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

Benjamin Bushong
Michigan State University

Tristan Gagnon-Bartsch*
Harvard University

June 24, 2021

Abstract

We demonstrate that people's impressions of a real-effort task are shaped by the elation or disappointment they felt when first working on the task. In our experiments, participants learned from experience about one of two unfamiliar tasks, one clearly more onerous than the other. We manipulated participants' initial expectations about which task they would face: some were assigned their task by chance just prior to their initial experience, while others knew in advance which task they would face. In a second session conducted hours later, we elicited their willingness to work again on their previously assigned task. Participants assigned the less-onerous task by chance were more willing to work than those who faced it with certainty; conversely, those assigned the more-onerous task by chance were less willing to work than those who faced it with certainty. These qualitative results—and the fact that differences in willingness to work were observed hours after first impressions were formed—are consistent with a form of attribution bias wherein participants wrongly ascribed sensations of positive or negative surprise to the underlying disutility of their assigned task.

JEL Classification: C91, D03.
Keywords: Attribution bias; Reference dependence; Real-effort experiment

*E-mails: bbushong@msu.edu and gagnonbartsch@fas.harvard.edu. 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, and seminar audiences at the ASSA annual meeting (2020), Boston College, Boston University School of Management, Caltech, Cornell, Harvard, HBS, Michigan, Michigan State, Norwegian School of Economics (NHH), the North American ESA Conference, Purdue, Stanford, the Stanford Institute for Theoretical Economics, Tennessee, and UCSD Rady for comments. We thank Jeffrey Yang for excellent research assistance. We thank Alexander Millner for sharing the annoying audio stimulus used in these experiments. We gratefully acknowledge financial support from the Eric M. Mindich Research Fund for the Foundations of Human Behavior.

1 Introduction

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 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

[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.

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 2 (following Gagnon-Bartsch and Bushong 2021) that guides our experimental designs. We then describe Experiment 1 (conducted online) in Section 3. 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 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% chance of either task). In a final treatment, participants were assigned to a task with near certainty (i.e., they faced a 99% chance of one task and a 1% 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: $\{$*control, coin-flip, high-probability*$\} \times \{$*noise, 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.

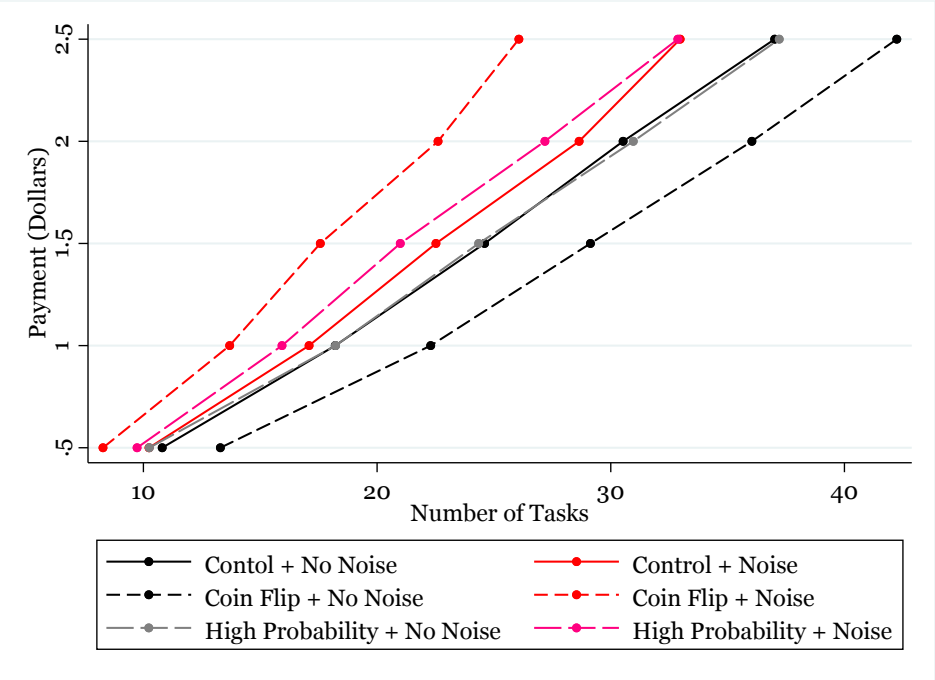

Figure 1: *Labor supply curves across treatments.* Each point represents the average WTW for a fixed payment as elicited using the BDM mechanism. Those assigned to their task after facing the most uncertainty (*coin-flip* groups) demonstrated greater WTW when assigned the noiseless task and less WTW when assigned the noisy task than either the *control* and *high-probability* 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% 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% 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* drawing different inferences from the experimental design itself.[3]

Experiment 2, presented in Section 4, 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.

[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.

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]

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 elations or disappointments loom 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]

[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 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 under-powered 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 et al. 1969; Kimball and

Along these lines, our evidence and theoretical framework provide a new lens for understanding existing empirical results. For example, Backus et al. (2020) 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 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 2003) 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) provides 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

Patterson 1997). Likewise, marketing has emphasized the role of expectations on 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.

predictions. For example, Haggag et al. do not speak to the notion of expectations management highlighted above; in contrast, this is an immediate implications 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 was the first outcome. We find suggestive evidence for such a contrast effect in Experiment 2 (see Section 4.3).

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; and 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 2020), domestic violence resulting from unexpected football losses (Card and Dahl 2011), decisions in game shows and sports (Post et al. 2008; Pope and Schweitzer 2011; Allen et al. 2015; Markle et al. 2015). 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 (2009), Abeler et al. (2011), Gill and Prowse (2012) and Karle et al. (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 suggest 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.

2 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 3.2 and 4.2 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 utility is distributed according to CDF $\widehat{F}_V$ with a mean value $\widehat{\mathbb{E}}[V]$, then gain-loss utility from outcome v is

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

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

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

[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 $u_C : \mathbb{R}_+ \to \mathbb{R}$ 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.

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 under-appreciates 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 $\hat{u}\left(\hat{v} \middle| \widehat{\mathbb{E}}[V]\right) = \hat{v} + \hat{\eta} n\left(\hat{v} \middle| \widehat{\mathbb{E}}[V]\right)$, and thus $\hat{v}$ solves $\hat{u}\left(\hat{v} \middle| \widehat{\mathbb{E}}[V]\right) = u\left(v \middle| \widehat{\mathbb{E}}[V]\right)$. Equations 1 and 2 imply that this misencoded outcome, $\hat{v}$, takes the following form:

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

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., $\lambda > 1$).

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]

[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.

[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: (1) the agent

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 $\widehat{\mathbb{E}}[V^k]$ along each dimension $k \in \{m,e\}$, the agent's total utility from realization $v = (v^m, v^e)$ is

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

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 $\hat{v}^k + \hat{\eta} n\left(\hat{v}^k \mid \widehat{\mathbb{E}}[V^k]\right) = v^k + \eta n\left(v^k \mid \widehat{\mathbb{E}}[V^k]\right)$.

3 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.

3.1 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 total depending on their willingness to work and chance.

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

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); (2) 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).

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

[12] We used digital-voice software to "read" reviews collected from Amazon.com. Unbeknownst to participants, all reviews were either 1-star reviews or 5-star reviews to make the task straightforward (though tedious). Reviews were edited to last approximately 20 seconds, to remove any specific references to author names or book titles, and for

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 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; (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 get reassigned without the noise, we blocked multiple logins and required unique email authentication to access each session of the experiment.

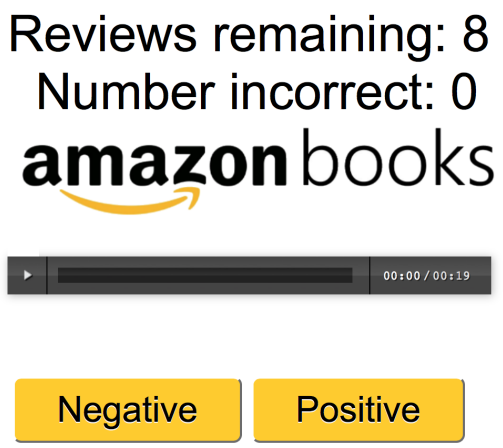

Figure 2: *Screenshot of the classification task from Experiment 1.* Buttons appeared after 10 seconds. Participants clicked the appropriate button to classify whether a review was positive (i.e., endorsing the book) or negative.

The two sessions of the experiment were conducted as follows.

Session 1: Initial-Learning Phase. Participants were instructed them 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

grammar. See the Appendix for sample text of the reviews.

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.

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 $1/2$ chance of doing the task without noise and a $1/2$ chance of doing the task with noise. They were then given a sample task (without noise) and a short sample of the unpleasant noise (8s 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.

Lastly, participants in the *high-probability* 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 Appendix H for full text with highlighted differences). Half of participants were assigned to a "$p = .99$" treatment and the other half were assigned to a "$p = .01$" 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 $p = .99$ arm were assigned the task without noise if $z = 100$; otherwise, they faced the task with noise. Participants in the $p = .01$ 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

[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.

[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

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.

3.2 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 3.3).

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 $p_i \in \{0, .01, .5, .99, 1\}$ denote the participant's ex-ante belief that she will be assigned to task $a = h$. Participant i completes 8 trials of her assigned task a in 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 her 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

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

where $c(\cdot)$ is an increasing function with $c(0) = 0$, $\theta_i(a)$ is a cost parameter that depends on

completed the second session.

$a \in \{h,l\}$, and $\varepsilon_{i,t}$ are i.i.d. mean-zero random cost shocks that are independent of $\theta_i(a)$. The structure of $v^e_{i,t}$ (Equation A.52) 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 $\varepsilon_{i,1}$, 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:

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

where $\widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$ 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 $\widehat{\mathbb{E}}_{i,0}[V^e_{i,1}] = -[p_i\hat{\theta}_{i,0}(h) + (1-p_i)\hat{\theta}_{i,0}(l)]c(8)$.

As described in Section 2 (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., $v^e_{i,1} > \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$), 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 $\varepsilon_{i,t}$ are normally 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

[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, $\hat{v}^e_{i,1} = v^e_{i,1}$.

[17] If the agent believes that $\theta_i(a) \sim N\left(\hat{\theta}_{i,0}(a), \rho^2\right)$ and $\varepsilon_{i,t} \sim N(0,\sigma^2)$, 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: $\hat{v} > \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$ implies that $\hat{\theta}_{i,1}(a|\hat{v}) < \hat{\theta}_{i,0}(a)$ and $\hat{v} < \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$ 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 $\varepsilon_{i,t}(a)$ are independent and normally distributed. See Chambers and Healy (2012) for general sufficient conditions for Bayesian updating in the direction of the signal.

$e_i^*(a|p_i)$ 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

$$\widehat{\mathbb{E}}_{i,1}\left[u_{i,2} \mid e_{i,2}\right] = \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right] + \eta \widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right]\right)\right] + m = 0. \tag{7}$$

This equation demonstrates that uncertainty over $\theta_i(a)$ and hence the disutility of effort—captured in the subjective expectation $\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right]$—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 Appendix C that $e_i^*(a|p_i)$ in fact solves

$$h\left(\hat{\theta}_{i,1}(a)\right) c(e_i^*) = m, \tag{8}$$

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 $v_{i,1}^e$, nor her beliefs about $\theta_i(a)$. Therefore, p_i does not influence $e_i^*(a|p_i)$.[18]

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 tends 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.

[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.

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 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 interpre-

[19]By focusing Observations 1 and 2 on $e_i^*(a|p_i)$, 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 $e_i^*(a|p_i)$ highlights that our predictions do not stem from some interaction between the BDM mechanism and reference dependence.

tation of signals nor 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, and 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 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 3.3). If participants did not (at least partially) incorporate the coin flip into their expectations, we would expect no differences across treatments; our data suggests otherwise.

3.3 Results

To test the theoretical predictions above, we first take a simple non-parametric 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 which 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 the *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 (i) did not answer all five elicitations of WTW (three participants), or (ii) stated a WTW equal to the maximum amount (100 tasks) for every payment level, which prevented us from estimating their responsiveness to payment (six participants).[21] Additionally, we omit participants who did not return for the second session—and whose WTW we therefore did not measure—though we present their demographics where applicable. With this set of restrictions, we are left with a sample of 803 participants.[22]

[20]There is a significant age difference between the first two treatments and the *high-probability* treatment. The first two treatments were run approximately 1 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. Of the six participants dropped due to the second restriction, three were from *control + no noise*, two were from *coin-flip + no noise*, and one was from *coin-flip + noise*. We believe these statements likely result 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.

[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.

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	33.29	33.61
	(12.04)	(12.30)	(11.45)	(11.96)	(9.35)	(9.78)
$\mathbb{1}$(Male)	.468	.464	.428	.387	.488	.529
	(.501)	(.500)	(.496)	(.489)	(.489)	(.501)
Income	2.71	2.58	2.90	2.61	2.46	2.36
	(1.009)	(1.092)	(1.066)	(1.103)	(1.069)	(1.011)
$\mathbb{1}$(Return)	.921	.882	.862	.944	.932	1
	(.271)	(.323)	(.346)	(.231)	(.253)	(0)
Observations	139	153	152	142	160	140

Notes: Standard deviations are in parentheses. Income is coded as a discrete variable which takes values 1-5, 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

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 to 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 ($p = .039$ for difference; statistical results obtained via two-sided t-test 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 ($p = .025$ for difference).[23]

While Table 2 provides a rough sense of the treatment effect, Figure 3 further disaggregates WTW by payment level. Figure 3 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).

[23]In Appendix Figure A1, 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

	Control		Coin Flip		High Prob.	
Variable	noise=0	noise=1	noise=0	noise=1	noise=0	noise=1
Willingness to Work (WTW)	24.23	22.29	28.60	17.64	24.20	21.34
	(1.354)	(1.570)	(1.618)	(1.358)	(1.292)	(1.267)
Observations	615	665	645	665	690	740

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 $p < .05$.

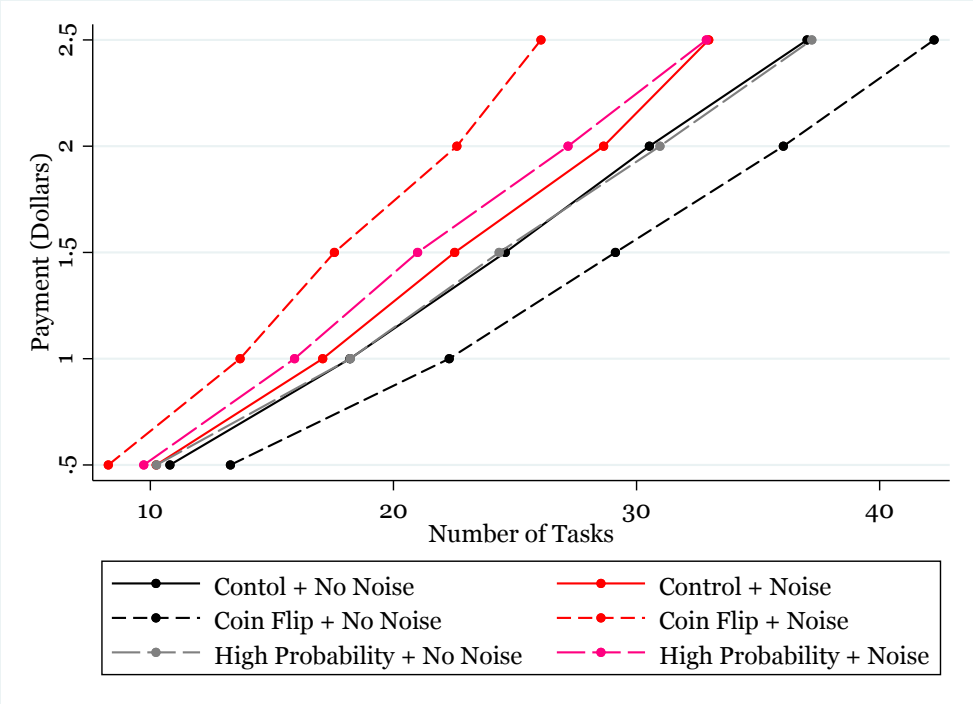

Figure 3: *Labor supply curves across all treatments.* Each point represents the average willingness to work (WTW) for a fixed payment as elicited using the BDM mechanism.

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 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% for the noisy task and decreased by 20% 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 at 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 and clustering standard errors at the individual level; $p = .064$). This validates that participants perceived a difference in the onerousness of the two tasks.

Comparing across treatments, Columns 3 to 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 ($p = .034$ 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 ($p = .047$ 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%. However, probability weighting—people's tendency to overweight small probabilities (e.g. Kahneman and Tversky 1979; Prelec 1998; Gonzalez and Wu 1999)—implies that behavior in the *high-probability* treatment may deviate from the corresponding *control* by more than a 1% chance would suggest. If this 1% 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 A4 in the appendix, we demonstrate that the observables we collect (e.g., task assignment and demographics) do not predict attrition.[24] Overall, attrition was quite low. We accordingly 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 Supplemental Tables A1 and A2 we reproduce Table 2 but divide the sample in two: those who returned after the mandatory 8-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 login times suggests 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 4.3.

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 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

[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.

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 non-linear 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 3.2, 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 $p \in \{0, .01, .5, .99, 1\}$ 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

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

[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 (2012) 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 A3, 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 Appendix C, we present a closed form solution for $h(\cdot)$ under specific distributional assumptions (e.g., normal priors and noise; see Equation A.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:

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

where $\mathbb{D}_i(\textit{treatment})$ is a dummy variable for each treatment (*control*, *coin-flip*, or *high-probability*) and $\mathbb{I}_i(\textit{noise})$ 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 $\gamma = \frac{1}{\beta_1}$ and $\hat{\theta}_1(a|p) = \exp\left(\frac{-\beta_j}{\beta_0}\right)$. 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 $\mathbb{D}_i(\textit{control})\mathbb{I}_i(\textit{noise})$ 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 (i) observed WTW is censored at a minimum value of 0 tasks and a maximum value of 100, and (ii) we have five observations for each person.

Table 3, Column (1) presents the estimates of the baseline specification in Equation 10. We find that 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 ($\hat{\theta}_1(h|.5) - \hat{\theta}_1(h|.99) = .0142$; $\chi^2(1) = 4.13, p = .042$) or faced no uncertainty prior to task assignment ($\hat{\theta}_1(h|.5) - \hat{\theta}_1(h|1) = .0149$; $\chi^2(1) = 4.27, p = .039$). 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 ($\hat{\theta}_1(l|.5) - \hat{\theta}_1(l|.01) = -.0087$; $\chi^2(1) = 4.06, p = .044$) or faced no uncertainty prior to task assignment ($\hat{\theta}_1(l|.5) - \hat{\theta}_1(l|0) = -.0064$; $\chi^2(1) = 2.49, p = .115$).

