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

Benjamin Bushong
Michigan State University

Tristan Gagnon-Bartsch*
Harvard University

Revision in Progress

December 9, 2019

Abstract

In this paper, we experimentally investigate whether participants exhibit a previously undocumented form of attribution bias stemming from reference-dependent preferences. In our baseline experiment, participants learned from experience about one of two unfamiliar tasks, one more onerous than the other. Some participants 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 those participants' willingness to work again at that same task. Participants assigned the less-onerous task by chance were more willing to work than those who faced it with certainty (or high probability). Conversely, participants 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 remained hours after initial impressions were formed, are consistent with the idea that participants mistakenly attributed sensations of positive or negative surprise (relative to expectations) to the effort cost of their assigned task.

JEL Classification: C91, D03.
Keywords: attribution bias, reference dependence, learning from experience, loss aversion, experiment, real effort.

*E-mails: bbushong@msu.edu and gagnonbartsch@fas.harvard.edu. We thank Ned Augenblick, Katherine Coffman, Ben Enke, Christine Exley, David Laibson, Muriel Niederle, Devin Pope, Matthew Rabin, Joshua Schwartzstein, Andrei Shleifer, Charles Sprenger, and seminar audiences at 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 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 both the lab and field emphasizes that our experiences are reference dependent. In particular, 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 2008; Abeler et al. 2011). As expectations and reference points may change over time, learning about our intrinsic preferences is complicated because it demands that we disentangle these two sources of utility. Guided by research in psychology, we designed experiments to explore whether people incorrectly attribute sensations of surprise—either elation or disappointment—to their intrinsic (dis)utility of a real-effort task.[1]

To motivate our experimental design, consider a worker completing a series of short-term tasks. Suppose that each day the worker is randomly assigned to one of two tasks—one more desirable than the other—meaning the job she faces each day comes as either a positive surprise or a disappointment. If she fails to account for these sensations as she forms her impressions of the jobs, she will develop incorrect beliefs about how much she enjoys each. Concretely, when she is fortunate and assigned to the desirable job she may incorrectly attribute the positive feelings arising from surprise to the intrinsic enjoyment of the task, and hence become too willing to perform that duty in the future. By contrast, when she is unfortunately assigned to the less desirable task she may misattribute the sensation of disappointment to the disutility of that task, and become too hesitant to work in that role. In both cases, the worker may form biased impressions of the task because she neglects the degree to which her (past) experienced utility depended on her expectations.

In this paper, we present two experiments that explore whether people wrongly attribute sensations of surprise or disappointment to the intrinsic value of a real-effort task. In Experiment 1, participants completed one of two previously-unexperienced tasks. Hours later, we elicited their willingness to complete additional work on their assigned task. Their responses are consistent with the idea that participants incorrectly learned the disutility associated with their task as a function of the exogenously imposed expectations they held before first encountering that task. We discuss how our results are inconsistent with rational learning models that assume either classical preferences or reference-dependent preferences without misattribution. In our second experiment, we manipulate initial expectations within subjects and examine how willingness to work changes over the course of a week as participants' expectations change. We again find that surprise and disap-

[1] Studies in both psychology and economics demonstrate that memories are imprecise and people may make mistakes when attributing the sources of their feelings. 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. 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 discuss additional evidence for such mistakes in attribution in Section 2.

pointment shape participants' willingness to work. In addition to contributing to the growing body of evidence on expectations-based reference points, our contribution is therefore the identification of a specific, previously unstudied form of attribution bias.

We first present an abridged version of the model from our companion paper which informs our experimental design and helps us derive behavioral predictions. Following Bell (1985) and Kőszegi and Rabin (2006), we assume the decision maker experiences expectations-based reference-dependent utility composed of two parts: *consumption* utility, which corresponds to the classical notion of payoffs, and *gain-loss* utility, which is proportional to the difference between the consumption utility earned and what the person expected. As alluded to above, a "misattributor" correctly recalls how she felt after each experience, but wrongly attributes sensations of surprise or disappointment—the gain-loss component of her utility—to the underlying outcomes—the consumption component. She thus forms biased impressions of the outcomes she faced. We describe the model's predictions in greater detail below.

Guided by our model, we designed a pair of experiments. Experiment 1 involved 886 subjects recruited from Amazon's Mechanical Turk (MTurk). In an initial learning session, each subject listened to Amazon book reviews read by a computer, 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.

Our primary experimental manipulation stemmed from varying subjects' beliefs over which task they would complete in the experiment. One third of participants were assigned to a task from the onset of the experimental instructions. Another third of participants flipped a coin to determine which task they would face moments before their first experience with that task. Finally, the remaining participants faced near-certain task assignment. 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, if applicable, resolving any uncertainty), participants completed eight trials of their assigned task. Knowing that they would later be asked about their willingness to work on this task, these initial trials gave participants an opportunity to learn about their preferences. More than eight hours later, we elicited their willingness to complete more trials of their assigned task for additional pay.

We first examine how willingness to work differed between the *control* and *coin-flip* treatments. Our misattribution model (sketched above) predicts that participants who face the task without noise as a result of the coin flip will form overly positive beliefs about that task, since their initial impressions will be influenced by a sense of positive surprise. Thus, we predict that participants

in the *coin flip + no noise* group will be more willing to work than those in the *control + no noise* group, despite the two groups ultimately completing the same task. By contrast, we predict those assigned the noisy task via the coin flip will form overly negative impressions of the task, as their initial experience was colored by their disappointment. Therefore, participants in the *coin flip + noise* group who suffer misattribution will be less willing to work than those in the *control + noise* group. Matching the predictions above, we find that participants who faced the task without noise by chance were more willing to work than those who faced that task with certainty. This effect is clearly seen in Figure 1 where the labor supply curve for the *coin flip + no noise* group is shifted outward relative to the *control + no noise* group.[2] In contrast, those who faced to the noisy task by chance were less willing to work than those who faced the noisy task for certain—manifesting in an inward shift of the labor supply curve for the *coin flip + noise* group relative to the *control + noise* group. We interpret this result as evidence that participants formed different beliefs about the task as a result of misattributing sensations of positive and negative surprise that arose during the initial learning session.[3] Additionally, we demonstrate that neither a classical model nor a reference-dependent model without misattribution predicts this differential willingness to work between participants assigned by coin flip versus certain assignment. These results are also robust to a modified experiment where all participants are exposed to both tasks.[4]

In Experiment 2, we elicited each participant's willingness to work during two different sessions, separated by one week. Our identification of misattribution in this setting stems from changes in a participant's reference point over the course of the week. In a first session, each participant flipped a coin to determine whether she faced the good or bad task and then immediately completed five trials of that task. Directly after this learning phase, we elicited the participant's willingness to continue working at that task. One week later, the same participants returned and repeated the

[2] For ease of visual presentation, we omit error bars but we discuss the significance of these differences in detail in the text.

[3] The time gap between participants forming their impressions and our elicitation of willingness to work helps distinguish our effect from that of short-term "transient moods", as any influence of the coin flip on mood should fade over more than eight hours. This was a central design concern because transient factors have, for example, been demonstrated to influence investor sentiment. Fluctuations in the weather (Saunders 1993, Hirshleifer and Shumway 2003) and sports outcomes (Edmans, Garcia and Norli 2007) both lead to systematic changes in stock returns. In psychology, the more general idea that positive or negative affect can distort unrelated behavior is well documented. For example, Isen and Levin (1972) showed that participants were more likely to help others after they themselves experience positive, unrelated events.

[4] In this alternative treatment, we followed the same instructions as the coin-flip version of Experiment 1, but replaced the coin flip with a *high-probability* environment: participants were nearly certain to face one of the tasks, but were told of both. We then compared the willingness to work of the coin-flip groups with that of participants in each *high-probability + noise* and *high-probability + no noise*. Our initial findings replicate for participants facing no noise ($p = .033$), and our results are directionally consistent but not significant ($p = .103$) for participants who face the task with noise. This high-probability treatment addresses concerns that might arise due to the fact that participants who were assigned their task with certainty only knew about that one task. For instance, the knowledge of an alternate task could lead to widened priors through some form of (plausibly rational) inference. Our robustness treatment rules out this and other alternative explanations that we discuss in greater detail in Section 3.1.

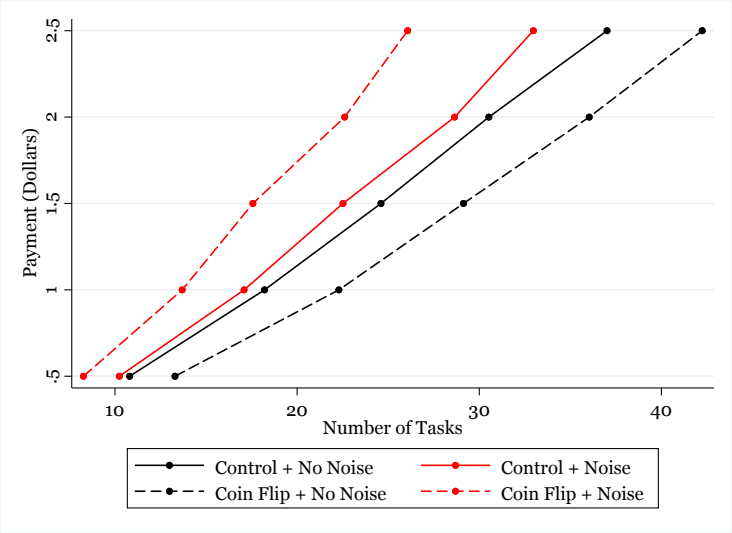

Figure 1: Labor supply curves across four treatments—each point represents the average willingness to work for a fixed payment as elicited under the BDM mechanism. Relative to groups whose assignment did not induce surprise, those assigned by chance (*coin flip* groups) demonstrate greater willingness to work when assigned to the task without noise and less willingness to work when assigned to the task with noise.

process above, except there was no coin flip: each knew with certainty that she would face the same task in the second session as she did in the first session. That is, each participant again completed five trials of her assigned task and then revealed her willingness to continue working.

Given we elicited willingness to work twice for each subject, our variable of interest is the difference in a participant's willingness to work between week one—when her task came as a surprise—and week two—when that same task was completely expected. We find that participants who where were pleasantly surprised in the first session were *less* willing to work in the second week than in the first, while those who were negatively surprised in the first session were *more* willing to work in the second week than the first. This result is consistent with our model: a participant who was, say, positively surprised (i.e., faced the good task) may have attributed this sensation to a quality of the underlying task and thus worked too much in the first week. Upon trying that same task again a week later—when it was no longer pleasantly surprising—the task may have failed to live up to the previous experience, and thus willingness to work decreased.

Attribution bias applied to the gain-loss element of reference dependence can generate a number of known biases in belief updating.[5] For instance, a misattributor relies too heavily on her personal experience—in particular, recent experiences—when making decisions. Additionally, when comparing her outcomes against past experiences, a misattributor may exhibit sequential contrast

[5] We explore the implications of misattribution of reference dependence in greater detail in Gagnon-Bartsch and Bushong (2019).

effects, whereby she perceives today's outcomes as better the worse was yesterday's. Finally, fixing the outcomes she faces, a misattributor forms the most optimistic beliefs after experiencing a sequence of increasing outcomes.[6]

Misattribution of reference dependence also provides a novel explanation for the findings of Gneezy and List (2006) and similar experiments. In their field experiment, surprisingly high earnings lead workers to increase efforts, but those increased efforts diminish over time. The evidence we provide herein speaks to both why these payments increase effort in the short-term and why these surprising wages fail to motivate longer-term changes in behavior as workers learn the underlying difficulty of their assigned task.

2 Theoretical Framework and Motivation

In this section, we present a streamlined version of our model of misattribution of reference dependence (Bushong and Gagnon-Bartsch 2019), which guides our experimental design. We also discuss motivating evidence for the central assumptions underlying this model.

Preferences and Misattribution. Following Kőszegi and Rabin (2006; henceforth KR), we assume that the agent's overall utility from an outcome 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 or a "reference point". Following Bell (1985), we take this reference point to be the agent's expectation of v, 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 any potential loss aversion. Accordingly, given expectations $\widehat{\mathbb{E}}[V]$,

[6] Evidence suggests that people both prefer improving sequences and form the most optimistic evaluations thereafter. For example, Ross and Simonson (1991) allow participants to sample two video games and find that willingness to pay for the bundle is significantly higher among those who sampled the better game second. Similarly, Haisley and Loewenstein (2011) show that advertising promotions are most effective when sequenced in increasing order of value—that is, the high-value promotional item is given last. Several authors argue that such assessments follow a mechanism like ours (e.g. Tversky and Griffin 1990; Loewenstein and Prelec 1993; Baumgartner, Sujan and Padgett 1997). Other forms of sequential contrast effects have been documented in decisions made by teachers (Bhargava 2007), speed daters (Bhargava and Fisman 2014) judges assessing asylum seekers, reviewers of loan applications, baseball umpires (Chen, Moskowitz, and Shue 2016) and in stock returns (Hartzmark and Shue 2016).

[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.

the person's total utility is

$$u\left(v\,\middle|\,\widehat{\mathbb{E}}[V]\right) = v + \eta n\left(v\,\middle|\,\widehat{\mathbb{E}}[V]\right), \tag{2}$$

where $\eta > 0$ is the weight given to sensations of gain and loss relative to absolute outcomes. Although we assume the reference point is expectations, we highlight in Section 3.2 that our experimental predictions are robust to many different specifications.[8]

Our notion of misattribution arises in environments where an individual attempts to learn about the consumption utility v she derives from an unfamiliar alternative. We assume a misattributor only observes her total utility, and when doing so she under-appreciates the extent to which her past experiences were influenced by reference dependence.[9] Thus, she infers v using a misspecified model that weights the gain-loss component of her utility by a diminished factor $\hat{\eta} \in [0, \eta)$. That is, a misattributor infers as if her utility function were $\hat{u}\left(\hat{v}\,\middle|\,\widehat{\mathbb{E}}[V]\right) = \hat{v} + \hat{\eta} n\left(\hat{v}\,\middle|\,\widehat{\mathbb{E}}[V]\right)$: she correctly recalls how happy she felt following outcome v, but she fails to fully account for how sensations of surprise or disappointment affected her total utility. Specifically, she infers $\hat{v}$ such that $\hat{u}\left(\hat{v}\,\middle|\,\widehat{\mathbb{E}}[V]\right) = u\left(v\,\middle|\,\widehat{\mathbb{E}}[V]\right)$. Equations 1 and 2 imply that this misencoded outcome, $\hat{v}$, takes the following simple 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 the outcome falls short of expectations. This bias is proportional to the deviation between the true outcome and expectations. Furthermore, the bias increases in the degree of misattribution—i.e., as $\hat{\eta}$ decreases. When the agent is loss averse—i.e., $\lambda > 1$—losses are misencoded by a greater extent than gains. Finally, the agent uses this wrongly encoded outcome $\hat{v}$ to update her beliefs about the underlying consumption utility.

To illustrate how misattribution leads to biased beliefs, recall the example from the introduction

[8] While we assume that the reference point corresponds to *mean* expectations, the predictions we examine do not substantially depend on whether we assume a deterministic reference point (à la Bell and Equation 1, above) or stochastic reference point (à la KR, where an outcome is compared to each possible alternative outcome under $\widehat{F}_V$ and every such comparison is weighted by the likelihood of the alternative). This is because misattribution in our setting does not influence the planning stage of actions. This planning stage—where the person forms plans and, accordingly, expectations—is crucial for many of KR's predictions. Given any particular model of the reference point, misattribution is applied *after* the resolution of uncertainty.

[9] There are at least two plausible interpretations of how these biased perceptions are formed. (1) The agent improperly encodes each outcome as they happen—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—which seems most plausible in settings where outcomes are easily observed (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).

wherein a worker's daily task is assigned at random: some days she faces a relatively task and other days she faces a more-onerous one. When the worker is assigned the onerous task, she simultaneously experiences both a bad material outcome and a sensation of disappointment—her task is worse than average. If she fails to properly disentangle this sensation of disappointment and wrongly attributes it to the underlying disutility of the task, she will recall 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 may recall the task as even better than it really was.

