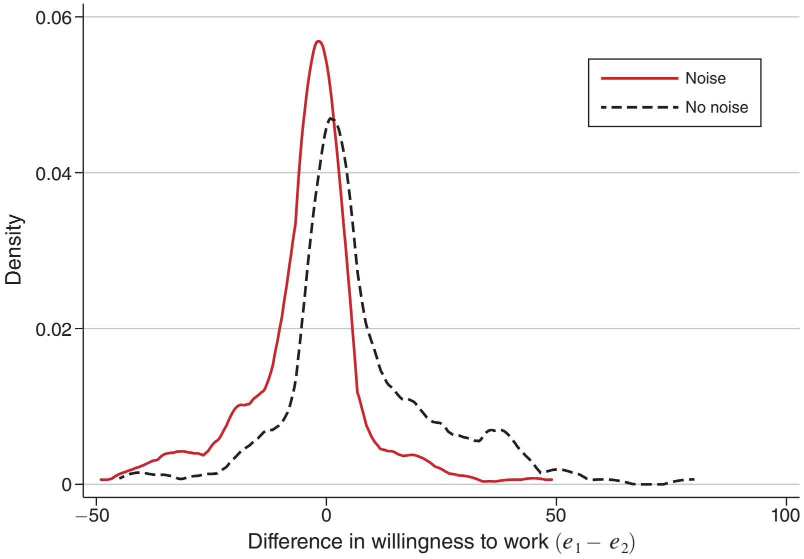


Figure 4. Kernel Density of the Difference in Willingness to Work (WTW) between the First and Second Sessions, Separated by Task Faced

Notes: Each underlying observation from this figure is the change in a participant's WTW for a fixed payment between Sessions 1 and 2 of the experiment. The dotted black curve represents participants who were assigned to the no-noise task; the solid red curve represents participants who were assigned to the noisy task.

assessment of the task is no longer confounded with a sense of disappointment. These effect sizes have similar magnitudes to those in Experiment 1.[35]

Parametric Analysis.—We now present a more structured approach, following the logic in Section IIB. Given that the experiment closely follows the approach from Experiment 1, the decision problem in each session can be modeled in the same way as the previous experiment. Thus, adopting our previous notation, equation (9) implies that for each period, $\log(e^*_{i,t}) = \frac{\log(m)}{\gamma} - \frac{\log[\hat{\theta}_{i,t}(a|p)]}{\gamma}$. Since we

[35] There are two important caveats to consider before comparing this calibration exercise to the results of Experiment 1. First, because the task in Experiment 2 was more time consuming than that of Experiment 1 and because these participants were paid more, the magnitudes of payments are quite different across experiments. Second, because the sample size in Experiment 2 is much smaller, the estimated effect size is very imprecise.

observe WTW for each individual in two periods, we examine the difference $\log(e_{i,1}) - \log(e_{i,2})$. Our econometric model is thus

$$[\log(e_{i,1}) - \log(e_{i,2})] = \beta \mathbb{I}_i(\text{noise}) + \epsilon_i. \tag{11}$$

From this specification we can recover aggregate estimates $\frac{\hat{\theta}_1(a|p)}{\hat{\theta}_2(a|p)} = \exp(-\gamma\beta)$. Since the cost curvature parameter γ is not identified in this specification, we separately model the first session (following equation (10)) to generate an in-sample estimate of $\gamma \approx 1.19$; note this falls close to our estimate from Experiment 1.[36] We use this value (and the equation above) to numerically estimate the ratio of interest.

As in Experiment 1, we estimate equation (11) using a random-effect Tobit model. The results are shown in Table 5. We find that $\frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})} = 1.36$ (column 1 of Table 5). This is very close to the analogous ratio we found in Experiment 1, $\frac{\hat{\theta}(\text{noise}\,|\,\text{coin flip})}{\hat{\theta}(\text{noise}\,|\,\text{control})} = 1.28$ (column 1 of Table 3). Likewise, the ratio $\frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})} = 0.80$ falls close to $\frac{\hat{\theta}(\text{no noise}\,|\,\text{coin flip})}{\hat{\theta}(\text{no noise}\,|\,\text{control})} = 0.84$. Thus, in both experiments and across all specifications, we find that uncertain assignment via the coin flip distorts WTW in the range of approximately 17–40 percent relative to certain assignment.