It is worth noting that we estimate $\gamma = 1.207$ (0.023), and 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 not weakly increase across the five payment levels. This drops

[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 $H_0 : \gamma(h) = \gamma(l); \chi^2(1) = 0.24; p = 0.624$.

Table 3: PARAMETRIC ANALYSIS, EXPERIMENT 1

	Dep var: $\log(e_i)$. Estimated w/ Random-Effects Tobit Regression		
	(1)	(2)	(3)
Cost curvature parameter, γ	1.199	1.197	1.159
	(.018)	(.017)	(.016)
$\hat{\theta}_1(\text{noise} \mid p = 0.5)$	.0673	.0635	.0728
	(0.006)	(.0120)	(.007)
$\hat{\theta}_1(\text{noise} \mid p = 0.99)$	.0531	.0510	.0573
	(.005)	(.009)	(.005)
$\hat{\theta}_1(\text{noise} \mid p = 1)$	.0524	.0493	.0553
	(.004)	(.008)	(.006)
$\hat{\theta}_1(\text{no noise} \mid p = 0)$	.0408	.0385	.0441
	(.004)	(.007)	(.004)
$\hat{\theta}_1(\text{no noise} \mid p = 0.01)$	.0431	.0416	.0468
	(.004)	(.007)	(.004)
$\hat{\theta}_1(\text{no noise} \mid p = 0.5)$	.0344	.0325	.0384
	(.003)	(.006)	(.004)
$H_0 : \hat{\theta}_1(\text{noise} \mid p = 0.5) = \hat{\theta}_1(\text{noise} \mid p = 0.99)$	$\chi^2(1) = 4.13$	$\chi^2(1) = 2.90$	$\chi^2(1) = 3.92$
	$(p = .042)$	$(p = .089)$	$(p = .048)$
$H_0 : \hat{\theta}_1(\text{no noise} \mid p = 0.5) = \hat{\theta}_1(\text{no noise} \mid p = 0.01)$	$\chi^2(1) = 4.06$	$\chi^2(1) = 4.77$	$\chi^2(1) = 2.73$
	$(p = .044)$	$(p = .029)$	$(p = .098)$
Joint test of above	$\chi^2(2) = 8.18$	$\chi^2(2) = 7.47$	$\chi^2(2) = 6.64$
	$(p = .017)$	$(p = .024)$	$(p = .036)$
Observations	4020	4020	3470
Clusters	804	804	694
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 (in parentheses) are clustered at the individual level and recovered via delta method. 18 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.

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.

3.4 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 B; here, we briefly overview our findings.

Critically, we find a significant effect of initial expectations on WTW when participants faced noiseless task. Workers were significantly more willing to work when assigned by coin flip versus the high-probability assignment ($p = .0424$ 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 ($p = .1391$ 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. 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 ($p = .0269$ for difference when facing noisy task; $p = .0017$ for difference when facing the noiseless task. See Table A9 in Appendix B for details). Moreover, despite the caveats above, the magnitude of the effect we observe in the replication—a

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

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

distortion of WTW around 20%—is similar to that we observed in Experiment 1.

4 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 allows 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.

4.1 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 nine groups completed eighteen sessions over the course of one 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' willingness to work (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.[31] 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 4 for a screenshot. Each session had the same structure: participants first

[31] 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.

completed an initial-learning phase which consisted of five mandatory trials, and then we elicited their WTW on additional trials for a bonus payment.

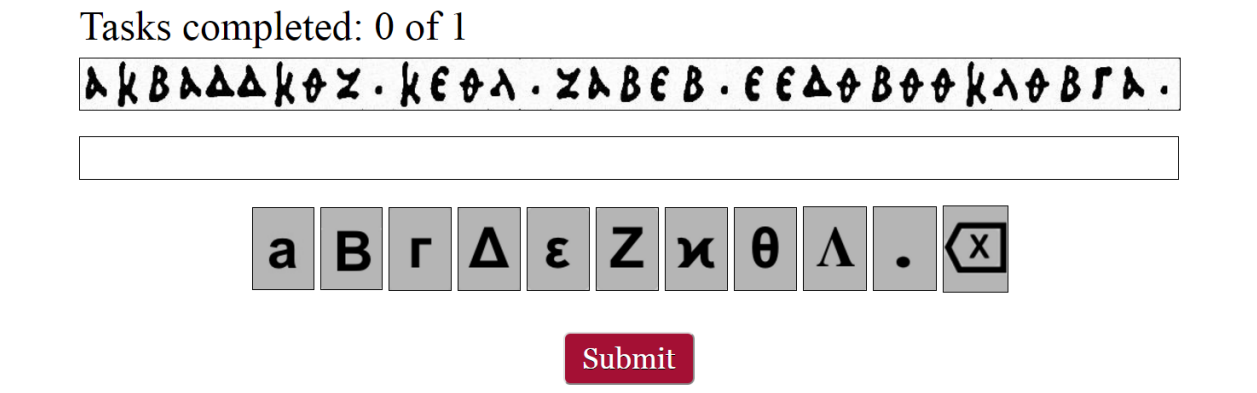

Figure 4: *Screenshot of the transcription tasks from Experiment 2.* Participants clicked the gray button that matched the handwritten letter to "transcribe" the text. Participants were required to achieve 80% accuracy to advance to the next transcription. Each participant randomly faced one of the two depicted alphabets (Greek or Cyrillic) during their first session and faced the other during their second session.

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 $1/2$ 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 U.S. quarter to determine their assignment. Immediately after the coin flip, participants completed 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 faced a Greek alphabet during the first session and Cyrillic during the second, while the other half faced 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.

4.2 Theoretical Predictions

Building from the same theoretical setup from Experiment 1 (Section 3.2), 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 $v^e_{i,t} = [\theta_i(a) + \varepsilon_{i,t}]c(5)$. 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 Appendix E, reference dependence absent misattribution creates an incentive for those facing the noisy task to *decrease* effort over time, and those facing the noiseless task to *increase* it.[32]

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% 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 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 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

[32]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 Appendix E for further details.

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 lest 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 mid-way 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.

4.3 Results

Our primary analysis considers participants who completed both sessions. Thus, our data comes from 72 participants. For completeness, we present an analysis of participant attrition in Table A6.

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 how reciprocity toward the experimenter and mood effects, both of which might plausibly explain the results in Experiment 1. We argue how 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 ($p = .004$). 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 ($p = .014$). Figure 5 depicts this result by plotting the density of $e_{i,1} - e_{i,2}$ for each task, averaged over the five payment levels.[33]

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)	31.39	25.93	25.97	26.41	7.47	-4.25
	(3.679)	(3.526)	(3.001)	(3.575)	(2.397)	(1.645)
Observations	215	220	180	185	180	185

Notes: Standard errors (in parentheses) are clustered at the individual level. Difference between Columns (1)-(3) significant at $p = .026$; Columns (2)-(4) at $p = .865$; Columns (5)-(6) at $p < .001$. Columns (5)-(6) both significantly different from zero at $p = .004$ and $p = .014$, respectively.

[33] We present these densities here using kernel smoothing (Epanechnikov kernel); in Appendix A, we show the raw data in Figure A2 and unsmoothed histograms in Figure A3.

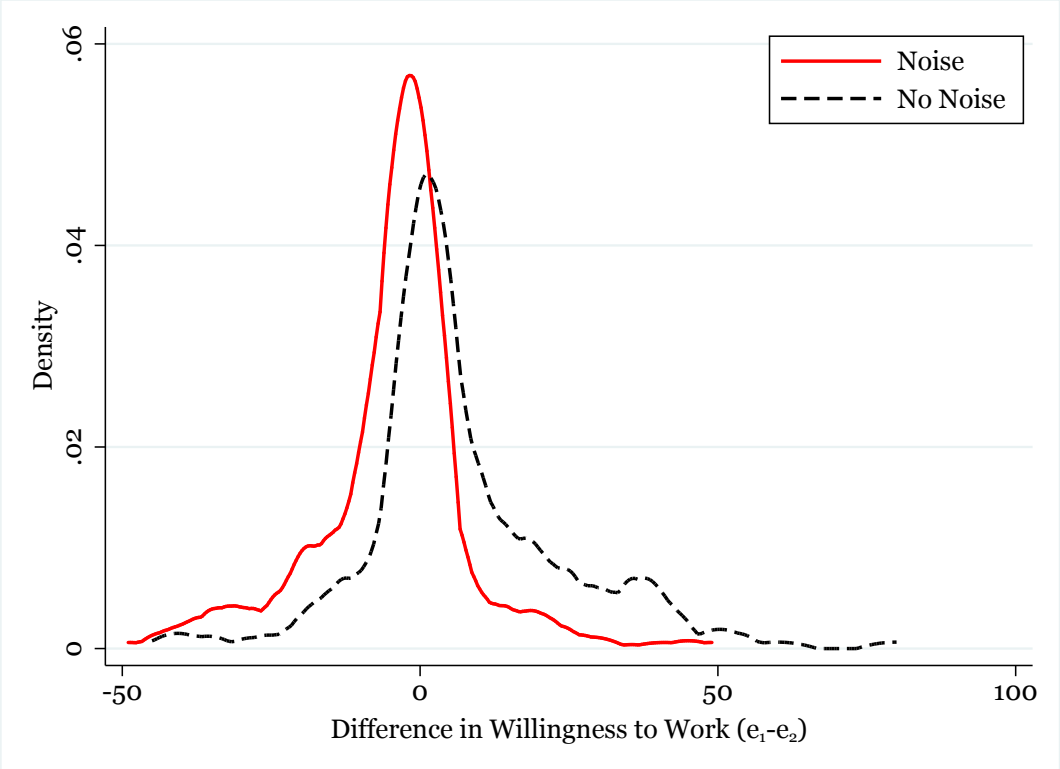

Figure 5: *Kernel density of the difference in willingness to work (WTW) between the first and second sessions, separated by task faced.* Each underlying observation from this figure is the change in a participant's WTW for a fixed payment between Sessions 1 and 2 of the experiment. The black curve represents participants who were assigned to the no-noise task; the red curve represents participants who were assigned to the noisy task.

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 assessment of the task is no longer confounded with a sense of disappointment. These effect sizes have similar magnitude to those in Experiment 1.[34]

[34]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

Parametric Analysis

We now present a more structured approach, following the logic in Section 3.2. 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 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

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

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 $\gamma \approx 1.11$; note this falls close to our estimate from Experiment 1.[35] 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 $\frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})} = 1.29$ (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 $\frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})} = 0.78$ 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% to 40% relative to certain assignment.

Discussion. As with Experiment 1, we suspect attrition is an unlikely explanation for our results. In Supplemental Table A6 (Appendix A) 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

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.

[35] As in Experiment 1, we tested whether $\gamma(h) = \gamma(l)$. Using data from the first session, we fail to reject the null $H_0 : \gamma(h) = \gamma(l); \chi^2(1) = 0.59, p = 0.442$.

Table 5:
PARAMETRIC ANALYSIS, EXPERIMENT 2

	Dep. var: $\log\left(\frac{e_{i,1}}{e_{i,2}}\right)$ Estimated w/ Random-Effects Tobit Regression
	(1)
Estimated ratio $\frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})}$	1.278
	(0.141)
Estimated ratio $\frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})}$	0.780
	(0.075)
$H_0 : \frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})} \geq 1$	$\chi^2(1) = 5.15$
	$p = .023$
$H_0 : \frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})} \leq 1$	$\chi^2(1) = 6.37$
	$p = .011$
Observations	348
Clusters	70

Notes: Standard errors (in parentheses) are clustered at the individual level and derived via the delta method. 12 observations are left-censored and 26 are right-censored. 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 $\gamma = 1.11$.

empirically in Supplemental Table A5 (Appendix A). There, we demonstrate that our qualitative results from Table 4 are robust to (i) controlling for extra tasks using OLS; (ii) controlling for extra tasks via two-stage least squares (utilizing the BDM randomness for identification); and (iii) 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 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 8 hours; for it to explain our results from Experiment 2, this desire must disappear over a week.[36] 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.

5 Conclusion

In this paper, we provide evidence consistent with misattribution of reference dependence—that people retrospectively 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

[36]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.

literature: productivity is rather inelastic (DellaVigna et al. 2019). 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.

References

ABELER, J., A. FALK, L. GOETTE, AND D. HUFFMAN (2011): "Reference Points and Effort Provision." *American Economic Review*, 101(2), 470–492.

ADHVARYU, A., NYSHADHAM, A., AND H. XU (2020): "Hostel takeover: Living conditions, reference dependence, and the well-being of migrant workers." *Working Paper*.

ALLEN, E., P. DECHOW, D. POPE, AND G. WU (2017): "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science*, 63(6), 1657–1672

AUGENBLICK, N., M. NIEDERLE, AND C. SPRENGER (2015): "Working Over Time: Dynamic Inconsistency in Real Effort Tasks." *Quarterly Journal of Economics*, 130(3), 1067–115.

AUGENBLICK N. AND M. RABIN (2019): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *Review of Economic Studies*, 86(3), 941–75.

BACKUS, M., T. BLAKE, D. MASTEROV, S. TADELIS (2020): "Expectation, Disappointment, and Exit: Evidence on Reference Point Formation from an Online Marketplace." *Working Paper*.

BELL, D. (1985): "Disappointment in Decision Making under Uncertainty." *Operations Research*, 33(1), 1–27.

BERNHEIM, D. AND C. SPRENGER (2020): "On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting." *Econometrica*, 88(4), 1363–1409.

BERTRAND, M. AND S. MULLAINATHAN (2001): "Are CEOs Rewarded for Luck? The Ones Without Principals Are." *Quarterly Journal of Economics*, 116(3), 901–32.

BOULDING, W., A. KALRA, R. STAELIN, AND V. ZEITHAML (1993): "A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions." *Journal of Marketing Research*, 30, 7–27.

BROWNBACK, A. AND M. KUHN (2019): "Attribution Bias, Blame, and Strategic Confusion in Punishment Decisions." *Games and Economic Behavior*, 117, 342–360.

BUFFAT, J. AND J. SENN. (2015): "Testing the Speed of Adjustment of the Reference Point in Models of Expectation-Based Reference-Dependent Preferences." *Working Paper.*

CAMERER, C., L. BABCOCK, G. LOEWENSTEIN, AND R. THALER (1997): "Labor Supply of New York City Cabdrivers: One Day at a Time," *Quarterly Journal of Economics*, 112(2), 407–441.

CARD, D. AND G. DAHL (2011): "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior." *Quarterly Journal of Economics*, 126(1), 103–143.

CHAMBERS, C. AND P. HEALY (2012): "Updating towards the signal." *Economic Theory*, 50, 765–786.

COLE, S., A. HEALY, AND E. WERKER (2012): "Do voters demand responsive governments? Evidence from Indian disaster relief," *Journal of Development Economics*, 97(2), 167–181.

CRAWFORD, V. AND J. 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), 1912–1932.

DELLAVIGNA, S., LIST, J.A., MALMENDIER, U., AND G. RAO (2019): "Estimating Social Preferences and Gift Exchange with a Piece-Rate Design." *Working Paper*.

DE QUIDT, J. (2018): "Your Loss Is My Gain: A Recruitment Experiment with Framed Incentives" *Journal of the European Economic Association*, 16(2), 522–559.

DUTTON, D. AND A. AARON (1974): "Some Evidence for Heightened Sexual Attraction Under Conditions of High Anxiety." *Journal of Personality and Social Psychology*, 30, 510–517.

EDMANS, A., D. GARCIA, AND O. NORLI (2007). "Sports Sentiment and Stock Returns." *Journal of Finance*, 62(4), 1967-98.

ERICSON, K. AND A. FUSTER (2011): "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *Quarterly Journal of Economics*, 126(4), 1879–907.

ERKAL, N., L. GANGADHARAN, AND B.H. KOH (2019): "Attribution Biases in Leadership: Is it Effort or Luck?" *Working Paper*.

FRYER, R., P. HARMS, AND M. 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, 10826–10829.

GAGNON-BARTSCH T. AND B. BUSHONG (2021): "Learning with Misattribution of Reference Dependence." *Working Paper.*

GILBERT, D. AND P. MALONE (1995): "The Correspondence Bias." *Psychological Bulletin*, 117(1): 21–38.

GILL, D. AND V. PROWSE (2012): "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review*, 102(1), 469–503.

GNEEZY, U. AND J. LIST (2006): "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica*, 74(5), 1365–1384.

GONZALEZ, W. AND G. WU (1999): "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38, 129–66.

GREENE, W. (2003): *Econometric Analysis*. Prentice-Hall.

HAGGAG, K., D. POPE, K. BRYANT-LEES, AND M. BOS (2019): "Attribution Bias in Consumer Choice." *Review of Economic Studies*, 86(5), 2136–83.

HEFFETZ, O. AND J. LIST (2014): "Is the Endowment Effect an Expectations Effect?" *Journal of the European Economic Association*, 12(5), 1396–422.

HEFFETZ, O. (2018): "Are Reference Points Merely Lagged Beliefs Over Probabilities?" *Working Paper*.

HIGHHOUSE, S. AND A. GALLO (1997): "Order Effects in Personnel Decision Making." *Human Performance*, 10(1), 31–46.

HIRSHLEIFER, D. AND T. SHUMWAY (2003): "Good Day Sunshine: Stock Returns and the Weather." *Journal of Finance*, 58(3), 1009–32.

IMAS, A., SADOFF, S., AND A. SAMEK (2017): "Do People Anticipate Loss Aversion?" *Management Science*, 63(5), 1271–1284.

KAHNEMAN, D. AND A. TVERSKY (1979): "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 1979; 47(2), 263–291.

KARLE, H., G. KIRCHSTEIGER, AND M. PEITZ (2015): "Loss Aversion and Consumption Choice: Theory and Experimental Evidence." *American Economic Journal: Microeconomics*, 7(2), 101–120.

KIMBALL, D. AND S. PATTERSON (1997): "Living Up to Expectations: Public Attitudes Toward Congress." *Journal and Politics*, 59, 701–728.

KOPALLE, P. AND D. LEHMANN (2006): "Setting Quality Expectations When Entering a Market: What Should the Promise Be?" *Marketing Science*, 25(1), 8–24.

KŐSZEGI, B. AND M. RABIN (2006): "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics*, 121(4), 1133–65.

MALMENDIER, U. AND S. NAGEL (2011): "Depression Babies: Do Macroeconomic Experiences Affect Risk-Taking?" *Quarterly Journal of Economics*, 126, 373–416.