For ease of presentation, we have thus far discussed the case where consumption, v, is unidimensional. However, our experiments examine a setting with two dimensions—money and effort. To accommodate this, we extend the model outlined above: 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

$$u\left(v \mid \widehat{\mathbb{E}}[V]\right) = \sum_{k\in\{m,e\}} \left\{ \underbrace{v^k}_{\text{Consumption utility}} + \underbrace{\eta\, n\left(v^k \mid \widehat{\mathbb{E}}[V^k]\right)}_{\text{Gain-loss utility}} \right\}. \tag{4}$$

The misencoded outcomes due misattribution, $\hat{v}^k$, are then defined as above (Equation 3) along each dimension. That is, a misattributor recalls an outcome $\hat{v}^k$ such that $\hat{v}^k + \hat{\eta} n\left(\hat{v}^k \mid \widehat{\mathbb{E}}[V^k]\right) = v^k + \eta n\left(v^k \mid \widehat{\mathbb{E}}[V^k]\right)$.

Motivating Evidence. Research in both economics and psychology showcases empirical evidence consistent with the basic idea of reference dependent utility. Early studies by Kahneman and Tversky (e.g., 1979) demonstrated that changes in wealth relative to some reference point lead to sensations of positive surprise and disappointment, which shape behavior. More recently, studies have demonstrated that reference dependence affects behavior across a wide range of contexts. This evidence spans labor supply among taxi drivers (Camerer et al. 1997; Crawford and Meng 2011), domestic violence resulting from unexpected football losses (Card and Dahl 2011), decisions in game shows and sports (Post, van den Assem, Baltussen and Thaler 2008; Pope and Schweitzer 2011; Allen et al. 2015; Markle et al. 2015), and even the behavior of capuchin monkeys (Chen et al. 2006).

While the general idea of a reference point that shapes behavior is well-established, the evidence that the reference point corresponds to forward-looking *expectations* is much less clear. In favor of expectations, Abeler et al. (2011) study real-effort provision in the presence of stochastic wages and demonstrate that varying expectations over these wages changes effort. Gill and Prowse (2012) look at a two-person sequential game in which players exert real effort and the probability of winning a prize depends on the total effort exerted. Importantly, the probability of winning in their experiment is linear in effort, meaning that the second player's behavior should not depend

on that of the first. However, the authors find a discouraging effect of low first-player effort that is consistent with a model of expectations-based reference dependence. Sprenger (2015) provides evidence that choices are driven by stochastic reference points—that is, the reference point depends on the full distribution of a lottery. Exploring a prediction of Kőszegi and Rabin (2007), Sprenger demonstrates that participants choose risky options more often when expecting a risky lottery rather than a sure payoff. Finally, Karle et al. (2015) show that food choices depend on the realization of uncertain prices in a way that is consistent with expectations-based reference dependence. Although these studies provide evidence in favor of expectations as the reference point, a growing literature contradicts the evidence above. In a similar experiment to that of Karle et al., Wenner (2015) finds no evidence for the KR model, but attributes his results to non-equilibrium behavior. And while Ericson and Fuster (2009) demonstrate that the endowment effect is at least partially driven by expectations of future endowments, Heffetz and List (2014), Heffetz (2018), and others provide contradictory evidence.[10] Our particular experimental setting does not require any specific model of expectations—many reference points and solution concepts will generate the patterns of behavior we observe. Therefore while our paper is closely related to this literature, we cannot address the mixed evidence contained therein.

In contrast to reference-dependent preferences, misattribution has received little attention in the economics literature, though various errors in attribution have been explored in psychology. Often referred to as the "fundamental attribution error" or "correspondence bias" in that literature (e.g., Ross 1977; Gilbert and Malone 1995), these errors may share a common psychology with that of misattribution of reference-dependent utility: transient sensations (e.g., sensations of surprise or disappointment) are incorrectly attributed to an underlying, stable source.[11] In the economics literature, Simonsohn (2007, 2009) explores the effect of a transient shock (weather) on the subsequent preferences of would-be college students and admissions officers. Simonsohn (2007) demonstrates that college applicants with particularly strong academic qualities were evaluated higher by admis-

[10] Both Ericson and Fuster (2009) and Abeler et al. (2011) were included in replication studies by Camerer et al. (2016) and both studies were replicated with smaller effect sizes narrowly outside of the $p = .05$ standard.

[11] Although Kahneman and Tversky's (1979) "Prospect Theory" supposes that people behave *as if* they experience reference-dependent sensations or hedonics, those authors do not take a strong stand on whether this behavior truly reflects hedonic sensations. Many studies provide suggestive evidence that sensations of positive and negative surprise are a hedonic phenomenon. More directly, Rutledge et al. (2014) shows that a reference-dependent model predicts self-reported happiness during a simple gambling experiment. Additionally, the authors use fMRI to find a neural signal in the midbrain that follows this reference-dependent model. These signals are commonly interpreted as stemming from a non-hedonic reinforcement-learning model that is encoded by midbrain dopamine neurons (Schultz, Dayan and Montague 1999). These reinforcement-learning models predict a signal very similar to that of the gain-loss function (absent loss aversion). Accordingly, previous neuroscience evidence on reinforcement learning, when reinterpreted through this lens, may provide some evidence on reference-dependence. Finally, recent papers show these reference-dependent signals extend beyond the midbrain to higher levels of cortex in both humans (e.g., Hayden et al. 2011; Hill, Boorman and Fried 2016) and other primates (e.g., Bayer and Glimcher 2005). Such signals in the ventral medial prefrontal cortex (an area associated with experienced utility) may suggest a neural basis for reference-dependent hedonics.

sions officers when the weather on that evaluation day was poor. Simonsohn (2009) shows that incoming freshman are more likely to matriculate at an academically rigorous school when the weather on their visit day to that school was cloudy versus sunny. Relatedly, a series of papers show that CEOs (Bertrand and Mullainathan 2003) and politicians (Wolfers 2007; Cole, Healy, and Werker 2012) are rewarded for luck.

Closely related in motivation, Haggag et al. (2019) provide evidence of a different form of attribution error: wrongly attributing state-dependent fluctuations in utility to underlying quality. In an experiment, the authors show that participants value an unfamiliar beverage more if they first drink it while thirsty rather than sated. Likewise, using field evidence they show that good weather during a person's visit to a theme park increases the likelihood that person plans to return. Our model and evidence departs from theirs in a number of ways. In terms of our predictions, errors in attribution in their model leads decision-makers to *underestimate* the utility difference between two outcomes. Our model predicts—and we observe—the opposite. Moreover, biased forecasts that result from Haggag et al.'s formulation may, in some settings, wash out with ample experience. These errors can persist under misattribution of reference dependence.[12] This distinction stems from the fact that Haggag et al. rule out complementaries where past experiences influence today's consumption utility. Reference dependence clearly introduces this complementarity, as past experiences form the reference point against which today's consumption is evaluated.

3 Experiment 1

In this section, we present our between-subject experiment, which we conducted on Amazon's Mechanical Turk (MTurk). We first describe the experimental design. Next, we provide theoretical predictions of both rational-learning models and our model of misattribution. Finally, we analyze our experimental data, noting throughout how the results are consistent with our notion of misattribution yet inconsistent with various rational-learning models with or without reference dependence.

[12] For example, misattribution of state-dependent utility can cause an agent to mislearn the mean outcome in the short run, but the bias will vanish in the long-run if states are independent from the timing of consumption. To illustrate, consider a diner learning about the quality of a restaurant. If she visits the restaurant when she's both hungry and not hungry, she will correctly learn the *average* quality. Accordingly, the framework from Haggag et al. (2019) may best apply to situations where choices are based on limited experience. In contrast, misattribution of reference-dependent utility in this example will lead a loss-averse agent to develop persistent misperceptions about the average quality.

3.1 Design

We recruited approximately 900 participants on MTurk to complete a two-session experiment.[13] The first session was an "initial-learning phase" designed to give participants experience with a new real-effort task. During the second session, we elicited participants' willingness to work on that task for additional pay. It took participants an average of 10 minutes to complete the first session and 15 minutes to complete the second. We paid participants a fixed fee of $4 for successfully completing both sessions, which translates to an hourly wage of approximately $9.60. Participants could earn up to $2.50 additional money for completing additional work depending on their willingness to work and chance.

Each participant worked on one of two tasks. Both involved listening to reviews of books read aloud. Specifically, we used digital-voice software to "read" reviews collected from Amazon.com. Participants had to guess whether each review was positive or negative.[14] In order to classify a review as positive or negative, participants pressed one of two buttons after listening to it, and they were given a warning if their classification was incorrect. Figure 2 depicts this interface. Our two versions of the task differed in a single way: some participants listened to unaltered audio, while others listened to audio with an annoying noise played in the background. This noise was a composite of a fork scraping against a record and a high-frequency tone.[15] The noise played approximately 15 decibels lower than the peak levels of the audio in the review. Hence, the noise was annoying but did not hinder participants' ability to classify the audio reviews.[16]

We now outline each of the two sessions of the experiment.

Session 1: Initial-Learning Phase. Participants completed eight reviews in the first session of the experiment, which we call the "initial-learning session." We instructed participants that the goal of this session was to learn about how much they enjoy the task, since they would later have an opportunity to complete additional rounds of that task for extra pay.

In order to examine how initial expectations altered subsequent evaluations, we randomly assigned participants into three groups: known assignment ($n = 292$), coin-flip assignment ($n = 294$), and high-probability assignment ($n = 300$). Participants in the known-assignment group were told from the start which task they would face, while participants in the coin-flip and high-probability

[13] Participants were required to be U.S. residents and to have completed at least 100 prior jobs on MTurk with a 95% approval rating.

[14] Reviews were edited to last approximately 20 seconds, to remove any specific references to author names or book titles, and for grammar. Unbeknownst to participants, all reviews were either 1-star reviews or 5-star reviews to make the task straightforward if tedious. See the Appendix for sample text of the reviews.

[15] The Nock Lab at Harvard generated this noise and 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 studies found the sound unpleasant, but there were no lasting effects (e.g., ringing ears).

[16] We ran a small pilot ($n = 12$) with reduced stakes (show-up fee of $1.50) to check the programming and to verify that participants in the noise and no-noise groups could both successfully complete the task. All participants in that pilot successfully completed the task, regardless of whether they faced the noise or not.

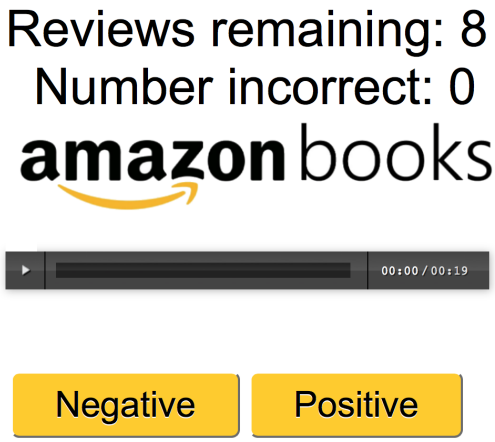

Figure 2: *Screenshot of the classification task from Experiment 1.* Buttons appeared after 10s. Participants clicked the appropriate button to classify whether a review was positive (i.e., endorsing the book) or negative.

groups were initially uncertain. We call these groups *control*, *coin flip*, and *high probability*, respectively.

Participants in the *control* treatment were randomly assigned—unbeknown to them—to one of two subgroups prior to entering the experiment: *noise* or *no noise*. Participants in the *control* + *noise* group were told that they would hear audio reviews with an annoying noise played overlaid. Participants in the *control* + *no noise* group completed classifications without the overlaid noise and were never told about the 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.[17]

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 audio (8s in duration; repeatable if desired). After the description and samples, each participant "flipped" a digital coin to determine whether she would ultimately face the task with noise or without. Each participant then completed the eight mandatory classifications prescribed by the result of their coin flip.

Lastly, participants in the *high-probability* treatment were told that they were very likely to face a given task (either *noise* or *no noise*). 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. For each participant, we uniformly drew a random integer z from $[1, 100]$. Participants in the $p = .99$ arm were assigned the task without noise if $z = 100$; otherwise,

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

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.

Session 2: Eliciting Willingness to Continue Working. In each group, the first session concluded after a participant completed their eight mandatory trials of their assigned task. We emailed each participant a link to the second session exactly eight hours after they finished the first.[18] In the second session, participants were reminded of their prior task assignment (noise or no noise) and given the option to complete additional trials (of that same task) for a bonus payment. Conditional on a participant's assigned task, the second session was identical across all treatment groups. Hence, the key difference across treatments is simply the different ex-ante likelihoods of being assigned the noisy task.

We elicited participants' willingness to continue working in exchange for five different payment values: {\$0.50, \$1.00, \$1.50, \$2.00, \$2.50}. 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 any integer $e^* \in [0, 100]$, which we call "willingness to work". We then uniformly drew a random integer $e \in [0, 100]$. If $e \leq e^*$, then the participant completed e additional tasks and received \$$m$. If $e > e^*$, then the participant completed no additional tasks and earned no bonus pay.

Our overlaid-audio design has an important feature: participants who faced the annoying noise could not avoid the noise and still successfully complete the task. We ensured participants actually listened to the audio reviews using three techniques. First, 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. Additionally, we hid the response buttons for the first ten seconds of each review, which required participants to listen to a substantial portion of the review before guessing. Finally, many of the reviews featured important details in the late part of the review.[19] To prohibit participants from reloading the web session (and thus generate new random numbers) in attempt to avoid the noise, we blocked multiple logins and required unique email authentication to access each session of the experiment.

[18] Fourteen subjects emailed the authors stating that they had not received an invitation to the second session after more than eight hours. All were sent an additional invitation and are therefore included in our main analyses. However, we suspect that others may have faced the same issue (due to emails getting caught in spam filters or participants providing an old or incorrectly-entered email address), leading to slightly higher attrition than desirable. Nevertheless, more than 90% of participants returned for the second session.

[19] Given this feature of the reviews, we may have helped participants answer correctly by withholding the response buttons. In our data, patient responders tend to be more accurate. However, there were very few mistakes overall: only two participants were dropped for inaccurate classifications.

3.2 Theoretical Predictions

In this section, we derive theoretical predictions for how a participant's willingness to work may depend on her initial expectations about the task she will face. Misattribution predicts that the participant's willingness to work will depend on her initial expectations about her task assignment. Rational learning—with or without reference dependence—does not generate such dependence. To clearly illustrate our predictions, we make several simplifying assumptions below. Many of these are unnecessary, and we conclude this section with a discussion of the key assumptions and robustness.

Setup. 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. Mirroring 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 chance 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 $m > 0$ (incentivized via the BDM mechanism).

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

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

where $c(\cdot)$ is an increasing function with $c(0) = 0$, $\theta_i(a)$ is a cost parameter that depends on the task $a \in \{h, l\}$, and $\varepsilon_{i,t}(a)$ are i.i.d. mean-zero random cost shocks that are independent of $\theta_i(a)$. We assume participant i knows that $v^e_{i,t}$ has the structure presented in Equation 5 and knows $c(\cdot)$. However, she is initially uncertain about the cost parameter, $\theta_i(a)$. Let $\pi_{i,0}(a)$ denote her prior over $\theta_i(a)$. We assume that the participant correctly anticipates that $\theta_i(h) > \theta_i(l) > 0$—i.e., the noisy task is more onerous than the noiseless one—and that her prior is independent of her treatment group—i.e., $\pi_{i,0}(a)$ is independent of p_i.

Belief Updating. Since the participant must decide how much to work on task a in period 2, she seeks to learn about her cost parameter $\theta_i(a)$ based on her experience working in period 1. We assume the participant cannot separately observe $\theta_i(a)$ and $\varepsilon_{i,1}(a)$, so she uses her experienced utility in period 1 as a signal to update her beliefs about $\theta_i(a)$. Importantly, this experienced utility may depend on the participant's initial expectations due to reference dependence.