Discussion: As with Experiment 1, we suspect attrition is an unlikely explanation for our results. In Table B6 (in online Appendix B) we demonstrate that attrition is independent of whether a participant faced noise, their average WTW in Session 1, and whether the participant first faced Cyrillic or Greek. All participants who completed extra tasks in Session 1 returned for Session 2.

However, a potential concern is that those participants who completed additional tasks during the first session may have held systematically different beliefs entering Session 2 than those who did not complete additional tasks. Theoretically, comparing participants who completed additional tasks with those who did not may introduce complications, as the two groups accumulated different amounts of experience. One-third of participants completed additional tasks in the first session, and we included these participants in our analyses above. We explore whether this distinction matters empirically in Table B5 (in online Appendix B). There we demonstrate that our qualitative results from Table 4 are robust to controlling for extra tasks using OLS, controlling for extra tasks via two-stage least squares (utilizing the BDM randomness for identification), and simply dropping participants who completed extra tasks. While statistical power decreases when dropping participants, our estimates remain similar.

A seemingly compelling alternative explanation for our results (across both experiments) is that they stem from reciprocity toward the experimenter: after a positive

[36] As in Experiment 1, we tested whether $\gamma(h) = \gamma(l)$. Using data from the first session only, we fail to reject the null $H_0 : \gamma(h) = \gamma(l); \chi^2(1) = 0.00, p = 0.957$.

Table 5—Parametric Analysis, Experiment 2

	Dependent variable: $\log\left(\frac{e_{i,1}}{e_{i,2}}\right)$ Estimated via OLS (1)
Estimated ratio $\frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})}$	1.359 (0.127)
Estimated ratio $\frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})}$	0.800 (0.091)
$H_0: \frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})} \geq 1$	$\chi^2(1) = 7.98$ $p = 0.005$
$H_0: \frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})} \leq 1$	$\chi^2(1) = 4.85$ $p = 0.028$
Observations	353
Clusters	71

Notes: Standard errors (in parentheses) are clustered at the individual level. Dropped observations result from taking logs under the assumption that $\omega = 0$. Each estimate $\frac{\hat{\theta}_1(a)}{\hat{\theta}_2(a)}$ is derived assuming that $\gamma = 1.19$.

surprise, participants may have "rewarded the experimenter" with high WTW, while after a negative surprise they may have "punished the experimenter" with low WTW. Note that for reciprocity to explain our results from Experiment 1, this desire to reciprocate must persist well over eight hours; for it to explain our results from Experiment 2, this desire must disappear over a week.[37] The evidence from Experiment 2, however, points toward a different mechanism. Specifically, WTW in Session 2 suggests more than a simple fading of reciprocity: we detect no difference in WTW between the noise and no-noise groups in Session 2, where a difference would be natural absent misattribution (see columns 3 and 4 in Table 4). We interpret this lack of difference in WTW across tasks as suggestive evidence of the contrast effect predicted by our model.

Similar arguments to those above may speak against other mood effects beyond the desire to reciprocate. Namely, for a coin-flip-induced mood to explain our results in both Experiments 1 and 2, it must persist over a few days but then disappear before a week. Furthermore, the strength of this mood effect must depend on the probability of the facing each task. Finally, such a mood effect would not explain the similar WTW across groups in Session 2, as discussed above.

[37] The differential WTW from Experiment 1 across the *high-probability* and *coin-flip* treatments further suggests that, fixing the outcome received, the ex ante probability of receiving that outcome must have altered the degree to which a person was motivated by reciprocity. We are unaware of a model or direct evidence of this form of reciprocity, but we concede that it is plausible.

IV. Conclusion

In this paper we provide evidence consistent with a specific form of attribution bias wherein people fail to account for their reference-dependent utility when learning about an unfamiliar real-effort task. In a series of experiments, we manipulated participants' expectations prior to their initial experiences. These initial expectations shaped participants' WTW both in the moment (Experiment 2) and hours later, when participants' task assignment was fully anticipated (Experiment 1). We now briefly discuss some benefits of our experimental design, some reasons for caution in interpreting our results, and directions for future research.

By focusing on the extensive margin (i.e., whether to complete additional work) rather than the intensive margin (i.e., how hard to work), our design sidestepped a challenge highlighted in the literature: productivity is rather inelastic (DellaVigna et al. 2022). By allowing people to choose how many tasks to do (rather than, say, working over a fixed period of time), our design was well powered to detect attribution bias and may serve as a guide for future experiments.

