Abstract
A growing body of evidence suggests that utility is reference dependent and outcomes are compared against recent expectations. When learning from experience, errors may arise if people neglect the degree to which their experienced utility was shaped by their expectations. In this paper, we use a pair of experiments to examine whether biased beliefs arise from a failure to retrospectively account for sensations of positive and negative surprise. Participants learn 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 working, while others knew in advance which task they would face. In a second session conducted hours later, we elicit participants’ willingness to work at that same task. Relative to participants who knew with certainty which task they would face, participants assigned to the less-onerous task by chance were more willing to work, while participants assigned to the more-onerous task by chance were less willing to work. 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 to the effort cost of their assigned task.

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

*Department of Economics and Business School, Harvard University. E-mails: bushong@fas.harvard.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 Caltech, Harvard, HBS, Norwegian School of Economics (NHH), the North American ESA Conference, Stanford, the Stanford Institute for Theoretical Economics, and UCSD Rady for comments. We thank Alexander Millner for sharing the noisy stimulus used in these experiments. We gratefully acknowledge 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 utility is reference dependent and depends on both the intrinsic value of an outcome 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). For instance, a diner at a new restaurant may greatly enjoy a meal when it beats expectations and may be disappointed with that same meal when she expected better. When predicting how much she will enjoy the restaurant in the future, a rational decision maker should therefore account for how those expectations shaped her experience. In doing so, a rational decision maker disentangles the intrinsic value of consumption from the sensation of surprise that it generated. When a person fails to separate these sources of utility, surprising outcomes may distort her beliefs. Guided by this intuition and research in psychology, we designed an experiment to explore whether decision makers incorrectly attribute sensations of surprise or disappointment to the intrinsic value of outcomes.¹

To further illustrate our notion of misattribution of reference dependence, consider a research assistant completing a series of short-term tasks for a professor. Suppose that each day the worker is assigned to one of two tasks, one more desirable than the other. If each morning the worker thinks she may fill either role, then the job she completes that day comes as either a positive or negative surprise. If she fails to account for these sensations as she forms her impressions of the job, she will develop incorrect beliefs about how much she enjoys it. Concretely, when she is assigned to the desirable job—say, field research—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 assigned to the less desirable task—say, data entry—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 distorted beliefs because she neglects the degree to which her expectations influenced her experienced utility.

In this paper, we present evidence from real-effort experiments consistent with the idea that people misattribute sensations of positive and negative surprise to the underlying disutility of effort of an unfamiliar task. In an online experiment (Experiment 1), we assign participants to complete one of two previously-unexperienced tasks. Our evidence suggests that participants incorrectly learn the effort cost associated with their task as a function of the exogenously imposed expectations.

¹ Research in both psychology and economics demonstrates that memories are imprecise and people may make mistakes when attributing the sources of their feelings. Dutton and Aron (1974) demonstrate that judgments of a newly-met person depend on unrelated factors (e.g., current state of excitement or fear). Meston and Frohlich (2003) replicate this seminal result in different settings. Recent evidence in economics (Simonsohn 2007, 2010; Haggag and Pope 2016) 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 2.1.
they held before first encountering that task. We show that our results cannot be explained by classical models or by a model of reference dependence without misattribution (Kőszegi and Rabin 2006), and we demonstrate how our model of reference dependence with misattribution (Bushong and Gagnon-Bartsch 2016) predicts the observed behavior. In a laboratory experiment (Experiment 2), we manipulate beliefs within subjects and examine how willingness to work changes over the course of a week as participants’ expectations change. We again find that initial expectations shape participants’ willingness to work in a manner predicted by our model of misattribution. Although we contribute to the growing body of evidence on expectations-based reference points, our main contribution is therefore the identification of a specific, previously unstudied form of misattribution. As we show in our theoretical companion paper, Bushong and Gagnon-Bartsch (2016), this form of misattribution generates a number of known biases in belief updating. For instance, a misattributor may rely 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 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.\(^2\)

In Section 2, we present an abridged version of the model from our companion paper. Following Bell (1985) and Kőszegi and Rabin (2006), 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. A “misattributor” correctly recalls how she felt after each experience, but neglects the extent to which her utility was affected by reference dependence. She thus misremembers the exact outcome that generated her experienced utility. Specifically, she encodes a wrong outcome that would have generated her experienced utility if she weighted gains and losses by a diminished factor. When the true outcome \(v\) exceeds her expectations, the misattributor encodes a perceived outcome greater than \(v\). In contrast, when \(v\) falls short of expectations, the misattributor encodes a perceived outcome less than \(v\). In both

\(^2\) Evidence demonstrates 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 (Bhangava 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). Importantly, our mechanism is unlikely to be the primary explanation for the subset of these examples that do not have utility consequences (e.g., baseball umpires likely do not care whether each pitch is a ball or a strike). Our model applies in situations in which a person directly cares about the stochastic outcomes she faces.
cases, the misattributor forms biased beliefs by updating based on wrongly encoded outcomes.

Our experiments share a common setup: participants learn about a previously unexperienced real-effort task and, after this initial learning phase, report their willingness to do additional work for payment. Our identification of misattribution comes from manipulating participants’ expectations held prior to the initial learning phase and observing subsequent willingness to work. In Experiment 1, we observe participants’ willingness to work approximately eight hours after their initial impressions were formed. In Experiment 2, we observe each participant’s willingness to work twice: once immediately after initial impressions were formed and once a week later after expectations had changed. In both experiments, we contrast the willingness to work of participants who knew in advance which task they would complete with those who formed initial impressions immediately after they flipped a coin to determine their task assignment.

Section 3 describes our online experiment, which we conducted on Amazon’s Mechanical Turk (MTurk). Participants ($n = 586$) were recruited to a two-session experiment. In an initial learning session, subjects listened to audio recordings of Amazon book reviews and determined whether each review was either endorsing or criticizing the book. This simple and tedious classification task came in two variants. The first, which we call noise, had an annoying sound layered on top of the audio review. The less onerous variant, which we call no noise, had no additional sound. Our primary experimental manipulation came from varying subjects’ beliefs over which task they would complete in the experiment. Half of participants were assigned to a task from the onset of the experimental instructions. The other half of participants flipped a coin to determine which task they would face, and did so moments before their first experience with that task. Thus, half of participants were endowed with a stochastic reference point consisting of a chance of the noisy task and a chance of the no-noise task, while the other half held a deterministic reference point. Put together, this $2 \times 2$ design generates four groups: $\{\text{coin flip, no coin flip}\} \times \{\text{noise, no noise}\}$. After reading the initial instructions (and, if applicable, flipping the coin), participants completed eight classifications allowing them to learn about their assigned task. Eight hours later, we elicited their willingness to work at that same task for additional pay. Specifically, we asked participants the maximum number of tasks they were willing to complete for each of five possible fixed payments (and we chose one at random to count for real). We examine how willingness to work differed between groups, where willingness to work operates as a revealed-preference measure of beliefs over effort costs.

Our model predicts that participants who face the no-noise task 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 gain. Thus, we predicted that participants in the coin flip + no noise group will be more willing to work than those in the no coin flip + no noise group. By contrast, those stochastically assigned the noisy task will form overly negative impressions of the task, as their initial beliefs
Figure 1: Preview of Results. Labor supply curves by group assignment—each point represents the average willingness to work for a fixed payment, as elicited by a BDM mechanism. Relative to groups whose assignment did not induce surprise, those assigned by chance (treatment 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.

will be influenced by their disappointment. Therefore, we predicted that participants in the *coin flip + noise* group will be less willing to work than those in the *no coin flip + noise* group.

Matching the predictions above, participants who faced the task without noise by chance were more willing to work than those who faced that task with certainty (*p* = .031, random-effects model with standard errors clustered at the individual level). Figure 1 plots the labor supply curves for each of the four experimental groups in Experiment 1.³ The effect described above is clearly seen in the outward shift of the labor supply curve for the *coin flip + no noise* group as compared to the *no coin flip + no noise* group. 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 *no coin flip + noise* group (*p* = .026, random effects model with standard errors clustered at the individual level).

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. Additionally, we discuss how our design rules out a few intuitive explanations. Specifically the time gap between participants forming their impressions and our elicitation of willingness to work helps distinguish our effect from that of “transient moods”—e.g., that the coin flip induced short-term mood effects distinct from our mechanism.⁴ Additionally, this gap gives

---

³ For ease of visual presentation, we omit error bars but note the significance of these results in the text.

⁴ For instance, transient factors have 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
time for reference points to adapt.\footnote{In a laboratory experiment, Song (2016) finds that reference points over monetary outcomes adapt within approximately ten minutes. Participants in our experiment completed the second session—where we elicit willingness to work—at least eight hours after forming their initial impressions. The average participant completed the second session a little more than ten hours after the first session. We discuss the role of reference point adaptation in greater detail in Sections 3.2 and 4, and demonstrate that—even with sluggish reference points—existing theories of reference-dependent preferences do not predict our experimental results.}

In Section 3.2, we contrast the predictions of our model with a classical model and a reference-dependent model without misattribution. Fixing the task a participant faced, neither alternative predicts differential willingness to work between participants assigned by coin flip versus certain assignment. Furthermore, modifications and extensions of reference-dependent models without misattribution fail to generate our effects. Specifically, our results are inconsistent with a reference-dependent model without misattribution even if we assume that reference points do not adapt over the course of the experiment. However, we faced a remaining confound: participants who were assigned their task with certainty only knew about that one task.\footnote{The knowledge of an alternative task could have distorted behavior through a number of mechanisms. For instance, participants could have formed opinions over the experimenter and exhibited some form of reciprocity as a result of the outcome of the coin flip. Alternatively, the knowledge of an alternate task could lead to widened priors through some form of (plausibly rational) inference. Our robustness treatment rules out simple forms of both alternative explanations.} To equalize participants’ information sets prior to task assignment, we ran an additional treatment, described in Section 4.1. We followed the same instructions as the coin-flip version of Experiment 1, but replaced the coin flip with a 99\% treatment: 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 99\% + noise and 99\% + 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.

In Section 5, we describe our laboratory experiment, which measures each participant’s willingness to work during two different sessions, separated by one week. Participants in the first session were assigned to a task via coin flip and returned one week later to face that same task a second time. Our identification of misattribution 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 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. 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 pleasantly surprised in the first session exhibited declining willingness to work, while those who were negatively surprised in the first session exhibited increasing willingness to work. This result is consistent with our model: for instance, a participant who received a positive surprise (and faced the good task) may have attributed this sensation to the quality of the task and thus underestimated its cost during the first session. Upon trying that same task again a week later—when it was no longer surprising—the task may have failed to live up to the previous experience, and thus willingness to work decreased.

Our model of misattribution of reference dependence sheds light on a number of important domains. A basic implication of the model is that a person’s beliefs may overreact to surprising personal experiences. A number of recent studies highlight how beliefs arising from personal experience tend to be over-reactive, including work on IPO subscription (Kaustia and Knüpf 2008), risk taking and stock-market participation (Malmendier and Nagel 2011), insurance take-up (Cai and Song 2017), and compliance to deadlines (Haselhuhn et al. 2012). Taking the example from Malmendier and Nagel (2011), our model predicts that individuals who personally suffer surprising losses in the stock market are more likely to exit the market or invest in low-risk assets because they overreact to this experience and form overly-pessimistic views of stock returns. Misattribution further implies a form of overreaction and over-extrapolation based on recent events. This accords with a range of empirical findings, including Greenwood and Shleifer (2014) and Gennaioli, Ma, and Shleifer (2015) who find that investors and managers predictions of their future earnings are overly-volatile and exhibit forecast errors that negatively relate with past performance. More directly, the results of our experiment speak to the literature on realistic job previews (see Phillips 1998 for a review). A pervasive finding is that employees who face a realistic job preview perform better and are less likely to leave their job than their peers who do not have a preview that calibrates their expectations. Our model and the evidence presented here may shed light on the underlying mechanism that generates this effect and the importance of expectations management.

2 A Model of Misattribution of Reference-Dependent Utility

We now present an abridged version of our model of misattribution of reference dependence (Bushong and Gagnon-Bartsch 2016), which guides both our experimental design and main predictions. While we apply the model to our experimental setting in Section 3.2, this section provides an abstract overview of the model to help understand our experimental design, and a discussion concerning the model’s main assumptions.
Following Kőszegi and Rabin (2006) (henceforth KR), we assume the total utility from a choice consists of two parts: “consumption utility”, which corresponds to the traditional outcome-based notion of payoffs, and “gain-loss utility”, which corresponds to the difference between the realized outcome and expectations. Loosely, the decision maker’s utility from outcome \( v \in \mathbb{R} \) when expecting \( r \in \mathbb{R} \) is

\[
 u(v|r) = x + \eta n(v|r),
\]

where her gain-loss utility \( n(v|r) \) is proportional to the difference between \( v \) and \( r \) and parameter \( \eta > 0 \) measures the impact of gains and losses on utility.