To describe how reference dependence may influence the participant's experienced utility, we must fully specify her reference point. Because she is assigned task $a = h$ with probability p_i,

the participant's expected consumption value on the effort dimension entering period 1 (and thus her reference point) is $\widehat{\mathbb{E}}_{i,0}[V^e_{i,1}] = -[p_i\hat{\theta}_{i,0}(h) + (1-p_i)\hat{\theta}_{i,0}(l)]c(8)$, where $\hat{\theta}_{i,0}(a)$ denotes the expected value of $\theta_i(a)$ under her prior. This follows from Equation 5 along with the fact that each participant must complete exactly eight rounds of the task in period 1, so $e_{i,1} = 8$. Participant i's total experienced utility in period 1 (Equation 4) is thus

$$u_{i,1} = v^e_{i,1} + \eta n\left(v^e_{i,1} \,\middle|\, \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]\right).^{20} \tag{6}$$

Upon realizing $u_{i,1}$, let $\hat{v}^e_{i,1}$ denote the participant's perceived value of $v^e_{i,1}$. As described in Section 2, a misattributor encodes an "exaggerated" value $\hat{v}^e_{i,1}$ according to Equation 3. In particular, if $v^e_{i,1} > \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$, then $\hat{v}^e_{i,1} > v^e_{i,1}$—she overestimates the signal—and if $v^e_{i,1} < \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$, then $\hat{v}^e_{i,1} < v^e_{i,1}$—she underestimates the signal. In contrast, a rational agent who fully appreciates the extent to which her utility depends on expectations (i.e., $\hat{\eta} = \eta$) encodes the correct value, $\hat{v}^e_{i,1} = v^e_{i,1}$.

Given the participant's perception of her period-1 (dis)utility of effort, $\hat{v}^e_{i,1}$, we denote her updated expectation of $\theta_i(a)$ by $\hat{\theta}_{i,1}(a|\hat{v}^e_{i,1})$. We do not require that updating precisely follows Bayes' rule, but we do assume that these revised expectations, $\hat{\theta}_{i,1}(a|\cdot)$, obey two basic properties consistent with Bayesian updating. First, updating is monotonic in the signal: for any two encoded values of consumption utility $\hat{v}, \hat{v}' \in \mathbb{R}_+$, $\hat{\theta}_{i,1}(a|\hat{v}) > \hat{\theta}_{i,1}(a|\hat{v}')$ if and only if $\hat{v} < \hat{v}'$. Second, the participant updates in the direction of her signal: if $\hat{v} > \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$, then $\hat{\theta}_{i,1}(a|\hat{v}) < \hat{\theta}_{i,0}(a)$; if instead $\hat{v} < \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]$, then $\hat{\theta}_{i,1}(a|\hat{v}) > \hat{\theta}_{i,0}(a)$. That is, when effort is less onerous than expected, beliefs about $\theta_i(a)$ revise downward; otherwise, they revise upward.[21,22]

Effort Choice in Period 2. In period 2, the participant announces how many additional tasks she is willing to do for a bonus payment of m dollars. Our main question is whether this willingness to work in period 2 depends on the likelihood that the participant was assigned to the noisy task, p_i. This likelihood is irrelevant in the rational model given that the participant was told well in advance that her period 2 task will exactly match her period 1 task. Under misattribution, however, sensations of elation or disappointment experienced in period 1 are wrongly attributed to the underlying task, and these sensations of surprise naturally depend on the chance of facing the noisy task, p_i. To allow for such an effect, let $e^*_i(a|p_i)$ denote participant i's willingness to work as

[20] Recall that there is no opportunity to earn additional pay in period 1. Hence, total utility in period 1 depends solely on the effort dimension.

[21] The second assumption is not required for analysis of Experiment 1, but we present here for clarity and ease of interpretation. Updating in the direction of the signal results from Bayesian learning for some commonly assumed distributions for $[\theta_i(a) + \varepsilon_{i,t}(a)]\,c(e)$, including the case where $\theta_i(a)$ and $\varepsilon_{i,t}(a)$ are independent and normally distributed. See Chambers and Healy (2012) for details.

[22] To simplify some additional analysis (specifically in Appendix A), we also assume that for any value of $\hat{v}$, the participant's posterior over θ corresponds to the random variable $\hat{\theta}_{i,1}(a|\hat{v}) + Z_{i,1}$, where the expectation term $\hat{\theta}_{i,1}$ is a constant and $Z_{i,1}$ is symmetric and independent of $\hat{v}$—that is, the person's updated expectation of θ depends on $\hat{v}$ but the residual noise around this expectation is invariant of $\hat{v}$.

a function of p_i.

Throughout our primary analyses in the main text, we assume that the participant's response, $e_i^*(a|p_i)$, represents the number of tasks that renders her indifferent between completing $e_i^*(a|p_i)$ tasks for a payment of \$$m$ and not working at all. That is, $e_i^*(a|p_i)$ is the number of tasks such that her expected total effort cost is equal to m.[23] For sake of a complete analysis, we take an alternative approach in Appendix B, where we assume that a participant with reference dependence incorporates the uncertainty induced by the BDM into her reference points along the effort and money dimensions and best responds accordingly. Importantly, *either* approach gives rise to the same key predictions: under misattribution, $e_i^*(a|p_i)$ depends on the participant's expectations over her initial task assignment, captured by p_i, while under rational updating $e_i^*(a|p_i)$ is independent of p_i. We discuss these predictions in greater detail next.

We first consider predictions under rational learning (i.e., no misattribution). To build intuition using the simplest case, suppose utility is not reference dependent (i.e., $\eta = 0$). As described above, the participant chooses effort $e_i^*(a|p_i)$ so that she is indifferent between completing $e_i^*(a|p_i)$ tasks for m dollars and not working at all; hence $e_i^*(a|p_i)$ solves

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

where $V_{i,2}^e = -[\theta_i(a) + \varepsilon_{i,2}(a)]c\left(e_{i,2}\right)$ and thus $\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right] = -\hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)c(e_{i,2})$. Condition 7 implies that e_i^* solves

$$\hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)c(e_i^*) = m, \tag{8}$$

and $e_i^*(a|p_i)$ is therefore a decreasing function of her expected cost parameter, $\hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)$. Critically, the only channel for p_i to influence $e_i^*(a|p_i)$ is through $\hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)$. However, a rational agent's beliefs are independent of p_i since the rational agent's signal, $\hat{v}_{i,1}^e$, is independent of her expectations—and hence p_i.

This result holds for a rational agent with reference-dependent preferences as well. In this case, indifference between completing $e_i^*(a|p_i)$ tasks for m dollars and not working at all implies that

[23] This effort level is the optimal response to the BDM mechanism when the participant has a "standard" utility function that does not exhibit expectations-based reference dependence. However, given that the BDM mechanism creates additional uncertainty over how much the participant will eventually work, the mechanism can conceivably alter the optimal response of a participant with reference dependence who incorporates this BDM-specific uncertainty into her reference point. Despite this caveat, we assume the participant's effort choice $e_i^*(a|p_i)$ ignores the uncertainty induced by the BDM mechanism. This approach simplifies the exposition, and it is most consistent with the wording of our survey, which asked participants to truthfully report the maximum number of tasks they are willing to do for each payment level. This approach additionally highlights that our main predictions hold when participants answer in this "intuitive" way even when it's not perfectly optimal under the BDM, and that our predictions do not stem from some interaction between the BDM mechanism and reference dependence.

$e_i^*(a|p_i)$ solves

$$\widehat{\mathbb{E}}_{i,1}\left[u_{i,2} \mid e_{i,2}\right] = \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right] + \eta \widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right]\right)\right] + m = 0, \tag{9}$$

Building on Equation 9, we show in Appendix A that $e_i^*(a|p_i)$ solves

$$h\left(\hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)\right) c(e_i^*) = m, \tag{10}$$

where $h(\cdot)$ is an increasing function of $\hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)$ that depends on η and the participant's subjective distribution of $V_{i,2}$, but which does not depend on p_i.[24] Thus while the condition characterizing $e_i^*(a|p_i)$ is more complicated due to reference dependence, the same punchline applies: the only way for p_i to influence e_i^* is through beliefs about $\theta_i(a)$. That said, these beliefs are independent of the chance that task was assigned under rational updating.

Observation 1. *Rational Learning with or without Reference-Dependent Preferences.* Let $e^*(a|p)$ denote the average effort choice among participants who face task a in period 2 and held prior beliefs that there was chance p of facing the noisy task in period 1. If participants' reference points adapt between periods 1 and 2, then both the classical and reference-dependent model without misattribution predict that average effort is independent of p: $e^*(a|p) = e^*(a|p')$ for all p, p'.

Note that Observation 1 does not say that a participant behaves the same with or without reference-dependent preferences. Rather, it says that—regardless of the underlying preferences—behavior should not depend on the prior probability of facing each task.

We now describe $e_i^*(a|p_i)$ under misattribution. As in the case above, $e_i^*(a|p_i)$ solves Equation 10. However, the misattributor makes this choice based on her (potentially) biased assessment of $\theta_i(a)$. Since she wrongly attributes sensations of elation or disappointment to $\theta_i(a)$, the misattributor errs when inferring her disutility of effort—$v_{i,1}^e$—from her total experienced utility in period 1. In particular, she encodes an overly optimistic signal $\hat{v}_{i,1}^e$ whenever the true signal $v_{i,1}^e$ beats expectations, and she encodes an overly pessimistic signal whenever the true signal falls short of expectations. Thus, fixing the outcome, raising initial expectations leads to a more pessimistic view of the underlying task, and lowering expectations leads to a rosier view of the underlying task. We therefore predict that for each option $a \in \{h, l\}$, participants' average willingness to work, $e^*(a|p)$, is increasing in p.

To illustrate more concretely, consider participant i who faces a chance $p > 0$ of being assigned the noisy task. For simplicity, assume that cost shocks are negligible, $\varepsilon_{i,t}(a) \approx 0$, and that the participant initially holds "unbiased" priors about the cost of effort: $\hat{\theta}_{i,0}(a) = \theta_i(a)$ for both $a \in$

[24] Given this statement, we have implicitly assumed that the person chooses according to her *true* η; however, our qualitative results are robust to the agent choosing according to her misspecified model $\hat{\eta}$.

$\{h,l\}$. Hence, before realizing her assigned task, she expects an effort cost in period 1—in which she completes eight mandatory trials of the task—equal to $\widehat{\mathbb{E}}_{i,0}[V^e_{i,1}] = -[p\theta_i(h)+(1-p)\theta_i(l)]c(8)$. Suppose the participant is assigned the no-noise task. Her total utility in period 1 is $u_{i,1} = v^e_{i,1} + \eta n\left(v^e_{i,1} \,|\, \widehat{\mathbb{E}}_{i,0}[V^e_{i,1}]\right)$ where $v_{i,1} = -\theta_i(l)c(8)$. Since $\theta_i(l) < p\theta_i(h)+(1-p)\theta_i(l)$, it follows that this outcome beats expectations and her gain-loss utility is positive. Thus, $u_{i,1} > v^e_{i,1}$—the participant experiences a utility higher than the intrinsic consumption utility associated with the task—and misattribution leads her to think that her consumption utility was higher than it actually was (i.e., $\hat{v}^e_{i,1} > v^e_{i,1}$). As such, she wrongly infers that the noiseless task is less onerous than it really is and forms an inappropriately low estimate of its cost parameter, $\hat{\theta}_{i,1}(l|\hat{v}^e_{i,1}) < \theta_i(l)$. Furthermore, if p is larger—the noisy task is more likely—then the noiseless task generates greater elation and the misattributor's estimate of $\theta_i(l)$ is biased downward by more. The converse is true if the misattributor were instead assigned the noisy task: her estimate of $\theta_i(h)$ is biased upward by more when the noisy task comes as a greater disappointment; that is, the lower is p.

Observation 2. *Learning With Misattribution of Reference Dependence.* Let $e^*(a|p)$ denote the average utility-maximizing effort choice among participants in period 2 who face task a and held prior beliefs that there was chance p of facing the noisy task in period 1. Suppose $\hat{\eta} < \eta$ and suppose each participant's prior beliefs over $\theta(a)$ are independent of treatment with $\hat{\theta}_{i,0}(l) < \hat{\theta}_{i,0}(h)$. If participants' reference points adapt between periods 1 and 2, then elicited effort $e^*(a|p)$ is increasing in p for each $a \in \{h,l\}$.

The two observations together highlight our empirical strategy. Fixing the task participants ultimately faced, we compare willingness to work across treatment groups to test whether the prior probability affects the resulting willingness to work and whether it matches the comparative static discussed in Observation 2.

3.2.1 Discussion of Assumptions

We now discuss some of the assumptions underlying the results above. First, we clarify the extent to which they rely on a participant's reference point changing between the two sessions of the experiment. Second, we discuss robustness to participant's prior beliefs and highlight the relationship between these priors and the motivation behind our *high-probability* treatment.

Adjustment of the Reference Point Across Periods. The observations above assume that the participant's reference point does not adapt between the coin flip and her initial work. In assuming this, we leveraged the fact that the participant begins working immediately after the coin flip and thus there is almost no time for a reference point to adapt. We do, however, assume that the participant's reference point adapts between sessions 1 and 2; that is, the lottery over task assignment that determines expectations in period 1 no longer influences expectations in period 2, which comes at

least 8 hours after the resolution of the gamble. 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 muddle these distinctions. In particular, if participants have sluggish reference points (i.e., expectations still depend on the lottery in session 2) *and* experience reference-dependent utility over effort and *not* money, then reference dependence without misattribution predicts effort patterns similar to those predicted by our model of misattribution. We find this particular constellation of assumptions unlikely; moreover, it is inconsistent with existing evidence demonstrating reference-dependent preferences over money.

Our design utilizes a relatively long gap between sessions to help ensure that reference points adapt by the time Session 2 begins. The evidence to date supports the idea that reference points adapt over modest time periods (and indeed informed our experimental design). As mentioned previously, Song (2016) demonstrates that reference points incorporate new information over the course of approximately ten minutes. Likewise, Smith (2012) and Buffat and Senn (2015) both provide evidence of relatively quick reference-point changes in laboratory settings with small stakes.[25] Taken together, we share Song's (2016) interpretation of the broader literature: for small-stakes laboratory experiments, reference points seem to adjust on the scale of tens of minutes. Furthermore, we empirically explore this concern in the analysis below. Recall that subjects could choose when to complete session 2 so long as they waited at least 8 hours after session 1. We find no difference between participants who completed session 2 relatively soon after the mandatory 8-hour waiting period and those who waited longer.

Robustness to Poorly-Calibrated Priors. The observations above do not require subjects to have well-calibrated priors about the tasks (i.e., about the $\theta_i(a)$'s). If prior beliefs are biased on average, our observations will hold so long as these participants' priors do not systematically vary across treatment groups. In this case, fixing the task a participant faced, rational learning will lead to the same posterior beliefs regardless of the treatment—the treatment does not influence in the interpretation of signals nor priors. In contrast, misattribution creates an interaction between poorly calibrated priors and the treatment. However, so long as those priors are reasonable—specifically, participants believe the noisy task is more onerous than the noiseless one—then the prediction from

[25] Smith (2012) endows participants with a lottery to receive a water bottle. Some participants face a low-probability of winning while others face a higher chance. Once prizes are awarded, winners reveal their willingness-to-accept (WTA) to sell their bottle, and losers are asked their willingness-to-pay (WTP) for the water bottle. The author highlights that WTA and WTP for the bottle should increase in the probability of winning the water bottle—however, he does not find evidence of such an "attachment effect". Smith interprets this as evidence that reference points adjust quickly. Buffat and Senn (2015) examine preferences after the resolution of sequential lotteries over money. In that study, all participants face one of three possible gambles and, after the realization of that gamble, participants give their WTP for a 50/50 chance to gain CHF 10. In this setting, a slowly-adapting reference point would lead participants to react differently to the three initial gambles—however, the authors find no evidence of this for small stakes. For larger stakes, there is some evidence of a *house-money* effect, wherein risk attitudes depend on the outcome of the initial gamble.

Observation 2 still holds. We believe that such "reasonable" priors are likely given that participants sampled each task during the instructions, and our data indeed suggests that participants disliked the noisy task relative to the task without noise.