Our model predicts that loss-averse participants will form more distorted perceptions of bad outcomes than good ones. In our first experiment we find weak but suggestive evidence of loss aversion reflected through misattribution: the average WTW for those assigned to the noisy task by chance was more distorted than the willingness to work of those assigned to the no-noise task by chance. However, both our replication and our aggregate results in Experiment 2 do not demonstrate signs of loss aversion. It is possible that we are unable to see loss aversion in Experiment 2 because of an overall diminished WTW (among all participants) in the second session. Additionally, asymmetric distortion of bad outcomes (relative to good outcomes) may be difficult to observe in both Experiment 2 and our replication of Experiment 1 due to compression of the response scales at low values. With low WTW, participants may utilize the response scale differently than those with higher WTW, which may make detecting loss aversion more difficult. Loosely, choices may be more finely tuned near the bottom of the scale and hence less susceptible to big changes. As loss aversion is central to models of reference-dependent preferences, future work should address the extent to which losses drive asymmetric belief updating.

More broadly, our results suggest that organizations (e.g., firms or political parties) can shape short-run impressions by managing expectations. For instance, our results suggest that employees would form more favorable impressions of undesirable tasks if they knew well ahead of time that they would have to complete them. This accords with evidence from firms that give realistic job previews prior to hiring. As Phillips (1998) shows, employees who face a realistic job preview perform better and are less likely to leave the job than their peers who do not experience a job preview. Misattribution may provide the underlying mechanism for such effects.

Appendix A. Experiment 1 Replication

In this Appendix we present results from our replication study discussed in Section IID. Given that the objective was to eliminate concerns about nonrandom

Table A1—Demographics and Summary Statistics, Experiment 1 Replication

	Control		Coin flip		High prob.	
Variable	noise = 0	noise = 1	noise = 0	noise = 1	noise = 0	noise = 1
Age	40.84	37.60	39.11	42.23	38.78	38.01
	(13.11)	(10.66)	(12.74)	(11.50)	(11.16)	(12.23)
1(Male)	0.509	0.510	0.528	0.514	0.524	0.553
	(0.501)	(0.502)	(0.501)	(0.502)	(0.501)	(0.499)
Income	2.863	2.703	3.120	3.007	3.028	2.787
	(1.137)	(1.212)	(1.127)	(1.202)	(1.190)	(1.246)
1(Return)	0.832	0.884	0.901	0.950	0.917	0.893
	(0.375)	(0.321)	(0.299)	(0.219)	(0.276)	(0.310)
Observations	161	155	142	140	145	150

Notes: Standard deviations are in parentheses. Income is coded as a discrete variable that takes a value from one to five, corresponding to the following income brackets: (1) less than $15,000, (2) $15,000–$29,999, (3) $30,000–$59,999, (4) $60,000–$99,999, (5) $100,000 or more.

Table A2—Baseline Results, Experiment 1 Replication

	Control		Coin flip		High prob.	
Variable	noise = 0	noise = 1	noise = 0	noise = 1	noise = 0	noise = 1
Willingness to work (WTW)	19.18	16.63	22.54	15.95	18.77	18.59
	(1.144)	(1.229)	(1.371)	(1.103)	(1.256)	(1.383)
Observations	670	675	635	655	660	660

Notes: Willingness to work is averaged over five payment levels. Standard errors (in parentheses) are clustered at the individual level. The difference between columns 3–5 is significant at $p = 0.0425$; the difference between columns 4–6 is not significant ($p = 0.1352$).

assignment to *coin-flip* versus *high-probability* treatments, we focus our discussion on results from those two groups.

We first present demographic characteristics in Table A1 for comparison to Table 1 in the main text. Our replication sample is more male and slightly older than our original sample. We note that Table A1 suggests similar levels of attrition across the replication and the original experiment.

We implemented the identical data-cleaning procedures as in Experiment 1 when forming our primary dataset. We removed participants who did not answer all five elicitations of WTW (zero participants); who stated a WTW equal to the maximum amount (100 tasks) for every payment level, which prevented us from estimating their responsiveness to payment (three participants); or who did not return for the second session (and whose WTW we therefore did not measure). With this set of restrictions, we are left with a sample of 796 participants. We present the main results in Table A2, which is a direct analogue to the original Table 2. We discuss these results in the main text. Figure A1 also shows the labor supply curves for the two critical treatments.