In our experiment, we explore reference-dependent utility along two dimensions: money and effort. The decision maker’s utility thus takes the following form:

\[
 u(v|r) = \sum_{k \in \{\text{money, effort}\}} \left\{ \frac{v_k}{\text{Consumption utility}} + \frac{\eta n(v_k|r_k)}{\text{Gain-loss utility}} \right\},
\]

(1)

We now define the reference point more precisely. If the decision-maker believes that \( v_k \) is drawn from distribution \( F_k \), we assume that her reference point is given by that full distribution. As KR note, when a person expects a gamble between $0 or $100, receiving $50 feels like a gain relative to $0 and a loss relative to $100. This mixed sensation of gains and losses is captured by this “stochastic reference point” whereby each outcome \( v_k \) is compared against each hypothetical outcome along that dimension, and each comparison is weighted by the probability of the hypothetical outcome:

\[
 n(v_k|F_k) \equiv \int_{\tilde{v}_k < v_k} (v_k - \tilde{v}_k) dF_k(\tilde{v}_k) + \lambda \int_{\tilde{v}_k \geq v_k} (v_k - \tilde{v}_k) dF_k(\tilde{v}_k),
\]

(2)

where parameter \( \lambda \geq 1 \) captures loss aversion—losses are weighted more than equally-sized gains.

Turning back to our experiment, if a participant with well-calibrated priors knew her task with certainty, her reference distribution would collapse to the effort cost of that specific task, and thus her gain-loss utility would be zero. In contrast, when task assignment comes as the result of chance, the person’s reference point contains a chance of both the good and bad tasks. Thus, when assigned the good task, the person’s realized effort cost is less than expected, and she experiences positive gain-loss utility. Conversely, when assigned the bad task, she experiences negative gain-loss utility.\(^7\)

To introduce misattribution, we consider a person learning about the potential distribution of outcomes she may face. In such situations, we assume that outcomes are drawn from a c.d.f. \( F \) that depends on an unknown parameter \( \theta \), and the decision maker uses past experience to update
her beliefs about θ. Here, we diverge from previous models and consider a decision maker who neglects the extent to which her past experiences were influenced by reference dependence. The decision maker infers from past utility as if she weighted the gain-loss component by a diminished factor $\hat{\eta} \in [0, \eta)$; that is, as if her utility function were $\hat{u}(v|r) = v + \hat{\eta} n(v|r)$. Although she correctly encodes how happy she felt following some episode, she misinterprets the underlying source of her utility. Specifically, she thinks she experienced an outcome with consumption value $\hat{v}$ such that $\hat{u}(\hat{v}|r) = u(v|r)$. We assume that the misattributor uses the wrongly encoded values $\hat{v}$ to learn about the unknown parameter, but we impose no mistake in updating on top of this wrong encoding—if the person is Bayesian, then she updates her beliefs according to Bayes’ Rule using $\hat{v}$ instead of $v$.

Returning to the example from the introduction, consider a worker whose daily task is random: some days she faces easy tasks and other days she faces more onerous tasks. When the worker is assigned a difficult task, she faces both the bad outcome and a sensation of loss—her task is worse than expected. If she fails to account for this sensation of disappointment and instead attributes it to the underlying disutility of the task, she will wrongly believe her assigned task was more onerous than it really was. That is, the decision maker ultimately misinterprets her outcomes because she incorrectly attributes gain-loss utility to her consumption utility. We explore a scenario closely matching this example in our experiments.

2.1 Discussion and Motivating Evidence

In this section, we discuss evidence supporting the primary assumptions of our model: (1) that utility is reference dependent, (2) that reference points are shaped by expectations, and (3) that people incorrectly attribute sources of utility or incorrectly interpret the cause of outcomes. We also briefly discuss the interpretation of the model. Importantly, to our knowledge, there are no direct studies of the role of reference dependence in learning from experience.

Evidence of reference-dependent utility has been prominent in economics and psychology. Early work by Kahneman and Tversky (1979) proposed that changes in wealth relative to some reference point lead to sensations of gains and losses, and the authors supported this intuition with evidence from simple choices. More recently, studies have demonstrated forms of reference dependence.

---

8 Throughout the paper, we present intuitions that convey our mistake as one in which a person improperly encodes outcomes as they happen. However, in many situations—particularly those in which people can directly observe the objective outcomes they faced—our mistake may operate through a bias in memory. We discuss these interpretations at greater length in our companion paper.

9 Somewhat relatedly, Cohn (1999)—cited in Kahneman and Tversky (2000)—demonstrates that people exaggerate the duration of unhappiness following a bad life event. This could stem from people failing to anticipate a changing reference point—a phenomenon that generates similar predictions to ours in settings in which a person’s beliefs are not directly measured. With reference-point projection, behavior may be consistent with exaggerated perceptions of previous outcomes that deviated from expectations. However, if beliefs were directly measured, a model of reference-point projection would not predict that outcomes are misremembered or mis-encoded (as our model predicts).
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).

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). Recent papers, however, 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.

A recent literature in economics has examined expectations-based reference points—a notion posited by KR and assumed throughout this paper when examining the predictions of that model. 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 a clear demonstration that choice is 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),

---

10 Formal models have taken different stances on the hedonic nature of gains and losses. Although Kahneman and Tversky’s (1979) “Prospect Theory” supposes that people behave as if they experience reference-dependent hedonics, the authors do not take a strong stand on whether this behavior truly reflects hedonic sensations. Bell (1985) and Loomes and Sugden (1982) both posit that regret and rejoice (relative to a reference point) are real hedonic sensations. Recent authors—particularly Koszegi and Rabin (2009)—have endorsed the hedonic interpretation of sensations of gain and loss.

11 Kőszegi and Rabin (2009) interpret all such monetary experiments as evidence that changes in beliefs, even absent consumption, can generate the sensation of gains and losses. Our model does not include this, and extending the model to include belief-based gain-loss utility is an avenue for future research.

12 Reinforcement-learning models predict a signal that matches that of the gain-loss function (absent loss aversion). Thus, previous neuroscience evidence on reinforcement learning, when reinterpreted through this lens, may provide some evidence on reference-dependence.
Sprenger demonstrates that participants choose risky options more often when expecting a risky lottery rather than a sure payoff. 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. In a similar experiment, Wenner (2015) finds no evidence for the KR model, but attributes his results to non-equilibrium behavior. Ericson and Fuster (2009) demonstrate that the endowment effect is at least partially driven by expectations of future endowments, but Heffetz and List (2014) provide contradictory evidence. Although the evidence remains incomplete, on the whole it suggests that expectations play a key role in reference dependence and this paper adds to the growing body of evidence.

In our first experiment, we manipulate beliefs via exogenously imposed uncertainty. While not necessary, our model makes the cleanest predictions assuming expectations adjust between sessions—approximately eight hours. A recent literature addresses how quickly reference points adjust. Song (2016) demonstrates that reference points largely adjust over the course of ten minutes. Indirectly, Smith (2012) and Buffat and Senn (2015) both provide evidence of relatively quick reference-point changes in laboratory settings with small stakes. We share Song’s (2016) interpretation of the existing evidence: for small-stakes laboratory experiments, reference points seem to adjust on the scale of minutes.

In contrast to reference-dependent preferences, misattribution has received little attention in the economics literature. Haggag and Pope (2016) provide evidence of misattribution of state-dependent utility from both experimental and field data. First, the authors show that experimental participants value an unfamiliar beverage more if they first drink it while thirsty rather than sated. Likewise, their field evidence shows 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 two ways: first, misattributors in their model tend to underestimate the utility difference between two outcomes, while our model predicts (and we observe) exaggerated differences. Furthermore, unlike mistakes driven by misattribution of reference dependence, biased forecasts in Haggag and

---

13 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.

14 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.
Pope’s formulation may wash out with experience in some settings.\textsuperscript{15} This distinction stems from the fact that Haggag and Pope rule out complementaries where past experiences influence todays consumption utility. Reference dependence clearly introduces this complementarity, as past experiences form the reference point against which todays consumption is evaluated. In two papers, 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 admissions 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. This form of misattribution is often referred to as the “fundamental attribution error” or “correspondence bias” in psychology (e.g., Ross 1977; Gilbert and Malone 1995). Although these latter studies do not involve misattribution of utility, these errors may share a common psychology with that of misattribution of reference-dependent utility: transient sensations (e.g., sensations of gain or loss) are incorrectly attributed to an underlying, stable source.

3 Experiment 1: Design and Predictions

This section introduces our between-subject experiment.\textsuperscript{16} Our identification of misattribution stems from manipulating participants’ expectations prior to their first experience completing an unfamiliar task. More than eight hours after this initial learning session, we elicited their willingness to work at that same task. By eliciting willingness to work, we indirectly measured the beliefs about the effort cost (or underlying disutility of effort) that participants formed based on their initial experiences with their assigned task. We begin by describing the specifics of our experimental design. We then discuss the theoretical predictions of (1) a classical model, (2) a model of reference dependence without misattribution, and (3) our model of misattribution.

\textsuperscript{15} For example, when states are independent from consumption experiences, mislearning of mean outcomes caused by misattribution of state-dependent utility will vanish in the long-run. For example, consider a diner learning about the quality of a restaurant. If she returns to the restaurant when she’s both hungry and not hungry, she will correctly learn the average quality. As such, the framework from Haggag and Pope (2016) may best apply to situations where choices are based on limited experience.

\textsuperscript{16} Throughout this paper, we refer to our two experiments in reverse order of their implementation. This allows us to outline the theoretical predictions of our model in the between-subject version of the experiment, which is more straightforward.
3.1 Design

Our first experiment was conducted on Amazon’s Mechanical Turk (MTurk). We recruited approximately 600 workers to complete a two-session experiment between August 1 and August 7, 2016. The first session was an initial learning phase, designed to give participants experience with a novel real-effort task. During the second session, we elicited participants’ willingness to work at that task for additional pay. Participants were required to be U.S. residents and must have completed at least 100 prior tasks with 95% approval rating. Our data comes from 586 people who completed the first session—522 people completed both sessions.\textsuperscript{17} It took participants an average of 10 minutes to complete the first session and 15 minutes to complete the second. We paid participants $4 for successfully completing both sessions, which translates to an hourly wage of approximately $9.60.

Each participant worked on one of two tasks, which we now describe in detail. Both tasks involved listening to audio reviews of books and classify the review as either positive or negative. In order to make a classification, subjects pressed one of two buttons and were given a warning if their classification was incorrect. Figure 2 depicts the interface participants used to classify reviews. We generated the audio files using digital-voice software. This voice “read” aloud book reviews collected from Amazon.com. 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 noise of a fork scraping against a record and a high-frequency tone.\textsuperscript{18} 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.\textsuperscript{19} Finally, in both versions of the task, the book reviews were edited to last approximately 20 seconds, to remove any specific references to author names or book titles, and for grammar. See the Appendix for sample text of the reviews.

Participants were required to complete 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, as they would later have an opportunity to complete additional work at that same task for extra pay. In order to examine the effect of expectations on subsequent evaluations, we randomly assigned participants into a treatment group ($n = 294$) and a control group ($n = 292$). Participants in the treatment group were assigned a task by chance, while

\textsuperscript{17} We did not reach six hundred workers because fourteen would-be participants seemed to have agreed to work on the HIT, but did nothing thereafter. It is possible that some dropped out after reading the informed consent form, but we do not have any evidence speaking to the cause of these drop-outs.

\textsuperscript{18} The Nock Lab at Harvard generated this noise and used the stimuli in forthcoming work unrelated to our own. In their studies, this sound was played at modest volume (slightly louder than we played the noise). Participants found the sound unpleasant, but there were no lasting effects (e.g., ringing ears).

\textsuperscript{19} 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.
participants in the control group faced one of two tasks with certainty. We call these groups no coin flip and coin flip, respectively, and discuss the implementation of each below.

Those participants in the no coin flip group were randomly assigned into two subgroups: noise and no noise. Participants in the no coin flip + noise group were told that they would hear audio reviews with an annoying noise played over the top. Participants in the no coin flip + no noise group were never told of the existence of the noise and simply completed classifications without the overlaid audio. Likewise, participants in the no coin flip + noise group were never aware of the possibility of facing the task without the added noise. Participants in each group were given one “sample task”—matching their assignment—to teach them how to use the interface. Participants then completed the eight mandatory classifications.

In contrast, participants in the coin flip group 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 small sample of the audio (8s in duration). After the sample, each participant “flipped” a digital coin to determine which task she would ultimately face. Participants then completed the eight mandatory classifications according to the result of the coin flip (either noise or no noise).

After completing the eight mandatory classifications, the first session concluded and we instructed participants that we would email them a link to the second session in eight hours. In the second session, participants were reminded of their task assignment (noise or no noise) and given

---

Footnote 20: 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 a link to the second session on receipt of their email. We suspect that others may have faced the same issue (due to emails getting caught in spam filters, or participants using an old or incorrectly-entered email address), leading to slightly higher attrition than desirable. Nevertheless, more than 90% of participants returned for a second session.
the option to complete additional trials (of the same task) for supplemental pay. We asked subjects their willingness to continue working for five different payment values: \( \{ \$0.50, \$1.00, \$1.50, \$2.00, \$2.50 \} \). Participants responded by using a slider to select any integer \( e \in [0, 100] \), which we call “willingness to work”. In order to incentivize participants to truthfully reveal their willingness to work, we utilized the Becker-Degroot-Marshak (BDM) mechanism. The mechanism operated as follows: for each possible payment \( y \in \{ \$0.50, \$1.00, \$1.50, \$2.00, \$2.50 \} \), we asked participants how many tasks they would complete in order to receive that amount of money. We then drew a random integer \( z \) between 0 and 100 from a uniform distribution. If \( z \leq e \), the participant completed \( z \) additional tasks and received \( $y \). If \( z > e \), the participant received no additional money and completed no additional tasks.\(^{21}\)

Our overlaid-audio design has an important feature: participants who faced the annoying noise had to listen to it in order to complete the task. We ensured participants listened to the audio reviews using two techniques. First, all participants were told that they could not get more than two out of eight classifications wrong or else they would be removed from the study without pay. Additionally, we hid the response buttons for ten seconds, which required participants to listen to longer portions of the reviews. Finally, many of the reviews featured important details in the late part of the review.\(^{22}\)

To keep participants from reloading the session in attempt to avoid the noise, we blocked multiple logins and required unique email authentication to access each session of the experiment.