MARKLE, A., G. WU, R.J. WHITE, AND A.M. SACKETT (2015): "Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence," *Working Paper*.

MEDVEC, V.H., S.F. MADEY, AND T. GILOVICH (1995): "When less is more: counterfactual thinking and satisfaction among Olympic medalists." *Journal of Personality and Social Psychology*, 69(4), 603–10.

MESTON, C.M. AND P.F. FROHLICH (2003): "Love at First Fright: Partner Salience Moderates Roller-Coaster-Induced Excitation Transfer." *Archives of Sexual Behavior*, 32(6), 537-44.

OLIVER, R. (1977): "Effect of Expectation and Disconfirmation of Post-Exposure Product Evaluation: An Alternative Interpretation." *Journal of Applied Psychology*, 62, 480–486.

OLIVER, R. (1980): "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions." *Journal of Marketing Research*, 17, 460–469.

PATTERSON, S., G. BOYNTON, AND R. HEDLUND (1969): "Perceptions and Expectations of the Legislature and Support for It." *American Journal of Sociology*, 75(1), 62–76.

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

POPE, D. AND M. SCHWEITZER (2011): "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review*, 101(1), 129–157.

POST, T., M. VAN DEN ASSEM, G. BALTUSSEN, AND R. THALER (2008): "Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show." *American Economic Review*, 98(1), 38–71.

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

RABIN, M. AND J. SCHRAG (1999): "Inference by Believers in the Law of Small Numbers." *Quarterly Journal of Economics*, 114(1): 37–82.

ROSS, L. (1977): "The Intuitive Psychologist and his Shortcomings: Distortions in the Attribution Process." In Berkowitz, L. *Advances in Experimental Social Psychology*, Academic Press, 173–220.

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

SIMONSOHN, U. (2007): "Clouds Make Nerds Look Good: Field Evidence of the Impact of Incidental Factors on Decision Making." *Journal of Behavioral Decision Marking*, 20(2), 143–152.

SIMONSOHN, U. (2010): "Weather to Go to College." *The Economic Journal*, 120(543), 270–280.

SMITH, A. (2012): "Lagged Beliefs and Reference-Dependent Preferences." *Working Paper.*

SONG, C. (2016): "An Experiment on Reference Points and Expectations." *Working Paper.*

THAKRAL, N. AND T. TÔ (2020): "Daily Labor Supply and Adaptive Reference Points." *American Economic Review. Forthcoming.*

WENNER, L. (2015): "Expected Prices as Reference Points—Theory and Experiments." *European Economic Review*, 75, 60-79.

WOLFERS, J. (2007): "Are Voters Rational? Evidence from Gubernatorial Elections," *Working Paper.*

Appendix

FOR ONLINE PUBLICATION

A Supplemental Tables and Figures

In this appendix we provide additional empirical results that supplement the main text and provide robustness checks for our primary results.

Figure A1 shows the CDFs of WTW for the *control* and *coin-flip* treatments from Experiment 1, aggregated over all payment levels (and smoothed using the Epanechnikov kernel). As a validation of our basic setup, a Kolmogorov-Smirnov equality-of-distributions test verifies that control participants were more willing to work on the noiseless task than the noisy one ($D = .1225; p < .001$).[37] Speaking to our main hypotheses, the figure highlights that WTW in the *contol + no noise* group was lower than the *coin flip + no noise* group—the latter almost first-order stochastically dominates the former. By contrast, the cumulative distribution of WTW in the *control + noise* group first-order stochastically dominates that of the *coin flip + noise* group.

We next show that dividing the Experiment 1 sample in half according to the total amount of time participants spent on the experiment (from the start of Session 1 to completion) does not have a large effect on our nonparametric results. This is demonstrated in Tables A1 and A2 below. However, these comparisons based on duration are limited due to unequal group sizes. Regression analysis (included in Table 3 in the main text) demonstrates that this effect does not alter the results of our parametric analysis.

We further show that the results of our parametric analysis are robust to changing the Stone-Geary background parameter that appears in the effort-cost function. Although our numerical estimates vary with this parameter, we show in Table A3 that our qualitative results hold for two alternative specifications of the background parameter which vary by an order of magnitude. For the reader's ease, we omit such analysis for Experiment 2.

Next, we utilize a logit model to explore whether any observables predict attrition in Experiment 1 (Table A4). Although we have overall lower attrition in the high-probability treatment, we do not find that other factors influenced attrition. This effect is easily seen in Table 1 in the main text. We suspect this is due to the fact that we ran the high-probability session at a slightly different time of day.

We then turn to the Experiment 2. To address potential concerns about differential experience

[37] While this test fails to account for redundancy in the data stemming from multiple observations from each individual, we calculated a more conservative version of the statistic by running individual K-S tests for each payment level. Three out of five payment levels showed significant differences between the cumulative distributions of WTW for *control + noise* and *control + no-noise*; the five p values were .024, .189, .041, .019, .090.

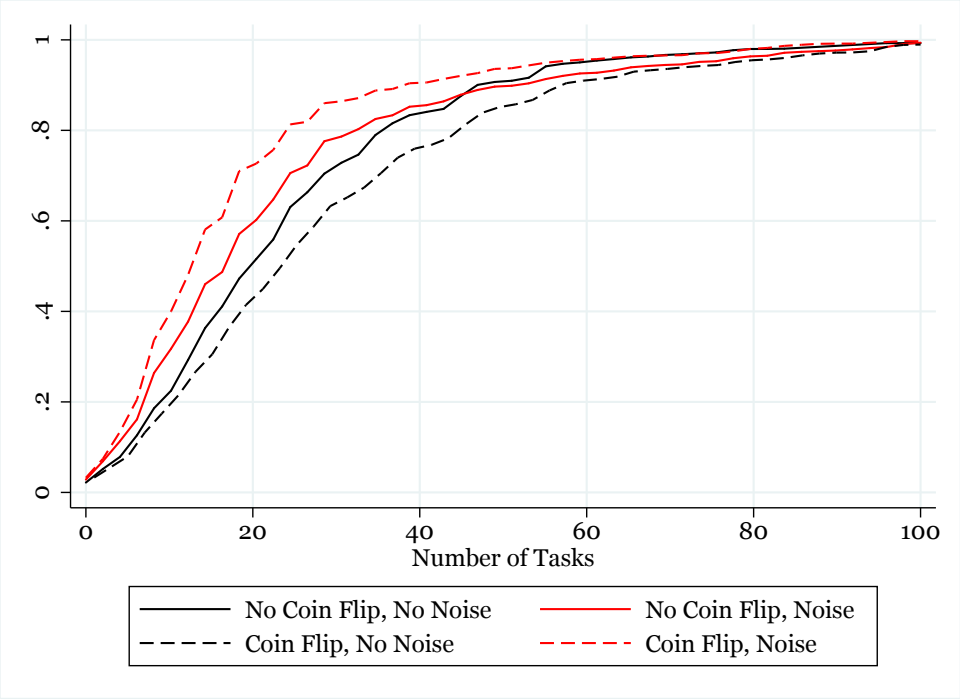

Figure A1: *Cumulative bid distribution by group.* Cumulative distribution curves are over all five payment levels and are smoothed using the Epanechnikov kernel.

and learning, Table A5 presents non-parametric results for Experiment 2 (analogous to the final two columns of Table 4) in which we drop any participants who completed extra tasks in the first session. This analysis leaves far fewer participants in our sample, but our qualitative results hold. We then utilize the random numbers from the BDM in our experiment to instrument for whether a person completed extra tasks. This analysis verifies that, while doing extra tasks may have changed WTW in Session 2, our primary conclusions remain for those who did not complete extra tasks.

Finally, following the robustness exercise in Experiment 1 concerning attrition, we estimate a similar logit model for Experiment 2 (Table A6). We did not collect demographic information from participants in Experiment 2, and thus we have fewer potential explanatory variables. That said, we find no convincing link between observables and attrition.

Table A1:
EXPERIMENT 1. BASELINE RESULTS (LESS THAN MEDIAN TOTAL DURATION)

	Control		Coin Flip		High Prob.	
Variable	noise=0	noise=1	noise=0	noise=1	noise=0	noise=1
Willingness to Work (WTW)	22.38	20.69	27.67	17.43	23.27	21.25
	(1.458)	(1.876)	(2.034)	(1.689)	(2.203)	(2.759)
Observations	430	385	365	390	245	195

Notes: Willingness to work is averaged over five payment levels. Standard errors (in parentheses) are clustered at the individual level with 402 clusters.

Table A2:
EXPERIMENT 1. BASELINE RESULTS (GREATER THAN MEDIAN TOTAL DURATION)

	Control		Coin Flip		High Prob.	
Variable	noise=0	noise=1	noise=0	noise=1	noise=0	noise=1
Willingness to Work (WTW)	28.53	24.5	29.81	17.941	24.71	21.38
	(2.846)	(2.670)	(2.617)	(2.253)	(1.596)	(1.412)
Observations	185	280	280	275	445	545

Notes: Willingness to work is averaged over five payment levels. Standard errors (in parentheses) are clustered at the individual level with 402 clusters.

Table A3:
EXPERIMENT 1. ROBUSTNESS OF PARAMETRIC ANALYSIS

	Estimated w/ Random-Effects Tobit Regression	
	($\omega = 1$)	($\omega = 10$)
Cost curvature parameter, γ	1.327 (.018)	2.168 (.031)
$\hat{\theta}_1(\text{noise} \mid p = 0.5)$	.0420 (.004)	.0013 (.0002)
$\hat{\theta}_1(\text{noise} \mid p = 0.99)$	.0329 (.004)	.0010 (.0001)
$\hat{\theta}_1(\text{noise} \mid p = 1)$	.0324 (.003)	.0099 (.0001)
$\hat{\theta}_1(\text{no noise} \mid p = 0)$	.0255 (.002)	.0008 (.0001)
$\hat{\theta}_1(\text{no noise} \mid p = 0.01)$	.0267 (.002)	.0008 (.0001)
$\hat{\theta}_1(\text{no noise} \mid p = 0.5)$	.0213 (.002)	.0006 (.0001)
$H_0 : \hat{\theta}_1(\text{noise} \mid p = 0.5) = \hat{\theta}_1(\text{noise} \mid p = 0.99)$	$\chi^2(1) = 4.59$ ($p = .032$)	$\chi^2(1) = 5.00$ ($p = .025$)
$H_0 : \hat{\theta}_1(\text{no noise} \mid p = 0.5) = \hat{\theta}_1(\text{no noise} \mid p = 0.01)$	$\chi^2(1) = 4.25$ ($p = .039$)	$\chi^2(1) = 4.65$ ($p = .031$)
Joint test of above	$\chi^2(2) = 8.83$ ($p = .012$)	$\chi^2(2) = 9.45$ ($p = .009$)
Observations	4020	4020
Clusters	804	804

Notes: Standard errors (in parentheses) are clustered at the individual level and recovered via delta method. 18 observations are left-censored and 43 are right-censored.

Table A4:
EXPERIMENT 1. DETERMINANTS OF RETURNING FOR SECOND SESSION

	Logit. Dependent variable: 𝟙(return)					
	Raw	AMEs	Raw	AMEs	Raw	AMEs
𝟙(Noise)	-0.096	-0.007	-0.122	-0.008	-0.132	-0.010
	(0.251)	(0.018)	(0.248)	(0.017)	(0.255)	(0.021)
𝟙(Coin Flip)			0.003	0.000	0.031	0.002
			(0.261)	(0.018)	(0.279)	(0.022)
𝟙(High Probability)			1.069**	0.062**	1.101**	0.079**
			(0.289)	(0.018)	(0.371)	(0.026)
Demographics					X	X
Constant	2.523***		2.270***		2.096***	
	(0.184)		(0.229)		(0.586)	
Observations	887	887	887	887	887	887

Notes: Standard errors in parentheses. The control treatment forms the baseline comparison group; demographics includes dummies for income of respondent and gender, and age; none are significant.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5:
EXPERIMENT 2. THE EFFECT OF EXTRA TASKS IN SESSION 1

	Dependent variable: $(e_{i,1} - e_{i,2})$
Dropping Extra Tasks	
$\mathbb{1}$(Noise)	-5.443***
	(1.885)
$\mathbb{1}$(No Noise)	4.896*
	(2.685)
Observations	240
OLS	
$\mathbb{1}$(Noise)	-5.443***
	(1.882)
$\mathbb{1}$(No Noise)	4.896*
	(2.681)
$\mathbb{1}$(Extra Tasks)*$\mathbb{1}$(Noise)	3.351
	(3.730)
$\mathbb{1}$(Extra Tasks)*$\mathbb{1}$(No Noise)	8.431
	(5.164)
IV using BDM as instrument	
$\mathbb{1}$(Noise)	-8.010*
	(4.629)
$\mathbb{1}$(No Noise)	11.181**
	(4.428)
$\mathbb{1}$(Extra Tasks)*$\mathbb{1}$(Noise)	10.459
	(11.765)
$\mathbb{1}$(Extra Tasks)*$\mathbb{1}$(No Noise)	-12.137
	14.176
Observations	360

Notes: Standard errors (in parentheses) clustered at individual level. All regressions include random effects at individual level. Instruments are random number from BDM and dummies for the randomly selected question (five such variables).
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A6:
EXPERIMENT 2. DETERMINANTS OF RETURNING FOR SECOND SESSION

	Logit. Dependent variable: $\mathbb{1}$(return)			
	Raw	AMEs	Raw	AMEs
$\mathbb{1}$(Noise)	-0.134	-0.019	-0.129	-0.015
	(0.568)	(0.080)	(0.659)	(0.080)
Avg WTW, Session 1			-0.006	-0.001
			(0.014)	(0.003)
$\mathbb{1}$(Russian, Session 1)			0.436	0.052
			(0.674)	(0.092)
Constant	1.638***		0.803	
	(0.413)		(1.134)	
Session Dummies			X	X
Observations	87	87	87	87

Notes: Standard error in parentheses. $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

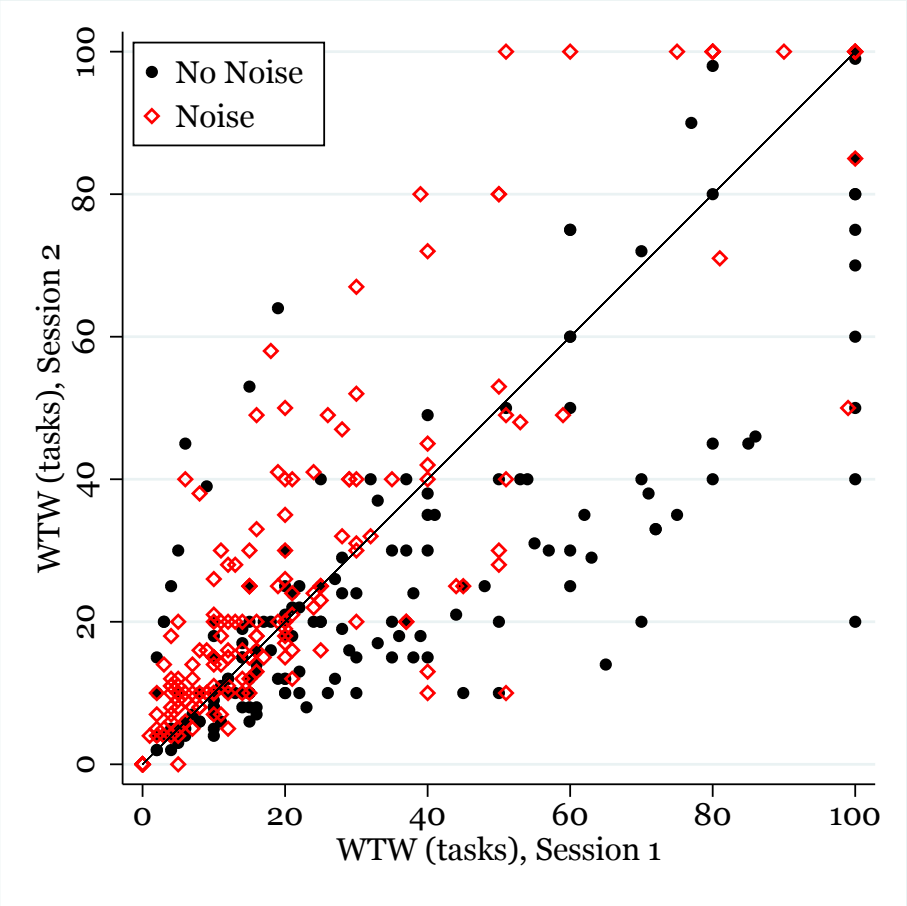

Figure A2: *Raw willingness to work (WTW) data from Experiment 2.* Each observation in this figure represents a participant's WTW for a fixed payment in sessions one and two of the experiment. Black dots represent participants who faced the no-noise task; red diamonds represent participants who faced the noisy task.

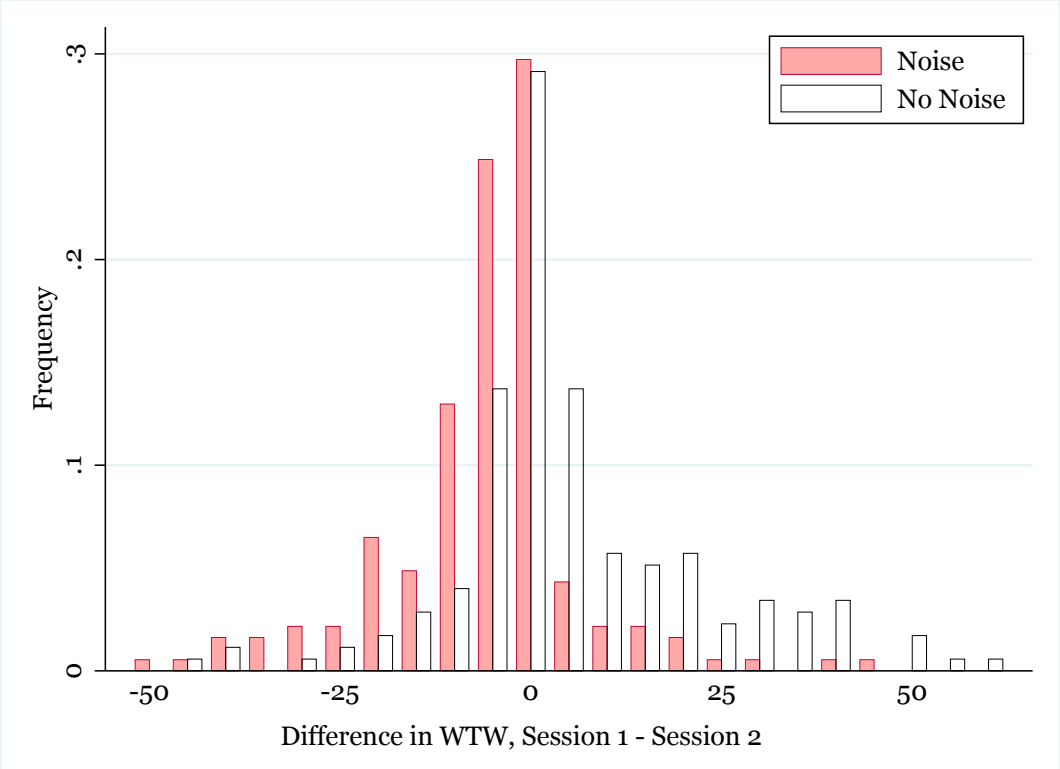

Figure A3: *Histogram of the difference in willingness to work (WTW) between the first and second sessions in Experiment 2.* Each observation in this figure represents the change in a participant's WTW for a fixed payment between sessions one and two of the experiment. Clear bars represent participants who faced the no-noise task; solid red bars represent participants who faced the noisy task.