Priors Independent of Treatment-Group Assignment. The claims above (specifically, Observations 1, 2 and robustness to poorly calibrated priors) rely on independence between a participant's priors about the tasks and her treatment group (i.e., the likelihood she is assigned the noisy task). However, it is is plausible that participants in the *coin flip* treatment—who are exposed to both tasks during the instructions—form beliefs about a given task that systematically differ from participants in the control who were exposed to only the that task they will face. (For instance, the existence of both an easy and hard version of the task might lead a participant in the *coin-flip* group to infer that the noisy task is particularly onerous, while a participant in the control group is only aware of the noisy task and might expect it to resemble a "typical" MTurk task.) This would violate our assumption that priors are independent of treatment. This (plausibly) rational-inference story could generate willingness to work that is more exaggerated across tasks for those in the *coin flip* group, much like our misattribution theory predicts. Our *high-probability* treatment addresses this concern: exposure to the two tasks in this treatment exactly match the *coin flip* treatment, mitigating concerns about differential inference. In this sense, we use the *high-probability* group (i.e., participants very likely to face task a) as a less-confounded version of the associated *control* group (i.e., participants certain to face task a). In both groups, participants strongly expect to face task a, but in the *high-probability* version they are perfectly aware of the alternative task.

3.3 Results

In this section, we analyze the results of Experiment 1. Guided by the theoretical discussion above, we first present a non-parametric analysis demonstrating that willingness to work in Session 2 strongly depends on participants' initial expectations regarding their task assignment. We then estimate parameters of our model and demonstrate that behavior is consistent with participants wrongly learning about 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.[26] Note that vari-

[26] 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 age and time-of-day effects.

ability 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.

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)
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)
𝟙(Return)	.921	.882	.862	.944	.932	1
	(.271)	(.323)	(.346)	(.231)	(.253)	(0)
Observations	139	153	152	142	160	140

Notes: Standard errors 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

We implemented some data-cleaning procedures to form our primary dataset. We removed participants who either (i) did not answer all five elicitations of willingness to work[27] (three participants), or (ii) stated a willingness to work equal to the maximum amount (100 tasks) for every payment level, which prevented us from estimating their responsiveness to payment (six participants).[28] Additionally, we omit participants who did not return for the second session—and whose willingness to work 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.

[27] 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.

[28] Of the six participants dropped due to the latter criterion, 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.

3.3.1 Nonparametric Analysis

Our main hypothesis was that participants' willingness to work on a given task would depend on their expectations regarding their task assignment prior to the initial-learning session. As a first step to investigate this hypothesis, we compare the average willingness to work in the *control* and *coin-flip* treatments, where we average 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 whether uncertainty over task assignment in the initial-learning session affected subsequent behavior. Relative to the control group, participants who faced the noiseless task were willing to work significantly more when their initial impressions were formed after the resolution of the coin flip ($p = .039$ for difference; standard errors clustered at individual level). In contrast, participants who faced the noisy task were willing to work significantly *less* (relative to control) when their initial impressions were formed after the resolution of the coin flip ($p = .025$ for difference; standard errors clustered at individual level).

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	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 $p < .05$.

While Table 2 gives a rough sense of the treatment effect, we further disaggregate willingness to work by payment level in Figure 3. Continuing our comparison of the *control* and *coin-flip* treatments, the top panel of Figure 3 shows the average willingness to work at each of the five payment levels {\$0.50, \$1.00, \$1.50, \$2.00, \$2.50} for each of the four groups (crossing treatment with task assignment). At all payment levels, we find that those who formed initial impressions of the noiseless task when it came as a positive surprise were less willing to work than those who faced the same task with certainty. In contrast, those who formed initial impressions of the noisy task when it came as a negative surprise were less willing to work than those who faced the same task with certainty.

The bottom panel of Figure 3 shows the cumulative distributions of willingness to work in each

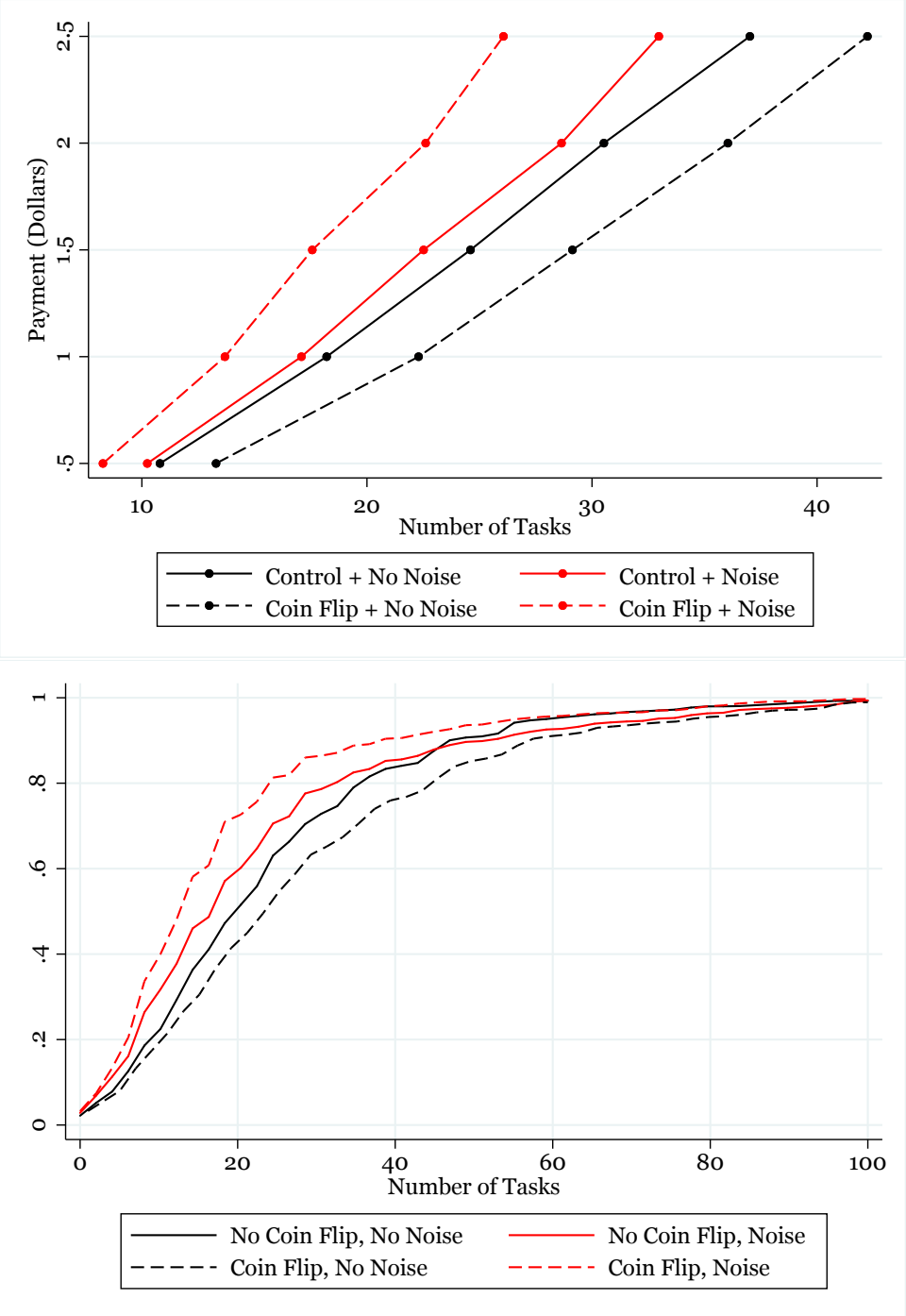

Figure 3: *(a) Labor supply curves and (b) cumulative bid distribution by group assignment.* Cumulative distribution curves aggregate over all five payment levels and are smoothed using the Epanechnikov kernel.

of the four groups, aggregated over all payment levels (and smoothed using the Epanechnikov kernel). As a simple check, a Kolmogorov-Smirnov equality-of-distributions test reveals that control participants were more willing to work at the noiseless task than the noisy one—verifying that the noisy task was, in fact, more onerous.[29] Speaking to our main hypotheses, the figure highlights that willingness to work in the *contol + no noise* was significantly lower than the *coin flip + no noise* group—the latter almost first-order stochastically dominates the former. By contrast, the cumulative distribution of willingness to work in the *control + noise* group first-order stochastically dominates that of the *coin flip + noise* group.

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.

We now address three plausible alternative explanations for these baseline results: differential information across treatments, reciprocity toward the experimenter, and attrition. For each, we discuss how we can limit the scope for the alternative explanation.

Independent Priors Across Treatments. As discussed above (Section 3.2.1), the observed differences between the *control* and *coin-flip* groups may reflect differences in information rather than misattribution: recall that the a participant in the control group was told only about the task she worked on, while a participant in the *coin-flip* group was told about both tasks regardless of her assignment. This differential exposure to the possible tasks may create priors about task difficulty that differ across groups (as previous noted in the discussion concluding Section 3.2). The *high-probability* treatment helps address this potential confound: participants in the *control* vs *coin-flip* treatments had different exposure to the two tasks and thus they may have formed different prior beliefs. The *high-probability* treatment helps rule out such concerns. A participants in that group was very likely to face a particular task (and therefore had expectations about assignment that were similar to a participant in the *control* group) yet was exposed to both tasks in the experimental instructions (similar to a participant in the *coin-flip* group). By comparing willingness to work in the

[29] While this test fails to account for redundancy in the data stemming from multiple observations from each individual, we calculated a 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 willingness to work for *control + noise* and *control + no-noise*; the five p values were .024, .189, .041, .019, .090.

coin-flip and *high-probability* treatments, we can assess the effect of different expectations about task assignment while eliminating the differences in information that exist between the *control* and *coin-flip* groups

Columns 3 to 6 of Table 2 summarize behavior under the *high-probability* and *coin-flip* treatments. As before, we find a significant difference in willingness to work depending on expectations during the initial-learning session. Participants who were assigned the noiseless task based on the coin flip were, on average, willing to work significantly more than those who strongly expected the noiseless task ($p = .034$ for difference; standard errors clustered at individual level). In contrast, participants who were assigned the noisy task based on the coin flip were willing to work significantly *less* than those who strongly expected the noisy task ($p = .047$ for difference; standard errors clustered at individual level).

In the comparison above, we use our *high-probability* treatment as a replacement for the *control* group in order to equalize information across treatments. However, it is not a direct replacement: because of 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 willingness to work across the two tasks than those in the *control* groups (e.g., relative to being assigned the noisy task *for sure*, the noisy task is slightly more disappointing when expecting a high, but not certain, chance of that task). Thus, fixing the assigned task, our model predicts that the average willingness to work of those in the *high-probability* group should fall in between that of the *control* and *coin-flip* groups. Indeed, we find suggestive evidence to this effect (see Figure 4) although we are underpowered to properly compare these treatments.

Furthermore, 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* groups may substantially deviate from the corresponding control group. Probability weighting would suggest that the 1% chances presented in the *high-probability* treatment loom much larger than the objective probability. If this is the case, participants may treat the *high probability* as closer to the *coin flip* than is merited by the objective probabilities, hindering our ability to detect differences across these treatments. Thus, although we do find significant differences between the *high-probability* and *coin-flip* treatments, the statistical tests are perhaps overstating the likelihood of the null hypothesis being true.

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 no observables (e.g., task assignment, treatment, nor demographics) predict attrition. However, an alternative type of attrition is possible given the MTurk setting: some partic-

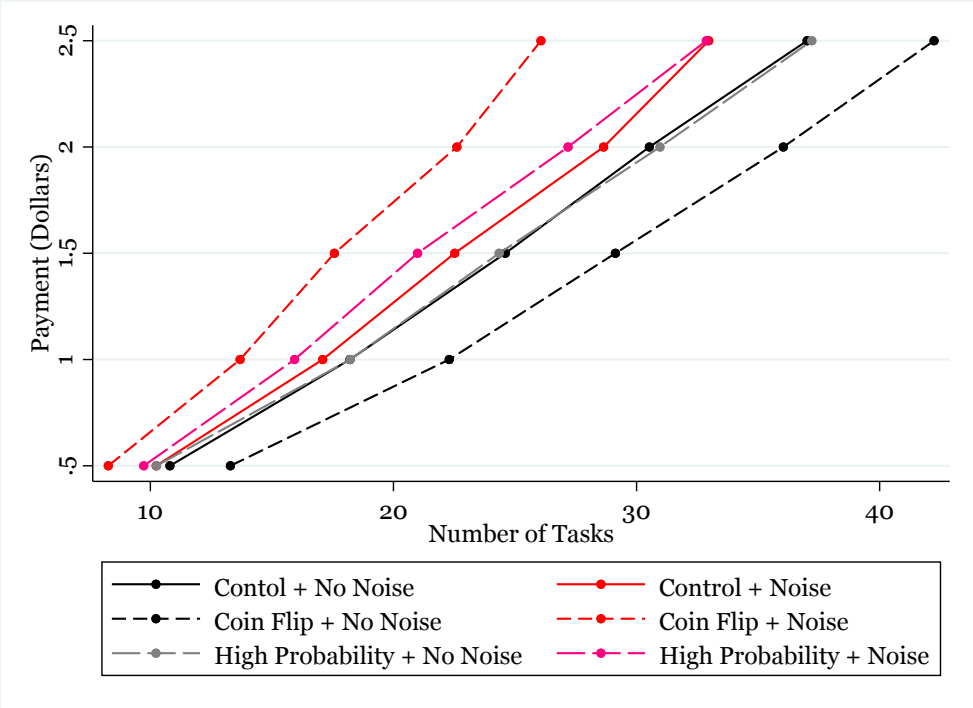

Figure 4: Labor supply curves across all treatments. Each point represents the average willingness to work for a fixed payment as elicited under the BDM mechanism.

ipants may have exited the survey when assigned to the noise 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 no-noise task and three were assigned to the noisy task. We accordingly reject attrition-based explanations for the observed effects.

Reciprocity. Instead of misattribution, our baseline findings could plausibly result from reciprocity toward the experimenter: a positive surprise encourages participants to work hard to reward the experimenter, yet a negative surprise leads participants to punish the experimenter through low effort. However, such an explanation requires a set of assumptions that is, in fact, similar to our notion of misattribution. Specifically, it must be the case that the ex-ante probability of task assignment alters the degree to which a person feels reciprocity towards the experimenter.[30] Furthermore, this explanation requires that the participant continues to feel positively toward the experimenter more than eight hours later. Given the relatively small stakes involved in this experiment, we suspect this is an unlikely explanation but such probability-dependent reciprocity is not directly ruled out by our design.

[30] Experimental demand effects are similar to the reciprocity argument above. Recently, de Quidt, Haushofer and Roth (2019) directly estimated and bounded those demand effects at approximately .17 standard deviations. This magnitude is insufficient to account for the effects we observe.

Transient Moods. Finally, we note that our experiment was designed with the concern that short-term moods induced by resolving uncertainty might explain our effects.[31] Specifically, the time gap between participants forming their impressions and our elicitation of willingness to work helps distinguish our effect from that of short-term "transient moods", as any influence of the coin flip on mood should fade over more than eight hours. We provide an additional empirical test in Supplemental Tables A1 and A2. There, we reproduce Table 2 but divide the sample in two: those who returned after more than the median amount of time between Sessions 1 and 2, and those who returned after less than the median return time. We find qualitatively similar results, though our statistical power is greatly diminished.

3.3.2 Parametric Analysis

Motivated by our simple nonparametric results, we now consider a more structured, regression approach. We follow the model outlined and discussed in Section 3.2. This allows us to properly account for the fact that effort costs in our experiment may be non-linear. In doing so, we provide better estimates of the aggregate effort-supply curves illustrated in Figure 3 while supplying appropriate confidence intervals.

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 group.[32] 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 will estimate the average value of this expectation, denoted $\hat{\theta}_1(a|p)$, among participants in each subgroup; that is, for each relevant combination of ex-ante probability of task assignment, p, and assigned task, a.

