Ambiguity in Performance Pay: An Online Experiment

June 28, 2013

David J. Cooper
Florida State University and University of East Anglia

David B. Johnson
Florida State University

Abstract: Many incentive plans are inherently ambiguous, lacking an explicit mapping between performance and compensation. Using an online labor market, Amazon Mechanical Turk, we study the effect of ambiguity on willingness to accept contracts to do a real-effort task as well as completion and performance of this task. Ambiguity about the relationship between performance and compensation affects the willingness of individuals to accept contracts and the likelihood of quitting before completion, but not performance. These effects are non-monotonic in the level of ambiguity. Information about ability at the task reduces willingness to accept and increases quitting, but does not affect performance.

Acknowledgements: We would like to thank Javier Cano-Urbina, Yan Chen, Emel Filiz Ozbay, Anastasia Semykina, and seminar participants at Chapman University, EWEBE, Florida State University, and the University of Maryland for helpful comments.

Keywords: ambiguity, incentives, performance pay, quitting, online experiment

JEL Classification: C99, D81, J33
1. **Introduction:** Incentive contracts tying pay to performance are ubiquitous in both the public and private sectors. In some cases these contracts are extremely explicit, making the relationship between performance and incentive pay unambiguous. Sales commissions are a well-known example of this sort of incentive scheme. For instance, car salesmen are normally paid a fixed percentage of the “front end” profit on a car sale (sales price – dealer overhead – invoice price).\(^1\) At the opposite end of the spectrum are incentive schemes where employees are completely uninformed about the relationship between their performance and their incentive pay. The “black box” compensation schemes used by some large law firms (i.e. Jones Day) provide a striking example of this sort of incentive pay. Hermann (2012) describes black box compensation as follows:

> “Under this system, the managing partner (or a small committee) sets compensation for each partner in the firm. There is no specific formula for allocating the spoils, and partners are forbidden from discussing their compensation with each other. Each partner is told what he’ll make in the coming year … and the process is over.”

Most incentive schemes fall somewhere between these extremes. While employees have an idea about what constitutes good performance and what the relationship is between performance and incentive pay, the precise definitions of good performance and the relationship between performance and pay may be ambiguous. An example that academics are painfully familiar with is the tenure system. While senior faculty can usually give general guidance about what is necessary for tenure, could they state a precise rule? More publications are better than less, but exactly how many are needed to get tenure? Better quality publications count for more, but exactly how is quality measured? How do publications compare across different field or subfields?

Continuing the tenure example, novice hires often face uncertainty about not just about the incentive pay system, but also about their likely performance. Suppose a newly hired economics professor is told that they need six publications in top field journals or better. It seems unlikely that a newly minted PhD knows precisely how difficult this task is or exactly how good they are at this task.

The preceding examples suggest that workers often face both risk and ambiguity. Beyond not knowing what their incentive pay will be, they don’t know the process generating their incentive pay. Put in a more technical fashion, workers face Knightian uncertainty about the distribution of incentive pay subject to effort. Our primary goal in this paper is to study how ambiguity in incentive contracts affects workers. Specifically, how does ambiguity affect the willingness of workers to take a job, the likelihood that they quit a job, and their performance at this job?

\(^1\) The percentage varies from dealership to dealership, and it is common to have incentive plans beyond the commission. (http://carsalesprofessional.com/car-salesman-commission/)
To address the effects of ambiguity, we conduct an online experiment. Subjects were recruited from Amazon Mechanical Turk (AMT), an online labor market specializing in workers who do short jobs for relatively low amounts of compensation. Subjects must choose between two real effort tasks: coding messages drawn from Charness and Dufwenberg (2006) or transcribing printed words into a text box. Coding messages is a challenging task where it is difficult for a worker to tell how well they are doing, making it a good task for inducing ambiguity. Subjects that choose the coding task earn a bonus which varies based on their performance. The transcription task is a sure thing. It takes a fixed amount of time to complete, identical to the time for the coding task, and yields a fixed and known bonus.

Because our experiment is conducting within a real labor market where participation is truly voluntary, quitting is a natural feature of the environment. Subjects working on AMT aren’t in a lab where they cannot easily leave and would face psychological discomfort if they did so.² If they don’t like the experiment or think they can make more money doing something else, our subjects can always switch to another webpage or task on AMT. About 11% of the subjects who choose the coding task quit before completing it even though the instructions and materials heavily stress that this means forfeiting their pay. We view quitting as a natural feature of work environments, and a number of our results involve the effects of our treatments on the likelihood of quitting.

We vary subjects’ information about the coding task along two dimensions: (1) the relationship between their performance and their bonus and (2) likely performance at the coding task.

There are five treatments varying what subjects know about the relationship between their performance and bonus. The first two involve no ambiguity while the latter three introduce increasing levels of ambiguity. The true bonus system for all but the first treatment was the 9¢ piece rate system.

1. (Fixed) Subjects receive a fixed and known amount for the coding task.
2. (Piece Rate) Subjects receive a known piece rate (9¢) for each message coded correctly. They know the rule for determining whether a coding is correct.
3. (Low Ambiguity) Subjects know the rule for determining whether a message is coded correctly. They know the minimum ($0.00) and maximum ($1.35) possible bonus and that their bonus depend on how well they do at coding, but do not know the relationship between the number of messages coded correctly and the bonus (low ambiguity).