B Experiment 1 Replication

In this appendix, we present results from our replication study discussed in Section 3.4. Given that the objective was to eliminate concerns about non-random 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 A7 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 A7 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 (i) did not answer all five elicitations of WTW (zero participants); (ii) 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 (iii)

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 A8, which is a direct analog to the original Table 2. We discuss these results in the main text. Figure A4 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 regression, since our data is 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.

Table A7:
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	37.6	39.11	42.23	38.01	38.78
	(13.11)	(10.66)	(12.74)	(11.50)	(12.23)	(11.16)
𝟙(Male)	.509	.510	.528	.514	.553	.524
	(.501)	(.502)	(.501)	(.502)	(.499)	(.501)
Income	2.86	2.70	3.12	3.01	2.79	3.03
	(1.137)	(1.212)	(1.127)	(1.202)	(1.246)	(1.190)
𝟙(Return)	.832	.884	.901	.95	.893	.917
	(.375)	(.321)	(.299)	(.219)	(.310)	(.276)
Observations	161	155	142	140	150	145

Notes: Standard deviations are in parentheses. Income is coded as a discrete variable which takes values 1-5, 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 A8: BASELINE RESULTS, EXPERIMENT 1

	Control		Coin Flip		High Prob.	
Variable	noise=0	noise=1	noise=0	noise=1	noise=0	noise=1
Willingness to Work (WTW)	19.17	16.50	22.54	15.71	18.78	18.31
	(1.223)	(1.141)	(1.367)	(1.096)	(1.253)	(1.373)
Observations	134	136	127	133	133	134

Notes: Willingness to work is averaged over five payment levels. Standard errors (in parentheses) are clustered at the individual level. Difference between Columns (3)-(5) significant at $p = .0424$; difference between (4)-(6) not significant ($p = .1391$).

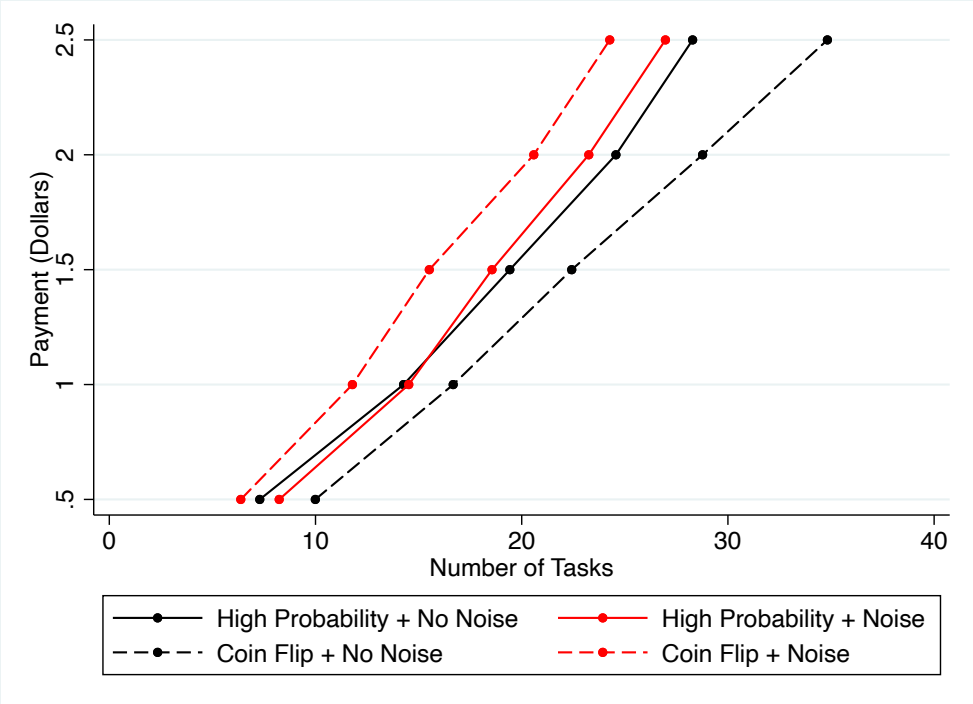

Figure A4: *Labor supply curves across key treatments.* Each point represents the average willingness to work (WTW) for a fixed payment as elicited using the BDM mechanism.

Table A9:
POOLED RESULTS: REPLICATION + MAIN EXPERIMENT

	Dep var: WTW Estimated w/ Random-Effects Tobit Regression
$\mathbb{1}$(coin flip + noise)	18.87 (1.016)
$\mathbb{1}$(high probability + noise)	21.83 (1.031)
$\mathbb{1}$(control + noise)	21.89 (1.146)
$\mathbb{1}$(control + no noise)	23.76 (1.014)
$\mathbb{1}$(high probability + no noise)	23.46 (0.989)
$\mathbb{1}$(coin flip + no noise)	27.94 (1.196)
$\mathbb{1}$(replication)	-4.60 (0.797)
$H_0 : \mathbb{1}$(coin flip + noise) $= \mathbb{1}$(high probability + noise)	$\chi^2(1) = 4.90$ $(p = .0269)$
$H_0 : \mathbb{1}$(coin flip + no noise) $= \mathbb{1}$(high probability + no noise)	$\chi^2(1) = 9.84$ $(p = .0017)$
Observations	7820
Clusters	1564

Notes: Standard errors (in parentheses) are clustered at the individual level and recovered via delta method. 104 observations are left-censored and 76 are right-censored.

C Experiment 1: Derivation of Optimal Effort

In this appendix we show that, under reasonable assumptions, a rational participant with reference-dependent utility will choose an effort level in Experiment 1 that is decreasing in her expected value of her cost parameter, $\theta_i(a)$. This analysis formalizes the predictions regarding effort stated in Observations 1 and 2.

Recall that the predicted effort of a participant with reference-dependent utility assigned to task a solves Equation 7 in the main text: indifference between completing $e_i^*(a|p_i)$ tasks for m dollars and not working at all implies that $e_i^*(a|p_i)$ is the value of $e_{i,2}$ that solves

$$\begin{aligned} & \widehat{\mathbb{E}}_{i,1}\left[u_{i,2}|e_{i,2}\right] = m + \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right] + \eta\widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right]\right)\right] = 0 \\ \Rightarrow \quad & \widehat{\mathbb{E}}_{i,1}\left[u_{i,2}|e_{i,2}\right] = m - \hat{\theta}_{i,1}(a)c(e_{i,2}) + \eta\widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right)\right] = 0. \end{aligned} \tag{A.1}$$

Recall that, conditional on $e_{i,2}$, the participant's effort cost in period 2 is a random variable $V^e_{i,2} = -[\theta_i(a) + \varepsilon_{i,2}]c(e_{i,2})$. Define the random variable $X_{i,2}(a) = \theta_i(a) + \varepsilon_{i,2}$ and let $\widehat{F}_{i,1}$ denote the participant's subjective CDF over $X_{i,2}$ conditional on any information obtained in period 1. Let $x_{i,2}$ denote the realization of $X_{i,2}$. Furthermore, note that $n\left(V^e_{i,2} \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right) = -[x_{i,2}(a) - \hat{\theta}_{i,1}(a)]c(e_{i,2})$ if $x_{i,2}(a) \leq \hat{\theta}_{i,1}(a)$, and otherwise $n\left(V^e_{i,2} \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right) = -\lambda[x_{i,2}(a) - \hat{\theta}_{i,1}(a)]c(e_{i,2})$. Thus,

$$\widehat{\mathbb{E}}_{i,1}\left[n\left(V^e_{i,2} \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right)\right] = -c(e_{i,2})\Bigg(\widehat{F}_{i,1}\left(\hat{\theta}_{i,1}(a)\right)\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) \leq \hat{\theta}_{i,1}(a)\right] + \lambda\left[1 - \widehat{F}_{i,1}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]\Bigg), \quad \text{(A.2)}$$

and thus

$$\widehat{\mathbb{E}}_{i,1}\left[n\left(V^e_{i,2} \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right)\right] = - c(e_{i,2})(\lambda - 1)\left[1 - \widehat{F}_{i,1}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]. \quad \text{(A.3)}$$

Substituting Equation A.3 back into Equation A.1 yields:

$$\begin{aligned}\widehat{\mathbb{E}}_{i,1}\left[u_{i,2}\middle|e_{i,2}\right] &= m - \hat{\theta}_{i,1}(a)c(e_{i,2}) \\ &\quad -\eta(\lambda - 1)\left[1 - \widehat{F}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]c(e_{i,2}) \\ &= m - h\left(\hat{\theta}_{i,1}(a)\right)c(e_{i,2}), \quad \text{(A.4)}\end{aligned}$$

where

$$h\left(\hat{\theta}_{i,1}(a)\right) \equiv \hat{\theta}_{i,1}(a) + \eta(\lambda - 1)\left[1 - \widehat{F}_{i,1}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]. \quad \text{(A.5)}$$

Recall that we assumed the participant's prior over $\theta_i(a)$ and the distribution over $\varepsilon_{i,2}$ are both normal. Thus, according to the participant, $X_{i,2}$ is normally distributed with mean $\hat{\theta}_{i,1}(a)$; let ξ^2 denote the variance of $X_{i,2}$. We can then write $X_{i,2} = \hat{\theta}_{i,1}(a) + \delta$ where $\delta \sim N(0, \xi^2)$. Substituting this into Equation A.5 yields

$$h\left(\hat{\theta}_{i,1}(a)\right) = \hat{\theta}_{i,1}(a) + \eta(\lambda - 1)\frac{1}{2}\mathbb{E}[\delta\big|\delta > 0], \quad \text{(A.6)}$$

where we have additionally used the fact that $\widehat{F}_{i,1}(\hat{\theta}_{i,1}(a)) = \frac{1}{2}$ given that $X_{i,2}$ is symmetric about $\hat{\theta}_{i,1}(a)$. It is well known that $\mathbb{E}[\delta|\delta > 0] = 2\xi\phi(0) = 2\xi/\sqrt{2\pi}$, where ϕ is the standard normal

PDF (see, e.g., Greene 2003). Hence,

$$h\big(\hat{\theta}_{i,1}(a)\big) = \hat{\theta}_{i,1}(a) + \eta(\lambda - 1)\frac{\xi}{\sqrt{2\pi}}. \tag{A.7}$$

From Equation A.7, it is immediate that h is increasing in $\hat{\theta}_{i,1}(a)$. Given that e_i^* is chosen such that $\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i^*] = 0$, Equation A.4 implies that the participant will select e_i^* such that $h(\hat{\theta}_{i,1}(a))c(e_i^*) = m$. Therefore, e_i^* is decreasing in $\hat{\theta}_{i,1}(a)$. It then follows that if $\hat{\theta}_{i,1}(h)$ tends to be higher than $\hat{\theta}_{i,1}(l)$, then participants assigned the noisy task will exhibit lower effort levels than those assigned the noiseless task (i.e., fixing p, $e^*(h|p) < e^*(l|p)$ on average).

D Reference Points that Incorporate the BDM Mechanism

In this appendix we consider how our theoretical predictions of Experiment 1 extend when a participant's reference point incorporates the uncertainty introduced by the BDM mechanism. In particular, we show that the the optimal effort of a participant with reference-dependent utility is still decreasing in her estimate of the effort-cost parameter, $\hat{\theta}_{i,1}(a)$. Consider participant i who has been assigned to task a. Recall her willingness to work (WTW) on additional trials is elicited via a BDM mechanism: the participant announces $e_i \in [0,100]$ and then a number e is uniformly drawn from $[0,100]$ at random. If $e < e_i$, the participant completes e tasks in exchange for a bonus of m dollars. Otherwise, she does no additional work and does not earn a bonus. Thus, conditional on submitting e_i to the mechanism, the participant will do additional work with probability $G(e_i)$, where G denotes the CDF of a uniform random variable on $[0,100]$ (and g denotes the associated PDF). Furthermore, upon submitting e_i, the participant's expected consumption utilities on the money and effort dimensions are, respectively, $r^m(e_i) \equiv G(e_i)m$ and $r^e(e_i;\hat{\theta}_{i,1}(a)) \equiv G(e_i)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e < e_i]$ where $\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e < e_i] = \hat{\theta}_{i,1}(a) \cdot \int_0^{e_i} c(e)\frac{g(e)}{G(e_i)}de$. Thus, the values $r_i^m(e_i)$ and $r_i^e(e_i;\hat{\theta}_{i,1}(a))$ serve as the participant's reference points along each dimension in period 2. As such, she chooses e_i^* to maximize

$$\begin{aligned}\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i] = G(e_i)\Big\{\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e + \eta n\left(V_{i,2}^e \,\middle|\, r^e(e_i;\hat{\theta}_{i,1}(a))\right)\middle| e < e_i\right] + m + \eta\left(m - r^m(e_i)\right)\Big\} \\ + [1 - G(e_i)]\Big\{\eta\left(0 - r^e(e_i;\hat{\theta}_{i,1}(a))\right) + \eta\lambda\left(0 - r^m(e_i)\right)\Big\},\end{aligned} \tag{A.8}$$

where the expectation $\widehat{\mathbb{E}}_{i,1}$ is with respect to the random number e drawn by the mechanism, $\varepsilon_{i,2}(a)$, and the participant's updated beliefs over $\theta_i(a)$. The first term in braces in Equation A.8 is the participant's expected utility conditional on the BDM assigning additional work. In this contingency, her disutility of effort will (on average) come as a loss relative to her expected value on this di-

mension, $r^e(e_i;\hat{\theta}_{i,1}(a))$, since this expectation incorporates a chance of no extra work and hence zero effort. Similarly, the monetary bonus comes as a gain relative to her expected monetary gain, $r^m(e_i)$, which incorporates a chance of no extra work and hence no bonus. The second term in braces is the participant's expected gain-loss utility conditional on the BDM assigning no additional work. In this contingency, she experiences a gain on the effort dimension but a loss on the monetary dimension.

Similar to the analysis in the main text, the treatment probability p may influence e_i is through its affect on the participant's perception of $\theta_i(a)$ (i.e., via misattribution). Thus, we will examine how the optimal effort choice, e_i^*, depends on this perception, $\hat{\theta}_{i,1}(a)$. To simplify the analysis below, we assume the participant forms certain beliefs about $\theta_i(a)$ following period 1, and thus the contingency in which she is assigned additional work necessarily comes as a loss on the effort dimension.

First consider the case without reference dependence (i.e., $\eta = 0$). The objective function in Equation A.8 reduces to

$$\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i] = G(e_i)\left(\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e \middle| e < e_i\right] + m\right) = \hat{\theta}_{i,1}(a)\cdot\int_0^{e_i} c(e)g(e)de + G(e_i)m, \tag{A.9}$$

and the first-order condition implies an optimal choice of $e_i^* = c^{-1}\left(m/\hat{\theta}_{i,1}(a)\right)$. Clearly e_i^* is decreasing in $\hat{\theta}_{i,1}(a)$.

We now consider the case with reference dependence (i.e., $\eta > 0$). It is helpful to rewrite the objective function in Equation A.8 as the sum of two components: the expected monetary benefit from statement e_i, which we denote by

$$B(e_i) \equiv G(e_i)\Big\{m + \eta\left(m - r^m(e_i)\right)\Big\} - \eta\lambda[1-G(e_i)]r^m(e_i), \tag{A.10}$$

and the expected effort cost from e_i, which we denote by

$$K(e_i;\hat{\theta}_{i,1}(a)) \equiv -G(e_i)\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e + \eta n\left(V_{i,2}^e \,\middle|\, r^e(e_i;\hat{\theta}_{i,1}(a))\right)\middle| e < e_i\right] + \eta[1-G(e_i)]r^e(e_i;\hat{\theta}_{i,1}(a)). \tag{A.11}$$

Thus, the objective in Equation A.8 reduces so that the person chooses e_i to maximize

$$\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i] = B(e_i) - K(e_i;\hat{\theta}_{i,1}(a)). \tag{A.12}$$

Given the objective above, we now analyze when the optimal effort choice, e_i^*, is a decreasing function of $\hat{\theta}_{i,1}(a)$. Let $L(e_i;\hat{\theta}_{i,1}(a))$ denote the first derivative of the objective function with

respect to e_i:

$$L(e_i;\hat{\theta}_{i,1}(a)) \equiv \frac{\partial B(e_i)}{\partial e_i} - \frac{\partial K(e_i;\hat{\theta}_{i,1}(a))}{\partial e_i}, \tag{A.13}$$

so the first-order condition (FOC) requires $L(e_i^*,\hat{\theta}_{i,1}(a)) = 0$. Using the Implicit Function Theorem,

$$\frac{\partial e_i^*}{\partial \hat{\theta}_{i,1}(a)} = -\left(\frac{\partial L(e_i^*;\hat{\theta}_{i,1}(a))}{\partial e_i^*}\right)^{-1} \frac{\partial L(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)}. \tag{A.14}$$

Thus, so long as the second-order condition (SOC) holds and the FOC thus describes the optimum, then $\frac{\partial L(e_i^*;\hat{\theta}_{i,1}(a))}{\partial e_i^*} < 0$ and

$$\operatorname{sgn}\left(\frac{\partial e_i^*}{\partial \hat{\theta}_{i,1}(a)}\right) = \operatorname{sgn}\left(\frac{\partial L(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)}\right). \tag{A.15}$$

Furthermore, since only the cost component of the objective depends on $\hat{\theta}_{i,1}(a)$, we have

$$\frac{\partial L(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)} = -\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)\partial e_i^*}. \tag{A.16}$$

From Equation A.11 and the definition of $r^e(e_i;\hat{\theta}_{i,1}(a))$ (along with our assumption that the participant has resolved uncertainty over $\theta_i(a)$), we have

$$\begin{aligned} K(e_i;\hat{\theta}_{i,1}(a)) =& -G(e_i)\left\{\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i] + \eta\lambda\left(\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i] - G(e_i)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i]\right)\right\} \\ & - \eta[1-G(e_i)]G(e_i)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i]. \\ =& -\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i]G(e_i)\left\{1+\eta(\lambda-1)[1-G(e_i)]\right\}. \end{aligned} \tag{A.17}$$

Note that $\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i] = -\hat{\theta}_{i,1}(a)\frac{1}{G(e_i)}\int_0^{e_i} c(e)g(e)de$. Since g is a uniform PDF, it is constant. We denote this constant by $\bar{g}$, and thus $G(e) = \bar{g}e$. (Given that our experiment uses $e \sim \text{Uniform}[0,100]$, $\bar{g}$ in this case is $\frac{1}{100}$.) Furthermore, let $\bar{c}(e_i) \equiv \int_0^{e_i} c(e)de$, so $\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i] = -\hat{\theta}_{i,1}(a)\frac{\bar{g}}{G(e_i)}\bar{c}(e_i)$. From A.17, we thus have

$$K(e_i;\hat{\theta}_{i,1}(a)) = \hat{\theta}_{i,1}(a)\bar{g}\bar{c}(e_i)\left\{1+\Lambda[1-G(e_i)]\right\}, \tag{A.18}$$

where $\Lambda \equiv \eta(\lambda-1)$. Similar simplification of $B(e_i)$ in Equation A.10 yields

$$B(e_i) = mG(e_i)\left\{1-\Lambda[1-G(e_i)]\right\}. \tag{A.19}$$

From Equations A.18 and $A.19$, it is immediate that the solution depends on the reference-dependence

parameters only through the "composite" parameter $\Lambda = \eta(\lambda - 1)$. Furthermore, for any η, $\lambda = 1$ implies $\Lambda = 0$ and K and B reduce to the standard cost and benefit functions absent reference dependence, and hence the objective function reduces to the one in Equation A.9. Thus, without loss aversion, the optimal choice e_i^* is same regardless of whether the participant has reference-dependent utility or not; therefore, e_i^* is clearly decreasing in $\hat{\theta}_{i,1}(a)$.