In order to estimate these parameters, we impose a particular form of effort-cost function: following Augenblick, Niederle and Sprenger (2016) and others, we assume $c(e) = (e+\omega)^{\gamma}$, where ω is a Stone-Geary background parameter.[33] Given this functional form, identification of the common cost function and the relevant parameter $\theta(a)$ is straightforward. Utilizing Equation 8 along

[31] This was a central design concern because transient factors have, for example, been demonstrated to influence investor sentiment. Fluctuations in the weather (Saunders 1993, Hirshleifer and Shumway 2003) and sports outcomes (Edmans, Garcia and Norli 2007) both lead to systematic changes in stock returns. In psychology, the more general idea that positive or negative affect can distort unrelated behavior is well documented. For example, Isen and Levin (1972) showed that participants were more likely to help others after they themselves experience positive, unrelated events.

[32] As discussed in Appendix B, the reference to $\theta(a)$ above should technically be an increasing function of θ which we denote $h(\cdot)$. For ease of readability, we retain the notation from our theoretical section and direct the interested reader to the Appendix for appropriate additional details.

[33] For the analysis presented below, we take $\omega = 0$. Although numerical estimates of γ and the collection of parameters θ are sensitive to this assumption, our qualitative results are robust. Over a wide range of ω, we estimate significant differences in parameters across our treatments. We present this analysis in Table A3.

with our assumed cost function implies that participant i chooses e_i^* such that $\hat{\theta}_{i,1}(a|p)(e_i^* + \omega)^\gamma = m$. Rearranging, setting $\omega = 0$, and taking logs yields

$$\log(e_i^*) = \frac{\log(m)}{\gamma} - \frac{\log(\hat{\theta}_{i,1}(a|p))}{\gamma}. \tag{11}$$

Assuming an additive error structure, Equation 11 suggests the following regression model:

$$\log(e_i^*) = \beta_0 \log(m) + \sum_{j=1}^{6} \beta_j \big(\mathbb{D}_i(\text{treatment}) \times \mathbb{I}_i(\text{noise})\big) + \delta_i, \tag{12}$$

where $\mathbb{D}_i(\text{treatment})$ is a dummy variable indicating whether person i was in a particular treatment (*control*, *coin flip*, or *high probability*) and $\mathbb{I}_i(noise)$ is an indicator variable designating whether that person ultimately faced the task with or without noise. Variation in payouts, m, delivers identification of the curvature parameter, γ, and variation in treatment assignment crossed with the task the participant ultimately completed delivers identification of $\hat{\theta}_1(a|p)$. Thus mapping Equation 11 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 aggregate beliefs of participants in the *control* + *noise* subgroup—$\hat{\theta}_1(h|p=1)$—we combine the coefficient on $\mathbb{D}_i(control)\mathbb{I}_i(noise)$ with the coefficient on $\log(m)$ as prescribed above.

In Table 3, we present the results of 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 willingness to work is censored at a minimum value of 0 tasks and a maximum value of 100, and (ii) we have five observations for each person. Column (1) presents the estimates of the baseline specification in Equation 12. First, we estimate the cost-curvature parameter to be $\gamma = 1.207$ (0.023); we can accordingly reject a linear cost function despite the linear appearance of the aggregate data in Figure 3.[34]

Table 3, Column (1) demonstrates our main result: willingness to work—and accordingly our estimate of perceived effort costs—are shaped by participants' prior expectations over task assignment. For ease of interpretation, the rows of Table 3 (after the first) are ordered to match the ranking of cost perceptions predicted by our model of misattribution. Participants whose task assignment was determined by coin flip acted as if they formed the most extreme views of the underlying difficulty of the task. Specifically, when participants formed their initial impressions immediately after an unfavorable coin flip, they acted as if they formed more pessimistic views of the

[34] As a form of robustness check, we estimated a model that mirrored 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 $H_0 : \gamma(h) = \gamma(l); \chi^2(1) = 0.24; p = 0.624$.

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 ($\hat{\theta}_1(l|.5) - \hat{\theta}_1(l|0) = -.0064$; $\chi^2(1) = 2.49, p = .115$).

For robustness, Column (2) of Table 3 controls for demographic characteristics (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 responses were not weakly monotonic in payment—that is, their willingness to work did not weakly increase across all five payment levels. This drops a significant portion of the sample, but the point estimates of our effect remain similar.[35]

Perhaps most notable from Table 3 is that the ordering of parameter estimates closely matches the predictions of our misattribution model. Indeed, the hypothesis that $\hat{\theta}_1(a)$ does not depend on p is rejected ($\chi^2(4) = 9.88, p = .043$). Given our non-parametric results in combination with these structural estimates, we conclude that manipulating prior expectations had a significant effect on subsequent willingness to work in a pattern that is consistent with attribution bias of reference-dependent utility.

Finally, we note that the results above demonstrate a large and economically significant effect of expectations over task assignment on our estimates of perceived effort costs. For example, we estimate a roughly 20 percent difference in perceived effort costs between those participants facing the noisy task in the *coin-flip* treatment and those participants who ultimately faced the same task but began in the *high-probability* treatment. This finding mirrors our earlier non-parametric results.

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, highlighting how the approach allows us to rule out any interaction between treatment and priors that may have taken place in Experiment

[35] Although we observe a seemingly high number of non-monotonic responses, we believe that our response mode (slider) was conducive to small mistakes. There were a total of 111 total responses that were non-monotonic—that is, the willingness to work for some higher fixed payment was less than that at a lower fixed payment. The average mistake (that is, the magnitude of the deviation from responses that increase in stakes) was small.

Table 3: PARAMETRIC ANALYSIS, EXPERIMENT 1

	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	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 lef-censored and 43 are right-censored in the main sample; 11 are left-censored and 43 are right-censored in the "monotonic" sample.

1. We then discuss our theoretical predictions, which are similar to those from Experiment 1 when applied to a within-subject setting. Finally, we analyze the experimental data. Experiment 2 yields similar conclusions to Experiment 1, but extends our findings to a different experimental population, albeit with a greatly reduced sample size. Importantly, this design allows us to (noisily) estimate within-subject measures of misattribution, an exercise not possible with Experiment 1.

4.1 Design

We recruited participants from the Harvard student body for a two-session experiment, with sessions separated by a week. A total of eighteen sessions (nine groups) were conducted over the course of one month. Our primary sample consists of 87 subjects.[36] Participants were paid $7 for successfully completing each of two sessions in addition to any earnings from their choices. In order to prevent attrition, we paid participants contingent on completion of both sessions.

Before specifying the details of Experiment 2, we first provide a broad overview of the design to highlight how it 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. Thus, participants faced uncertainty over their task assignment in the first session, but not in the second. 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.

We measured participants' willingness to work in both sessions of the experiment. Assuming participants' expectations about task assignment change across sessions, then the change in participants' willingness to work across sessions allows us to identify misattribution. That is, our variable of interest is the difference in a participant's willingness to work in week one—when her task came as a surprise—and week two—when that same task was expected.

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.[37] Each trial of the task consisted of a string of 35 handwritten characters; participants "transcribed" each character by clicking the matching letter from an alphabet of

[36] Ex-ante power tests suggested that $n \approx 100$ would provide 80% power, assuming a modest effect size. We under-recruited because our sampling window coincided with the end of the academic school year. Additionally, two of the groups that we recruited later in the sampling window had higher-than-average attrition, which we suspect was due to final exams. One participant withdrew moments into the first session due to a scheduling conflict; a second withdrew in the middle of the first session because she did not want to take part in the study (and offered no further explanation). These two participants are excluded from all analyses.

[37] 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.

the relevant language. See Figure 5 for a screenshot. Participants were randomly divided into one of two language treatments: half transcribed Greek during the first session and Russian during the second, while the other half faced the opposite order.Aside from variation in the language, each session had the same structure: participants first completed an initial-learning phase which consisted of five mandatory trials, and then we elicited their willingness to complete additional trials for a bonus payment.

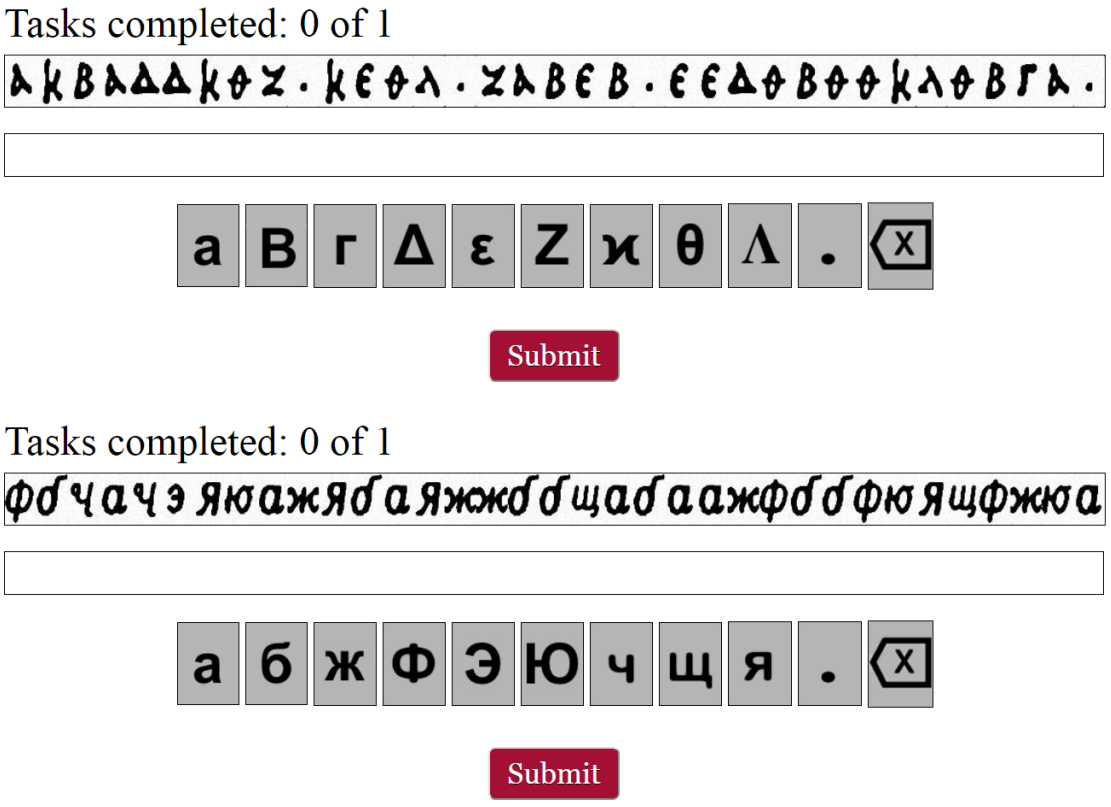

Figure 5: *Screenshot of the transcription task from Experiment 2.* Participants clicked the gray button that matched the handwritten letter to "transcribe" the text. Participants were required to achieve 80% accuracy to advance to the next transcription. Each participant randomly faced one language—Greek or Russian (Cyrillic)—during their first session, and then faced the other language 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 played through the headphones (calibrated to roughly 70-75 decibels); the noise was identical to Experiment 1, except it played on loop for the entire transcription time. In the noiseless version, no sound played through the headphones. In order to endow participants with reasonable priors about each task, the initial instructions included an interactive sample of the transcription task, and participants listened to an

eight-second sample of the annoying noise (repeatable if desired).

Session 1: Coin Flip and Eliciting Willingness to Continue Working. 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. In order to make this probability salient—and to enhance the sensation of surprise or disappointment—each participant flipped a U.S. quarter to determine their assignment. We instructed participants that a flip of heads would result in the noiseless task, while tails would result in the noisy one. After resolving the coin flip, each participant immediately started their initial-learning phase in which they 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. As in Experiment 1, we asked each participant how many additional tasks they were willing to complete for each of five payments: $\{\$4, \$8, \$12, \$16, \$20\}$. As in Experiment 1, participants responded by using a slider to select any integer $e \in \{0, \ldots, 100\}$, and we used the BDM mechanism to incentivize these responses.

Session 2: Different Language and Eliciting Willingness to Continue Working Again. 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' willingness to continue working on that task. The experiment concluded after participants completed any additional trials. Subjects were paid only upon completion of both sessions.

As noted above, participants transcribed a different language in the second session. We introduced this minor variation in the task across sessions so that participants could plausibly form different perceptions of the task over time and hence update their willingness to work. This design feature was intended to help reduce anchoring 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. That is, we provided subjects with a potential "cover story" for changing their desired amount of work across sessions.

4.2 Theoretical Predictions

We now sketch how our theoretical predictions from Experiment 1—presented earlier in Section 3.2—extend in this within-subject design. In contrast to Experiment 1, a participant in this setting receives two signals about her cost function, and we measure her willingness to work twice—once after receiving the first signal, and then again after receiving the second. These two signals derive from the participant's consumption utility of effort in the initial-learning phase of Sessions 1 and 2. We focus our analysis on participants who do not complete additional tasks during the first session.

Thus, aside from the experimental instructions, a participant's signals from the two initial-learning phases are her only information about the tasks.

We first describe the predictions of rational learning. Throughout this section, we maintain the same basic setup and assumptions introduced in the theoretical analysis of Experiment 1 (Section 3.2), and we further assume that each participant's priors over $\theta_i(a)$ are unbiased on average. In this case—where participants hold reasonable expectations about the difficulty of the tasks—rational learning without reference-dependent preferences predicts that a participant's willingness to work will not systematically vary across the two periods. In contrast, rational learning *with* reference dependence but without misattribution can lead a participant to systematically change her willingness to work across periods. Namely, as shown in the Online Appendix, 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.

We will now demonstrate how misattribution predicts an opposing effect. Specifically, our model predicts that those assigned the noisy task will typically increase effort between periods 1 and 2. In contrast, those assigned the noiseless task will *decrease* effort. As with the KR model, these behavioral changes stem from a participant's reference point evolving over the two periods. In the first period, her reference point puts a 50% chance on each of the two tasks. We assume that by the second period, the participant fully anticipates her assigned task and her reference point adapts accordingly. Thus, the participant's two experiences with her assigned task—the initial-learning phase at the start of each session—happen under different reference points. Misattribution will thus cause her to encode these similar experiences differently.

More formally, suppose that consumption utility takes the same form as the model underlying Experiment 1. Thus, following Equation 5, consumption utility from each initial-learning phase—in which the participant completes five trials of her assigned task $a \in \{h,l\}$—is $v^e_{i,t} = [\theta_i(a) + \varepsilon_{i,t}(a)]c(5)$.[38] On average, a participant assigned the noisy task ($a = h$) will encode these values such that $\hat{v}^e_{i,1} < \hat{v}^e_{i,2}$. This is because her first signal incorporates a sense of disappointment—in period 1, she anticipates a 50% chance of facing the better task. But her second signal comes with less disappointment—in period 2, she fully expects the worse task. Put differently, the participant's first experience falls short of expectations by a greater amount than the second and is the misattributor remembers it as worse. In contrast, an average participant assigned to the noiseless ($a = l$) task will encode values such that $\hat{v}^e_{i,1} > \hat{v}^e_{i,2}$: the first signal incorporates a sense of elation from the coin flip, but the second signal comes with less (if any) such elation.

To illustrate the logic at play above, first consider a participant in the no-noise condition. As-

[38] We do not assume that the cost function in Experiment 2 is the same as Experiment 1 given that the tasks in these two experiments are quite different. That said, we model the cost function in a similar way for both experiments. As such, we assume value of θ and functional form of $c(\cdot)$ vary across the two experiments.

suming priors about $\theta(a)$ are initially unbiased, a participant will form a distorted perception of her assigned task in the initial-learning phase immediately after the coin flip—since the no-noise task comes as pleasant surprise, the participant underestimates the true cost. Given that her stated willingness to work in Session 1 is based on this overly-optimistic perception of the underlying disutility of effort, her statement will be 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, and this experience in the learning phase of Session 2 tends to come as an *unpleasant* surprise. Since her expectations developed in Session 1 overestimate her enjoyment, 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 willingness to work in the second session. If this "contrast effect" between the first and second rounds is sufficiently strong, then the *no-noise* participant's revealed willingness to work will decrease over the two sessions. Similar logic extends to a misattributor in the noisy condition increases her effort across sessions: in the first session, the negative surprise of her unfavorable task assignment leads her to overestimate the disutility of effort. Her experience with that same task in the second session, however, will typically surpass her overly-pessimistic expectations. This positive surprise then increases her willingness to work in the second session.