### 3.2 Theoretical Predictions

In this section, we derive predictions for how willingness to work depends on a participant’s expectations over their potential task assignment. We focus on three models of individual preferences and belief updating: a classical model, a reference-dependent model without misattribution, and our misattribution model. We apply all three models to the following environment: participant \( j \) chooses an effort level \( e_j(i) \in \mathbb{R}_+ \) given that she faces task \( i \in \{ h, l \} \). Task \( h \) is the noisy, high-effort-cost task and task \( l \) is the noiseless, low-effort-cost task. We observe \( e_j(i) \) through willingness to work under the BDM mechanism, and it is our key variable of interest.

When facing task \( i \), exerting effort comes at cost \( -\theta_j(i)c(e_j(i)) + \varepsilon_j(i) \), where we assume \( c(\cdot) \) is an increasing function, \( \varepsilon_j(i) \) is a mean-zero random shock, and \( \theta_j(i) \) is a cost parameter that

\(^{21}\) This design yielded fewer censored observations than that of Augenblick, Niederle and Sprenger (2015). Because we offer a fixed payment rather than a piece-rate wage, almost all participants respond with a positive willingness to work. Additionally, since our highest payment ($2.50) is not particularly high, we observe only a minor amount of censoring at the high end of the scale.

\(^{22}\) By withholding the response buttons, we may have helped participants answer correctly. In our data, patient responders tended to be more accurate. Overall, there were very few mistakes: only two participants were dropped for inaccurate classifications.
depends on the task. The parameters $\theta_j(i)$ and the shocks $\varepsilon_j(i)$ are unknown to the participant, while we assume $c(\cdot)$ is known. We assume that $\theta_j(h) > \theta_j(l) > 0$ for all $j$, and that participants are aware of this relationship. Participants who do not face a coin flip face no uncertainty about their assigned task, and only face uncertainty about the difficulty of that task. Participants in the coin-flip group additionally believe they will face the noisy task with probability $\frac{1}{2}$ and the noiseless task with probability $\frac{1}{2}$.

For each model, we assume participants update their beliefs about $\theta(i)$ using their experienced utility as a signal of the underlying effort cost. Their signal comes from completing the eight mandatory tasks in the initial learning session, and we assume participants update in the direction of that signal. Bushong and Gagnon-Bartsch (2016) differs from the other two models by assuming that people update their beliefs using a wrongly-encoded signal, but otherwise follow the same updating as rational learners. We let $\hat{F}_{j,0}(\theta_j(i))$ denote participant $j$’s prior subjective probability distribution over $\theta_j(i)$, and we let $\hat{\theta}_{j,0}(i)$ denote her expected value of $\theta_j(i)$ given this belief. We assume that these prior beliefs are independent across participants and independent of a participant’s assignment to the control or treatment group. Finally, we let $\hat{\theta}_j(i)$ denote participant $j$’s updated belief about $\theta_j(i)$ following the initial learning session.

Recall that the choice variable $e_j(i)$ is the stated maximal number of tasks that participant $j$ is willing to complete for a fixed monetary payment. Let $m$ denote the participant’s utility from this payment, and we assume $m$ is known to the participant. Given the Becker-Degroot-Marchashak (BDM) mechanism we utilize for incentive-compatible reporting, the mechanism creates uncertainty over the number of tasks a person will complete. Thus, let $G_j(e)$ denote participant $j$’s (subjective) c.d.f. over the effort she will exert as a result of the BDM, which has support on $[0, 100]$, and let $g_j(e)$ be the associated p.d.f.

### 3.2.1 Classical Preferences and Behavior

First, we consider a classical model with unknown effort cost. Given the setup described above, we can write the expected utility of the rational agent facing task $i$ as:

$$\mathbb{E}[u_j^i(e)] = G_j(e) \cdot m - \hat{\theta}_j(i) \int_0^e c(\tilde{\varepsilon})dG_j(\tilde{\varepsilon}).$$

---

23 Although we assume a common cost function $c(\cdot)$, we could include additional between-subject variability in cost functions or the value of money (discussed below). Our results would extend in expectation if $c(\cdot)$ represented the population average cost.

24 This mild updating condition subsumes Bayesian updating for a number of commonly assumed distributions for $\theta_j(i)c(e) + \varepsilon_i(i)$. See Chambers and Healy (2012) for a complete characterization.

25 The distribution used in the mechanism is not given to participants. We suspect that for certain incorrect beliefs about the distribution in the mechanism, reference-dependent preferences could lead to substantial misreporting. However, this concern is independent from our experimental treatment. Thus, we leave a fuller discussion of this issue to future research.
Hence, the optimal willingness to work for task $i$ solves $\hat{\theta}_j(i)c(e^*_j(i)) = m$. Since we are giving a fixed payment rather than a marginal payment, the utility-maximizing decision sets total cost equal to total payment.

In a slight abuse of notation, denote by $e^*_j(i|p)$ the optimal willingness to work for task $i$ when the participant believed there was a $p$ chance of facing task $i$. Since expectations over task assignment do not affect learning in the classical model, willingness to work does not depend on whether the participant formed her initial impressions after the coin-flip or after knowing her task assignment with certainty.

Observation 1. Rational learners demonstrate willingness to work that is independent from the treatment: $e^*_j(i|p = \frac{1}{2}) = e^*_j(i|p = 1)$ for both $i \in \{h, l\}$.

Although the result above is stated at the individual level, we will observe results in the aggregate, as our design is between-subjects.

The observation above relies on prior beliefs that are independent of treatment. If, for instance, people who faced the coin-flip held more extreme prior beliefs as a result of the instructions in that treatment, this result may no longer hold. We account for a variant of treatment-dependent priors in our robustness experiment.

3.2.2 Reference-Dependent Preferences and Behavior

We now consider a reference-dependent model (KR) and demonstrate how the uncertainty over task assignment that we induce can influence willingness to work. As we discuss below, if the reference point adjusts between Sessions 1 and 2 of the experiment, reference dependence does not affect behavior differentially by treatment and thus leads to the same predictions as the classical model above. However, if the reference point is sluggish (i.e., slow to adapt), then willingness to work may be different across treatments. In this case, the coin flip induces reference-dependent participants to decrease the difference in their willingness to work across the good and bad tasks. As we show in the next section, our model of reference dependence with misattribution predicts the opposite: since sensations of gains and losses are incorrectly attributed to effort costs, the coin flip induces exaggerated differences in willingness to work across the good and bad tasks.

We specifically utilize the reference-dependent model introduced in Section 2, Equation 1. Translating our experiment to that framework yields a two-dimensional utility function with $v_{\text{money}} = m$ and $v_{\text{effort}} = -\theta_j(i)c(e_j(i)) + \epsilon_j(i)$. Perhaps the most important assumption for a reference-dependent model in this setting is a specification of a participant’s reference point entering the second session. If participants’ reference points adapt over the course of eight hours, then their willingness to work should be independent of whether they faced the coin flip or not. When reference points adapt, uncertainty that induced gain-loss utility initially is no longer present at time
of choice. Assuming that reference points adapt to the realized task over eight hours accords with existing empirical work on reference-point adaptation. In order to show we are robust to slowly-adapting reference points and to highlight that existing models of reference-dependent utility fail predict our observed behavior, we explore the implications of a sluggish reference point that—more than eight hours later—still incorporates the uncertainty stemming from the coin flip.

Following KR, we consider reference points that are formed by the planned effort and expected earnings for each realization of the coin flip. In that model, the person weights each possible contingency by the probability of it occurring: thus, the person’s reference point is a 50-50 mix of her plan for the noisy task and her plan for the no-noise task. Recall that a participant’s estimated cost parameter for each task is \( \hat{\theta}_j = (\hat{\theta}_j(h), \hat{\theta}_j(l)) \). Given these beliefs, participants form plans over effort levels in each contingency, which we denote by \( \tilde{\epsilon}_j = (\tilde{\epsilon}_j(h), \tilde{\epsilon}_j(l)) \). Thus her her expected disutility of effort in contingency \( i \in \{h, l\} \) is \( r_{j,\text{effort}}(i) = -\hat{\theta}_j(i) \mathbb{E}[G(\tilde{\epsilon}_j(i)) \cdot c(\tilde{\epsilon}_j(i))] \), where each entry corresponds to her plan conditional on facing the noise or no-noise task, respectively. Importantly, a participant’s plans also induce a reference point over expected earnings: \( r_{j,\text{money}}(i) = \mathbb{E}[G_j(\tilde{\epsilon}_j(i))] \cdot m \). Because each contingency happens with a 50% chance, the model assumes that a participant assesses her actual effort cost and expected earnings relative to a 50% chance of \( (r_{j,\text{effort}}(h), r_{j,\text{money}}(h)) \) and a 50% chance of \( (r_{j,\text{effort}}(l), r_{j,\text{money}}(l)) \).

To close the model, we must specify a person’s plans. Roughly put, the person chooses a plan that she would follow through with given the reference point implied by that plan. More formally, denote by \( \mathbb{E}[u(e_j(i)|\tilde{\epsilon}_j, \hat{\theta}_j)] \) the expected reference-dependent utility of taking action \( e(i) \) when facing task \( i \) given parameter estimate \( \hat{\theta}_j \) and plan \( \tilde{\epsilon}_j \).

**Definition 1.** A plan \( \tilde{\epsilon} \in \mathbb{R}_+ \) is a consistent plan if for each \( i \in \{h, l\} \), \( \mathbb{E}[u(\tilde{\epsilon}(i)|\tilde{\epsilon}, \hat{\theta})] \geq \mathbb{E}[u(e(i)|\tilde{\epsilon}, \hat{\theta})] \) for all \( e(i) \in \mathbb{R}_+ \).

There are often many consistent plans, but all of them share a common feature in this setting: relative to a certain reference point, all consistent plans decrease the difference in willingness to work across tasks when the reference point includes uncertainty.

**Observation 2.** Suppose participants have reference-dependent preferences as described in Equations 1 and 2. Under all consistent plans and for all \( j \):

1. If the reference point adapts between Sessions 1 and 2 of the experiment, \( e_j^*(l|p = \frac{1}{2}) = e_j^*(l|p = 1) \) and \( e_j^*(h|p = \frac{1}{2}) = e_j^*(h|p = 1) \).

---

26 In this specification, we do not have a fully stochastic reference point that incorporates the uncertainty over the parameters \( \theta_j(i) \) or the uncertainty induced by the mechanism. We collapse uncertainty in both cases to the expected value. Uncertainty in either case complicates the analysis, but does not qualitatively change the predictions presented in this section.
2. If the reference point is sluggish and $\lambda \geq 1$, either $e_j^*(l|p = \frac{1}{2}) = e_j^*(h|p = \frac{1}{2})$ or $e_j^*(l|p = \frac{1}{2}) \leq e_j^*(l|p = 1)$ and $e_j^*(h|p = \frac{1}{2}) = e_j^*(h|p = 1)$.

Intuitively, reference-dependent preferences induce a tendency to smooth willingness to work across the two tasks. The reference-dependent worker forms plans which diminish effort in the case of the good outcome and raise effort in the case of the bad outcome. By “compressing” the gap between plans in the good and bad states, the decision-maker faces smaller potential losses in each state.

These results show that the KR model fails to predict our observed behavior, even if we consider reference points that are slow to adapt. As noted, reference points likely adapt on a sufficiently short time scale (discussed above), and this analysis highlights that this assumption is not necessary for identification. For simplicity, we assume that reference points fully adapt in the subsection that follows. We return to issues pertaining to sluggish reference points in the discussion of our theoretical results.

### 3.2.3 Misattribution of Reference Dependence and Behavior

We now turn to our misattribution model introduced in Section 2. We assume that a participant infers her effort costs by using her experienced utility from the initial session as a signal. This utility is described in Equation 1. A rational agent $j$ would “subtract off” the $\eta \left( v_j^{effort}(i) \right)$ term and use the resulting value as a signal of $\theta(i)$. Contrastingly, a misattributor fails to properly account for these feelings of gains and losses. We describe how this distorts learning below.