We now consider the case with loss aversion, so $\Lambda > 0$. Together, Equations A.15 and A.16 imply that e_i^* is decreasing in $\hat{\theta}_{i,1}(a)$ if $\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)\partial e_i^*} > 0$. From A.18, $\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)\partial e_i^*} > 0$ iff

$$\begin{aligned} & c(e_i^*)\{1+\Lambda[1-G(e_i^*)]\} - \bar{g}\Lambda\bar{c}(e_i^*) > 0 \\ & \qquad \Leftrightarrow \{1+\Lambda[1-G(e_i^*)]\} > \bar{g}\Lambda\frac{\bar{c}(e_i^*)}{c(e_i^*)}. \end{aligned} \tag{A.20}$$

Furthermore, using Equations A.19 and A.18, the SOC implies that

$$\begin{aligned} & \left.\frac{\partial^2 B(e_i)}{\partial e_i^2}\right|_{e_i=e_i^*} - \left.\frac{\partial^2 K(e_i;\hat{\theta}_{i,1}(a))}{\partial e_i^2}\right|_{e_i=e_i^*} < 0 \\ & \qquad \Leftrightarrow 2m\bar{g}\Lambda < \hat{\theta}_{i,1}(a)\left[c'(e_i^*)\{1+\Lambda[1-G(e_i^*)\} - 2\bar{g}\Lambda c(e_i^*)\right]. \end{aligned} \tag{A.21}$$

Maintaining our implicit assumption that $\hat{\theta}_{i,1}(a) > 0$, Condition A.21 then holds only if

$$0 < c'(e_i^*)\{1+\Lambda[1-G(e_i^*)]\} - 2\bar{g}\Lambda c(e_i^*) \Leftrightarrow \{1+\Lambda[1-G(e_i^*)]\} > 2\bar{g}\Lambda\frac{c(e_i^*)}{c'(e_i^*)}. \tag{A.22}$$

Substituting inequality A.22 into A.20 establishes that $\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1}(a))}{\partial \hat{\theta}_{i,1}(a)\partial e_i^*} > 0$ if

$$2\frac{c(e_i^*)}{c'(e_i^*)} > \frac{\bar{c}(e_i^*)}{c(e_i^*)} \quad \Leftrightarrow \quad 2c(e_i^*)^2 > c'(e_i^*)\bar{c}(e_i^*). \tag{A.23}$$

Condition A.23 holds, for instance, for any $c(\cdot)$ that is a power function, as we assume in our parametric estimation. Under our specification of $c(e) = e^\gamma$ for $\gamma > 1$ (see Section 3.3), Condition A.23 is equivalent to

$$2e^{2\gamma} > \frac{\gamma}{\gamma+1}e^{2\gamma}. \tag{A.24}$$

We have therefore shown that, with a power-function cost specification (or any other specification that meets Condition A.23), the optimal action e_i^* is a decreasing function of $\hat{\theta}_{i,1}(a)$ when the participant's reference point is the expected value of the lottery induced by the BDM mechanism. Given that e_i^* is a decreasing function of $\hat{\theta}_{i,1}(a)$, the predictions of Observations 1 and 2 carry over to this setting. Namely: p does not directly influence a participant's objective function, but, under

misattribution, $e_i^*(a|p)$ is an increasing function of p because $\hat{\theta}_{i,1}(a)$ is a decreasing function of p.

E Experiment 2: Predictions of Reference Dependence

In this appendix we consider the predictions of the reference-dependent model absent misattribution in Experiment 2. In particular, we show that expectations-based reference dependence with a "forward looking" reference point (a la Kőszegi and Rabin) generates an effect that pushes effort in our experiment in the opposite direction as misattribution. Namely, reference-dependence causes participants assigned the noiseless task to (on average) increase effort across periods, and those assigned the noisy task to (on average) decrease effort across periods. In Section 4.2 we discussed how sufficiently strong misattribution generates the opposite pattern: participants assigned the noiseless task tend to *decrease* effort across periods, while those assigned the noisy task tend to increase it.

The analysis in this section builds on Appendix D, where we described how a participant with reference-dependent utility optimally chooses effort when her reference point incorporates the uncertainty introduced by the BDM mechanism. We now extend that analysis to the two-period setting of Experiment 2.

The key differentiating feature of Experiment 2 is that the participant's reference point when making her first effort decision might still reflect the lottery induced by the coin flip. That is because a participant's decision about effort comes roughly 10 minutes after the resolution of the coin flip, and this may be too little time for the participant's reference point to fully adapt to her assigned task. If her reference point does not adapt, then her expected utility on each dimension *prior* to the coin flip determines her reference point on that dimension. Thus, her reference point—and hence her behavior—depends on the utility she expects from *each* of the tasks, not just the one she is ultimately assigned. This contrasts with Experiment 1: in that design, a participant chooses effort only once, and that choice happens well after she learned her task assignment. This allows ample time for her reference point to adapt to her assigned task. Thus, in Experiment 1, a person's reference point at the time of choice depends solely on her realized task assignment and not on what "could have been" if the coin landed differently.

Before analyzing the case where the participant's reference point does not adapt to the assigned task by the time of her first effort decision, it is worth noting predictions for the case where it *does* adapt before the first decision. With a quickly adapting reference point, effort in each period is described by the single-decision solution derived in Appendix D. Thus, if we assume that, on average, participants have unbiased priors about $\theta(a)$—implying that the average of participants' expectations over $\theta(a)$ does not move in a systematic direction over time—then the average effort of those facing a given task is constant across periods. Fixing participants average beliefs over

$\theta(a)$, this effort level is given by the value $e^*(a)$ that maximizes Objective A.12 in Appendix D. As we showed there, $e^*(l) > e^*(h)$—effort by those assigned the noiseless task is predicted to be greater than those assigned the noisy task.

In contrast, even with unbiased priors (on average), reference dependence absent misattribution can generate systematic aggregate changes in effort across periods when participants' reference points do not adapt prior to the first decision. The basic intuition is as follows. Let $e_t^*(a)$ denote the optimal effort level a participant reports to the BDM mechanism in period t when assigned to task a.[38] In period 2—when the participant's task is fully anticipated—her optimal effort choice $e_2^*(a)$ will follow the derivation in Appendix D: she exhibits a high WTW if assigned the noiseless task, and a low WTW if assigned the noisy task. In period 1, however, her optimal strategy involves less effort on the noiseless task relative to period 2 (i.e, $e_1^*(l) < e_2^*(l)$), and *more* effort on the noisy task relative to period 2 (i.e,. $e_1^*(h) > e_2^*(h)$). In other words, the difference in effort across tasks is more compressed in period 1 than it is in period 2 (i.e., $e_1^*(l) - e_1^*(h) < e_2^*(l) - e_2^*(h)$).

The strategy described above is optimal because it mitigate losses. In particular, the participant chooses to work less on the noiseless task in period 1 (relative to what she would do in period 2) so that her expected payment and effort cost from the noiseless task in period 1 are more similar to what she expects to earn from the noisy task. By equalizing expected payments and effort costs across the tasks, neither assignment will generate large sensations of loss. For example, if the participant instead planned to work as much on the noiseless task in the first period as she would in the second period, then being assigned to the noisy task would come with a substantial loss on the money dimension: she would have earned more if she were assigned the noiseless task because she planned to work a substantial amount on that task. By planning to initially work less on the noiseless task (and more on the noisy task) relative to period 2, she can reduce such sensations of disappointment stemming from her assignment. Notice that this loss-mitigation strategy is only relevant when the participant compares her realized outcome to the expected outcomes from *each* possible task assignment—that is, when the participant's reference point has not adjusted to her assigned task. This is why such a strategy is irrelevant for our analysis of effort choices in Experiment 1 and in period 2 of Experiment 2.

We now formalize the intuition described above. As in Appendix D, we simplify the analysis by assuming the participant forms degenerate beliefs about $\theta(h)$ and $\theta(l)$ following the initial learning session; we denote these beliefs by $\hat{\theta}(h)$ and $\hat{\theta}(l)$, respectively.[39] Our approach follows the "per-

[38]The remainder of the analysis focuses on a single participant, and we will therefore drop subscripts denoting the participant's label (e.g., i) in order to reduce notational clutter.

[39]This allows us to derive predictions by focusing on a single participant. We could instead allow for uncertainty over $\theta_i(a)$ and derive aggregate predictions. Given our previous assumption that an individual's priors are unbiased on average (i.e., $\mathbb{E}[\hat{\theta}_{i,0}(a)|\theta_i(a)] = \theta_i(a)$), the average population beliefs should remain constant over time under rational updating.

son equilibrium" concept introduced by Kőszegi and Rabin (2006): prior to her task assignment, the participant forms effort plans for each possible assignment outcome and period, denoted by $e_t(a)$, and these plans determine her expectations in each period. A personal equilibrium requires that these plans are *consistent*: given the reference points induced by her plans, it is optimal for the participant to follow through with these plans. Furthermore, we will focus on the participant's *preferred* personal equilibrium, which is the consistent plan that provides the highest expected utility out of all consistent plans.

Reference Points. Let r_t^m and r_t^e denote the participant's reference point in period t over money and effort, respectively. We assume that these reference values are equal to the participant's expected monetary payment and effort cost in each round. In a personal equilibrium, these values are therefore endogenously determined by the participant's effort plans. To clarify, recall from our analysis of Experiment 1 (Appendix D) that the chance of working in period t conditional on announcing $e_t(a)$ to the BDM mechanism is $G(e_t(a)) = \bar{g}e_t(a)$ (where $\bar{g} = \frac{1}{100}$ since our experiment uses $e \sim \text{Uniform}[0,100]$). Furthermore, the expected disutility of effort from task a conditional on announcing $e_t(a)$ is $G(e_t(a))\widehat{\mathbb{E}}_t[V_t^e | e < e_t(a)] = -\bar{g}\hat{\theta}(a)\bar{c}(e_t(a))$, where $\bar{c}(e) = \int_0^e c(x)dx$. These formulae allow us to write r_t^m and r_t^e in terms of the participant's effort plans:

- In period 1, we assume the participant's reference point along each dimensions matches the expectations she forms *prior* to the coin flip. Thus, her expected disutility of effort is the average disutility she expects to face from the noiseless and noisy task, and her expected payment is the average earnings she expects from each task. These ex-ante expectations depend on the participant's effort plan contingent on each outcome of the coin-flip. Thus, if she plans to report $e_1(a)$ to the mechanism conditional on being assigned to task a, then her reference points are given by:

$$r_1^m = \frac{\bar{g}}{2}[e_1(h) + e_1(l)]m \text{ and } r_1^e = -\frac{\bar{g}}{2}[\hat{\theta}(h)\bar{c}(e_1(h)) + \hat{\theta}(l)\bar{c}(e_1(l))]. \tag{A.25}$$

- In period 2, we assume the participant's reference point along each dimension has adapted to her task: she expects to work on task a, and therefore she forms her expectations over her disutility of effort and payment conditional on working on task a. This matches our assumption for the reference point in the single-period analysis of Experiment 1. Thus, conditional on being assigned to task a, her reference points are given by:

$$r_2^m(a) = m\bar{g}e_2(a) \text{ and } r_2^e(a) = -\hat{\theta}(a)\bar{g}\bar{c}(e_2(a)), \tag{A.26}$$

We next analyze the optimal effort plans given these reference points they induce.

Objective Function. As in our analysis of Experiment 1, we can break down the objective

function in each period into the expected monetary benefit and expected effort cost. Following Equation A.10 from Appendix D, the expected monetary benefit from task a in period t is

$$\begin{aligned} B_t^a(e_t(a)|r_t^m) &\equiv G(e_t(a))\left\{m+\eta\,(m-r_t^m)\right\}-\eta\lambda[1-G(e_t(a))]r_t^m \\ &= (1+\eta)\bar{g}e_t(a)m-\eta\bar{g}e_t(a)r_t^m-\eta\lambda[1-\bar{g}e_t(a)]r_t^m \\ &= (1+\eta)\bar{g}e_t(a)m-\eta r_t^m-\Lambda[1-\bar{g}e_t(a)]r_t^m, \end{aligned} \tag{A.27}$$

where $\Lambda \equiv \eta(\lambda-1)$. Following Equation A.11 from Appendix D, the expected effort cost from task a in period t is

$$\begin{aligned} K_t^a(e_t(a)|r_t^e) &\equiv -G(e_t(a))\left\{\widehat{\mathbb{E}}_t\left[V_t^e|e<e_t(a)\right]+\eta\lambda\left(\widehat{\mathbb{E}}_t\left[V_t^e|e<e_t(a)\right]-r_t^e\right)\right\} \\ &\qquad +\eta[1-G(e_t(a))]r_t^e \\ &= (1+\eta\lambda)\hat{\theta}(a)\bar{g}\bar{c}(e_t(a))+\eta\lambda\bar{g}e_t(a)r_t^e+\eta[1-\bar{g}e_t(a)]r_t^e \\ &= (1+\eta\lambda)\hat{\theta}(a)\bar{g}\bar{c}(e_t(a))+\eta r_t^e+\Lambda\bar{g}e_t(a)r_t^e. \end{aligned} \tag{A.28}$$

Optimal Effort in $t=1$. In period $t=1$, the optimal effort choices, $e_1^*(h)$ and $e_1^*(l)$, jointly maximize $\frac{1}{2}[B_1^h(e_1(h)|r_1^m)-K_1^h(e_1(h)|r_2^e)]+\frac{1}{2}[B_1^l(e_1(l)|r_1^m)-K_1^l(e_1(l)|r_2^e)]$. The FOC with respect to $e_1(h)$ is thus

$$\begin{aligned} &(1+\eta)\bar{g}m-\{2\eta+\Lambda[2-\bar{g}(e_1(h)+e_1(l))]\}\frac{\partial r_1^m}{\partial e_1(h)}+\Lambda\bar{g}r_1^m \\ &\qquad -(1+\eta\lambda)\hat{\theta}(h)\bar{g}c(e_1(h))-\{2\eta+\Lambda\bar{g}(e_1(h)+e_1(l))\}\frac{\partial r_1^e}{\partial e_1(h)}-\Lambda\bar{g}r_1^e=0, \end{aligned} \tag{A.29}$$

and the FOC with respect to $e_1(l)$ is

$$\begin{aligned} &(1+\eta)\bar{g}m-\{2\eta+\Lambda[2-\bar{g}(e_1(h)+e_1(l))]\}\frac{\partial r_1^m}{\partial e_1(l)}+\Lambda\bar{g}r_1^m \\ &\qquad -(1+\eta\lambda)\hat{\theta}(h)\bar{g}c(e_1(l))-\{2\eta+\Lambda\bar{g}(e_1(h)+e_1(l))\}\frac{\partial r_1^e}{\partial e_1(l)}-\Lambda\bar{g}r_1^e=0. \end{aligned} \tag{A.30}$$

From the definitions of r_1^m and r_1^e in Equation A.25, we have

$$\frac{\partial r_1^m}{\partial e_1(a)}=\frac{m\bar{g}}{2} \quad \text{and} \quad \frac{\partial r_1^e}{\partial e_1(a)}=-\frac{\bar{g}}{2}\hat{\theta}(a)c(e_1(a)). \tag{A.31}$$

Hence, two FOCs above can be written, respectively, as

$$m\{1-\Lambda+\Lambda\bar{g}[e_1(h)+e_1(l)]\}-\hat{\theta}(h)\{1+\Lambda-\Lambda\frac{\bar{g}}{2}[e_1(h)+e_1(l)]\}c(e_1(h)) \\ +\Lambda\frac{\bar{g}}{2}[\hat{\theta}(h)\bar{c}(e_1(h))+\hat{\theta}(l)\bar{c}(e_1(l))]=0, \quad \text{(A.32)}$$

and

$$m\{1-\Lambda+\Lambda\bar{g}[e_1(h)+e_1(l)]\}-\hat{\theta}(l)\{1+\Lambda-\Lambda\frac{\bar{g}}{2}[e_1(h)+e_1(l)]\}c(e_1(l)) \\ +\Lambda\frac{\bar{g}}{2}[\hat{\theta}(h)\bar{c}(e_1(h))+\hat{\theta}(l)\bar{c}(e_1(l))]=0. \quad \text{(A.33)}$$