Finally, we present a simple regression analysis that pools the data from the original experiment and the replication. We include a fixed effect for all of the replication data and cluster standard errors at the individual level. As before, we utilize interval

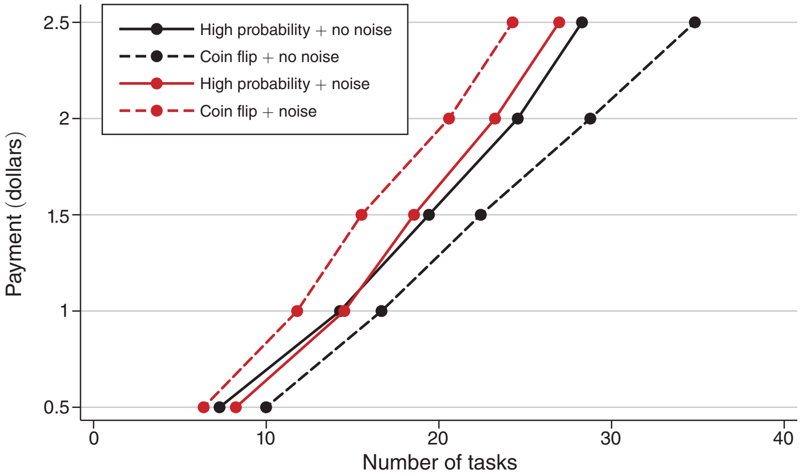


Figure A1. Labor Supply Curves across Key Treatments

Note: Each point represents the average willingness to work (WTW) for a fixed payment as elicited using the BDM mechanism.

Table A3—Pooled Results: Replication + Main Experiment

	Dependent variable: WTW estimated with Tobit Regression
1(coin flip + noise)	19.017 (1.001)
1(high probability + noise)	22.130 (1.018)
1(control + noise)	21.677 (1.122)
1(control + no noise)	23.946 (1.001)
1(high probability + no noise)	23.704 (0.994)
1(coin flip + no noise)	27.832 (1.176)
1(replication)	−4.540 (0.785)
H_0: **1**(coin flip + noise) = **1**(high probability + noise)	$\chi^2(1) = 5.67$ $p = 0.0173$
H_0: **1**(coin flip + no noise) = **1**(high probability + no noise)	$\chi^2(1) = 9.84$ $p = 0.0036$
Observations	7,970
Clusters	1,594

Notes: Standard errors (in parentheses) are clustered at the individual level and recovered via delta method. Seventy observations are left censored and 76 are right censored.

regression, since our data are censored below at 0 and above at 100. We include two special rows to highlight the hypothesis tests that compare the *coin-flip* and *high-probability* treatments; we discuss these results in the main text.

REFERENCES

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. “Reference Points and Effort Provision.” *American Economic Review* 101 (2): 470–92.

Adhvaryu, Achyuta, Anant Nyshadham, and Huayu Xu. 2020. “Hostel Takeover: Living Conditions, Reference Dependence, and the Well-Being of Migrant Workers.” Unpublished.

Allen, Eric J., Patricia M. Dechow, Dechow G. Pope, and George Wu. 2017. “Reference-Dependent Preferences: Evidence from Marathon Runners.” *Management Science* 63 (6): 1657–72.

Augenblick, Ned, and Matthew Rabin. 2019. “An Experiment on Time Preference and Misprediction in Unpleasant Tasks.” *Review of Economic Studies* 86 (3): 941–75.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. “Working over Time: Dynamic Inconsistency in Real Effort Tasks.” *Quarterly Journal of Economics* 130 (3): 1067–1115.

Backus, Matthew, Thomas Blake, Dimitriy Masterov, and Steven Tadelis. 2021. “Expectation, Disappointment, and Exit: Evidence on Reference Point Formation from an Online Marketplace.” *Journal of the European Economic Association* 20 (1): 116–49.

Bell, David E. 1985. “Disappointment in Decision Making under Uncertainty.” *Operations Research* 33 (1): 1–27.

Bernheim, B. Douglas, and Charles Sprenger. 2020. “On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting.” *Econometrica* 88 (4): 1363–1409.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. “Are CEOs Rewarded for Luck? The Ones without Principals Are.” *Quarterly Journal of Economics* 116 (3): 901–32.

Boulding, William, Ajay Kalra, Richard Staelin, and Valarie A. Zeithaml. 1993. “A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions.” *Journal of Marketing Research* 30 (1): 7–27.