For sake of example, consider a person having just completed the “learning phase” and who faced the task without noise. Suppose for simplicity that $\epsilon_j = 0$, and that the participant held correct priors: $\hat{\theta}_{j,0}(l) = \theta_j(l)$. Imagine further that this participant was in the coin-flip condition: she thought there was a chance she may face the noisy task. During the initial learning phase—in which she completes eight mandatory tasks—she experiences utility

$$-\theta_j(l)c(8) + .5 \eta \left( -\theta_j(l)c(8) \right) - \hat{\theta}_{j,0}(l) c(8) + .5 \eta \left( -\theta_j(l)c(8) \right) - \hat{\theta}_{j,0}(h) c(8).$$

Normalizing $c(8) = 1$, the gain-loss utility function in Equation 2 implies that the above expression reduces the above to $-\theta_j(l) - .5 \eta \left( \hat{\theta}_{j,0}(h) - \theta_j(l) \right)$.

A rational learner should subtract off these reference effects—the term $.5 \eta \left( \hat{\theta}_{j,0}(h) - \theta_j(l) \right)$ in the equation above—to extract a signal of $\theta_j(l)$. But a misattributor neglects of surprise on her utility, and therefore infers a biased signal of $\theta_j(l)$. If, for instance, $\hat{\eta} = 0$, the misattributor records a signal $\theta_j(l) - .5 \eta \left( \hat{\theta}_{j,0}(h) - \theta_j(l) \right)$. Since we assume the misattributor updates in the direction of her signal, she infers that the noiseless task is less onerous than it really is.

The converse is true for a misattributor in the coin-flip treatment who faces the noisy task $i =
Similar to the logic above, if $\hat{\theta}_{j,0}(h) = \theta_j(h)$, then the misattributor earns utility $-\theta_j(h) - .5\eta\lambda(\theta_{j,0}(l) - \hat{\theta}_j(h))$. Again, if $\hat{\eta} = 0$, then the misattributor wrongly uses this experienced utility as a signal of $\theta(h)$. Since she updates in the direction of her signal, she infers that the noisy task is more onerous than it really is.

However, a participant in the no-coin-flip condition faces no uncertainty about her assigned task. In that case, a misattributor only experiences sensations of gain and loss relative to her prior beliefs about that task. Even if the population has biased priors, misattribution leads to a systematic bias. This yields our main empirical hypothesis, which stands in opposition to that of the classical model and the reference-dependent model.

**Observation 3.** Let $\hat{\eta} < \eta$. If the reference point adapts between Sessions 1 and 2 of the experiment, then:

1. If priors are correct, $e^*_j(l|p = \frac{1}{2}) > e^*_j(l|p = 1)$ and $e^*_j(h|p = \frac{1}{2}) < e^*_j(h|p = 1)$

2. For all prior beliefs where priors are independent of treatment and $\hat{\theta}_{j,0}(l) < \hat{\theta}_{j,0}(h), E[e^*_j(l|p = \frac{1}{2})] > E[e^*_j(l|p = 1)]$ and $E[e^*_j(h|p = \frac{1}{2})] < E[e^*_j(h|p = 1)]$, where the expectation represents the average willingness to work of participants in each of the four groups.

Point 1 of Observation 3 is easily seen from the preceding examples and discussion. Put simply: those assigned to the no-noise task by chance will feel a sensation of positive surprise that they incorrectly attribute to the effort cost of the task. This leads the misattributor to incorrectly perceive the no-noise task as easier than it really is, which increases her willingness to work relative to those who face that task with certainty. Contrastingly, those assigned to the noisy task will incorrectly attribute their sensations of negative surprise to the task and form pessimistic views of the effort cost. This will diminish willingness to work relative to those who face that task with certainty. We discuss the latter point of Observation 3, which highlights robustness to wrong prior beliefs, with more detail in the Appendix.

### 3.2.4 Discussion of Theoretical Results

The previous sections demonstrated how uncertainty over task assignment may lead misattributors to wrong beliefs over the underlying disutility of effort. In this section, we briefly provide additional justification for some of our model’s assumptions and discuss a potential confound.

As noted above, we do not require well-calibrated prior beliefs in any of the observations. Even if aggregate beliefs are biased, the results presented above still hold as long as the priors are independent from treatment. Intuitively, fixing the task a participant faced, both the classical model and the reference dependent model lead to the same (mean) posterior beliefs if priors are independent...
of treatment. In contrast, misattribution creates an interaction between wrong beliefs and the treatment. So long as those miscalibrated beliefs are reasonable—namely, a person believes the noisy task is worse than the task without noise—misattribution still yields the prediction from Observation 3. Assuming reasonable priors seems warranted given participants received a representative sample of the task at the start of the experiment.

While we showed above that a reference-dependent model without misattribution leads to a different prediction than our model, when reference points are sluggish, two concerns arise. First, the KR predictions—that oppose those of our model—rely on the two dimensions of utility: money and effort. If participants have sluggish reference points and reference-dependent utility over only effort and not money, then the two models make similar predictions. Furthermore, the idea that reference points adapt over eight hours is consistent with the previous literature. As highlighted in the motivating evidence, Song (2016) demonstrates that reference points incorporate new information over the course of approximately ten minutes—especially for experiments with small stakes. Furthermore, we empirically explore this concern in the analysis below by comparing results between those who completed the second session quickly and later responders. As a second concern, if the reference point is sluggish, the “contracting” force described in Observation 2 will push against misattribution. Which force prevails depends on the specifics of the learning model that is assumed. We explore a more complete model of misattribution in the Appendix and provide some guidance to when misattribution will be overcome by reference dependence with a sluggish reference point. As we show there, a necessary condition for misattribution to overcome reference-dependence is that the difference in $\theta_j(h) - \theta_j(l)$ is sufficiently greater than $\lambda - 1$.

A final concern is that participants in the coin-flip condition may draw some inference about the experimental environment as a result of being exposed to both the no-noise and noise task. This violates our assumption that priors are independent of treatment. Suppose, for instance, that those who are exposed to both tasks infer that the two tasks must be significantly different in terms of disutility of effort. This plausibly rational inference could explain why willingness to work is more exaggerated across tasks for those facing the coin flip. Importantly, this explanation also requires that participants did not fully update their priors as a result of the eight mandatory tasks. A closely related concern is that participants in the coin-flip condition may express reciprocity to the experimenter as a result of their stochastic assignment. We address both of these concerns in a robustness treatment (Experiment 1b).

4 Experiment 1: Results

Our experimental design generates four groups: treatment status (i.e. whether participants faced a coin flip) crossed by eventual task assignment (i.e. noise or no noise). Of the 586 participants that
successfully completed the first session, 139 faced no coin flip and no noise; 153 faced no coin flip and noise; 152 faced a coin flip that resulted in no noise; and 142 faced a coin flip that resulted in noise. Table 1 shows the demographic characteristics and the probability of returning to the second session for participants in each of the four groups.

Table 1: 
SUMMARY STATISTICS, EXPERIMENT 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control noise=0</th>
<th>Control noise=1</th>
<th>Treatment noise=0</th>
<th>Treatment noise=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(Male)</td>
<td>.468 (.501)</td>
<td>.464 (.500)</td>
<td>.428 (.496)</td>
<td>.387 (.489)</td>
</tr>
<tr>
<td>Income</td>
<td>2.712 (1.009)</td>
<td>2.582 (1.092)</td>
<td>2.901 (1.066)</td>
<td>2.613 (1.103)</td>
</tr>
<tr>
<td>I(Return)</td>
<td>.921 (.271)</td>
<td>.882 (.323)</td>
<td>.862 (.346)</td>
<td>.944 (.231)</td>
</tr>
</tbody>
</table>

Observations 139 153 152 142

Note: Income is coded as a discrete variable with values 1-5, which correspond to the following:
(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 formed our primary dataset following a few cleaning procedures. We first removed participants who did not answer all five questions, which dropped three participants.\(^{27}\) We next removed participants who responded that their willingness to work was 100 for all tasks, which does not allow us to estimate responsiveness to payment and which we interpret as a form of confusion or lack of attention.\(^{28}\) This restriction dropped an additional six participants—two from \textit{coin flip + no noise}, three from \textit{no coin flip + no noise}, one from \textit{noise + coin flip}. Finally, for the purposes of our main results, we omit people who did not return for the second session (and whose willingness to work we therefore do not measure). We are left with a sample of 519 participants.

Our main hypothesis is that the mean willingness to work—fixing the task that participants

\(^{27}\) The first restriction was the result of a faulty Qualtrics restriction that should have forced all participants to answer all questions, but did not function properly on some browsers.

\(^{28}\) Additionally, a participant who responded 100 to the question with stakes of $0.50 is indicating an extremely low hourly wage rate, which we think more likely reflects the concerns above rather than a true preference. We present supplemental results on the raw dataset in the Appendix. Our point estimates of effects do not change much, though our power goes down because the extreme data inflates the standard errors.
ultimately faced—depends on participants’ expectations entering the initial learning session. In order to investigate this, we averaged participants’ willingness to work over all five fixed payments for each of the four groups described above.

**Aggregate Result, Experiment 1:** Averaging over the five fixed payments, participants who faced the task without noise were willing to do 28.59 tasks when their initial impressions were formed after the resolution of the coin flip. Participants who faced the task without noise and did not face the coin flip were willing to do 24.03 tasks ($p = .032$ for difference; random-effects model and standard errors clustered at individual level). In contrast, participants who faced the task with noise are willing to do 17.64 tasks when their initial impressions were formed after the resolution of the coin flip. Participants who faced the task with noise and did not face the coin flip were willing to do 22.29 tasks ($p = .025$ for difference; random-effects model and standard errors clustered at individual level).

We find a strong effect of prior expectations on willingness to work. This effect is consistent with misattribution of reference-dependent utility to the underlying effort cost of the assigned task. Although this result gives a general idea of the magnitudes, we further disaggregate the payment levels to better demonstrate the effect. Figure 3 shows the average labor supply curves of each of the four groups. That is, we plot the average willingness to work (as elicited from the BDM mechanism) for each of the five payments \{\$0.50, \$1.00, \$1.50, \$2.00, \$2.50\} for each of the four groups. At all payment levels, we find that those whose initial impressions of the noisy task were formed when it came as a bad surprise were less willing to work than those who faced the noisy task for certain. In contrast, those whose initial impressions of the no-noise task were formed when it came as a good surprise were more willing to work than those who faced the no-noise task for certain.

While the top panel of Figure 3 suggests a small difference between willingness to work at the no-noise and noise tasks when both were assigned with certainty, the second figure demonstrates that these average curves can be somewhat misleading. The bottom panel shows the cumulative distributions of willingness to work in each of the four groups, aggregating over all payment levels (and smoothed using the Epanechnikov kernel). This figure highlights that participants were more willing to work at the no-noise task versus the noisy task when assigned with certainty before

---

29 While we did not preregister our analyses, our experimental design and primary hypotheses follow directly from our model, which was developed *ex ante* in our companion paper. We highlight any instances in which we ran *ex post* empirical tests.

30 Since we offer a fixed payment for work (rather than a marginal wage), willingness to work reflects total costs. Thus averaging willingness to work over participants provides the average labor-supply function.
Figure 3: a) Labor supply curves and b) cumulative bid distribution by group assignment. Cumulative distribution curves are smoothed using the Epanechnikov kernel.
their initial learning. Figure 2b also highlights the universality of our main result: the cumulative distribution of willingness to work in the coin flip + no-noise group almost first-order stochastically dominates that of no coin flip + no noise. By contrast, the cumulative distribution of willingness to work in the no coin flip + noise group first-order stochastically dominates that of coin flip + noise group.

Table 2: Experiment 1. Effect of Coin Flip, No Noise

<table>
<thead>
<tr>
<th></th>
<th>Random-Effects Tobit. Dep Variable: WTW</th>
<th>Main Sample</th>
<th>Main Sample</th>
<th>Monotonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(Coin flip)</td>
<td>4.611**</td>
<td>4.703**</td>
<td>4.857**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.144)</td>
<td>(2.159)</td>
<td>(2.372)</td>
<td></td>
</tr>
<tr>
<td>Fixed payment ($)</td>
<td>13.703***</td>
<td>13.703***</td>
<td>14.360***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.303)</td>
<td>(0.303)</td>
<td>(0.326)</td>
<td></td>
</tr>
<tr>
<td>Session 1 duration (min)</td>
<td>0.009</td>
<td>0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.093)</td>
<td>(0.099)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.155</td>
<td>-0.165</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.095)</td>
<td>(0.102)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1(Male)</td>
<td>-0.899</td>
<td>-0.904</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.184)</td>
<td>(2.402)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>3.523**</td>
<td>9.827**</td>
<td>9.164*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.597)</td>
<td>(4.967)</td>
<td>(5.364)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors (in parentheses) are clustered at the individual level and calculated from observed information matrix. 23 observations are censored in both full and monotonic sample. Income fixed effects are dummies for each of the five income groups in demographic questionnaire.