Let $L_1^h(e_1(h),e_1(l))$ and $L_1^l(e_1(h),e_1(l))$ denote the functions defined by the left-hand side of Equations A.32 and A.33, respectively. Restricting attention to interior solutions, the optimal effort plans in period 1, $e_1^*(h)$ and $e_1^*(l)$, must solve the system of equations given by $L_1^h(e_1^*(h),e_1^*(l))=0$ and $L_1^l(e_1^*(h),e_1^*(l))=0$. Notice, however, that if $L_1^h(e_1^*(h),e_1^*(l))=L_1^l(e_1^*(h),e_1^*(l))=0$, then it is immediate from Equations A.32 and A.33 that $\hat{\theta}(h)c(e_1^*(h))=\hat{\theta}(l)c(e_1^*(l))$; that is, the optimal effort levels are chosen to equalize the "consumption utility" of effort across the two tasks. To summarize:

Lemma A.1. *Given the setup formalized above, the optimal effort choices in period 1, $e_1^*(h)$ and $e_1^*(l)$, are such that the participant's effort cost is the same regardless of her task assignment; that is, $\hat{\theta}(h)c(e_1^*(h))=\hat{\theta}(l)c(e_1^*(l))$. Furthermore, given that $\hat{\theta}(h)>\hat{\theta}(l)$, the participant's initial effort on the noiseless task exceeds her initial effort on the noisy task; that is, $e_1^*(l)>e_1^*(h)$.*

Optimal Effort in $t=2$. In period $t=2$, the optimal effort choice conditional on being assigned to task a, $e_2^*(a)$, maximizes $B_2^a(e_2(a)|r_2^m(a))-K_2^a(e_2(a)|r_2^e(a))$, and thus solves the following FOC:

$$(1+\eta)\bar{g}m-\{\eta+\Lambda[1-\bar{g}e_2(a)]\}\frac{\partial r_2^m(a)}{\partial e_2(a)}+\Lambda\bar{g}r_2^m(a) \\ -(1+\eta\lambda)\hat{\theta}(a)\bar{g}c(e_2(a))-\{\eta+\Lambda\bar{g}e_2(a)\}\frac{\partial r_2^e(a)}{\partial e_2(a)}-\Lambda\bar{g}r_2^e(a)=0. \quad \text{(A.34)}$$

From the definition of $r_2^m(a)$ and $r_2^e(a)$ in Equation A.26, we have

$$\frac{\partial r_2^m(a)}{\partial e_2(a)}=m\bar{g} \text{ and } \frac{\partial r_2^e(a)}{\partial e_2(a)}=-\hat{\theta}(a)\bar{g}c(e_2(a)), \quad \text{(A.35)}$$

and thus the FOC in Equation A.34 can be written as

$$m\{1-\Lambda+2\Lambda\bar{g}e_2(a)\}-\hat{\theta}(a)\{1+\Lambda-\Lambda\bar{g}e_2(a)\}c(e_2(a))+\hat{\theta}(a)\Lambda\bar{g}\bar{c}(e_2(a))=0. \quad \text{(A.36)}$$

Let $L_2^a(e)$ denote the function on the left-hand-side of the FOC above. Thus, $e_2^*(a)$ is such that $L_2^a(e_2^*(a))=0$.

Given that this FOC takes the same form as the one in Appendix D, the analysis from Appendix D implies that $e_2^*(a)$ is decreasing in $\hat{\theta}(a)$. Thus, given that $\hat{\theta}(h)>\hat{\theta}(l)$, we have that $e_2^*(l)>e_2^*(h)$: similar to period 1, the optimal plan in period 2 calls for greater effort when facing the noiseless task than the noisy one.

Changes in Optimal Effort Across Periods. We have so far shown that a participant who believes $\hat{\theta}(h)>\hat{\theta}(l)$ will, within each period, exert more effort if she's assigned the noiseless task rather than the noisy one. But, fixing the task she ultimately faces, how will her effort change *across* periods? The final step of our analysis compares $e_1^*(a)$ with $e_2^*(a)$ for each $a\in\{h,l\}$. For this step, we will simplify matters by assuming—as in previous sections—that effort costs follow a power function; that is, $c(e)=e^\gamma$ for some $\gamma>1$. We will first consider changes in effort for the noiseless task ($a=l$) and then consider the noisy task ($a=h$).

1. *Willingness to Work on the Noiseless Task Increases Across Periods*. We now show that $e_1^*(l)<e_2^*(l)$. From Lemma A.1, we have $\hat{\theta}(l)c(e_1^*(l))=\hat{\theta}(h)c(e_1^*(h))$ and thus

$$e_1^*(h)=c^{-1}\left(\frac{\hat{\theta}(l)}{\hat{\theta}(h)}c(e_1^*(l))\right)=\left(\frac{\hat{\theta}(l)}{\hat{\theta}(h)}\right)^{\frac{1}{\gamma}}e_1^*(l)=\psi_L e_1^*(l), \quad \text{(A.37)}$$

where $\psi_L\equiv\left(\frac{\hat{\theta}(l)}{\hat{\theta}(h)}\right)^{\frac{1}{\gamma}}<1$. In light of Equation A.37, we can write the FOC $L_1^l(e_1^*(h),e_1^*(l))$ characterizing effort in period 1 entirely in terms of $e_1^*(l)$ by substituting the expression for $e_1^*(h)$ from Equation A.37 into Equation A.33, which yields[40]

$$m\{1-\Lambda+\Lambda\bar{g}[1+\psi_L]e_1^*(l)\}-\hat{\theta}(l)\{1+\Lambda-\Lambda\frac{\bar{g}}{2}[1+\psi_L]e_1^*(l)\}c(e_1^*(l))$$
$$+\Lambda\frac{\bar{g}}{2}[1+\psi_L]\hat{\theta}(l)\bar{c}(e_1^*(l))=0. \quad \text{(A.38)}$$

Letting $\widetilde{L}^l(e_1(l);\psi)$ denote the left-hand side of the equation above as a function of $e_1(l)$ and the parameter ψ, we have that $e_1^*(l)$ must satisfy $\widetilde{L}^1(e_1^*(l);\psi_L)=0$. Notice, however,

[40]Recall that $\bar{c}(e)=\frac{e^{\gamma+1}}{\gamma+1}$. Hence, $e_1(h)=\psi_L e_1(l)$ implies that $\hat{\theta}(h)\bar{c}(e_1(h))=\hat{\theta}(h)\left(\frac{\hat{\theta}(l)}{\hat{\theta}(h)}\right)^{1+1/\gamma}e_1(l)^{\gamma+1}/(\gamma+1)=\hat{\theta}(h)\frac{\hat{\theta}(l)^{1/\gamma}}{\hat{\theta}(h)^{1+1/\gamma}}\bar{c}(e_1(l))=\psi_L\bar{c}(e_1(l))$.

that the FOC characterizing $e_2^*(l)$ (Equation A.36) is identical to the FOC above except the parameter ψ takes value 1 instead of $\psi_L < 1$; that is, $e_2^*(l)$ solves $\widetilde{L}^l(e_2^*(l);1) = 0$. Thus, to show that $e_1^*(l) < e_2^*(l)$, it suffices to show that the solution to $\widetilde{L}^l(e^*;\psi) = 0$ is increasing in ψ. Using the implicit function theorem,

$$\frac{\partial e^*}{\partial \psi} = -\left(\frac{\partial \widetilde{L}^l(e^*;\psi)}{\partial e^*}\right)^{-1} \frac{\widetilde{L}^l(e^*;\psi)}{\partial \psi}. \tag{A.39}$$

Focusing on interior solutions, the second-order condition implies that $\frac{\partial \widetilde{L}^l(e^*;\psi)}{\partial e^*} < 0$, and hence $\frac{\partial e^*}{\partial \psi} > 0 \Leftrightarrow \frac{\widetilde{L}^l(e^*;\psi)}{\partial \psi} > 0$. From Equation A.38, we have

$$\frac{\widetilde{L}^l(e^*;\psi)}{\partial \psi} = m\Lambda\bar{g}e^* + \Lambda\frac{\bar{g}}{2}\hat{\theta}(l)[e^*c(e^*) + \bar{c}(e^*)] > 0, \tag{A.40}$$

which therefore establishes that $e_2^*(l) > e_1^*(l)$.

2. *Willingness to Work on the Noisy Task Decreases Across Periods.* We now show that $e_1^*(h) > e_2^*(h)$. The argument is symmetric to the one above. Namely, from Lemma A.1, the optimal effort in period 1 must satisfy $e_l^*(l) = \psi_H e_1^*(h)$, where $\psi_H = 1/\psi_L > 1$. Substituting this expression for $e_l^*(l)$ into the FOC in Equation A.32 implies that $e_1^*(h)$ must solve

$$\begin{aligned} m\{1-\Lambda+\Lambda\bar{g}[1+\psi_H]e_1^*(h)\} - \hat{\theta}(h)\{1+\Lambda-\Lambda\frac{\bar{g}}{2}[1+\psi_H]e_1^*(h)\}c(e_1^*(h)) \\ +\Lambda\frac{\bar{g}}{2}[1+\psi_H]\hat{\theta}(h)\bar{c}(e_1^*(h)) = 0. \end{aligned} \tag{A.41}$$

Again letting $\widetilde{L}^h(e_1(h);\psi)$ denote the left-hand side of the equation above as a function of $e_1(h)$ and the parameter ψ, we have that $e_1^*(h)$ must satisfy $\widetilde{L}^h(e_1^*(h);\psi_H) = 0$. Equation A.36 reveals, however, that $e_2^*(h)$ must solve $\widetilde{L}^h(e_1^*(h);1) = 0$. The implicit-function-theorem argument from the previous part implies that the solution to $\widetilde{L}^h(e^*;\psi) = 0$ is also increasing in ψ. Thus, since $\psi_H > 1$, we have that $e_2^*(h) < e_1^*(h)$.

F Additional Theoretical Details

In this appendix we provide details on how misattribution interferes with belief updating when a participant misencodes signals but otherwise follows Bayes' Rule. These details provide a more formal derivation of the predictions generated by misattribution described in Sections 3.2 and 4.2. We consider two rounds of learning; that is, the participant receives two signals in sequence. We

examine beliefs and behavior after the first round to address predictions for Experiment 1, and then consider both the first and second round to address predictions for Experiment 2. Regarding Experiment 1, we demonstrate that a misattributing participant i will form a systematically distorted estimate of her underlying effort-cost parameter, $\theta_i(a)$. This estimate undershoots the true value when she is assigned the noiseless task and overshoots it when she is assigned the noisy one. We then demonstrate that the second signal is likely to move the participant's estimate in the *opposite* direction of her initial error: her estimated cost of the noiseless task following the second round tends to increase while that of the noisy task tends to decrease.

As in previous sections, we assume that each participant i has prior beliefs over $\theta_i(a)$ that follow a normal distribution: $\theta_i(a) \sim N\left(\hat{\theta}_{i,0}(a), \rho^2\right)$, where $\hat{\theta}_{i,0}(h) > \hat{\theta}_{i,0}(l)$. We also assume that participants' initial estimates of $\theta_i(a)$ are unbiased in the population so that for all i and a, $\mathbb{E}[\hat{\theta}_{i,0}(a)|\theta_i(a)] = \theta_i(a)$; hence, the initial estimates of $\theta_i(a)$ averaged across individuals is equal to the true value. A participant's signals about $\theta_i(a)$ stem from either the single initial learning session in Experiment 1, or from both learning sessions in Experiment 2.[41] Assuming the participant is assigned task a, each learning session $t \in \{1,2\}$ provides a signal $X_{i,t}(a) = \theta_i(a) + \varepsilon_{i,t}$, where $\varepsilon_{i,t} \sim N\left(0, \sigma^2\right)$.[42] Let $x_{i,t}(a)$ denote the realized signal and let $\hat{x}_{i,t}(a)$ denote the misattributor's encoded value of this signal.

Recall that $\hat{\theta}_{i,t}(a)$ denotes the participant's estimate of $\theta(a)$ entering period $t = 1,2$. Given our normality assumptions, a participant who is Bayesian (aside from misencoding signals) updates her beliefs as follows:

$$\hat{\theta}_{i,t}(a) = \hat{\theta}_{i,t-1}(a) + \alpha_t \left[\hat{x}_{i,t}(a) - \hat{\theta}_{i,t-1}(a)\right], \tag{A.42}$$

where $\alpha_t = \frac{\rho^2}{t \cdot \rho^2 + \sigma^2}$. The encoded signal, $\hat{x}_{i,t}$, is defined by Equation 3, and thus depends on how the true signal, $x_{i,t}(a)$, compares to the participant's expected effort cost entering round t. This expectation in turn depends on the participant's treatment group: those who are initially uncertain which task they will work on will hold different expectations about their eventual effort cost than those who are certain. We will therefore examine how updating differs depending on the participant's treatment group, p, where where p denotes the participant's ex ante likelihood of being assigned the noisy task. Recall that $p = 1$ and $p = 0$ correspond to the control group, and $p = 1/2$ corresponds to the coin-flip group.

We now analyze the encoded signal that a misattributing participant forms based on her assigned task and treatment group. For this exercise, we fix the true signal the participant receives in a given period, $x_{i,t}(a)$, and consider how she would encode this signal if she were in the coin-flip

[41] We focus on a participant who is not assigned to do additional work in the first session of Experiment 2. Hence, the initial learning sessions comprise the participant's only experience with the task.

[42] Notice that $X_{i,t}(a) = -V^e_{i,t}(a)/c(e_{i,t})$, where $V^e_{i,t}(a)$ is defined in Equation A.52 and $e_{i,t}$ is the required number of trials the participant must complete in learning-session t.

group versus the control group. We let $\hat{x}_{i,t}(a|p)$ denote this misencoded signal as a function of the treatment, p. Once we establish the direction in which signals are biased across treatment groups, the updating rule in Equation A.42 will then immediately reveal how the average estimate of the cost parameter in a given period should differ across treatments under misattribution.

Biased Updating in Period 1. We begin by analyzing how the treatment distorts signals in the first period. The predictions we obtain here primarily relate to Experiment 1. Consider participant i whose treatment group is such that she expects to face the noisy task with probability p. Her expected cost signal entering period 1 is thus $\widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p] \equiv p\hat{\theta}_{i,0}(h) + (1-p)\hat{\theta}_{i,0}(l)$. Upon realizing signal $x_{i,1}(a)$, Equation 3 implies that her misencoded signal is

$$\hat{x}_{i,1}(a|p) = \begin{cases} x_{i,1}(a) + \left(\frac{\eta-\hat{\eta}}{1+\hat{\eta}}\right)\left(x_{i,1}(a) - \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p]\right) & \text{if} \quad x_{i,1}(a) \leq \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p] \\ x_{i,1}(a) + \lambda\left(\frac{\eta-\hat{\eta}}{1+\hat{\eta}\lambda}\right)\left(x_{i,1}(a) - \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p]\right) & \text{if} \quad x_{i,1}(a) > \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p]. \end{cases} \tag{A.43}$$

Letting $\kappa^G \equiv \left(\frac{\eta-\hat{\eta}}{1+\hat{\eta}}\right)$ and $\kappa^L \equiv \lambda\left(\frac{\eta-\hat{\eta}}{1+\hat{\eta}\lambda}\right)$, we can define the following random variable that measures the extent to which a signal is misencoded:

$$K_{i,1}(a|p) \equiv \begin{cases} \kappa^L & \text{if} \quad x_{i,1}(a) > \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p] \\ \kappa^G & \text{if} \quad x_{i,1}(a) \leq \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p]. \end{cases} \tag{A.44}$$

Thus, we can write the misencoded signal in Equation A.43 more simply as

$$\begin{aligned} \hat{x}_{i,1}(a|p) &= x_{i,t}(a) + k_{i,1}(a|p)\left[x_{i,1}(a) - \widehat{\mathbb{E}}_{i,0}[X_{i,1}(a)|p]\right] \\ &= x_{i,t}(a) + k_{i,1}(a|p)\left[x_{i,1}(a) - p\hat{\theta}_{i,0}(h) - (1-p)\hat{\theta}_{i,0}(l)\right], \end{aligned} \tag{A.45}$$

where $k_{i,1}(a|p)$ is the realization of $K_{i,1}(a|p)$. Notice that if the participant does not suffer misattribution (i.e., $\hat{\eta} = \eta$), then $k_{i,1}(a|p)$ is always equal to zero. Furthermore, if the participant does suffer missatribution and is loss averse, then $\kappa^L > \kappa^G$, implying that high cost signals are distorted upward more than low cost signals are distorted downward.[43]

First consider how signals about the noiseless task in particular are differentially distorted depending on whether the participant is in the coin-flip group (i.e., $p = 1/2$) or the control group

[43]Recall that $x_{i,t}(a)$ reflects the participant's *cost* in period t. Hence, the participant experiences a positive surprise when her signal is *less* than expected; she experiences a negative surprise when it is greater than expected. This is why the signs in Equation A.43 are flipped relative to 3—the latter was written in terms of benefits rather than costs.

(i.e., $p = 0$). From Equation A.45, we have

$$\begin{aligned}\hat{x}_{i,1}(l|p=1/2) - \hat{x}_{i,1}(l|p=0) \\ = k_{i,1}(l|1/2)\left[x_{i,1}(l) - .5\hat{\theta}_{i,0}(h) - .5\hat{\theta}_{i,0}(l)\right] - k_{i,1}(l|0)\left[x_{i,1}(l) - \hat{\theta}_{i,0}(l)\right] \quad \text{(A.46)}\end{aligned}$$

There are three cases to consider, depending on the values of $k_{i,1}(l|1/2)$ and $k_{i,1}(l|0)$:

i. $x_{i,1}(a) > .5\hat{\theta}_{i,0}(h) + .5\hat{\theta}_{i,0}(l)$, in which case $k_{i,1}(l|1/2) = k_{i,1}(l|0) = \kappa^L$;

ii. $x_{i,1}(a) \in \left[\hat{\theta}_{i,0}(l), .5\hat{\theta}_{i,0}(h) + .5\hat{\theta}_{i,0}(l)\right]$, in which case $k_{i,1}(l|1/2) = \kappa^G$ and $k_{i,1}(l|0) = \kappa^L$;

iii. $x_{i,1}(a) < \hat{\theta}_{i,0}(l)$, in which case $k_{i,1}(l|1/2) = k_{i,1}(l|0) = \kappa^G$.