These systematic changes in the participant's encoding of her experiences have direct implications for her perceptions of the task and effort choices across periods. We continue to assume that a participant's updated expectation of $\theta_i(a)$ following each of her encoded signals, denoted by $\hat{\theta}_{i,t}(a|\hat{v}^e_{i,t})$, has the two properties introduced in Section 3.2: it is monotonic in $\hat{v}^e_{i,t}$, and it updates in the direction of the signal. Under our maintained assumptions, these encoding patterns imply that participants assigned to the noisy task will typically find their task less onerous in period 2 than period 1—that is, on average $\hat{\theta}_{i,2}(h) < \hat{\theta}_{i,1}(h)$. In contrast, those assigned the noiseless task will typically find it more onerous in period 2 than period 1—that is, on average $\hat{\theta}_{i,2}(l) > \hat{\theta}_{i,1}(l)$. These are the main predictions we empirically test.

The behavioral implications of these predictions, however, are met by a countervailing force stemming from rational reference dependence noted above.[39] Accordingly, mapping these predictions to effort choices—the observable in our experiment—is not trivial. If the countervailing force is small (for instance, if loss aversion is relatively small) then the predictions extend to effort choices. Let $e^*_{i,t}(a)$ denote the observed effort choice for participant i facing task a in period t.

[39] This countervailing force—an incentive for those facing the noisy task to decrease effort over time and those facing the noiseless task to increase it—stems from a (rather sophisticated) forward-thinking equilibrium notion introduced in Kőszegi and Rabin (2006). (See the Online Appendix for details.) If, alternatively, participants did not make such forward-thinking plans, the countervailing force would not be present and the results above are an immediate extension of the theoretical results from Experiment 1.

With either strong misattribution or sufficiently weak loss aversion, we predict $e^*_{i,1}(l) > e^*_{i,2}(l)$ and $e^*_{i,1}(h) < e^*_{i,2}(h)$ on average.

Discussion. Our analysis above assumed that priors were unbiased. If priors are systematically biased in a specific way—namely, they significantly overestimate the disutility of the task with noise and underestimate the disutility of the task without noise—then changes in willingness to work across sessions may result from rational learning, even absent reference-dependent preferences or misattribution. We believe our assumption of correct priors (on average) is justified from the experimental design: participants were exposed to both versions of the task before commencing work, and therefore should have reasonably well-calibrated priors. Fortunately, as highlighted previously, this limitation does not apply to Experiment 1, where comparing our *high probability* and *coin flip* treatments removes any scope for biased priors.

Finally, the discussion above assumed that reference points adapt to the assigned task by the beginning of Session 2. This seems warranted given that there was no uncertainty in task assignment in the second session, and participants knew about their task assignment a week in advance. Furthermore, they were reminded by email midway through the week. Before beginning Session 2, all participants were required to verbally state which task they had faced in Session 1, and all participants did so successfully. This suggests that the assignment was salient and memorable.

4.3 Results

For our primary analysis, we only consider participants who returned to both sessions. Thus, our data comes from 70 participants who completed the experiment across a total of nine different experimental groups. For completeness, we present a (simple) analysis of participant attrition in Table A6.

We first present nonparametric analyses demonstrating that willingness to work systematically changes over time depending on the resolution of the coin flip in Session 1. We then structurally estimate the parameters of a model similar to Experiment 1, but utilizing the within-subject nature of this design. We conclude by demonstrating that our results are robust to informational explanations stemming from those participants who completed additional tasks in the first session.

4.3.1 Nonparametric Analysis

Relative to Experiment 2, this experiment consisted of only one "treatment": all participants faced uncertain assignment (via coin flip). However, we elicited each participant's willingness to work twice—once when the participant had very recently resolved uncertainty over task assignment and once when assignment was known for a long period. Accordingly, Sessions 1 and 2 of this experiment mirror the coin-flip and control treatments from Experiment 1, respectively, where

the difference stemmed from either a stochastic or deterministic reference point. Table 4 presents participants' average willingness to work—averaged over the five payment levels—in each session.

Our design is not intended to detect between-subject differences in WTW, and the aggregate results in Table 4 obscure important within-subjects variation. Instead, our design allows us to account for individual differences in overall willingness to work by examining *changes* in WTW across Sessions 1 and 2. Accordingly, our variable of interest in this case is the change in an individual's willingness to work conditional on her task assignment in Session 1. First, we find that willingness to work significantly changes across sessions (see Columns (5)-(6) of Table 4). Furthermore, consistent with our theoretical predictions, participants' assigned the noiseless task tend to decrease their willingness to work across sessions while those assigned the noisy task tend to increase it. When assigned the noiseless task, participants were on average willing to complete 7.1 more tasks in Session 1 than in Session 2 ($p = .0014$, standard errors clustered at individual level). 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 ($p = .006$, standard errors clustered at the individual level.)

Figure 6 depicts this result by plotting the density of $e^*_{i,1} - e^*_{i,2}$ for each task, averaged over the five payment levels.[40] Figure 6 demonstrates that the difference in willingness to work is primarily positive for participants assigned the noiseless task; it is primarily negative for those assigned the noisy task.

To assess the economic magnitudes of these results, we again consider a hypothetical firm seeking workers to complete 25 transcriptions (as done the discussion of Experiment 1, Section 3.3). To incent the average participant to complete 25 noiseless transcriptions, a firm would have to pay \$7.75 right after the worker forms her initial impression (i.e., just after the positive outcome of the coin flip); this increases to \$11 when the participant returns and 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 forms her initial impression (i.e., just after the negative outcome of the coin flip); this decreases to \$10.50 when the participant returns and her assessment of the task is no longer confounded with disappointment. These effect sizes have similar magnitude to those in Experiment 1.[41]

[40] We present these densities using kernel smoothing (Epanechnikov kernel) for readability; raw histograms appear in the Appendix as Figure A2.

[41] 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 is more time-consuming than that of Experiment 1 and because the lab subjects are paid more in general, the magnitudes of payments are significantly different across experiments. Second, because the sample size in Experiment 2 is much smaller, the estimated effect size is quite imprecise, and we cannot make claims across experiments with much confidence.

Table 4:
BASELINE RESULTS, EXPERIMENT 2

	Session 1		Session 2		$\left(e^*_{i,1} - e^*_{i,2}\right)$	
Variable	noise=0	noise=1	noise=0	noise=1	noise=0	noise=1
Willingness to Work	30.95	25.93	26.01	26.41	7.14	-4.25
	(3.672)	(3.526)	(3.092)	(3.575)	(2.429)	(1.645)
Observations	215	220	175	185	175	185

Notes: Standard errors (in parentheses) are clustered at the individual level. Differences between Columns (1)-(3) significant at $p = .026$; between Columns (2)-(4): $p = .865$. Columns (5)-(6) both significantly different from zero: $p = .0014$ and $p = .006$, respectively.

4.3.2 Parametric Analysis

We now present our qausi-structural estimation. Given the experiment closely follows the approach from Experiment 1, the decision problem in each session is the same as in the previous experiment, and is thus described by the logic in Section 3.2. Mirroring our parametric approach to Experiment 1—and adopting the previous notation—Equation 11 implies that for each period, $\log(e^*_{i,t}) = \frac{\log(m)}{\gamma} - \frac{\log(\hat{\theta}_{i,t}(a|p))}{\gamma}$. However, now that we have multiple observations for each individual, we can examine the difference $\log(e^*_{i,1}) - \log(e^*_{i,2})$. This difference is independent of m, thereby eliminating a potential source of (unmodeled) heterogeneity. Our econometric model is thus

$$\left(\log(e^*_{i,1}) - \log(e^*_{i,2})\right) \;=\; \beta\, \mathbb{I}_i(\text{noise}) + \varepsilon_i. \tag{13}$$

Given this specification, we can recover aggregate estimates $\frac{\hat{\theta}_1(a|p)}{\hat{\theta}_2(a|p)} = \exp(-\gamma\beta)$. Since γ is not identified in this specification, we separately model the first session only (following Equation 12) to generate an in-sample estimate of $\gamma \approx 1.14$; note this estimate falls close to our estimate from Experiment 1.[42] We then plug this estimate into the recovery equation above to numerically approximate the ratio of interest.

As with Experiment 1, we estimate Equation 13 using a random-effect Tobit model. The results are shown in Table 5. Our structural estimates align closely with those of Experiment 1. Compare the ratio $\frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})} = 1.29$ to the analogous ratio implied by Column (1) of Table 3: $\frac{\hat{\theta}(\text{noise}|\text{coin flip})}{\hat{\theta}(\text{noise}|\text{control})} = 1.28$. Likewise the ratio $\frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})} = 0.79$ falls close to that implied by Column

[42] As in Experiment 1, we tested whether $\gamma(h) = \gamma(l)$. Testing the first session only, we fail to reject the null $H_0 : \gamma(h) = \gamma(l); \chi^2(1) = 0.13, p = 0.722$. Aggregating data across both sessions, we fail to reject the null $H_0 : \gamma(h) = \gamma(l); \chi^2(1) \approx 0, p = 0.946$.

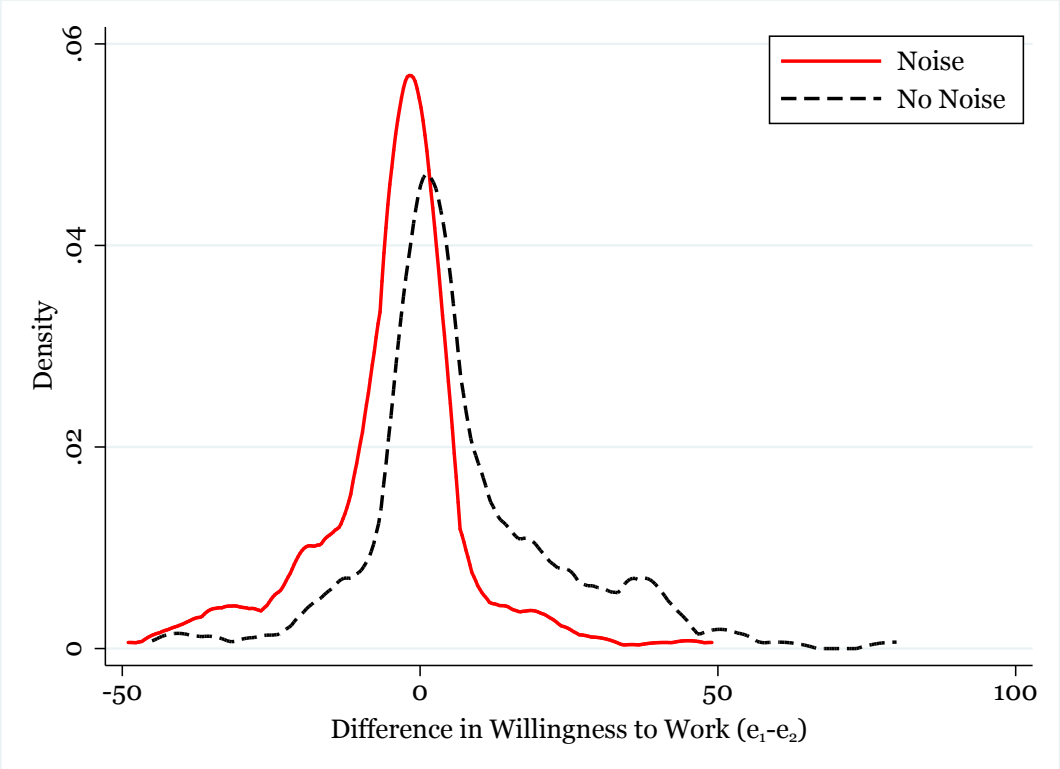

Figure 6: *Kernel density of the difference in willingness-to-work between the first and second sessions, separated by task faced.* Each underlying observation from this figure is the change in a participant's willingness to work for a fixed payment between Sessions 1 and 2 of the experiment. The black curve represents participants who were assigned to the no-noise task; the red curve represents participants who were assigned to the noisy task.

(1): $\frac{\hat{\theta}(\text{no noise}|\text{coin flip})}{\hat{\theta}(\text{no noise}|\text{control})} = 0.84$. In both studies and across all specifications, we find that uncertain assignment via coin flip distorts willingness to work 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—presented in Appendix C—we demonstrate that participants who face the noise and no-noise tasks do not exhibit differential attrition rates. Additionally, that table demonstrates that attrition is independent of both mean willingness to work in Session 1 and whether a participant first faced Russian or Greek.

A potential concern in this setting is that the participants who completed additional tasks during the first session formed different beliefs than those who did not complete additional tasks. The

Table 5:
PARAMETRIC ANALYSIS, EXPERIMENT 2

	Dep. var: $\log\left(\frac{e_{t=1}}{e_{t=2}}\right)$	
	(1)	(2)
Estimated ratio $\frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})}$	1.288	1.451
	(0.124)	(0.252)
Estimated ratio $\frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})}$	0.786	0.883
	(0.091)	(0.113)
$H_0 : \frac{\hat{\theta}_1(\text{noise})}{\hat{\theta}_2(\text{noise})} \geq 1$	$\chi^2(1) = 5.43$	$\chi^2(1) = 3.19$
	$p = .010$	$p = .037$
$H_0 : \frac{\hat{\theta}_1(\text{no noise})}{\hat{\theta}_2(\text{no noise})} \leq 1$	$\chi^2(1) = 5.51$	$\chi^2(1) = 1.08$
	$p = .009$	$p = .149$
Observations	348	348
Clusters	70	70
Controls	No	Yes

Notes: Standard errors (in parentheses) are clustered at the individual level and derived via delta method. 12 observations are left-censored and 26 are right-censored in Columns (1)-(2). 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 $\gamma = 1.14$.

BDM mechanism induces randomness in whether a person will actually complete any additional tasks, and only one third of participants completed additional tasks in the first session. Comparing participants who completed additional tasks with those who did not complicates the analysis, as the two groups have accumulated different amounts of experience. Column (2) of Table 5 demonstrates that controlling for these additional tasks does not qualitatively change our main result. Furthermore, Supplemental Table A5 demonstrates that the non-parametric result is robust if we simply drop participants who completed extra tasks. While statistical power decreases when dropping participants, our overall estimates remain quite close.

5 Conclusion

In this paper we provide evidence 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 manipulate participants' expectations prior to their initial experiences. Consistent with our model, we observe systematic and persistent changes in subsequent willingness to work depending on subject's initial expectations, despite the fact that these initial beliefs are no longer relevant. We now briefly discuss some reasons for caution in interpreting our results as well as directions for future research.

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 willingness to work 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. Although the aggregate results in Experiment 2 do not demonstrate signs of loss aversion, it is possible that we are unable to see loss aversion because of an overall diminished willingness (among all participants) to work in the second session. Additionally, asymmetric distortion of bad outcomes (relative to good outcomes) may be difficult to observe in our paradigm due to compression of the response scales at low values. With low willingness to work, participants may utilize the response scale differently than those with higher willingness to work, which may make detecting loss aversion more difficult. Loosely, choices may be more finely tuned near the bottom of scale and hence less susceptible to big changes. Finally, as noted previously, loss aversion may act against our results in Experiment 2 and we may thus be unable to detect loss aversion in that paradigm. As loss aversion is central to our theoretical model and drives a number of predictions for long-run beliefs, future work should address the extent to which losses drive asymmetric belief updating.

Future experimental work could explore an additional theoretical prediction of the model: sequential contrast effects. Taking our Experiment 2 design as an example, our model predicts that

the disutility of effort on Session 2 is compared against the wrongly-encoded disutility from Session 1. For participants facing the noisy task, this should to an increase in willingness to work that may even "overshoot" the willingness to work of a participant who knew their task assignment all along. Contrastingly, for those facing the no-noise task, willingness to work may "undershoot" the willingness to work of a participant who knew their task assignment all along. Our results hint in this direction—e.g., in Table 4 the difference in willingness to work between noise and no noise on the second session is small and the sign does not match intuition. However, we would need more data to make such claims with statistical power.