*p < 0.10, **p < 0.05, ***p < 0.01

31 We use a Kolmogorov-Smirnov equality-of-distributions test. While this fails to account for redundancy in the data stemming from multiple observations from each individual, we calculated 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 noise and no-noise; the five p-levels were .024, .189, .041, .019, .090.
### Table 3:
**EXPERIMENT 1. EFFECT OF COIN FLIP, NOISE**

<table>
<thead>
<tr>
<th></th>
<th>Random-Effects Tobit. Dep Variable: WTW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main Sample</td>
</tr>
<tr>
<td>(Coin flip)</td>
<td>-4.657**</td>
</tr>
<tr>
<td></td>
<td>(2.085)</td>
</tr>
<tr>
<td>Fixed payment ($)</td>
<td>10.196***</td>
</tr>
<tr>
<td></td>
<td>(0.283)</td>
</tr>
<tr>
<td>Session 1 duration (min)</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>(0.132)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.016</td>
</tr>
<tr>
<td></td>
<td>(0.088)</td>
</tr>
<tr>
<td>(Male)</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>(2.181)</td>
</tr>
<tr>
<td>Constant</td>
<td>7.029***</td>
</tr>
<tr>
<td></td>
<td>(1.534)</td>
</tr>
<tr>
<td>Observations</td>
<td>1330</td>
</tr>
<tr>
<td>Clusters</td>
<td>266</td>
</tr>
<tr>
<td>Income FE</td>
<td>No</td>
</tr>
</tbody>
</table>

**Notes:** Standard errors (in parentheses) are clustered at the individual level and calculated from observed information matrix. 13 observations are censored in both full and monotonic sample. Income fixed effects are dummies for each of the five income groups in demographic questionnaire.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tables 2 and 3 present our main results in regression form, and demonstrates the statistical significance of our treatment and controls for a proxy for the task difficulty and individual characteristics. Our preferred specification is a Tobit regression with random effects at the individual level, since observations are censored at a maximum willingness to work of 100 tasks and we take five observations for each person. Our tables split the data by those assigned to no noise (Table 2) versus noise (Table 3), akin to comparing the solid black line to the dashed black line (no-noise) and the solid red line to the dashed red line (noise) from Figure 3. For simplicity and given the aggregate results from Figure 3, we control for the fixed payment without an interaction term, which constrains the elasticity of supply to depend solely on the task assignment and not treatment. The first column of both tables considers a baseline specification that regresses willingness to work on
treatment, controlling only for the payment. The effect of the coin flip on subsequent willingness to work is significant for no noise ($p = .029$) and noise ($p = .026$) when clustering standard errors at the individual level. The second column of each table controls for demographic characteristics (age, gender, and income) and for the length of the first session, which we view as a rough proxy for subjective task difficulty. Finally, the third column removes any participants whose responses were not weakly monotonic. That is, we dropped any participant who did not demonstrate increasing willingness to work across all five fixed payments. This drops a significant portion of the sample, but the point estimates of our effect remain very similar. As a final form of robustness for these results, we drop all censored observations and replicate the analysis in Tables 2 and 3 with a baseline random-effects model rather than the Tobit model presented above. We present this analysis in Appendix C (Tables 6 and 7).

Our results demonstrate economically-meaningful magnitudes. As a calibrational exercise, consider a firm seeking to hire workers to complete 25 of our classification tasks. Workers who faced no uncertainty when they formed their initial impressions required $1.70 payment to complete 25 noisy tasks and $1.50 to complete 25 no-noise tasks. This difference is significantly exaggerated by sensations of positive and negative surprise when forming initial impressions: workers with contaminated initial impressions required $2.30 to complete 25 noisy tasks and $1.20 to complete 25 no-noise tasks. Thus average costs decreased by 20% for no noise and increased by 35% for noise, and the payment gap—or the difference in payments required to do the task with noise over the task without noise—increased from $0.20 to $1.10.

A natural concern is that some form of differential attrition may drive our results. Table 1, which presents summary statistics for each of the four groups, suggests that attrition is fair similar across the four groups. In Table 9, we test if attrition is predictable given the observable data (task assignment, treatment, and so on). None of the demographic variables predict attrition. Additionally, group assignment is not predictive of attrition either on its own or when controlling for demographics and first-session length. However, a more subtle version of attrition is possible, given the MTurk setting: some participants 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.

Importantly, our design allows us to rule out (plausibly rational) explanations based on poorly calibrated priors. Participants in both the coin flip and no-coin-flip conditions had the same amount

---

32 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 91 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. These mistakes were spread amongst 80 participants, and the average mistake was small.
of exposure to their assigned task. Therefore, if priors are independent of the treatment, participants should have similar posterior beliefs after completing the initial eight reviews and should thus exhibit similar willingness to work. However, it is possible that those who were exposed only to one task—which was the case with the no-coin-flip condition—help different prior beliefs than those who faced the coin flip. Relatedly, exposure to both tasks may have formed a version of a “contrast effect” whereby the perceived difference in tasks was exaggerated. This concern led us to run an additional robustness treatment which addresses both concerns.

4.1 Experiment 1b: High-Probability Treatment

In order to address the concern raised above, we designed a treatment in which participants were assigned to the noise task with either probability \( p = .99 \) or \( p = .01 \). Below we briefly describe this additional treatment and compare the results to our previous treatment.

Participants \((n = 300)\) were recruited using the same MTurk HIT information as in Experiment 1. All procedures were the same as in Experiment 1 with one exception: rather than using a digital coin flip to determine the task assignment, participants either faced a very high or very low chance of the noisy task. Half of participants were assigned to a \( p = .99 \) treatment and the other half were assigned to a \( p = .01 \) treatment. For each participant, we drew an integer at random between 1 and 100. Participants in the \( p = .99 \) arm completed the task without noise if that random number was 100; otherwise, they faced the task with noise. Participants in the \( p = .01 \) treatment completed the task with noise if the random number was 100; otherwise, they faced the task without noise. In a slight abuse of terminology, we refer to these two treatments collectively as “high probability” conditions, as this term reflects the high degree of certainty participants had about which task they would face.\(^{33}\) Otherwise, all procedures and tasks were identical to Experiment 1.

If our original result is driven by the fact that only participants in the treatment groups were exposed to both tasks, then comparing our previous coin-flip condition to this new high-probability condition should address the concern. Our specific null hypothesis is that, fixing the task faced, there is no difference between the willingness to work among participants in the high-probability condition versus the coin-flip condition. Since participants in all treatments have been exposed to both tasks, any remaining difference in willingness to work can be attributed to the different probabilities that participants faced prior to forming their initial impressions.

Figure 4 plots the mean willingness to work in both the coin-flip and high-probability groups.\(^{34}\)

\(^{33}\) We were aware in advance that we would not have statistical power for those participants assigned to the unlikely outcome, and we therefore decided to ignore their data. Three participants were thus omitted from the analysis.

\(^{34}\) As with Experiment 1, we remove participants from our analysis sample if their willingness to work was 100 for all payment levels. This removed three participants: two who faced the task without noise and one who faced the task with noise. As with Experiment 1, we ignore those who fail to come back to the second session except in our analysis of attrition. This left us with \( n = 287 \) participant in the high-probability conditions.
Figure 4: Labor supply curves by group assignment in Experiment 1b, presented together with curves from Coin Flip groups in Experiment 1.

As before, we find that those who faced the no-noise task were more willing to work when their initial impressions were formed after resolving a large amount of uncertainty (coin flip) versus a small amount (high probability). In contrast, those who faced the noisy task are less willing to work when their initial impressions were formed after resolving a large amount of uncertainty versus a smaller amount.

Our main analyses in this experiment exactly mirror that of Experiment 1, replacing the no-coin-flip groups with their appropriate analogues from the high-probability treatment. We first consider aggregate results, along the lines of Result 1. Participants who faced the no-noise task demonstrate significantly higher willingness to work (4.40 tasks) in the coin-flip condition versus the high-probability condition ($p = .033$, random-effects model with standard errors clustered at the individual level). In contrast, participants who faced the noisy task were less willing to work (3.05 fewer tasks), but this result is not statistically significant ($p = .103$, random-effects model with standard errors clustered at the individual level). In Table 4 we control for additional variables as in Experiment 1. We find that workers facing the no-noise task were significantly more willing to work as a result of the coin flip as compared to the high-probability treatment. In contrast, workers facing the noisy task were directionally less likely to work, but the result is not statistically significant.
Table 4:  
EXPERIMENT 1B. COIN-FIP VS 99% TREATMENT

<table>
<thead>
<tr>
<th></th>
<th>Random-Effects Tobit. Dep Variable: WTW</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Noise</td>
<td>Noise</td>
<td>No Noise</td>
<td>Noise</td>
<td></td>
</tr>
<tr>
<td>1 (Coin flip)</td>
<td>4.483**</td>
<td>5.248**</td>
<td>-3.035†</td>
<td>-2.601</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.075)</td>
<td>(2.200)</td>
<td>(1.862)</td>
<td>(1.933)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed payment ($)</td>
<td>13.888***</td>
<td>13.888***</td>
<td>10.273***</td>
<td>10.310</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.292)</td>
<td>(0.292)</td>
<td>(0.326)</td>
<td>(0.251)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1 duration (min)</td>
<td>-0.015</td>
<td>.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.092)</td>
<td>(0.120)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.116</td>
<td>-0.049</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.103)</td>
<td>(0.088)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (Male)</td>
<td>0.248</td>
<td>-0.078</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.089)</td>
<td>(1.918)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>3.375**</td>
<td>5.895</td>
<td>5.288***</td>
<td>9.572</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.507)</td>
<td>(4.340)</td>
<td>(1.333)</td>
<td>(4.186)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1335</td>
<td>1335</td>
<td>1410</td>
<td>1410</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clusters</td>
<td>267</td>
<td>267</td>
<td>282</td>
<td>282</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Standard errors (in parentheses) are clustered at the individual level and calculated from observed information matrix. 15 observations are censored in no-noise regression and 10 are censored in noise regression. Income fixed effects are dummies for five income groups in demographic questionnaire.

† p = .103 * p < 0.10, ** p < 0.05, *** p < 0.01

Because of the (albeit small) uncertainty, our model predicts that participants in the high-probability treatment will demonstrate greater differences in willingness to work across the two tasks than those in the no-coin-flip condition. Therefore, we would expect diminished effects relative to our initial results in Experiment 1. Additionally, our power issues may result from probability weighting. If people overweight small probabilities (e.g. Kahneman and Tversky 1979; Prelec 1998; Gonzalez and Wu 1999), then the 1% chance may loom much larger than its objective probability. If this is the case, then participants would view the high-probability condition as closer to the coin-flip condition, hindering our ability to detect differences across these treatments. We see some evidence of this, as participants in the high-probability treatment exhibit exaggerated labor responses relative to the deterministic-expectations condition in Experiment 1. However, we do not have statistical power to properly examine this issue.
5 Experiment 2

We now turn to our within-subject experiment, which was conducted in the lab. We follow a similar two-session design, but elicit willingness to work in both sessions. We first discuss the design and our theoretical predictions, which closely resemble those from Experiment 1. We then present the experimental results and some robustness checks.

5.1 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 at the Harvard Decision Science Lab between April 15 and May 5, 2016. Our primary sample consists of 87 subjects.\(^{35}\) Participants were paid $7 for successfully completing each of two sessions. In order to prevent attrition, we paid participants contingent on completion of both sessions.

Participants in the first session were assigned to one of two tasks via coin flip and returned one week later to face that same task a second time. We measured participants’ willingness to work during both sessions of the experiment. Our identification of misattribution stems from changes willingness to work that result from differences in expectations between the first and second sessions. In the first session, we endow participants with uncertainty over their task assignment (through the coin flip). In the second session, this uncertainty is no longer present. We both instructed participants that their coin flip would count for both days and sent an email reminder of this approximately two days before their second session. Therefore, participants knew exactly which task they would face in the second session prior to showing up. Thus, the manipulation of expectations across sessions but within subjects mirrors the manipulation of expectations between subjects in Experiment 1. Our variable of interest is the difference in a participant’s willingness to work between week one—when her task comes as a surprise—and week two—when that same task is perfectly expected.

During each session, participants completed a real-effort task similar to that of Augenblick, Neiderle, and Sprenger (2015) and Augenblick and Rabin (2016). Participants were tasked to “transcribe” handwritten Greek and Russian letters. Each task consisted of a string of 35 handwritten characters; participants “transcribed” them by clicking the appropriate button matching the character. See Figure 5 for a screenshot of the task. Participants were randomly divided into one of

---

\(^{35}\) Ex-ante power tests suggested that \(n \approx 100\) would provide 80% power, assuming a modest effect size. Since we conducted this experiment first, we did not have the results of Experiment 1 to inform our sample 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 finals week. 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 participants are excluded from all analyses.
the two language treatments: half transcribed Greek during the first session and Russian during the second, while the other half faced the opposite order. In each session, participants were required to complete an initial learning phase which consisted of five mandatory tasks. After the initial learning phase, participants had the opportunity to complete supplemental tasks for additional pay.\textsuperscript{36} Importantly, we asked participants their willingness to work in both of the two sessions.

Figure 5: \textit{Screenshot of the transcription task from Experiment 2}. Participants click the gray button that matches the handwritten letter in order to “transcribe” the text. Participants were required to get 80\% accuracy to advance to the next transcription. Participants faced both languages—Greek or Russian (Cyrillic)—but faced only one language per session.

As in the coin-flip condition of Experiment 1, we presented each participant with two variants of the task: a baseline, no-noise task and a noisy task. In the noisy task, we played an annoying background noise to participants via headphones while they completed transcriptions. Headphones were calibrated so the audio played between 70 and 75 decibels. We used the same audio stimulus as in Experiment 1, except it played without stopping throughout the entire transcription time.