Brownback, Andy, and Michael A. Kuhn. 2019. “Understanding Outcome Bias.” *Games and Economic Behavior* 117: 342–60.

Buffat, Justin, and Julien Senn. 2015. “Testing the Speed of Adjustment of the Reference Point in Models of Expectation-Based Reference-Dependent Preferences.” Unpublished.

Bushong, Benjamin, and Tristan Gagnon-Bartsch. 2023. "Replication Data for: Attribution Bias and Reference Dependence: Evidence from Real-Effort Experiments.” American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E149701V1.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler. 1997. “Labor Supply of New York City Cabdrivers: One Day at a Time.” *Quarterly Journal of Economics* 112 (2): 407–41.

Card, David, and Gordon B. Dahl. 2011. “Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior.” *Quarterly Journal of Economics* 126 (1): 103–43.

Chambers, Christopher P., and Paul J. Healy. 2012. “Updating towards the Signal.” *Economic Theory* 50: 765–86.

Cole, Shawn, Andrew Healy, and Eric Werker. 2012. “Do Voters Demand Responsive Governments? Evidence from Indian Disaster Relief.” *Journal of Development Economics* 97 (2): 167–81.

Crawford, Vincent P., and Juanjuan Meng. 2011. “New York City Cabdrivers’ Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income.” *American Economic Review* 101 (5): 1912–32.

de Quidt, Jonathan. 2018. “Your Loss Is My Gain: A Recruitment Experiment with Framed Incentives.” *Journal of the European Economic Association* 16 (2): 522–59.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2022. “Estimating Social Preferences and Gift Exchange at Work.” *American Economic Review* 112 (3): 1038–74.

Dutton, Donald G., and Arthur P. Aaron. 1974. “Some Evidence for Heightened Sexual Attraction under Conditions of High Anxiety.” *Journal of Personality and Social Psychology* 30: 510–17.

Edmans, Alex, Diego García, and Øyvind Norli. 2007. “Sports Sentiment and Stock Returns.” *Journal of Finance* 62 (4): 1967–98.

▶**Ericson, Keith M., and Andreas Fuster.** 2011. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Quarterly Journal of Economics* 126 (4): 1879–1907.

▶**Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh.** 2022. "By Chance or By Choice? Biased Attribution of Others' Outcomes When Social Preferences Matter." *Experimental Economics* 25 (2): 413–33.

▶**Fryer, Roland G., Philipp Harms, and Matthew O. Jackson.** 2019. "Updating Beliefs When Evidence is Open to Interpretation: Implications for Bias and Polarization." *Journal of the European Economics Association* 17 (5): 1470–1501.

▶**Fudenberg, D., and D. Levine.** 2014. "Learning with Recency Bias." *Proceedings of the National Academy of Sciences* 111: 10826–29.

Gagnon-Bartsch, Tristan, and Benjamin Bushong. 2021. "Learning with Misattribution of Reference Dependence." Unpublished.

▶**Gilbert, Daniel T., and Patrick S. Malone.** 1995. "The Correspondence Bias." *Psychological Bulletin* 117 (1): 21–38.

▶**Gill, David, and Victoria Prowse.** 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review* 102 (1): 469–503.

▶**Gneezy, Uri, and John A. List.** 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica* 74 (5): 1365–84.

▶**Gonzalez, Richard, and George Wu.** 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology* 38 (1): 129–66.

▶**Haggag, Kareem, Devin G. Pope, Kinsey B. Bryant-Lees, and Maarten Bos.** 2019. "Attribution Bias in Consumer Choice." *Review of Economic Studies* 86 (5): 2136–83.

Heffetz, Ori. 2018. "Are Reference Points Merely Lagged Beliefs over Probabilities?" Unpublished.

▶**Heffetz, Ori, and John A. List.** 2014. "Is the Endowment Effect an Expectations Effect?" *Journal of the European Economic Association* 12 (5): 1396–1422.

▶**Highhouse, Scott, and Andrew Gallo.** 1997. "Order Effects in Personnel Decision Making." *Human Performance* 10 (1): 31–46.

▶**Hirshleifer, David, and Tyler Shumway.** 2003. "Good Day Sunshine: Stock Returns and the Weather." *Journal of Finance* 58 (3): 1009–32.

▶**Ho, Teck H., and Yu-Sheng Zheng.** 2004. "Setting Customer Expectation in Service Delivery: An Integrated Marketing-Operations Perspective." *Management Science* 50 (4): 479–88.