Indeed our theoretical paper describes a number of further avenues for experimental work. For instance, our model predicts that as bad outcomes become less common, a misattributor will perceive those outcomes as worse. In contrast, as good outcomes become less common, a misattributor will perceive those outcomes as better. This basic comparative static has important implications for product evaluation and firm strategy. A straightforward test of this comparative static would involve manipulating prior expectations such that participants face a wide range of probabilities of facing the bad task. While the varied treatments in Experiment 1 provide a first look at the role of probabilistic assignment in subsequent evaluations, future research should explore this more completely.

Our results suggest that firms can shape employees' evaluations in the short run by managing expectations. For instance, consider a firm in which employees must complete a number of short-term tasks—some less desirable than others. Our results suggest that employees would form the most favorable impressions of an undesirable tasks if they knew well ahead of time that they would have to complete it, rather than facing uncertainty when forming impressions. This accords with evidence on firms that give realistic job previews prior to hiring. As Phillips (1998) shows, employees that face a realistic job preview are higher performing and less likely to leave their job than their peers who do not experience a job preview. Misattribution along the lines discussed in this paper may provide an underlying mechanism for this effect.

More broadly, we believe that this paper provides the first direct evidence of misattribution of reference dependence. Misattribution has been well-documented in psychology and nascent research in economics has explored some implications of other forms of misattribution in other domains. In our companion paper, we provide a portable, tractable model of a specific form of misattribution that has broad implications. Here, we provide direct evidence of this mistake.

References

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

ALLEN, E., P. DECHOW, D. POPE, AND G. WU (2016): "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science*, Forthcoming.

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.

BARTOV, E., D. GIVOLY, AND C. HAYN (2002): "The Rewards to Meeting or Beating Earnings Expectations." *Journal of Accounting and Economics,* 33(2), 173–204.

BAUMGARTNER, H., M. SUJAN, AND D. PADGETT (1997): "Patterns of Affective Reactions to Advertisements: The Integration of Moment-to-Moment Responses into Overall Judgments." *Journal of Marketing Research*, 34(2), 219–32.

BAYER, H.M. AND P.W. GLIMCHER (2005): "Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal." *Neuron*, 47(1), 129–41.

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

BENJAMIN, D., M. RABIN, AND C. RAYMOND (2016): "A Model of Non-Belief in the Law of Large Numbers." *Journal of the European Economic Association*, 14(2), 515–44.

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

BHARGAVA, S. (2007): "Perception is Relative: Contrast Effects in the Field." *Working Paper*.

BHARGAVA, S. AND R. FISMAN (2014): "Contrast Effects in Sequential Decisions: Evidence from Speed Dating." *Review of Economics and Statistics*, 96(3), 444–57.

2165–211.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2016): "Diagnostic Expectations and Credit Cycles." *Working Paper*.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2019): "Memory, Attention and Choice." *Working Paper*.

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.*

GAGNON-BARTSCH T. AND B. BUSHONG (2019): "Learning with Misattribution of Reference Dependence." *Working Paper.*

CAI, J. AND C. SONG (2017): "Do disaster experience and knowledge affect insurance take-up decisions?" *Development Economics*, 124, 83–94.

CAMERER C., A. DREBER, E. FORSELL, T. HO, J. HUBER, M. JOHANNESSON, M. KIRCHLER, J. ALMENBERG, A. ALTMEJD, T. CHAN, E. HEIKENSTEN, F. HOLZMEISTER, T. IMAI, S. ISAKSSON, G. NAVE, T. PFEIFFER, M. RAZEN, AND H. WU (2016): "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351(6280), 1433–6.

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.

CHEN, D., T. MOSKOWITZ, AND K. SHUE (2016): "Decision-making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires." *Quarterly Journal of Economics*, 131(3), 1181–1241.

CHEN, M.K., V. LAKSHMINARAYANAN, AND L.R. SANTOS (2006): "How Basic Are Behavioral Biases? Evidence from Capuchin Monkey Trading Behavior." *Journal of Political Economy*, 114(3), 517-47.

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.

COHN, B. (1999): "The Lay Theory of Happiness: Illusions and Biases in Judging Others." *Princeton University*, Undergraduate Dissertation.

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.

DE QUIDT, J., J. HAUSHOFER, AND C. ROTH (2018): "Measuring and Bounding Experimenter Demand." *American Economic Review*, 108(11), 3266–3302.

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.M.M. 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.

GENNAIOLI, N., Y. MA, AND A. SHLEIFER (2015): "Expectations and Investment." *NBER Macroeconomics Annual*, 30, 379–442.

GILBERT, D., 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.A. 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.

GREENWOOD, R. AND A. SHLEIFER (2014): "Expectations of Returns and Expected Returns." *Review of Financial Studies*, 2014; 27(3): 714–46.

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

HASELHUHN, M., D. POPE, AND M. SCHWEITZER (2012): "Size Matters (and so Does Experience): How Personal Experience with a Fine Influences Behavior." *Management Science*, 58(1), 35–51.

HARTZMARK, S. AND K. SHUE (2016): "A Tough Act to Follow: Contrast Effects in Financial Markets." *Working Paper*.

HAISLEY, E. AND G. LOEWENSTEIN (2011): "It's Not What You Get But When You Get It: The Effect of Gift Sequence on Deposit Balances and Customer Sentiment in a Commercial Bank." *Journal of Marketing Research*, 48(1), 103–15.

HAYDEN, B.Y., S.R. HEILBRONNER, J.M. PEARSON, AND M.L. PLATT (2011): "Surprise Signals in Anterior Cingulate Cortex: Neuronal Encoding of Unsigned Reward Prediction Errors Driving Adjustment in Behavior." *Journal of Neuroscience*, 31(11), 4178–87.

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

HILL, M.R., E.D. BOORMAN, AND I. FRIED (2016): "Observational Learning Computations in Neurons of the Human Anterior Cingulate Cortex." *Nature Communications*, 7(12722), 1–12.

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

HOGARTH, R. AND H. EINHORN (1992): "Order Effects in Belief Updating: The Belief-Adjustment Model." *Cognitive Psychology*, 24, 1–55.

HSEE, C. AND E. WEBER (1997): "A fundamental prediction error: Self-other discrepancies in risk preference." *Journal of Experimental Psychology, General*, 1997; 126, 45–53

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

KAHNEMAN, D. AND A. TVERSKY (1972): "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology*, 1972; 3(3), 430–454.

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.

KAUSTIA, M. AND S. KNÜPFER (2008): "Do Investors Overweight Personal Experience? Evidence from IPO Subscriptions." *Journal of Finance*, 63(6), 2679–702.

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

KŐSZEGI, B. AND M. RABIN (2009): "Reference-Dependent Consumption Plans." *American Economic Review*, 99(3), 909–36.

KUHNEN, C. (2015): "Asymmetric Learning from Financial Information." *Journal of Finance*, 70(5), 2029–62.

LEVIN, P. AND A. ISEN (1972): "Further Studies on the Effect of Feeling Good on Helping." *Sociometry* 38(1), 141-7.

LOEWENSTEIN, G. AND D. PRELEC (1993): "Preferences for Sequences of Outcomes." *Psychological Review*, 100(1), 91–108.

LOOMES, G. AND R. SUGDEN (1986): "Disappointment and Dynamic Consistency in Choice under Uncertainty." *Review of Economic Studies*, 53(2), 271–82.

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.

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.

RICHARDSON, S., S. TEOH, P. WYSOCKI (2004): "The Walk-down to Beatable Analyst Forecasts: The Role of Equity Issuance and Insider Trading Incentives." *Contemporary Accounting Research,* 21(4), 885–924.

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.

ROYCHOWDHURY, S. (2006) "Earnings Management Through Real Activities Manipulation." *Journal of Accounting and Economics,* 42, 335–370.

RUTLEDGE, R.B., N. SKANDALI, P. DAYAN, AND R.J. DOLAN (2014): "A Computational and Neural Model of Momentary Subjective Well-Being." *Proceedings of the National Academy of Sciences*, 111(33), 12252-7.

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

SCHULTZ W., P. DAYAN, P.R. MONTAGUE (1997): "A Neural Substrate of Prediction and Reward." *Science*, 275(5306), 1593-–9.

SIMONSONHN, 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.

SIMONSONHN, U. (2009): "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.*

SPRENGER, C. (2015): "An Endowment Effect for Risk: Experimental Tests of Stochastic Reference Points." *Journal of Political Economy*, 123(6), 1456–99.

TEOH, S., Y. YANG AND Y. ZHANG (2009): "The Earnings Numbers Game: Rewards to Walk Down and Penalties to Walk Up Of Analysts' Forecasts of Earnings." *Working Paper.*

TVERSKY, A. AND GRIFFIN, D. (1991): "On the Dynamics of Hedonic Experience: Endowment and Contrast in Judgments of Well-Being." In Strack, F., Argyle, M. and Schwartz, N. (Eds.), *Subjective Well-Being*, Pergamon Press, 101–18.

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

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

A Derivation of Optimal Effort in Experiment 1

In this appendix we show that, under reasonable assumptions, a rational participant with reference-dependent preferences will choose an effort level in Experiment 1 that is decreasing in her expected value of her cost parameter, $\theta_i(a)$. Recall that this effort level solves Equation 9 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

$$\begin{aligned} & \widehat{\mathbb{E}}_{i,1}\left[u_{i,2}|e_{i,2}\right] = \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right] + \eta\widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e\right]\right)\right] + m = 0 \\ \Rightarrow \quad & \widehat{\mathbb{E}}_{i,1}\left[u_{i,2} \mid e_{i,2}\right] = -\hat{\theta}_{i,1}(a)c(e_{i,2}) + \eta\widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right)\right] + m = 0. \end{aligned} \tag{14}$$

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

$$\begin{aligned} \widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right)\right] = -c(e_{i,2})\Bigg(& \widehat{F}\left(\hat{\theta}_{i,1}(a)\right)\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) \leq \hat{\theta}_{i,1}(a)\right] \\ & + \lambda\left[1 - \widehat{F}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]\Bigg) \end{aligned} \tag{15}$$

and thus

$$\begin{aligned} & \widehat{\mathbb{E}}_{i,1}\left[n\left(V_{i,2}^e \mid \hat{\theta}_{i,1}(a)c(e_{i,2})\right)\right] = \\ & \qquad -c(e_{i,2})(\lambda - 1)\left[1 - \widehat{F}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]. \end{aligned} \tag{16}$$

Plugging Equation 16 back into Equation 14 yields:

$$\begin{aligned} \widehat{\mathbb{E}}_{i,1}\left[u_{i,2}\middle|e_{i,2}\right] &= -\Bigg(\hat{\theta}_{i,1}(a) + \eta(\lambda - 1)\left[1 - \widehat{F}\left(\hat{\theta}_{i,1}(a)\right)\right]\widehat{\mathbb{E}}_{i,1}\left[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \mid X_{i,2}(a) > \hat{\theta}_{i,1}(a)\right]\Bigg)c(e_{i,2}) \\ &= -h\left(\hat{\theta}_{i,1}(a)\right)c(e_{i,2}), \end{aligned} \tag{17}$$

where

$$h\big(\hat{\theta}_{i,1}(a)\big) \equiv \hat{\theta}_{i,1}(a) + \eta(\lambda - 1)\big[1 - \widehat{F}\big(\hat{\theta}_{i,1}(a)\big)\big]\widehat{\mathbb{E}}_{i,1}\big[X_{i,2}(a) - \hat{\theta}_{i,1}(a) \;\big|\; X_{i,2}(a) > \hat{\theta}_{i,1}(a)\big]. \quad (18)$$

Recall that we have assumed that for any value of $\hat{\theta}_{i,1}$, the participant's posterior over θ corresponds to the random variable $\hat{\theta}_{i,1} + Z_{i,1}$ where $Z_{i,1}$ is symmetric and independent of $X_{i,1}$. That is, the person's updated expectation of θ depends on $x_{i,1}$ but the noise about this expectation is invariant of $x_{i,1}$. This is the case, for instance, when the participant's priors follow a normal distribution over θ and $\varepsilon_{i,t}$ are normally distributed. Under this assumption, the expectation term in Equation 18 is independent of $\hat{\theta}_{i,1}(a)$ and thus h is increasing in $\hat{\theta}_{i,1}(a)$. Hence, the participant will select e_i^* such that $h(\hat{\theta}_{i,1}(a))c(e_i^*) = m$, and therefore e_i^* is decreasing in $\hat{\theta}_{i,1}(a)$.

B Reference Points that Incorporate the BDM Mechanism

In this section we consider how our theoretical predictions of Experiment 1 extend when a participant's reference point incorporates the uncertainty introduced by the BDM mechanism.

Recall that participant i's desired effort, e_i^*, 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 $r^e(e_i^*;\hat{\theta}_{i,1}) \equiv G(e_i^*)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e | e < e_i^*]$ where $\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e | e < e_i^*] = \hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e) \cdot \int_0^{e_i^*} c(e)\frac{g(e)}{G(e_i^*)}de$. Thus, the values $r_i^m(e_i^*)$ and $r_i^e(e_i^*;\hat{\theta}_{i,1})$ serve as the participant's reference points along each dimension in period 2. As such, she chooses e_i^* to maximize

$$\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i^*] = G(e_i^*)\Big\{\widehat{\mathbb{E}}_{i,1}\big[V_{i,2}^e + \eta n\left(V_{i,2}^e \,\big|\, r^e(e_i^*;\hat{\theta}_{i,1})\right)\big|e < e_i^*\big] + m + \eta\left(m - r^m(e_i^*)\right)\Big\}$$
$$+ [1 - G(e_i^*)]\Big\{\eta\left(0 - r^e(e_i^*;\hat{\theta}_{i,1})\right) + \eta\lambda\left(0 - r^m(e_i^*)\right)\Big\}, \quad (19)$$

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 19 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 dimen-

sion, $r^e(e_i^*;\hat{\theta}_{i,1})$, since this expectation incorporates a chance of not working at all. Similarly, the monetary bonus comes as a gain relative to her expected monetary gain, $r^m(e_i^*)$, which incorporates a chance of 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 (which assumes that the BDM does not influence the person's reference point), the only way for the treatment probability p to influence e_i^* is through its affect on the participant's perception of θ_i. Thus, we will examine how e_i^* depends on this perception, $\hat{\theta}_{i,1}$. To simplify the analysis below, we assume the participant forms certain beliefs about θ 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 from Equation 19 reduces to

$$\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i^*] = G(e_i^*)\left(\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e \middle| e < e_i^*\right] + m\right) = \hat{\theta}_{i,1}(a|\hat{v}_{i,1}^e)\cdot\int_0^{e_i^*} c(e)g(e)de + G(e^*)m, \tag{20}$$

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

We now consider the case with reference dependence (i.e., $\eta > 0$). Beginning from the objective function in Equation 19, it is helpful to rewrite it as the sum of two components: the expected monetary benefit from statement e_i^*, which we denote by

$$B(e_i^*) \equiv G(e_i^*)\Big\{m + \eta\left(m - r^m(e_i^*)\right)\Big\} - \eta\lambda[1 - G(e_i^*)]r^m(e_i^*), \tag{21}$$

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

$$K(e_i^*;\hat{\theta}_{i,1}) \equiv -G(e_i^*)\widehat{\mathbb{E}}_{i,1}\left[V_{i,2}^e + \eta n\left(V_{i,2}^e \,\middle|\, r^e(e_i^*;\hat{\theta}_{i,1})\right)\middle| e < e_i^*\right] + \eta[1 - G(e_i^*)]r^e(e_i^*;\hat{\theta}_{i,1}). \tag{22}$$

Thus, the objective from Equation 19 reduces so that the person chooses e_i^* to maximize expected monetary benefit minus expected effort cost:

$$\widehat{\mathbb{E}}_{i,1}[u_{i,2}|e_i^*] = B(e_i^*) - K(e_i^*;\hat{\theta}_{i,1}). \tag{23}$$