All participants were told that they faced a 1/2 chance of doing the task without noise and

\textsuperscript{36} Although our task is based on that of Augenblick, Niederle, and Sprenger (2015), we used a different set of letters which was easier to transcribe: Participants completed each row in approximately 40 seconds, compared to approximately 54 seconds in the first week of Augenblick, Niederle and Sprenger.
a $1/2$ chance of doing the task with noise. In order to induce participants to have reasonable beliefs about the tasks, the initial instructions included an interactive sample of the transcription task. Additionally, subjects were exposed to the audio stimulus, played for eight seconds (and repeatable if participants wished to listen again). Which task they actually faced was determined by a real coin flip. In order to make the probability salient, participants themselves flipped a U.S. quarter. We instructed participants that a flip of heads would result in the no-noise task, while a flip of tails would result in the noisy task. 44 participants faced the task without noise, while 43 faced the task with noise.

After completing the initial learning phase during Session 1, subjects were given the option to complete additional trials for supplemental pay. Departing from Experiment 1’s design, we elicited participants’ willingness to work during both the first and second session. To match common stakes in the lab setting, we asked willingness to work for the following fixed payments: \{4, 8, 12, 16, 20\}. As with the first experiment, participants used a slider to select any integer $e \in \{0, \ldots, 100\}$ and used the BDM mechanism to promote truthful responses.

The second session of the experiment followed the same timeline as the first session except there was no coin flip: participants faced whichever task they did on the first session. As noted above, participants faced a different language the second session. This allowed participants to plausibly form different beliefs across sessions and to change their willingness to work between sessions as a result of these updated beliefs. Additionally, this may have reduced anchoring effects: since participants faced a somewhat-different task, they may have been less likely to answer exactly the same as they did during the first session.

### 5.2 Theoretical Predictions

We sketch how our predictions from Experiment 1 extend to a within-subject design.\footnote{37 We focus our theoretical predictions on a participant who does not complete additional tasks during the first session. Thus, her only information comes from the two initial learning sessions at the beginning of each session. The logic we present in this section extends to those who also completed extra tasks, but receive an additional signal of the underlying disutility of effort of their assigned task.} We focus on the difference between a participant’s willingness to work in the first session—where she faces a surprise about her task assignment—and her willingness in the second session—where she faces no surprise. Thus, our outcome variable of interest is a participant’s change in willingness to work across the two sessions, $e_{j,1}(i) - e_{j,2}(i)$. Our predictions are similar to those of Experiment 1: fixing task assignment $i \in \{h, l\}$, the within-subject comparison in willingness to work across sessions in Experiment 2 mirrors the comparison between the coin flip and no coin flip groups in Experiment 1. Hence, analogous to Experiment 1, we describe below how our misattribution model predicts that participants facing the good task will be less willing to work in the second
session than the first; and that participants facing the bad task will be more willing to work in the second session than the first.

In order to apply the results from Experiment 1 to this setting, we require that participants have reasonably well-calibrated priors on aggregate. If priors are systematically biased in a specific way—namely, they significantly overestimate the cost of the noise task and underestimate the cost of the no-noise task—then changes in willingness to work across sessions may result from learning in the classical (i.e., non-reference-dependent) model. We believe this assumption is justified both by the data from Experiment 1 and from the experimental design: participants were exposed to both versions of the task before commencing work, and therefore should have reasonably well-calibrated priors. As such, we assume that priors are unbiased on aggregate.

We first describe how willingness to work is affected by reference-dependent utility without misattribution. In Experiment 1, participants waited at least eight hours to make decisions after learning which task they would face. In Experiment 2, however, participants waited only five to ten minutes. It is thus less likely that Experiment 2 provided sufficient time for reference-point adaptation in Session 1. As previously noted, even if reference points do not fully adapt, the Kőszegi and Rabin (2006) model does not predict decreasing effort among those facing the no-noise task and increasing effort among those facing the noisy task. This follows the logic of Observation 2, which shows that a participant with reference-dependent preferences and a stochastic reference point follows a strategy that “compresses” her willingness to work across the two tasks. That is, if the participant realizes the no-noise task in Session 1, she will work less than if her reference point was the no-noise task with certainty. And if she realizes the noisy task in Session 1, she will work more than if her reference point was the noisy task with certainty. While a participant’s reference point may incorporate the uncertainty induced by the coin flip when we elicit willingness to work in the first session, we assume the participant’s reference point entering Session 2 adjusts—it equals her perception of her assigned task and no longer depends on the task she was not assigned. This assumption seems warranted given that there is no uncertainty in task assignment in the second session, and subjects knew about their task assignment a week in advance. Thus following Observation 2, the reference-dependent model without misattribution predicts that a participant in the no-noise condition works more in Session 2 than in Session 1, and one in the noisy condition works less in Session 2 than in Session 1.

Misattribution acts in the opposite direction of the predictions above. To see this, first consider a participant in the no-noise condition. In Session 1, she will follow the strategy described above, but with respect to her biased perception about the underlying disutility of effort. Assuming priors are correct on average, this biased perception underestimates the true cost: as described in Section 2, the no-noise task will come as a positive surprise, leading the person to overestimate its enjoyability and hence underestimate its cost. Thus, her willingness to work in the first session will be biased
upward relative to a rational benchmark. In the learning phase of the second session, however, her experience will typically come as a loss: her expectations developed in Session 1 overestimate her enjoyment, and her second experience (now void of the positive surprise of avoiding the noisy task) will typically 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 experience is sufficiently strong, then the participant’s revealed willingness to work will decrease over the two sessions.

Similar logic predicts that a misattributor in the noisy condition may increase her effort across sessions. In the first session, the negative surprise of her poor 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 pushes her estimated cost downward, increasing her willingness to work in the second session. Hence, the misattributing participant’s revealed willingness to work likely increases over the two sessions.

5.3 Results

Our primary hypothesis concerns participants’ change in willingness to work across sessions. For those assigned to the task without noise, we predict their initial impressions will be “contaminated” by the sensation of positive surprise. Similarly, for those assigned to the task with noise, we predict their initial impressions will be contaminated by their disappointment. As highlighted above, our identification comes from the reference point adjusting in the second session. If initial beliefs were inflated because the participant misattributed sensations of positive surprise to the underlying effort cost, the new signal from the learning phase during the second session—which comes with no sensation of surprise—pushes those beliefs down. Contrastingly, if initial beliefs were pessimistic because the participant misattributed her disappointment to the effort cost, the new signal from the second session pushes those beliefs back up. We would expect this to manifest in changes in willingness to work, and we find such changes in the aggregate:

**Aggregate Result, Experiment 2:** Averaging over the five fixed payments, participants who faced the task without noise were willing to do 7.1 more tasks when their job assignment came as the result of chance versus when they faced the same task the second session ($p = .003$; random-effects model with standard errors clustered at individual level. Average willingness to work during first session: 30.9 tasks). In contrast, participants who faced the task with noise were willing to do 4.25 fewer tasks when their job assignment came as the result of chance versus when they faced the same task the second session ($p = .010$; random effects and clustered standard errors at individual level. Average willingness to work during first session: 25.9 tasks).
Consistent with our theoretical predictions, participants’ willingness to work tends to decrease across sessions when assigned the no-noise task and increase across sessions when assigned the noisy task. Given our small sample size, we present this result visually with a histogram of \( e_1 - e_2 \) for each of the tasks over all five payment (Figure 6). We use kernel smoothing to cleanly depict each of the curves (Epanechnikov kernel); raw histograms appear in the Appendix. Figure 6 demonstrates that the difference in willingness to work for participants assigned to the no-noise task is primarily positive. In contrast, the difference in willingness to work for participants assigned to the no-noise task is primarily negative.

![Kernel density of the difference in willingness-to-work between the first and second sessions, separated by task faced.](image)

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 one and two 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.

While the figure above suggests an effect, we establish statistical significance by running some simple regressions with \( e_1 - e_2 \) as the dependent variable. The results are shown in Table 5. As with our analyses of Experiment 1, we split the data by task assignment: noise versus no noise. Column (1) and (3) show a simple specification without controls, establishing that our effect is sta-
tistically significant while controlling for subject-level variability through a random-effects model with standard errors clustered at the individual level ($p = .003$ for no noise; $p = .010$ for noise). In columns (2) and (4), we add controls for the fixed payment, whether a participant completed extra tasks, the number of extra tasks a participant completed, and whether a participant faced Russian or Greek in the first session. In doing so, we find that the changes in willingness to work are just outside statistical significance, but directionally consistent with our model ($p = .077$ for no-noise; $p = .052$ for no noise).

Table 5:

<table>
<thead>
<tr>
<th>Dependent variable: $e_1 - e_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment effect</strong></td>
</tr>
<tr>
<td>No Noise</td>
</tr>
<tr>
<td>Treatment effect</td>
</tr>
<tr>
<td>7.143***</td>
</tr>
<tr>
<td>(2.442)</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td>Treatment effect</td>
</tr>
<tr>
<td>6.813*</td>
</tr>
<tr>
<td>(3.852)</td>
</tr>
<tr>
<td>-4.254***</td>
</tr>
<tr>
<td>(1.653)</td>
</tr>
<tr>
<td>-7.310*</td>
</tr>
<tr>
<td>(3.765)</td>
</tr>
<tr>
<td><strong>Fixed payment ($)</strong></td>
</tr>
<tr>
<td>No Noise</td>
</tr>
<tr>
<td>Fixed payment ($)</td>
</tr>
<tr>
<td>-0.046</td>
</tr>
<tr>
<td>(0.148)</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td>Fixed payment ($)</td>
</tr>
<tr>
<td>-0.186</td>
</tr>
<tr>
<td>(0.121)</td>
</tr>
<tr>
<td><strong>1(Extra tasks)</strong></td>
</tr>
<tr>
<td>No Noise</td>
</tr>
<tr>
<td>1(Extra tasks)</td>
</tr>
<tr>
<td>-3.985</td>
</tr>
<tr>
<td>(7.879)</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td>1(Extra tasks)</td>
</tr>
<tr>
<td>8.057</td>
</tr>
<tr>
<td>(5.937)</td>
</tr>
<tr>
<td><strong># Num extra tasks</strong></td>
</tr>
<tr>
<td>No Noise</td>
</tr>
<tr>
<td># Num extra tasks</td>
</tr>
<tr>
<td>0.533**</td>
</tr>
<tr>
<td>(0.237)</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td># Num extra tasks</td>
</tr>
<tr>
<td>-0.183</td>
</tr>
<tr>
<td>(0.216)</td>
</tr>
<tr>
<td><strong>1(Russian on day 1)</strong></td>
</tr>
<tr>
<td>No Noise</td>
</tr>
<tr>
<td>1(Russian on day 1)</td>
</tr>
<tr>
<td>-4.673</td>
</tr>
<tr>
<td>(4.917)</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td>1(Russian on day 1)</td>
</tr>
<tr>
<td>5.486</td>
</tr>
<tr>
<td>(3.745)</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>175</td>
</tr>
<tr>
<td>175</td>
</tr>
<tr>
<td>185</td>
</tr>
<tr>
<td>185</td>
</tr>
</tbody>
</table>

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Standard errors, clustered at individual level, in parentheses. All regressions include random effects at individual level.

In order to compare the economic significance of these results to those of Experiment 1, we conduct a similar calibration exercise to the one from Section 4. The magnitude of the mistake is similar. In order to incentivize the average participant to complete 25 transcriptions without noise, a firm would have to pay $7.75 if her initial impressions were formed immediately after the positive surprise from the outcome of the coin flip. This increases to $11 when the participant returns knowing her task assignment for certain. In contrast, a firm would have to pay $12 to

There are two caveats to this calibration exercise in Experiment 2. First, because the task in Experiment 2 is more time-consuming than that of Experiment 1, the magnitudes of payments are significantly different across experiments. Second, because the sample size is much smaller, these estimates are poorly estimated and are meant to be illustrative.
incentivize the average participant to do 25 transcriptions with noise if her initial impressions were formed immediately after her disappointment from the coin flip. This decreases to $10.50 when the participant returns knowing her task assignment for certain. Together with the estimates from Experiment 1, we find that the coin flip distorts willingness to work from approximately 17% to 40%. As the above highlights, our estimates from this experiment, while imprecise, fall in the same order of magnitude as those of Experiment 1.

This calibration exercise highlights a theoretical prediction of the model: sequential contrast effects. As described in the theoretical section above, our model predicts a form of contrast effect wherein the disutility of effort on Session 2 is compared against the wrongly-encoded disutility from Session 1. For participants facing the noisy task, this leads 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 calibration provides a hint of this (the difference in willingness to work between noise and no noise on the second session is extremely small), but we would need more data to make such claims with statistical significance.

As with Experiment 1, we present analyses demonstrating that attrition is an unlikely explanation for our results. In Supplemental Table 12, we show that participants who face the noise and no-noise tasks fail to show up for the second session at similar rates. Additionally, neither mean willingness to work from Session 1 nor whether a participant faced Russian or Greek the first session predicts attrition.

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. Our theoretical predictions become muddied when comparing participants who completed additional tasks with those who did not, as the two groups have accumulated different amounts of experience. Note, however, that only one-third of participants completed additional tasks in the first session. Columns (2) and (4) of Table 5 demonstrate that controlling for these additional tasks does not qualitatively change our main result. Additionally, Supplemental Table 11 demonstrates that our main result is robust if we simply drop participants who completed extra tasks. While statistical power decreases when dropping participants, our point-estimates remain quite close.