▶**Imas, Alex, Sally Sadoff, and Anya Samek.** 2017. "Do People Anticipate Loss Aversion?" *Management Science* 63 (5): 1271–84.

▶**Kahneman, Daniel, and Amos Tversky.** 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–91.

▶**Karle, Heiko, Goerg Kirchsteiger, and Martin Peitz.** 2015. "Loss Aversion and Consumption Choice: Theory and Experimental Evidence." *American Economic Journal: Microeconomics* 7 (2): 101–20.

▶**Kimball, David C., and Samuel C. Patterson.** 1997. "Living up to Expectations: Public Attitudes toward Congress." *Journal and Politics* 59 (3): 701–28.

▶**Kopalle, Praveen K., and Donal R. Lehmann.** 2006. "Setting Quality Expectations When Entering a Market: What Should the Promise Be?" *Marketing Science* 25 (1): 8–24.

▶**Kőszegi, Botond, and Matthew Rabin.** 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121 (4): 1133–65.

▶**Malmendier, Ulrike, and Stefan Nagel.** 2011. "Depression Babies: Do Macroeconomic Experiences Affect Risk-Taking?" *Quarterly Journal of Economics* 126 (1): 373–416.

▶**Markle, Alex, George Wu, Rebecca J. White, and Aaron Sackett.** 2018. "Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence." *Journal of Risk and Uncertainty* 56: 19–50.

▶**Medvec, Victoria Husted, Scott F. Madey, and Thomas Gilovich.** 1995. "When Less Is More: Counterfactual Thinking and Satisfaction among Olympic Medalists." *Journal of Personality and Social Psychology* 69 (4): 603–10.

▶**Meston, Cindy M., and Penny F. Frohlich.** 2003. "Love at First Fright: Partner Salience Moderates Roller-Coaster-Induced Excitation Transfer." *Archives of Sexual Behavior* 32 (6): 537–44.

▶**Oliver, Richard L.** 1977. "Effect of Expectation and Disconfirmation of Post-Exposure Product Evaluation: An Alternative Interpretation." *Journal of Applied Psychology* 62 (4): 480–86.

▶**Oliver, Richard.** 1980. "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions." *Journal of Marketing Research* 17 (4): 460–69.

Patterson, Samuel C., G. R. Boynton, and Ronald D. Hedlund. 1969. "Perceptions and Expectations of the Legislature and Support for It." *American Journal of Sociology* 75 (1): 62–76.

Phillips, Jean M. 1998. "Effects of Realistic Job Previews on Multiple Organizational Outcomes: A Meta-analysis." *Academy of Management Journal* 41 (6): 673–90.

Pope, Devin G., and Maurice Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review* 101 (1): 129–57.

Post, Thierry, Martijn J. van den Assem, Guido Baltussen, and Richard H. Thaler. 2008. "Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show." *American Economic Review* 98 (1): 38–71.

Prelec, Drazen. 1999. "The Probability Weighting Function." *Econometrica* 66 (3): 497–527.

Rabin, Matthew, and Joel L. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics* 114 (1): 37–82.

Ross, Lee. 1977. "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In *Advances in Experimental Social Psychology*, Vol. 10, edited by L. Berkowitz, 173–220. Amsterdam: Elsevier.

Saunders, Edward M. 1993. "Stock Prices and Wall Street Weather." *American Economic Review* 83 (5): 1337–45.

Simonsohn, Uri. 2007. "Clouds Make Nerds Look Good: Field Evidence of the Impact of Incidental Factors on Decision Making." *Journal of Behavioral Decision Marking* 20 (2): 143–52.

Simonsohn, Uri. 2010. "Weather to Go to College." *Economic Journal* 120 (543): 270–80.

Smith, Alec. 2019. "Lagged Beliefs and Reference-Dependent Preferences." *Journal of Economic Behavior and Organization* 167: 331–40.

Song, Changcheng. 2016. "An Experiment on Reference Points and Expectations." Unpublished.

Thakral, Neil, and Linh T. Tô. 2021. "Daily Labor Supply and Adaptive Reference Points." *American Economic Review* 111 (8): 2417–43.

Wenner, Lukas M. 2015. "Expected Prices as Reference Points—Theory and Experiments." *European Economic Review* 75: 60–79.

Wolfers, Justin. 2007. "Are Voters Rational? Evidence from Gubernatorial Elections." Unpublished.