Given the objective above, we now analyze when the maximizing value of e_i^* is a decreasing

function of $\hat{\theta}_{i,1}$. Let $L(e_i^*;\hat{\theta}_{i,1})$ denote the first derivative of the objective function:

$$L(e_i^*;\hat{\theta}_{i,1}) \equiv \frac{\partial B(e_i^*)}{\partial e_i^*} - \frac{\partial K(e_i^*;\hat{\theta}_{i,1})}{\partial e_i^*}, \tag{24}$$

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

$$\frac{\partial e_i^*}{\partial \hat{\theta}_{i,1}} = -\left(\frac{\partial L(e_i^*;\hat{\theta}_{i,1})}{\partial e_i^*}\right)^{-1} \frac{\partial L(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1}}. \tag{25}$$

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

$$\operatorname{sgn}\left(\frac{\partial e_i^*}{\partial \hat{\theta}_{i,1}}\right) = \operatorname{sgn}\left(\frac{\partial L(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1}}\right). \tag{26}$$

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

$$\frac{\partial L(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1}} = -\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1}\partial e_i^*}. \tag{27}$$

From 22 and the definition of $r^e(e_i^*;\hat{\theta}_{i,1})$ (along with our assumption of no uncertainty over θ), we have

$$\begin{aligned}
K(e_i^*;\hat{\theta}_{i,1}) =& -G(e_i^*)\left\{\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*] + \eta\lambda\left(\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*] - G(e_i^*)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*]\right)\right\} \\
&\qquad - \eta[1-G(e_i^*)]G(e_i^*)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*]. \\
=& -G(e_i^*)\left\{\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*] + \eta\lambda[1-G(e_i^*)]\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*]\right\} \\
&\qquad - \eta[1-G(e_i^*)]G(e_i^*)\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*]. \\
=& -\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*]G(e_i^*)\{1+\eta\lambda[1-G(e_i^*)] - \eta[1-G(e_i^*)]\} \\
=& -\widehat{\mathbb{E}}_{i,1}[V_{i,2}^e|e<e_i^*]G(e_i^*)\{1+\eta(\lambda-1)[1-G(e_i^*)]\}.
\end{aligned} \tag{28}$$

Note that $\widehat{\mathbb{E}}_{i,1}[v_{i,2}^e|e<e_i^*] = -\hat{\theta}_{i,1}\frac{1}{G(e_i^*)}\int_0^{e_i^*} c(e)g(e)de$. Since g is a uniform PDF, it is constant. We denote this constant by g, and thus $G(e) = ge$. (Given that our experiment uses $e \sim \text{Uniform}[0,100]$, g in this case is $\frac{1}{100}$.) Furthermore, let $\bar{c}(e_i^*) \equiv \int_0^{e_i^*} c(e)de$, so $\widehat{\mathbb{E}}_{i,1}[v_{i,2}^e|e<e_i^*] = -\hat{\theta}_{i,1}\frac{g}{G(e_i^*)}\bar{c}(e_i^*)$. From 28, we thus have

$$K(e_i^*;\hat{\theta}_{i,1}) = \hat{\theta}_{i,1}g\bar{c}(e_i^*)\{1+\Lambda[1-G(e_i^*)]\}, \tag{29}$$

where $\Lambda \equiv \eta(\lambda - 1)$. Similar simplification of $B(e_i^*)$ in Equation 21 yields

$$B(e_i^*) = mG(e_i^*)\left\{1 - \Lambda[1 - G(e_i^*)]\right\}. \tag{30}$$

From Equations 29 and 30, 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 as in Objective 20. Thus, without loss aversion, the optimal choice of e_i^* is same regardless of whether the agents has reference dependence preferences or not, and therefore e_i^* is clearly decreasing in $\hat{\theta}_{i,1}$.

We now consider cases with loss aversion, so $\Lambda > 0$. Together, Equations 26 and 27 imply that e_i^* is decreasing in $\hat{\theta}_{i,1}$ if $\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1} \partial e_i^*} > 0$. From 29, $\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1} \partial e_i^*} > 0$ iff

$$\begin{aligned} c(e_i^*)\left\{1 + \Lambda[1 - G(e_i^*)]\right\} - g\Lambda\bar{c}(e_i^*) > 0 \\ \Leftrightarrow \left\{1 + \Lambda[1 - G(e_i^*)]\right\} > g\Lambda\frac{\bar{c}(e_i^*)}{c(e_i^*)} \end{aligned} \tag{31}$$

Furthermore, using Equations 30 and 29, the SOC implies that

$$\begin{aligned} \frac{\partial B(e_i^*)}{\partial e_i^*} - \frac{\partial K(e_i^*;\hat{\theta}_{i,1})}{\partial e_i^*} < 0 &\Leftrightarrow 2mg\Lambda < \hat{\theta}\left[c'(e_i^*)\left\{1 + \Lambda[1 - G(e_i^*)\right\} - 2g\Lambda c(e_i^*)\right] \\ &\Leftrightarrow 0 < c'(e_i^*)\left\{1 + \Lambda[1 - G(e_i^*)]\right\} - 2g\Lambda c(e_i^*) \\ &\Leftrightarrow \left\{1 + \Lambda[1 - G(e_i^*)]\right\} > 2g\Lambda\frac{c(e_i^*)}{c'(e_i^*)}. \end{aligned} \tag{32}$$

Substituting inequality 32 into 31 establishes that $\frac{\partial^2 K(e_i^*;\hat{\theta}_{i,1})}{\partial \hat{\theta}_{i,1} \partial e_i^*} > 0$ if

$$\begin{aligned} 2\frac{c(e_i^*)}{c'(e_i^*)} &> \frac{\bar{c}(e_i^*)}{c(e_i^*)} \\ \Leftrightarrow 2c(e_i^*)^2 &> c'(e_i^*)\bar{c}(e_i^*). \end{aligned} \tag{33}$$

Condition 33 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 33 is equivalent to

$$2e^{2\gamma} > \frac{\gamma}{\gamma + 1}e^{2\gamma}. \tag{34}$$

We therefore have shown that under this cost structure (or any other that meets Condition 33), we have the optimal action, e_i^*, is a decreasing function of $\hat{\theta}_{i,1}$ when the participant's reference point

is expected value of the lottery induced by the BDM mechanism. Given that e_i^* is a decreasing function of $\hat{\theta}_{i,1}$, the predictions of Observations 1 through 3 carry over to this setting. Namely: p does not directly influence a participant's objective function. But under misattribution, e_i^* is an increasing function of p because $\hat{\theta}_{i,1}$ is a decreasing function of p.

C Supplemental Tables and Figures

In this Appendix, we provide additional results that supplement the main text and provide robustness checks for our results.

We first show that dividing the Experiment 1 sample in half according to the total amount of time the participant spent on the experiment (from the start of Session 1 to completion) does not drastically change our nonparametric results. This is demonstrated in Tables A1 and A2 below. However, this exercise is limited by unequal group sizes. Regression analysis (included in Table 3 in the main body) demonstrates that this effect does not alter the results of our parametric analysis.

We next include robustness analysis to changing the Stone-Geary background parameter (Table A3). Although our numerical estimates are not stable, our qualitative results hold for two alternative specifications of the background parameter.

Third, we utilize a simple logit model to explore whether any observables predict attrition in Experiment 1 (Table A4). We find no such observables across the three primary treatments.

We then turn to the second experiment. To address potential concerns about differential learning, we present simple, non-parametric results for Experiment 2 in which we have dropped any participants who completed extra tasks in the first session (Table A5). This analysis leaves far fewer participants in our sample, but our qualitative results hold.

Finally, following the robustness exercise in Experiment 1 concerning attrition, we present a logit model for Experiment 2 (Table A6). We did not collect demographic information from participants, and thus we have fewer potential explanatory variables. We analyze a few here without finding any convincing pattern of 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	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	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 in the main sample.

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

	Logit. Dependent variable: 𝟙(return)					
	Raw	AMEs	Raw	AMEs	Raw	AMEs
𝟙(Noise)	0.199	0.019	0.204	0.019	0.207	0.019
	(0.73)	(0.73)	(0.75)	(0.75)	(0.75)	(0.75)
𝟙(Coin Flip)			0.126	0.012	0.176	0.016
			(0.46)	(0.46)	(0.64)	(0.64)
Age					-0.001	-0.0001
					(-0.07)	(-0.07)
𝟙(Male)					0.514	0.047
					(1.77)	(1.76)
Constant	2.056***		1.992***		1.931**	
	(11.12)		(8.70)		(3.15)	
Observations	586	586	586	586	586	586

Notes: Standard errors in parentheses. Third regression includes fixed effects for income of respondent; no income variables significant.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5:
EXPERIMENT 2. DIFFERENCE IN WILLINGNESS TO WORK; NO EXTRA TASKS

	Dependent variable: e_1 - e_2			
	No Noise		Noise	
Constant	4.308	7.268*	-5.425***	-7.900**
	(2.754)	(3.846)	(1.653)	(4.025)
Fixed payment ($)		0.106		-0.258
		(0.148)		(0.170)
$\mathbb{1}$(Russian, Session 1)		-10.162*		7.433*
		(5.285)		(4.001)
Observations	120	120	120	120

Notes: Standard errors, clustered at individual level, in parentheses. All regressions include random effects at individual level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

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

	Logit, dependent variable: $\mathbb{1}$(return)			
	Raw	AMEs	Raw	AMEs
Avg WTW, Session 1	0.004	0.001	0.028	0.00389
	(0.015)	(0.002)	(0.025)	(1.17)
Avg WTW, Session 1*$\mathbb{1}$(Noise)	-0.022	-0.003	-0.063**	-0.009**
	(0.015)	(0.002)	(0.029)	(0.004)
$\mathbb{1}$(Noise)			1.802*	0.247*
			(0.984)	(0.130)
$\mathbb{1}$(Russian, Session 1)			0.289	0.040
			(0.606)	(0.083)
Constant	1.717***		0.586	
	(0.449)		(0.718)	
Observations	87	87	87	87

Note: Standard error in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

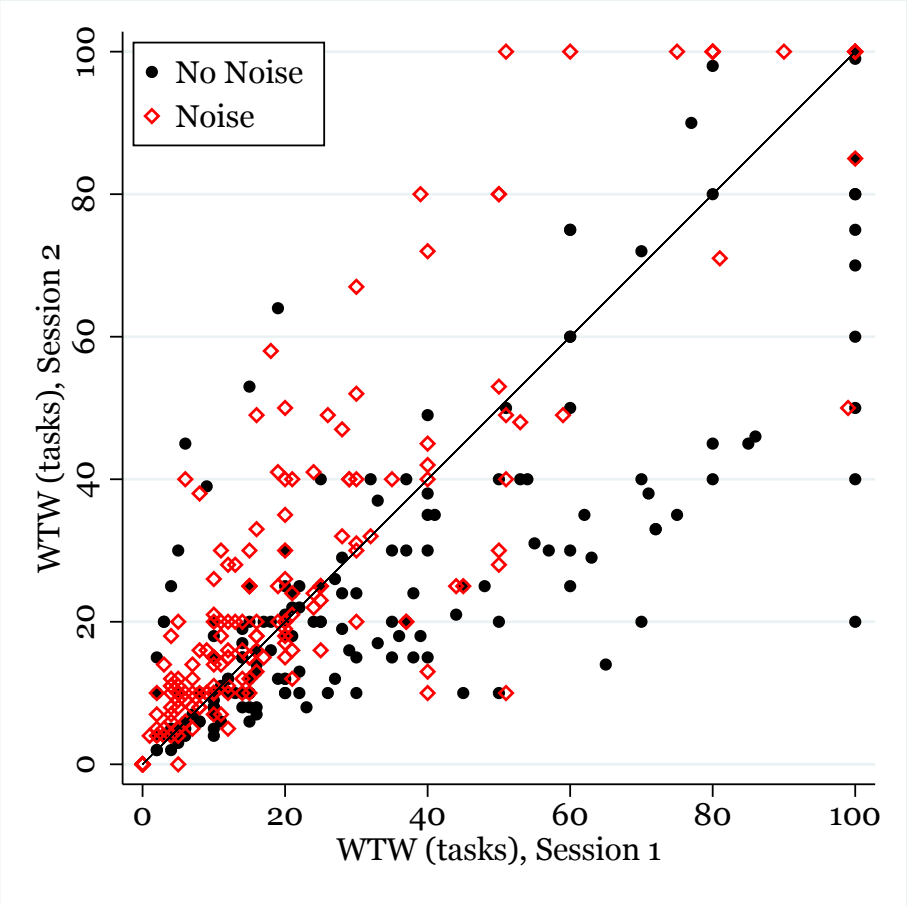

Figure A1: *Raw willingness-to-work data from Experiment 2.* Each observation in this figure represents a participant's willingness to work for a fixed payment in sessions one and two of the experiment. Black dots represent participants who faced the no-noise task; red diamonds represent participants who faced the noisy task.

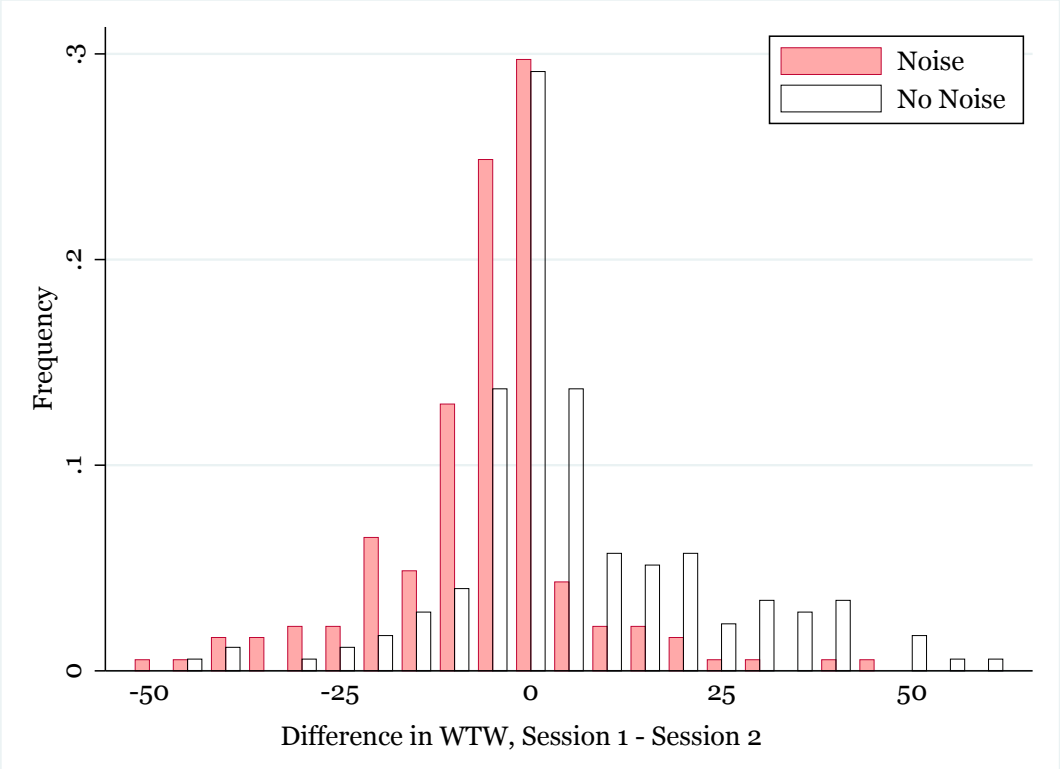

Figure A2: *Histogram of the difference in willingness-to-work between the first and second sessions in Experiment 2.* Each observation in this figure represents the change in a participant's willingness to work for a fixed payment between sessions one and two of the experiment. Clear bars represent participants who faced the no-noise task; solid red bars represent participants who faced the noisy task.

D Experimental Instructions

In this section, we provide the full text of experimental instructions. We use braces to denote alternative instructions corresponding to different treatments. All instructions commenced with an informed consent form.

D.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

D.2 Complete Experiment Instructions: Experiment 1

D.2.1 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.

D.2.2 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.

D.3 Experiment 1b Modified Lines

Experiment 1b used the same instructions as above, 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!]]

D.4 Full Experiment Instructions: Experiment 2

D.4.1 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 ran-

dom 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.

D.4.2 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.