In cases (i) and (iii), $k_{i,1}(l|0)$ and $k_{i,1}(l|1/2)$ are both equal to the same $\kappa^j \in \{\kappa^G, \kappa^L\}$, and hence A.46 reduces to

$$\hat{x}_{i,1}(l|p=1/2) - \hat{x}_{i,1}(l|p=0) = -\frac{\kappa^j}{2}[\hat{\theta}_{i,0}(h) - \hat{\theta}_{i,0}(l)] < 0. \quad \text{(A.47)}$$

In case (ii), $k_{i,1}(l|0)$ and $k_{i,1}(l|1/2)$ differ, leading to

$$\begin{aligned}\hat{x}_{i,1}(l|p=1/2) - \hat{x}_{i,1}(l|p=0) &= \kappa^G[x_{i,1}(l) - .5\hat{\theta}_{i,0}(h) - .5\hat{\theta}_{i,0}(l)] - \kappa^L[x_{i,1}(l) - \hat{\theta}_{i,0}(l)] \\ &< \kappa^G[x_{i,1}(l) - .5\hat{\theta}_{i,0}(h) - .5\hat{\theta}_{i,0}(l)] - \kappa^G[x_{i,1}(l) - \hat{\theta}_{i,0}(l)] \\ &= -\frac{\kappa^G}{2}[\hat{\theta}_{i,0}(h) - \hat{\theta}_{i,0}(l)] \\ &< 0, \quad \text{(A.48)}\end{aligned}$$

where the first inequality follows because $x_{i,1}(l) - \hat{\theta}_{i,0}(l) > 0$ given that $x_{i,1}(a) \in [\hat{\theta}_{i,0}(l), .5\hat{\theta}_{i,0}(h) + .5\hat{\theta}_{i,0}(l)]$.

All three cases together imply that participant i facing task l will always encode a lower signal if she were in the coin-flip group rather than the control group. An entirely symmetric argument (omitted) implies that participant i facing task h will always record a *higher* signal if she were in the coin-flip group rather than the control group. Thus, assuming participants update according to Equation A.42, the preceding results imply that $\hat{\theta}_{i,1}(l|p=1/2) < \hat{\theta}_{i,1}(l|p=0)$ and $\hat{\theta}_{i,1}(h|p=1/2) > \hat{\theta}_{i,1}(h|p=1)$, where $\hat{\theta}_{i,1}(a|p)$ denotes a participant's predicted expectation of $\theta_i(a)$ conditional on her treatment group, p.[44] The results of our parametric estimation in Section 3.3 mirror these predictions (see Table 3). Furthermore, given this ordering of beliefs across treatment groups, the analysis of Appendices C and D shows that these beliefs generate the predicted differences in aggregate effort described Observation 2.

[44]This prediction invokes our assumption that a participant's priors over the cost parameters are independent of their treatment assignment.

Biased Updating In Period 2. We now demonstrate how misattribution generates a predictable change in beliefs between periods 1 and 2 depending on a participant's task assignment. The predictions we obtain here relate exclusively to Experiment 2 since Experiment 1 had only one round of experience prior to the sole effort decision. Recall that in Experiment 2, each participant was assigned her task via a coin flip. Thus, in this analysis, a participant's period 1 beliefs, $\hat{\theta}_{i,1}(a)$, correspond to $\hat{\theta}_{i,1}(a|1/2)$ derived above.

We examine how these beliefs change after another round of learning. We maintain our assumption that the participant's reference point in the second period adjusts to her assigned task. That is, $\widehat{\mathbb{E}}_{i,1}[X_{i,2}(a)] = \hat{\theta}_{i,1}(a)$. Thus, the participant's encoded signal is

$$\begin{aligned} \hat{x}_{i,2}(a) &= x_{i,2}(a) + k_{i,2}(a)\left[x_{i,2}(a) - \widehat{\mathbb{E}}_{i,1}[X_{i,2}(a)]\right] \\ &= x_{i,2}(a) + k_{i,2}(a)\left[x_{i,2}(a) - \hat{\theta}_{i,1}(a)\right], \end{aligned} \tag{A.49}$$

where $k_{i,2}(a)$ is the realization of

$$K_{i,2}(a) \equiv \begin{cases} \kappa^L & \text{if} \quad x_{i,2}(a) > \widehat{\mathbb{E}}_{i,1}[X_{i,2}(a)] \\ \\ \kappa^G & \text{if} \quad x_{i,2}(a) \leq \widehat{\mathbb{E}}_{i,1}[X_{i,2}(a)]. \end{cases} \tag{A.50}$$

From the updating rule in Equation (A.42) and the expression above for the misencoded signal, the participant's expected change in beliefs between periods 1 and 2 (from an ex-ante perspective) is thus equal to

$$\begin{aligned} \mathbb{E}[\hat{\theta}_{i,2}(a) - \hat{\theta}_{i,1}(a)] &= \alpha_2 \mathbb{E}[\widehat{X}_{i,2}(a) - \hat{\theta}_{i,1}(a)] \\ &= \alpha_2 \mathbb{E}[(1 + K_{i,2}(a))(X_{i,2}(a) - \hat{\theta}_{i,1}(a))], \end{aligned} \tag{A.51}$$

where $\mathbb{E}[\cdot]$ denotes expectations with respect to the true underlying distributions. Notice that Equation A.51 is positive iff $\mathbb{E}[X_{i,2}(a)] > \mathbb{E}[\hat{\theta}_{i,1}(a)]$. Given that signals are independent across periods with mean $\theta_i(a)$, the previous condition holds iff $\theta_i(a) > \mathbb{E}[\hat{\theta}_{i,1}(a)]$. As we argued above, assignment to the noiseless task via the coin-flip implies that $\hat{\theta}_{i,1}(l)$ is biased downward.[45] Hence the previous inequality holds and thus, in expectation, $\hat{\theta}_{i,2}(l) > \hat{\theta}_{i,1}(l)$: the perceived effort cost of the noiseless task tends to increase across periods, reducing the participant's WTW on that task. Similarly, assignment to the noisy task via the coin-flip implies that $\hat{\theta}_{i,1}(h)$ is biased upward. Hence the inequality above *fails* to hold, and thus, in expectation, $\hat{\theta}_{i,2}(h) < \hat{\theta}_{i,1}(h)$: the perceived effort cost of the noisy task tends to decrease across periods, increasing the participant's WTW

[45] Our assumption that priors are unbiased in the population is relevant here: this assumption implies that, on average, $\hat{\theta}_{i,1}(l) < \theta_i(a)$; that is, that the average expectation among participants regarding $\theta(l)$ after one round of learning is lower than the true parameter value.

on that task. This pattern in beliefs generated by misattribution clearly runs against the predictions of reference-dependence absent misattribution (explored in Appendix E), where effort on the noiseless task tends to increase across periods while effort on the noisy task tends to decrease.

G Back-of-the-Envelope Parameter Estimates

This section presents the simple calculations that underlie our parameter estimates discussed at the end of Section 3.3.

We build from the belief-formation model presented in Section 3.2. Recall that Table 3 presents estimates of participants' perceptions of the cost parameter $\theta(a)$ for each task across the various treatment groups. The model in Section 3.2 can deliver a system of equations that characterize the predicted perceptions of $\theta(a)$ across groups of as a function of the underlying reference-dependence parameters, (η, λ), and the degree of misattribution, $\hat{\eta}$. Here, we substitute our estimated perceptions of $\theta(a)$ from Table 3 into this system of equations and solve for the implied values of η, λ, and $\hat{\eta}$.

This approach implicitly assumes that the values in Table 3 reflect the perceptions that a fixed representative agent would form in each of the treatment groups. This exercise therefore calculates the preference and bias parameters of this representative agent. Accordingly, we drop the participant label i from the subsequent notation.

Recall from Section 3.2 that the agent's consumption utility in the initial session is

$$v_1^e = -\left[\theta_i(a) + \varepsilon_1\right] c(8). \tag{A.52}$$

As above, we assume the agent believes that $\theta(a) \sim N\left(\hat{\theta}_0(a), \rho^2\right)$ and $\varepsilon_t \sim N(0, \sigma^2)$, which implies that her updated perception of $\theta(a)$ is

$$\hat{\theta}_1(a) = -\alpha\left(\frac{\hat{v}_1^e}{c(8)}\right) + (1-\alpha)\hat{\theta}_0(a) \quad \text{where} \quad \alpha \equiv \frac{\rho^2}{\rho^2 + \sigma^2}, \tag{A.53}$$

where $\hat{v}_1$ is agent's (mis)encoded consumption value. We simplify matters in two ways: (i) we consider the limit in which $\sigma \to 0$, and (ii) we assume that the agent's prior expectations match the true values, so $\hat{\theta}_0(a) = \theta(a)$ for $a = h, l$. The first simplification implies that the agent's consumption utility is approximately $v_1^e = -\theta(a)c(8)$, and that the agent's updated perception of $\theta(a)$ is approximately $\hat{\theta}_1(a) = -\hat{v}_1^e / c(8)$. The second simplification implies that the agent's beliefs are unbiased to begin with, and thus the perceptions we estimate from the control conditions reveal the agent's priors.

We take the estimated values in Table 3 to represent the values of $\hat{\theta}_1(a)$ resulting from each

possible treatment condition. To derive equations that characterize these values, we must consider the agent's misencoded consumption value in each condition. Using Equation 3, the $\hat{\theta}_1$ is given by

$$\hat{v}_1^e = \begin{cases} -\hat{\theta}(a)c(8) + \kappa^G\left(-\hat{\theta}(a)c(8) - \widehat{\mathbb{E}}[V_1]\right) & \text{if} \quad -\theta(a)c(8) \geq \widehat{\mathbb{E}}[V_1] \\ -\theta(a)c(8) + \lambda\left(\frac{\eta-\hat{\eta}}{1+\hat{\eta}\lambda}\right)\left(-\hat{\theta}(a)c(8) - \widehat{\mathbb{E}}[V_1]\right) & \text{if} \quad -\theta(a)c(8) < \widehat{\mathbb{E}}[V_1], \end{cases} \tag{A.54}$$

where $\kappa^G \equiv \frac{\eta-\hat{\eta}}{1+\hat{\eta}}$ and $\kappa^L \equiv \lambda\left(\frac{\eta-\hat{\eta}}{1+\hat{\eta}\lambda}\right)$. While we are not able to obtain separate estimates of η, λ, and $\hat{\eta}$, we will be able to solve for implied values of κ^G and κ^L. The magnitude of these two summary statistics help describe the extent of misencoding—since they would be both be zero absent misattribution—and the difference between them reveals the extent of asymmetric encoding of gains and losses due to loss aversion.

As a first step toward calculating κ^G and κ^L, note that $\widehat{\mathbb{E}}[V_1]$ in Equation A.43 varies across conditions. Given our assumption that priors are unbiased, the coin-flip condition induces $\widehat{\mathbb{E}}[V_1] = -.5c(8)(\theta(l) + \hat{\theta}(h))$. In contrast, the control condition facing task a induces $\widehat{\mathbb{E}}[V_1] = -\hat{\theta}(a)c(8)$. It is thus apparent from Equation A.43 that the control conditions involve no misencoding. Hence, the estimated value of $\hat{\theta}_1(\text{noise}|p=1) = .0493$ reported in Table 3 gives us $\theta(h)$. Similarly, the estimated value of $\hat{\theta}(\text{no noise}|p=0) = .0385$ gives us $\theta(l)$.[46]

Turning to the coin-flip condition, let $\hat{\theta}_1(a|p=.5)$ denote the agent's updated perception of $\theta(a)$ after facing task a in the coin-flip condition. Recall that $\hat{\theta}_1(a|p=.5) = -\hat{v}_1^e/c(e)$, where $\hat{v}_1^e$ is the misencoded value induced by the coin-flip condition; this value is obtained by substituting our expression for $\widehat{\mathbb{E}}[V_1]$ in coin-flip case into Equation A.43. This yields

$$\hat{\theta}_{i,1}(a|p=.5) = \begin{cases} \theta_i(l) - .5\kappa^G(\theta_i(h) - \theta_i(l)) & \text{if} \quad a = l \\ \theta_i(l) + .5\kappa^L(\theta_i(h) - \theta_i(l)) & \text{if} \quad a = h. \end{cases} \tag{A.55}$$

As described above, the control conditions yield numerical estimates of $\theta_i(a)$. And the two coin-flip estimates from Table 3 provide numerical estimates for the left-hand side of Equation A.55; namely, $\hat{\theta}_1(l|p=.5) = 0.0325$ and $\hat{\theta}_1(h|p=.5) = 0.0635$. Thus, the only unknowns in System A.55 are κ^G and κ^L. Solving these two equations for these values yields $\kappa^G = 1.1111$ and $\kappa^L = 2.6296$.

[46]For the calculations in this section, we use estimates from Column 2 of Table 3, which uses the full sample and includes controls.

H Experimental Instructions

In this section, we provide the full text of our experimental instructions. We use brackets to denote alternative instructions corresponding to different treatments. All instructions commenced with an informed consent form. The research in this study was reviewed by the Human Research Protection Program at Harvard University (protocol numbers: IRB15-0365 and IRB16-0944).[47]

H.1 Sample Reviews, Experiment 1

For a full text of the reviews used in Experiment 1, please contact the authors.

"To read this book is to go on a journey to places at once unexpected yet familiar; for example, one point is supported by reference to a diagram of nose shapes and sizes. His books teach rather than exposit; they do not lack for a direct thesis–they make arguments and reach conclusions."
Score: 5; Positive Review

"Sometimes you don't go out and find a book; the book finds you. Facing an impending loss without a foundation of faith to fall back on, I asked myself: 'What is the meaning of life if we're all just going to die?' The author answers that question in the most meaningful way possible."
Score: 5; Positive Review

"To be sure, this is a very quick read. The book is already very tiny, and the inside reveals large font and double spacing. It took me about two hours to finish this book. I believe I am an somewhat slow reader compared to other bookworms. On the other hand, I found many other books to be much more compelling and memorable takes on the meaning of life."
Score: 1; Negative Review

"Sometimes books like this are a real bore. Even worse, sometimes the science is terrible or inconsistent. I was pleased to find that this book is consistent with the established literature while also providing new insight."
Score: 5; Positive Review

"This book is nothing you expect it to be. I was looking forward to fun, witty tales of some of the author's romances. But no. He teamed up with a sociologist, and wrote a sociology textbook. It's bland and it's boring, with research percentages and the odd pie chart thrown in to liven things up."
Score: 1; Negative Review

[47] The Nock Lab at Harvard generated the noise used in our experiments. They used the stimuli in work unrelated to our own. In their studies, this sound was played at modest volume (slightly louder than we played the noise). Participants in their (more extensive) studies found the sound unpleasant, but with no lasting effects (e.g., ringing ears).

H.2 Complete Experiment Instructions: Experiment 1

Session 1

We will begin with some simple demographic questions.

What is your gender?
☐ Male ☐ Female
What is your annual income?
☐ less than $15,000
☐ $15,000 - $29,999
☐ $30,000 - $59,999
☐ $60,000 - $99,999
☐ $100,000 or more
What is your age (in years)?
What is your zip code? [Format: 00000]

We will not deceive you whatsoever in this experiment. All of the instructions provide examples and guidance for the actual tasks you will do. There will be no surprises or tricks. This study will consist of two sessions. You will do the first session now. You will sign in to do the second session later. In each session, you will do a simple job that takes roughly 3 to 5 minutes. You will earn a fixed payment of $4 for completing both sessions. In the second session, you will have the chance to earn extra pay if you elect to do extra work. You must complete both sessions to earn any pay for this study. There will be absolutely no exceptions to this rule. All payments will be credited to your MTurk account within one week of completing the study.

The second session will be unlocked 8 hours after the first session. In order to unlock the second session, a link will be emailed to you. We ask that you complete the second session as soon as you are able to. You must complete the second session within one week of the email in order to receive payment.

Your task in both sessions will be listening a series of audio recordings of book reviews (from Amazon) to determine whether each review is generally positive or negative.

You must wait at least 10 seconds before any buttons will appear. You must then decide if the review is positive or negative. A positive review means that the reviewer generally liked the book and is providing a recommendation. A negative review means that the reviewer generally disliked the book and is cautioning against reading it.

We will now give you a sample task to practice. Once you have listened to the review and correctly determined if it is a positive or negative review, please close the pop-up window and click the arrow below to continue. Please click the link below for a sample of the task. [LINK]

During each of the two participation sessions, you will have to complete eight tasks. Note: the average time of each recording is about 20 seconds.

During the eight required reviews, you cannot get more than two answers wrong. If you get more than two answers wrong, you will be dropped from the study and will not receive payment. However, if you listen to the entire audio recording, the answers should be quite easy.

During the second session, we will ask you about your willingness to do additional reviews for

extra pay. Your job in this first session is to learn about the difficulty of the task and think about your willingness to do additional reviews next session.

[*Coin flip*: Depending on chance, a background noise may be played on top of the audio review. We'll describe what determines whether you hear the noise in a moment. However, we'd like to make sure you know what the sound will be. Please click the play button below for a sample of the noise. When you are finished listening to the sample noise, click the arrow below to continue.]

[*Coin flip*: In a moment, you will begin the eight initial reviews. Before that, however, we must determine if you will have to hear the annoying noise over the audio review.In order to do this, you will flip a (digital) coin. If the coin lands Heads, you will not have to hear the noise. If it lands Tails, you will have to hear the noise.]

[*Coin flip*: Importantly, your flip today determines what you'll do on the second session of the experiment. If the coin flip lands Tails and you hear the annoying noise today, you will also hear it next session. If the coin flip lands Heads and you do not hear the annoying noise today, you will not hear it next session. So the result of this coin flip really matters!]

Click the button below to flip the coin: [BUTTON]

Sorry [Congratulations]. You will [not] have to hear the noise while you listen to the audio reviews. We will now begin the eight initial tasks. At the end of the task, you will see a code. You will need that code to continue. Click the words below to begin. [BEGIN TASK]

Remember - this experiment has two parts. The link to the second session will be emailed to you in 8 hours.

Since you heard [did not hear] the annoying noise today, you will also hear it next session. Please click the arrow to submit your work.

Modified Script for High-Probability Treatment

The *high probability* treatment used the same instructions as above for Session 1, except the paragraphs labeled *Coin flip* were replaced with the following:

[*High Probability:* In a moment, you will begin the eight initial reviews. Before that, however, we must determine if you will have to hear the annoying noise over the audio review. In order to do this, we will draw a random number from 1-100. If the random number is 100, you will not have to hear the noise. If it is any other number, you will have to hear the noise.]

[*High Probability*: Importantly, the random number today determines what you'll do on the second session of the experiment. If the number is 1-99 and you hear the annoying noise today, you will also hear it next session. If the random number is 100 and you do not hear the annoying noise today, you will not hear it next session. So the result of this random draw really matters!]]

Session 2

Welcome to the second session of the experiment.

As with the first session, if you choose not to participate in the study, you are free to exit. You must finish this session in order to receive payment. As a reminder: we will not deceive you whatsoever in this experiment. All of the instructions provide examples and guidance for the actual tasks you will do. There will be no surprises or tricks.