6 Conclusion

In this paper we provide evidence that people 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 at the tasks. We briefly discuss some reasons for caution in interpreting our results as well as directions for future research.

In our first experiment, we found weak but suggestive evidence of loss aversion: 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 a general diminished willingness to work across weeks. 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 use the response scale in different ways which makes detecting loss aversion more difficult. 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.

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 Experiment 1b provides a first look at the role of probabilistic assignment in subsequent evaluations, future research should explore this more completely.

Our results immediately 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 the undesirable tasks if they knew for certain they would have to complete it, rather than facing uncertainty each day. 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 its implications 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


A Discussion of Observation 2

In this Appendix, we provide a sketch proof of Observation 2 and the discussion of Experiment 2. In order to provide a more general discussion, we relax some of the specifics from the experiment. We describe this (more general) environment, then provide a proof of the result and a discussion of the implications across a number of domains.

We consider a worker (female) with reference-dependent preferences choosing her willingness to work $e \geq 0$ in a stochastic environment. With probability $1 - p$ she faces an “easy task” with effort-cost function $c(e)$, which we assume is an increasing, convex function that satisfies $c(0) = 0$. With probability $p$, the worker faces the “hard task” with effort-cost function $\theta c(e)$, where we assume $\theta > 1$.\footnote{We have simplified this example by normalizing $\theta(l) = 1$ and $\theta(h) = \theta$. This is done without loss since we are not considering the role of parameter uncertainty on behavior. Additionally, we have removed subscripts for time and person indices.}

We assume she is paid according to $m \cdot G(e)$; thus we are assuming that $m$ is the consumption utility of some fixed amount of dollars and that responding a higher willingness to work leads to a higher probability of payment. We make no functional form assumptions over $G(\cdot)$ other than to assume that it has support over some interval $[\underline{x}, \overline{x}]$; we will not consider how uncertainty from $G(\cdot)$ influences willingness to work. Additionally, to highlight the role of reference dependence and not mislearning or mistaken beliefs, we assume the parameters are all known to the worker. Following standard models in the field, we assume that the monetary dimension and the effort dimension are separate hedonic dimensions.\footnote{See K˝oszegi and Rabin (2006) for a discussion of how to identify whether this assumption is valid.}

Finally, we consider the preferences presented in Equations 1 and 2. For the purpose of discussion, we call personal equilibria with $e^*(l) \neq e^*(h)$ “separating” and those with equality “pooling”.

Consider when $p \in \{0, 1\}$. The worker maximizes such that:

\[ c(e^*(l)) = m \]
\[ \theta c(e^*(h)) = m \]

for maximizers $e^*(l)$ when the task is easy and $e^*(h)$ when the task is difficult (or $p = 0$ and $p = 1$, respectively).

Suppose $p \in (0, 1)$, and that the worker planned to exert effort such that $\tilde{e}(h|p) > \tilde{e}(l|p)$. Would she follow through on this plan? In order to assess this, we assume that she would and derive the optimal behavior, hoping to show consistency. Thus, the worker’s utility from effort $e(l|p)$ when
she realizes the easy task and had plan $\tilde{e}(l|p) > \tilde{e}(h|p)$ is given by:

$$u(e(l|p)|\tilde{e}, \theta) = m \cdot G(e) - \int_{0}^{e(l|p)} c(e')dG(e') + p \eta \left( m \cdot G(e(l|p)) - m \cdot G(\tilde{e}(h|p)) \right) + p \eta \lambda \left( \int_{0}^{e(l|p)} c(e')dG(e') + \theta \int_{0}^{\tilde{e}(h|p)} c(e')dG(e') \right).$$

Rearranging, taking the partial derivative with respect to $e$, and setting equal to zero yields:

$$\frac{1 + p \eta}{1 + p \eta} m = c(e^*(l|p))$$

which implies that $e^*(l|p) < e^*(l)$. However, in order for a plan to be in equilibrium, $\tilde{e}(l|p) = e^*(l|p)$. Therefore we have shown that there does not exist a (separating) personal equilibrium in which the worker plans to exert more effort in the easy task than she would if she faced that task with certainty. An analogous condition can be derived for a worker facing the hard task, which shows $e^*(h|p) > e^*(h)$. Thus, all separating equilibria are “compressing”: $e^*(l|p) - e^*(h|p) < e^*(l) - e^*(h)$.

There are personal equilibria in which the person exerts the same amount of effort in both cost realizations, and these include plans in which overall effort is higher. However, these equilibria are not preferred personal equilibria. In all separating equilibria, she will exert no more effort in either realization than the effort she would exert if she faced the easy task for certain.

Generically, it is not possible to say how the parameters affect when there is a unique personal equilibrium. However, if we apply a quadratic cost function, we can discuss the specific equilibria. Under the assumption that $c(e) = e^2$, there is a unique separating personal equilibrium when

$$\theta > \left( \frac{1 + p \eta \lambda}{1 + p \eta} \right)^2.$$ 

When $\theta < \left( \frac{1 + p \eta \lambda}{1 + p \eta} \right)^2$, there exist multiple pooling equilibria, with a unique pooling equilibrium where there is equality.

### B Additional Theoretical Predictions

In this Appendix, we describe a more complete model of misattribution in which participants update their beliefs according to Bayes’ Rule. We apply the model to two “rounds” of learning—that is, after participants have received two signals. We examine beliefs and behavior after the first round of learning to address Experiment 1 and both the first and second round to address Exper-
ment 2. With reference to Experiment 1, we demonstrate that for any prior belief that is independent of treatment, the misattributor forms exaggerated beliefs about the underlying disutility of effort. We then demonstrate that beliefs after the second signal—which is drawn with certainty from one of the underlying distributions—are likely to “contract” relative to the exaggerated perceptions after the first signal. We conclude with a note about how reference-dependence with a sluggish reference point decreases the observed difference in behavior.

For tractability, we assume that the participant’s prior estimates of \( \theta(i) \), denoted \( \hat{\theta}_{j,0}(i) \), are such that \( \hat{\theta}_{j,0}(i) \sim N(\theta(i), \gamma(i)) \). Conditional on \( \hat{\theta}_{j,0}(i) \), participant \( j \) has prior beliefs \( \theta(i) \sim N(\hat{\theta}_{j,0}(i), \rho(i)) \). Signals are generated from the initial learning session in Experiment 1 (or each of the two learning sessions in Experiment 2). Each learning session \( t \in \{1, 2\} \) in which a participant faces task \( i \) provide a signal \( v_{j,t}(i) = \theta(i) + \epsilon_{j,t}(i) \), where \( \epsilon_{j,t}(i) \sim N(0, \sigma(i)) \). Letting \( \hat{\theta}_{j,t}(i) \) denote the person’s estimate of \( \theta(i) \) entering elicitation \( t = 1, 2 \), it follows that participants who knew their assignment with certainty would update their beliefs according to the following:

\[
\hat{\theta}_{j,t}(i|p = 1) = \hat{\theta}_{j,t-1}(i) + \alpha_t(i) \left( \hat{v}_{j,t}(i) - \hat{\theta}_{j,t-1}(i) \right),
\]

where \( \alpha_t(i) = \frac{\rho(i)}{\tau \rho(i) + \sigma(i)} \) and \( \hat{v}_{j,t}(i) \) is the person’s encoded signal. The encoded signal of participants with certain task assignment are given by the following: \( \hat{v}_{j,t}(i|p = 1) = v_{j,t}(i) + \kappa_{j,t}(i) \left( v_{j,t}(i) - \hat{\theta}_{j,t-1}(i) \right) \), where \( \kappa_{j,t}(i) \) depends on whether the signal exceeded the participant’s prior expectations. If the signal exceeds the prior, \( \kappa_{j,t}(i) = \kappa^L \equiv \lambda \left( \frac{\eta - \eta}{1 + \eta} \right) \) and if it fell short of expectations \( \kappa_{j,t}(i) = \kappa^G \equiv \lambda \left( \frac{\eta - \eta}{1 + \eta} \right) \).

Rewriting the above yields

\[
\hat{\theta}_{j,t}(i|p = 1) = \hat{\theta}_{j,t-1}(i) + \alpha_t(i) \left( 1 + \kappa_{j,t}(i) \right) \left[ v_{j,t}(i) - \hat{\theta}_{j,t-1}(i) \right].
\]

In contrast, for participants who initially faced unknown task assignment, encoded signals are “contaminated” by beliefs about the task the participant ultimately didn’t face. Specifically, beliefs are given by

\[
\hat{\theta}_{j,t} \left( i|p = \frac{1}{2} \right) = \hat{\theta}_{j,t-1}(i) + \alpha_t(i) \left( \hat{v}_{j,t}(i) - .5 \hat{\theta}_{j,t-1}(i) - .5 \hat{\theta}_{j,t-1}(k) \right)
\]

for \( k \neq i \). We’ll now consider each possible state. First, we consider the case when \( i = l \) and \( t = 1 \):

\footnote{Note that since \( \theta_{j,t}(i) \) represents a negative cost parameter, the participant experiences a positive surprise when her signal is less than the disutility she expected; she experiences a negative surprise when her signal exceeds her expectations.}
that is, the participant faces the easy task on the first session. Thus
\[
\hat{\theta}_{j,t}(l|p = 1/2) = \hat{\theta}_{j,0}(l) + .5\alpha_1(i)(1 + \kappa_{j,1}(l))(\hat{v}_{j,1}(l) - \hat{\theta}_{j,0}(l)) + .5\alpha_1(i)(1 + \kappa_{j,1}(h))(\hat{v}_{j,1}(l) - \hat{\theta}_{j,0}(h))
\]
where we have abused notation with \(\kappa\): the signal \(v_{j,t}(l)\) is compared against \(\hat{\theta}_{j,0}(l)\) for \(\kappa_{j,t}(l)\) and the signal is compared against \(\hat{\theta}_{j,0}(h)\) for \(\kappa_{j,t}(h)\). Taking differences, \(\hat{\theta}_{j,t}(l|p = 1) - \hat{\theta}_{j,t}(l|p = 1/2)\) allows us to see how misattributors make a larger mistake in the coin-flip condition: \(^{42}\)

\[
\hat{\theta}_{j,1}(l|p = 1) - \hat{\theta}_{j,1}(l|p = 1/2) = .5\alpha_1(l)\left[(1 + \kappa_{j,1}(l))v_{j,1}(l) - \hat{\theta}_{j,0}(l)\right] - (1 + \kappa_{j,1}(h))\left[v_{j,1}(l) - \hat{\theta}_{j,0}(h)\right].
\]

Note that \(\hat{\theta}_{j,0}(l) < \hat{\theta}_{j,0}(h)\). We proceed by cases. First, suppose \(v_{j,t}(l) < \hat{\theta}_{j,0}(l) < \hat{\theta}_{j,0}(h)\). Then \(\kappa_{j,1}(h) = \kappa_{j,1}(l) = \kappa^G\), and it follows that \(\hat{\theta}_{j,1}(l|p = 1) > \hat{\theta}_{j,1}(l|p = 1/2)\). Likewise is true if \(\hat{\theta}_{j,0}(l) < \hat{\theta}_{j,0}(h) < \hat{\theta}_{j,0}(l)\). Finally, suppose \(\hat{\theta}_{j,0}(l) < \hat{\theta}_{j,0}(l) < \hat{\theta}_{j,0}(h)\). Then \(\kappa_{j,1}(l) = \kappa^L\) and \(\kappa_{j,1}(h) = \kappa^G\), and with some arithmetic it follows that \(\hat{\theta}_{j,1}(l|p = 1) > \hat{\theta}_{j,1}(l|p = 1/2)\). An analogous argument can be used to show that \(\hat{\theta}_{j,1}(h|p = 1) < \hat{\theta}_{j,1}(h|p = 1/2)\). Given these beliefs, it is trivial to show that behavior matches that of Observation 3.