As with last session, you will listen to an audio recording of a review and must determine whether the reviewer is giving a generally positive or negative review. Be careful to listen to

the whole review!

You heard [did not hear] the noise on top of the audio last session, and you will [not] hear it again this session. [*Noise only*: If you need a reminder of the noise, there is a sample below. To play, click the play button twice.]

As before you will have to complete eight reviews. However, this session you will have the option to complete extra reviews for additional payments. These extra tasks will come after the eight initial reviews. You will first decide how many extra reviews you would like to do on top of the eight initial reviews. You will then do the first eight reviews. Finally, you will have a chance to complete extra reviews if you were willing to do so. We will describe how this is determined on the next slides.

The method we use to determine whether you will complete extra reviews may seem complicated. But, we'll walk through it step-by-step. The punchline will be that it's in your best interest to just answer truthfully. First, we will ask you how many additional reviews you are willing to do for a fixed amount of money. For instance, we might ask: "What is the maximum number of extra reviews you are willing to do for $0.40?" This question means that we will give you $0.40 in exchange for you completing some amount of additional work.

On the decision screen, you will be presented a set of sliders that go between 0 and 100 tasks. You will also see an amount of money next to each slider. You will move each slider to indicate the maximal number of reviews you'd be willing to do for each amount of money. That is, if you would be willing to do 15 additional reviews but not 16, then you should move the slider to 15.

You will make five decisions, but only one will count for real. We will choose which decision counts for real using a random number generator. Therefore, it is in your best interest to take each question seriously and choose as if it were the only question.

Once we determine which question counts for real, we will draw a random number between 0 and 100. If your answer is less than that random number, you will not do additional reviews. However, if your answer is greater than or equal to that random number, you will do a number of additional tasks equal to the random number.

Example: Suppose you indicated you were willing to do 15 additional reviews for $0.40 and this question was chosen as the one that counts. If the random number was 16 or higher, you would do no additional tasks. However, if the random number was 12, you would do 12 additional reviews. The next pages have a short quiz to help clarify how this works.

Suppose you were asked "What is the maximum number of additional reviews you are willing to do for $0.80?" and you responded 60. If the random number is 17, how many reviews will you complete?

☐ 0 and I will be paid $0 in supplementary payments
☐ 60 and I will be paid $0.80 in supplementary payments
☐ 17 and I will be paid $0.80 in supplementary payments
☐ 17 and I will be paid $2.67 in supplementary payments

[On answering correctly] Correct. You will earn the extra payment if the random number is less than the number you indicated, and you will complete a number of additional reviews equal to the random number.

Suppose you were asked "What is the maximum number of additional reviews you are willing to do for $0.80?" and you responded 60. If the random number is 76, how many additional reviews will you complete?

☐ 0 and I will be paid $0 in supplementary payments

☐ 76 and I will be paid $0.80 in supplementary payments
☐ 60 and I will be paid $0.80 in supplementary payments
☐ 76 and I will be paid $0 in supplementary payments

[On answering correctly] Correct. If the random number is greater than your choice, you will complete zero reviews and you will not receive an extra payment.This method of selecting how many additional reviews you will do might seem very complicated, but as we previously highlighted, there's a great feature to it: your best strategy is to simple answer honestly. If, for example, you'd be willing to do 20 reviews for $0.40 but not 21, then you should answer 20. You may very well do less than 20 reviews (depending on the random number) but you certainly will not do more than 20. Put simply: just answer honestly.

Remember, you will decide whether to do additional reviews, then complete the eight initial reviews. Then we will draw a random number which determines if you will do extra reviews.

We will now ask you the questions about your willingness to do additional reviews for additional payment. Remember, we are using the method just described, so answer honestly. These are the real questions. One of the sliders will count for payment, so pay close attention.

What is the maximal number of additional reviews you're willing to complete for:
$2.50? [SLIDER]
$2.00? [SLIDER]
$1.50? [SLIDER]
$1.00? [SLIDER]
$0.50? [SLIDER]

We will determine whether you will do additional reviews after you complete the eight initial tasks. We will begin those on the next page.

Like last session, you will [not] have to hear the noise during the audio reviews. We will now begin the eight initial reviews. When you have completed these eight reviews, you will see a code. You will need that code to continue. Click the words below to begin. [BEGIN TASK]

We'll now draw the random number that determines which question counts for payment.

The random number selected the question where you were asked the maximum number of tasks you would do for [AMOUNT]. You answered [RESPONSE]. We'll now draw a second random number that determines whether you do additional tasks and, if so, how many.

The random number is: [RANDOM NUMBER]. You answered: [RESPONSE].

[*Random number too high*: Since the random number was higher than the number you were willing to do, you will not complete any extra reviews and you will not receive any extra payments.] Since the random number was lower than the number you were willing to do, you will complete extra reviews. You will do [RANDOM NUMBER] extra reviews and receive [AMOUNT]. In order to verify that you completed all the additional reviews, we will give you a code when you finish. [BEGIN SUPPLEMENTAL TASKS]

Thank you for participating. Your MTurk code is on the screen that follows. Payments will be processed within one week. Please click the final button below to submit your work.

H.3 Complete Experiment Instructions: Experiment 2

Session 1

In front of you is an informed-consent form to protect your rights as a participant. Please read it. If you choose not to participate in the study, you are free to leave at any point. If you have any questions, we can address those now. We will pick up the forms after the main points of the study are discussed.

We will not deceive you whatsoever in this experiment. All of the instructions provide examples and guidance for the actual tasks you will do. There will be no surprises or tricks. If you have any questions at any time, please raise your hand and we will do our best to clarify things for you.

In this experiment, you will have the chance to earn supplemental payments ranging from $2-$25/hour. It is very important for the study that you participate in both days. Unfortunately, if you miss one of your participation dates, you will forgo any completion payments and supplemental payments and will be removed from the study (you will receive the show-up fee). There will be absolutely no exceptions to this rule, regardless of the reason. Completion and supplemental payments will be made as one single payment in cash at the end of the study.

Your task will be transcribing a line of handwritten text in a foreign language. We will explain the task and then allow you to spend a few moments practicing this job on the computer. Note that the example text may not exactly match what you will face in the experiment.

Letters will appear in a Transcription Box on your screen. For each handwritten letter, you will need to enter the corresponding letter into the Completion Box. In order to enter a letter into the Completion Box, simply click the letter from the provided alphabet. We refer to one row of text is one task. In order to advance to the next task, your accuracy must be above 90%.

We will now give you a sample task to practice. You will see handwritten characters and must enter the corresponding character into the Completion Box by clicking on the appropriate button. When you have transcribed a whole row, press "Submit". You may spend as much time as you like transcribing the text. If you succeed, a new line of text will appear. Once you have transcribed one row successfully, please close the pop-up window and click the arrow below to continue. Please click the link below for a sample of the task. [SAMPLE TASK]

During each of the two participation days, you will have to complete five tasks (five lines of foreign text). Note: the average time to complete a similar task in a different experiment was about 52 seconds (about 70 tasks/hour).

After completing five initial tasks, you will have the option to complete additional supplementary tasks for supplementary payments. The number of supplementary tasks you must complete on each participation day and the supplementary payment will depend on your own willingness to work. The supplementary tasks will come shortly after the five initial tasks.

In order to determine whether you will complete additional tasks, we will ask you how many additional tasks you are willing to do for a fixed amount of money. For instance, we might ask: "What is the maximum number of additional tasks you are willing to do for $5?" This question means that we will give you $5 in exchange for you completing some amount of additional work. The next few screens describe a pretty complicated system that will determine how many additional tasks you actually do. But the point of this system is simple: there is no way to game the system. It is in your best interest to answer honestly.

On the decision screen, you will be presented a set of sliders that go between 0 and 100 tasks.

You will also see an amount of money next to each slider. You will move each slider to indicate the maximal number of tasks you'd be willing to do for each amount of money. That is, if you would be willing to do 15 additional tasks but not 16, then you should move the slider to 15. For example (you need not enter anything) What is the maximal number of additional tasks you're willing to complete for:

$1? [SLIDER]
$2? [SLIDER]
$3? [SLIDER]
$4? [SLIDER]
$5? [SLIDER]

You will make five decisions, but only one will count for real. We will choose which decision counts for real using a random number generator. Therefore, its in your best interest to take each question seriously and choose as if it was the only question.

Once we determine which question counts for real, we will draw a random number between 0 and 100. If your answer is less than that random number, you will do no additional tasks. However, if your answer is greater than or equal to that random number, you will do a number of additional tasks equal to the random number.

Example: Suppose you indicated you were willing to do 15 additional tasks for $5 and this question was chosen as the one that counts. If the random number was 16 or higher, you would do no additional tasks. However, if the random number was 12, you would do 12 additional tasks. The next page has a short quiz to help clarify this system.

Suppose you were asked "What is the maximum number of additional tasks you are willing to do for $10?" and you responded 30. If the random number is 8, how many tasks will you complete?

☐ 0 and I will be paid $0 in supplementary payments
☐ 30 and I will be paid $10 in supplementary payments
☐ 8 and I will be paid $10 in supplementary payments
☐ 8 and I will be paid $2.67 in supplementary payments

Correct. You will be paid the full amount regardless of the random number, and if the random number is less than the number you indicated, you will only need to complete a number of additional tasks equal to the random number.

Suppose you were asked "What is the maximum number of additional tasks you are willing to do for $10?" and you responded 30. If the random number is 46, how many additional tasks will you complete?

☐ 0 and I will be paid $0 in supplementary payments
☐ 46 and I will be paid $10 in supplementary payments
☐ 0 and I will be paid $10 in supplementary payments
☐ 30 and I will be paid $0 in supplementary payments

Correct. If the random number is greater than your choice, you will complete zero tasks and you will not get paid. This method of selecting how many additional tasks you will do might seem very complicated, but as we previously highlighted, there's a great feature to it: your best strategy is to simple answer honestly. If you'd be willing to do 20 tasks for $5 but not 21, then you should answer 20. You may very well do less than 20 tasks (depending on the random number) but you certainly will not do more than 20. Put simply: just answer honestly.

Depending on chance, a background noise may be played throughout the transcription process. We'll describe what determines whether you hear the noise in a moment. However, we'd like

to make sure you know what the sound will be. Please click the play button below twice for a sample of the noise. When you are finished listening to the sample noise, click the arrow below to continue.

In a moment, you will begin the five initial tasks. Before that, however, we must determine if you will have to hear that annoying noise during the whole transcription process. In order to do this, you will flip a coin. If the coin lands Heads, you will not have to hear the noise. If it lands Tails, you will have to hear the noise.

Importantly, your flip today determines what you'll do on the second day of the experiment. If the coin flip lands Tails and you hear the annoying noise today, you will also hear it next week. If the coin flip lands Heads and you do not hear the annoying noise today, you will not hear it next week. So the result of this coin flip really matters!

When you reach this screen, please put your hand up. You may remove your headphones for this stage of the instructions. One of the experimenters will come by and help you. We are using a standard U.S. Quarter. This is not a trick coin and we're going to ask you to flip it. Please flip it and let it land on the table in front of you. If the coin does not flip more than twice, we will ask you to flip again. You'll be asked to flip a practice flip, and then you'll flip the one that counts. Reminder: Heads $\rightarrow$ No Noise. Tails $\rightarrow$ Annoying Noise

The experimenter will the answer this question.

☐ Tails

☐ Heads

Enter Code to Advance

[*Noise*: You will have to hear the noise. Please put your headphones back on. We will now begin the five initial tasks.] You will not have to hear the noise. However, we ask that you please put your headphones on so that you do not hear others. At the end of the task, you will see a code. You will need that code to continue. Click the words below to begin. [BEGIN TASK] Please enter the code below to continue

We will now ask you some questions about your willingness to do additional tasks for additional payment. Remember, we are using the system described earlier, so answer honestly.One of the sliders will count for real payment, so pay close attention.

What is the maximal number of additional tasks you're willing to complete for:

$20? [SLIDER]

$16? [SLIDER]

$12? [SLIDER]

$8? [SLIDER]

$4? [SLIDER]

We'll now draw a random number to determine which question counts for payment.

The random number selected the question where you were asked the maximum number of tasks you would do for [AMOUNT]. You answered [RESPONSE]. We'll now draw a second random number that determines whether you do additional tasks and, if so, how many.

The random number is: [RANDOM NUMBER]. You answered: [RESPONSE].

[*Random number too high*: Since the random number was higher than the number you were willing to do, you will not complete any extra reviews and you will not receive any extra payments.] Since the random number was lower than the number you were willing to do, you will complete extra reviews. You will do [RANDOM NUMBER] extra reviews and receive [AMOUNT]. In order to verify that you completed all the additional reviews, we will give you a code when you finish.

[BEGIN SUPPLEMENTAL TASKS]

Thank you for participating. [*Noise*: REMINDER: Since you heard the annoying noise today, you will also hear it in a week.]

REMINDER: Since you did not hear the annoying noise today, you will not hear it in a week.

Day 1 of the experiment is complete. Please return at the same time one week from now.Please click the arrow to submit your work. When you have finished, you may exit the lab.

Session 2

Welcome to the second day of the experiment.

Please turn your cell phones off. If you have a question at any point in the experiment, please raise your hand and a lab assistant will be with you to help. There will be a short quiz once we have finished the instructions. If you do not understand the instructions after both the instruction period and the quiz, please raise your hand and ask for help.

As with the first day, if you choose not to participate in the study, you are free to leave at any point. If you have any questions, we can address those now.

As a reminder: we will not deceive you whatsoever in this experiment. All of the instructions provide examples and guidance for the actual tasks you will do. There will be no surprises or tricks.

Like last week, your task is to transcribe a line of handwritten letters from a foreign language. This week, you will do a different language. You will the task under the same conditions as last week.

[*Noise*: You heard the noise last week, and you will hear it again this week. If you need a reminder of the noise, there is a sample below. To play, click the play button twice.]

You did not hear the noise last week, and you will not hear it again this week.

As with last week, letters will appear in a Transcription Box on your screen. For each handwritten letter, you will need to enter the corresponding letter into the Completion Box. In order to enter a letter into the Completion Box, simply click the letter from the provided alphabet. We refer to one row of text is one task. In order to advance to the next task, your accuracy must be above 90%.

As before you will have to complete five tasks (five lines of foreign text) and then you will have the option to complete additional supplementary tasks for supplementary payments. The supplementary tasks will come shortly after the five initial tasks.

In order to determine whether you will complete additional tasks, we will ask you how many additional tasks you are willing to do for a fixed amount of money. For instance, we might ask: "What is the maximum number of additional tasks you are willing to do for $5?" This question means that we will give you $5 in exchange for you completing some amount of additional work. It is in your best interest to answer these questions honestly.

Recall we used a random number system to determine how many additional tasks you did (if any). We'll provide a quick reminder of that system now.

On the decision screen, you will be presented a set of sliders that go between 0 and 100 tasks. You will also see an amount of money next to each slider. You will move each slider to indicate the maximal number of tasks you'd be willing to do for each amount of money. That is, if you would be willing to do 15 additional tasks but not 16, then you should move the slider to 15.

You will make five decisions, but only one will count for real. We will choose which decision counts for real using a random number generator. Therefore, its in your best interest to take each

question seriously and choose as if it was the only question.

Once we determine which question counts for real, we will draw a random number between 0 and 100. If your answer is less than that random number, you will do no additional tasks. However, if your answer is greater than or equal to that random number, you will do a number of additional tasks equal to the random number.

Example: Suppose you indicated you were willing to do 15 additional tasks for $5 and this question was chosen as the one that counts. If the random number was 16 or higher, you would do no additional tasks. However, if the random number was 12, you would do 12 additional tasks. The next page has a short quiz to help clarify this system.

Suppose you were asked "What is the maximum number of additional tasks you are willing to do for $10?" and you responded 60. If the random number is 17, how many tasks will you complete?

☐ 0 and I will be paid $0 in supplementary payments

☐ 60 and I will be paid $10 in supplementary payments

☐ 17 and I will be paid $10 in supplementary payments

☐ 17 and I will be paid $2.67 in supplementary payments

Correct! You will be paid the full amount regardless of the random number, and if the random number is less than the number you indicated, you will complete a number of additional tasks equal to the random number.

Suppose you were asked "What is the maximum number of additional tasks you are willing to do for $10?" and you responded 60. If the random number is 76, how many additional tasks will you complete?

☐ 0 and I will be paid $0 in supplementary payments

☐ 76 and I will be paid $10 in supplementary payments

☐ 60 and I will be paid $10 in supplementary payments

☐ 76 and I will be paid $0 in supplementary payments

Correct. If the random number is greater than your choice, you will complete zero tasks and you will not get paid.This method of selecting how many additional tasks you will do might seem very complicated, but as we previously highlighted, there's a great feature to it: your best strategy is to simple answer honestly. If you'd be willing to do 20 tasks for $5 but not 21, then you should answer 20. You may very well do less than 20 tasks (depending on the random number) but you certainly will not do more than 20. Put simply: just answer honestly.

[*Noise*: Like last week, you will have to hear the noise. Please put your headphones back on.] Like last week, you will not have to hear the noise. However, we ask that you please put your headphones on so that you do not hear others. We will now begin the five initial tasks. At the end of the task, you will see a code. You will need that code to continue. Click the words below to begin. [BEGIN TASK] Please enter the code below to continue:

We will now ask you some questions about your willingness to do additional tasks for additional payment. Remember, we are using the system described earlier, so answer honestly.One of the sliders will count for real payment, so pay close attention.

What is the maximal number of additional tasks you're willing to complete for:

$20? [SLIDER]

$16? [SLIDER]

$12? [SLIDER]

$8? [SLIDER]

$4? [SLIDER]

We'll now draw a random number to determine which question counts for payment.

The random number selected the question where you were asked the maximum number of tasks you would do for [AMOUNT]. You answered [RESPONSE]. We'll now draw a second random number that determines whether you do additional tasks and, if so, how many.

The random number is: [RANDOM NUMBER]. You answered: [RESPONSE].

[*Random number too high*: Since the random number was higher than the number you were willing to do, you will not complete any extra reviews and you will not receive any extra payments.] Since the random number was lower than the number you were willing to do, you will complete extra reviews. You will do [RANDOM NUMBER] extra reviews and receive [AMOUNT]. In order to verify that you completed all the additional reviews, we will give you a code when you finish. [BEGIN SUPPLEMENTAL TASKS]

Thank you for participating. As you know, the experiment consisted of two days. Our main hypothesis was whether the chance of getting a different task on the first day changed your perceptions of the task difficulty that day. We did not highlight this specific hypothesis during the experiment in order to maintain the external validity of the study. We're excited to analyze the data and thank you again for your participation. Click the arrow to submit your work.