Beliefs in the second period—when the reference point has adjusted—likewise follow Equation (2) above. That is: \(\hat{\theta}_{j,2}(i) = \hat{\theta}_{j,1}(i) + \alpha_2(i)(1 + \kappa_{j,2}(i))[v_{j,2}(i) - \hat{\theta}_{j,1}(i)]\). We will compare these beliefs when first-period beliefs were formed under either \(p = 1/2\) or \(p = 1\). For shorthand, we will denote these beliefs \(\hat{\theta}_{j,2}(i|p = 1) \leq \hat{\theta}_{j,2}(i|p = 1/2)\). Our argument is straightforward and stems from our assumption that priors are unbiased. Although we did not require this assumption for the result above, we will leverage unbiased priors to demonstrate how willingness to work changes over days. Given unbiased priors, it’s clear that in ex-ante, unconditional expectation \(E[\hat{\theta}_{j,1}(i|p = 1)] = \theta_j(i)\). Thus \(E[\hat{\theta}_{j,2}(i|p = 1)] = \theta_j(i)\). In contrast, it can easily be shown that \(E[\hat{\theta}_{j,1}(l|p = 1/2)] < \theta_j(l)\) and \(E[\hat{\theta}_{j,1}(h|p = 1/2)] < \theta_j(h)\). We now want to show \(E[\hat{\theta}_{j,2}(l|\hat{\theta}_{j,1}(l|p = 1/2)] > \hat{\theta}_{j,1}(l)\). Since

\[
\hat{\theta}_{j,2}(l) = \alpha_2(l)v_{j,2}(l) + (1 - \alpha_2(l))\hat{\theta}_{j,1}(l)
\]

Thus \(E[\hat{\theta}_{j,2}(l)|\hat{\theta}_{j,1}(l|p = 1/2)] > \hat{\theta}_{j,1}(l) \Leftrightarrow E[v_{j,2}(l) + \kappa_2(l)v_{j,2}(l) - \hat{\theta}_{j,1}(l)]|\hat{\theta}_{j,1}(l)]\). Rearranging yields \(E[(1 + \kappa_{j,1}(l))(v_{j,2}(l) - \hat{\theta}_{j,1}(l))]> 0\). Since both possible values of \(\kappa\) are greater than zero, the above reduces to \(E[x - \hat{\theta}_{j,1}(l)] > 0 \Leftrightarrow \theta(l) > \hat{\theta}_{j,1}(l)\). We have thus shown that in ex-ante

\(^{42}\) In this approach, we fix the realized signal at some \(\nu\) and vary the expectations of a participant before forming her first-period beliefs. This approach provides the same result as taking expectations across participants but will simplify the presentation of the result. Additionally, it provides an indirect proof of Observation 3 Part 1 with appropriate substitution for the prior beliefs and taking \(\varepsilon \rightarrow 0\).
expectation with unbiased priors, beliefs change in a way that matches the observed behavior in Experiment 2.

We conclude with a discussion of how reference-dependence with sluggish reference points pushes against the results of Observation 3. In order to formalize this, we make some simplifying assumptions. As in the previous section of this Appendix, we will normalize $\theta(j(l)) = 1$. We will consider a person whose first-period signal exactly equals their prior: $v_{j,1}(i) = \hat{\theta}_{j,0}(i)$. Again considering a person who faces the good task, beliefs after the first round are given by

$$
\hat{\theta}_{j,1}(i|p = 1/2) = \hat{\theta}_{j,0}(l) + \frac{\alpha_1(l)}{2} \left( \frac{\eta - \hat{\eta}}{1 + \hat{\eta}} \right) \left( \hat{\theta}_{j,0}(l) - \hat{\theta}_{j,0}(h) \right)
$$

$$
= 1 + \frac{\alpha_1(l)}{2} \left( \frac{\eta - \hat{\eta}}{1 + \hat{\eta}} \right) \left( 1 - \hat{\theta}_{0,j}(h) \right)
$$

Utilizing the conditions from the previous section, we can see that misattribution overcomes the contracting force of Kőzegi and Rabin whenever

$$
\hat{\theta}_{j,1}(i|p = 1/2) < \left( \frac{2 + \eta}{2 + \eta \lambda} \right).
$$

Rearranging and simplifying yields

$$
\frac{\alpha_1(l)}{2} \left( \frac{\eta - \hat{\eta}}{1 + \hat{\eta}} \right) \left( \hat{\theta}_{0,j}(h) - 1 \right) > \eta (\lambda - 1)
$$

The above allows us to see how $\hat{\eta}$ must be small and the difference between $\hat{\theta}_{j,0}$ and 1 must be sufficiently large to overcome loss aversion. Put another way, the above clearly shows that the distortion caused by misattribution must be larger than the distortion caused by reference dependence, which is proportional to $\eta (\lambda - 1)$. Given the restrictions on the parameters $\alpha_1(l), \hat{\eta}$ we can see that a necessary condition for when misattribution overcomes reference-dependence is $\hat{\theta}_{j,0}(h) - 1 > 2(\lambda - 1)$. This condition highlights the important role of loss aversion. Taking the limit $\lambda \rightarrow 1$, reference-dependence does not distort behavior, and therefore only a very small difference in tasks is required to distort willingness to work. For participants assigned to the hard task, an analogous condition can easily be derived following the same steps as above.

\[\text{43} \] It is easy to see from this equation that without the normalization of the effort-cost parameter, this condition is replaced by $\theta_{j,0}(h) - \theta_{j,0}(l) > 2(\lambda - 1)$. 48
### Supplemental Tables and Figures

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

#### Table 6: Experiment 1. Effect of Coin Flip, No Noise - Dropped Censored Data

<table>
<thead>
<tr>
<th></th>
<th>Reduced Sample</th>
<th>Reduced Sample</th>
<th>Reduced+Monotonic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 (Coin Flip)</strong></td>
<td>3.520**</td>
<td>3.364*</td>
<td>3.145*</td>
</tr>
<tr>
<td></td>
<td>(1.695)</td>
<td>(1.724)</td>
<td>(1.819)</td>
</tr>
<tr>
<td><strong>Fixed payment ($)</strong></td>
<td>12.510***</td>
<td>12.510***</td>
<td>12.990***</td>
</tr>
<tr>
<td></td>
<td>(0.248)</td>
<td>(0.248)</td>
<td>(0.256)</td>
</tr>
<tr>
<td><strong>Session 1 duration (min)</strong></td>
<td>0.025</td>
<td>0.023</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.074)</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>-0.116</td>
<td>-0.104</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.077)</td>
<td>(0.079)</td>
<td></td>
</tr>
<tr>
<td><strong>1 (Male)</strong></td>
<td>0.726</td>
<td>1.001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.743)</td>
<td>(1.837)</td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>3.248**</td>
<td>6.920*</td>
<td>4.688</td>
</tr>
<tr>
<td></td>
<td>(1.262)</td>
<td>(4.031)</td>
<td>(4.184)</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>1195</td>
<td>1195</td>
<td>1010</td>
</tr>
<tr>
<td><strong>Clusters</strong></td>
<td>239</td>
<td>239</td>
<td>202</td>
</tr>
<tr>
<td><strong>Income FE</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Notes: Standard errors (in parentheses) are clustered at the individual level. All regressions include random effects.

* p < 0.05, ** p < 0.01, *** p < 0.001
Table 7: 
**EXPERIMENT 1. EFFECT OF COIN FLIP, NOISE - DROPPED CENSORED DATA**

<table>
<thead>
<tr>
<th>OLS, Dependent Variable: ( WTW )</th>
<th>Reduced Sample</th>
<th>Reduced Sample</th>
<th>Reduced+Monotonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) (Coin Flip)</td>
<td>-4.095**</td>
<td>-4.119**</td>
<td>-3.421*</td>
</tr>
<tr>
<td></td>
<td>(1.743)</td>
<td>(1.779)</td>
<td>(1.905)</td>
</tr>
<tr>
<td>Fixed payment ($)</td>
<td>9.523***</td>
<td>9.523***</td>
<td>9.943***</td>
</tr>
<tr>
<td></td>
<td>(0.254)</td>
<td>(0.254)</td>
<td>(0.258)</td>
</tr>
<tr>
<td>Session 1 duration (min)</td>
<td>-0.018</td>
<td>-0.144</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.111)</td>
<td>(0.152)</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.031</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.075)</td>
<td>(0.079)</td>
<td></td>
</tr>
<tr>
<td>(1) (Male)</td>
<td>-1.062</td>
<td>-1.053</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.854)</td>
<td>(1.946)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>6.018***</td>
<td>3.310</td>
<td>4.131</td>
</tr>
<tr>
<td></td>
<td>(1.295)</td>
<td>(4.041)</td>
<td>(4.279)</td>
</tr>
<tr>
<td>Observations</td>
<td>1290</td>
<td>1290</td>
<td>1075</td>
</tr>
<tr>
<td>Clusters</td>
<td>258</td>
<td>258</td>
<td>215</td>
</tr>
<tr>
<td>Income FE</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Notes: Standard errors (in parentheses) are clustered at the individual level. All regressions include random effects.  
\( \dagger \) \( p < 0.1 \), \( * \) \( p < 0.05 \), \( ** \) \( p < 0.01 \), \( *** \) \( p < 0.001 \)
Table 8:

**EXPERIMENT 1. DIVIDED SAMPLE BY TOTAL EXPERIMENT DURATION**

<table>
<thead>
<tr>
<th></th>
<th>No Noise</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Above median</td>
<td>Below median</td>
</tr>
<tr>
<td>( \dagger ) (Coin flip)</td>
<td>3.480</td>
<td>4.666*</td>
</tr>
<tr>
<td></td>
<td>(3.427)</td>
<td>(2.590)</td>
</tr>
<tr>
<td>Fixed payment ($)</td>
<td>14.470***</td>
<td>12.954***</td>
</tr>
<tr>
<td></td>
<td>(0.468)</td>
<td>(0.384)</td>
</tr>
<tr>
<td>Constant</td>
<td>5.207</td>
<td>2.458</td>
</tr>
<tr>
<td></td>
<td>(2.684)</td>
<td>(1.828)</td>
</tr>
<tr>
<td>Observations</td>
<td>630</td>
<td>635</td>
</tr>
<tr>
<td>Clusters</td>
<td>126</td>
<td>127</td>
</tr>
</tbody>
</table>

Notes: Standard errors (in parentheses) are clustered at the individual level. All regressions include random effects at the individual level.

** p < 0.1, * * p < 0.05, ** * * p < 0.01
Table 9: Experiment 1. Determinants of Returning for Second Session

<table>
<thead>
<tr>
<th></th>
<th>Logit. Dependent variable: ( \mathbb{I}(\text{return}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw AMEs</td>
</tr>
<tr>
<td>( \mathbb{I}(\text{Noise}) )</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td>(0.73)</td>
</tr>
<tr>
<td>( \mathbb{I}(\text{Coin Flip}) )</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>(0.46)</td>
</tr>
<tr>
<td>Age</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>(-0.07)</td>
</tr>
<tr>
<td>( \mathbb{I}(\text{Male}) )</td>
<td>0.514</td>
</tr>
<tr>
<td></td>
<td>(1.77)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.056***</td>
</tr>
<tr>
<td></td>
<td>(11.12)</td>
</tr>
<tr>
<td>Observations</td>
<td>586</td>
</tr>
</tbody>
</table>

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 10: Experiment 1B. Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>High-Probability Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>noise=0</td>
</tr>
<tr>
<td>Age</td>
<td>33.294</td>
</tr>
<tr>
<td></td>
<td>(9.352)</td>
</tr>
<tr>
<td>(1(\text{Male}))</td>
<td>.488</td>
</tr>
<tr>
<td></td>
<td>(.501)</td>
</tr>
<tr>
<td>Income</td>
<td>2.463</td>
</tr>
<tr>
<td></td>
<td>(1.069)</td>
</tr>
<tr>
<td>(1(\text{Return}))</td>
<td>.932</td>
</tr>
<tr>
<td></td>
<td>(.253)</td>
</tr>
<tr>
<td>Observations</td>
<td>160</td>
</tr>
</tbody>
</table>

Note: Income is coded as in Experiment 1.

Table 11: Experiment 2. Difference in Willingness to Work; No Extra Tasks

<table>
<thead>
<tr>
<th>Dependent variable: (e_1 - e_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>No Noise</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>4.308</td>
<td>7.268*</td>
</tr>
<tr>
<td></td>
<td>(2.754)</td>
<td>(3.846)</td>
</tr>
<tr>
<td>Fixed payment ($)</td>
<td>0.106</td>
<td>-0.258</td>
</tr>
<tr>
<td></td>
<td>(0.148)</td>
<td>(0.170)</td>
</tr>
<tr>
<td>(1(\text{Russian, Session 1}))</td>
<td>-10.162*</td>
<td>7.433*</td>
</tr>
<tr>
<td></td>
<td>(5.285)</td>
<td>(4.001)</td>
</tr>
<tr>
<td>Observations</td>
<td>120</td>
<td>120</td>
</tr>
</tbody>
</table>

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 12: **Experiment 2. Determinants of Returning for Second Session**

<table>
<thead>
<tr>
<th></th>
<th>Logit, dependent variable: $1$(return)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
</tr>
<tr>
<td>Avg $WTW$, Session 1</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
</tr>
<tr>
<td>Avg $WTW$, Session 1*$1$(Noise)</td>
<td>-0.022</td>
</tr>
<tr>
<td></td>
<td>(0.015)</td>
</tr>
<tr>
<td>$1$(Noise)</td>
<td>1.802*</td>
</tr>
<tr>
<td></td>
<td>(0.984)</td>
</tr>
<tr>
<td>$1$(Russian, Session 1)</td>
<td>0.289</td>
</tr>
<tr>
<td></td>
<td>(0.606)</td>
</tr>
<tr>
<td>Constant</td>
<td>1.717***</td>
</tr>
<tr>
<td></td>
<td>(0.449)</td>
</tr>
<tr>
<td>Observations</td>
<td>87</td>
</tr>
</tbody>
</table>

*Note: Standard error in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Figure 7: 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 expost; they do not lack for a direct thesis–they make arguments and reach conclusions.”

Score: 5; Positive Review

“Sometimes you dont 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 were 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.] 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.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, it’s in your best interest to take each question seriously and choose as if it was the only question.

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

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

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

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

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

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

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

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

Depending on chance, a background noise may be played throughout the transcription process. We’ll describe what determines whether you hear the noise in a moment. However, we’d like to make sure you know what the sound will be. Please click the play button below twice for a sample of the noise. When you are finished listening to the sample noise, click the arrow below to continue.

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

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

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