² As part of the consent process, subjects in lab experiments are routinely told that they can leave at any time without consequences. Our casual observation is that subjects virtually never take advantage of this. The procedures in lab experiments that come closest to quitting in AMT are real effort experiments where subject can leave the lab as a natural part of the experiment (i.e. Abeler, Falk, Götte, and Huffman, 2011; Rosaz, Slonim, and Villeval, 2012) or can earn money by not doing the experimental task (i.e. Cooper, D’Adda, and Weber, 2013).
4. (High Ambiguity, 135) Subjects do not know the rule for determining whether a message is coded correctly. They know the minimum ($0.00) and maximum ($1.35) possible bonus but have no information about how this bonus is determined.

5. (High Ambiguity, 100) This is the same as the preceding treatment, except subjects are told that the maximum possible bonus is $1.00.

All subjects coded five practice messages prior to selecting a task, providing a measure of their ability at the coding task. In the fixed, piece rate, and low ambiguity treatments we vary whether subjects receive feedback about their performance on the practice messages (i.e. how many practice messages they coded correctly). We never gave feedback in the high ambiguity treatment as this would have undone some of the ambiguity in this treatment.

To generate hypotheses about treatment effects, we adapted the smooth ambiguity model of Klibanoff, Marinacci, and Mukerji (2005) to our experiment. The model predicts that pure reductions in ambiguity (i.e. mean preserving contractions) will increase willingness to select the coding task. This implies that the take-up rate for the coding task will increase as we move from high ambiguity ($1.35 maximum bonus) to low ambiguity to the piece rate treatment. The model also predicts that the take-up rate will decrease in the high ambiguity treatment when the maximum possible bonus is reduced from $1.35 to $1.00. The predicted effect of feedback depends on whether the primary effect of feedback is to reduce ambiguity or to reduce subjects’ overconfidence about their ability at the coding task. In the former case the effect on take-up should be positive while in the latter case a negative effect is predicted. The theory does not make any clear predictions about performance at the coding task, but we conjectured that the number of quits would increase and the number of messages coded correctly (subject to choosing the coding task) would decrease with higher ambiguity.

Subjects’ choices over tasks responded strongly to varying their information, but not always as predicted. Relative to the piece rate treatment, subjects were significantly less likely to select the coding task in the low ambiguity treatment. Holding the maximum possible bonus fixed at $1.35, subjects were significantly more likely to select the coding task with high ambiguity than with low ambiguity. Willingness to select the coding task under high ambiguity responded strongly to the maximum possible bonus, as predicted, with the proportion picking the coding task dropping significantly when the maximum bonus was reduced from $1.35 to $1.00. Providing subjects with feedback about their performance on the practice messages also significantly lowered willingness to select the coding task, although only in the cases where performance is relevant for the bonus (i.e. not the fixed bonus treatment). This suggests the primary effect of giving feedback is to reduce subjects’ confidence about their abilities, a conjecture supported by additional analysis of the data.
Turning to performance, we consider both the likelihood that a subject quits without completing the coding task and performance at the coding task. Relative to the piece rate treatment, both high ambiguity treatments (but not the low ambiguity treatment) significantly increased the probability of quitting the coding task. Giving subjects feedback about their performance on the practice questions also significantly increased quits in the treatments where performance is payoff relevant. Both high ambiguity treatments lowered the number of messages coded correctly, but this effect falls below significance if the sample is limited to subjects who complete the coding task and completely vanishes with controls for the ability of subjects with controls for the ability of subjects. Like Dohmen and Falk (2011), selection plays a central role in generating treatment effects on performance. What is novel is the important role played by quitting as a determinant of performance.

The fixed payment treatment was included in the experimental design to confirm that subjects’ choices respond to the relatively low incentives used in an AMT experiment. As predicted, there is a strong negative relationship between the probability of choosing the coding task and the value of the outside option. The likelihood of choosing the coding task is significantly lower in the fixed payment than the piece rate treatment even though the piece rate is calibrated to yield almost identical payoffs across treatment. We attribute this to high levels of confidence among our subjects, leading them to overestimate their likely pay with a piece rate. The fixed payment treatment yielded by far the highest rate of quitting, although ambiguity averse subjects were significantly less likely to quit than others.

Our paper contributes to several existing literatures. There is a growing literature that uses real effort tasks to study the effects of labor contracts on selection and/or performance. Notable recent examples include Cadsby, Song, and Tapon (2007), Niederle and Vesterlund (2007), Eriksson, Teyssier, and Villeval (2009), Dohmen and Falk (2011), and Gill, Prowse, and Vlassopoulos (2012). Our paper differs from existing work along two dimensions. First, we are less interested in the form of the incentive contract (fixed vs. variable) than the information subjects have about the incentive contract. Our work establishes that changing subjects’ information about either the relationship between performance and pay or their likely performance strongly affects how well a contract works, both in terms of attracting and keeping workers. Some of these effects go in the anticipated direction, such as low ambiguity making a contract less attractive relative to a known piece rate. Others do not; for instance, we did not anticipate high ambiguity making a contract more attractive relative to low ambiguity. Second, our experiment examines an aspect of performance that matters a great deal for real firms but has received little previous

---

3 In the United States, somewhat more than two million individuals quit their jobs each month (Bureau of Labor Statistics, 2013). One recent survey reported that “... the number of employees making a New Year's resolution to get a new job jumped to 38 percent [for 2013].” (Perman, 2013). Further, a recent news article reports employee turnover costs the employer approximately 16 to 20 percent of the employee’s annual salary, with percentage being larger for high salaried workers (Lucas, 2012).
attention in the experimental literature: quitting the task.\footnote{There exist a number of papers in the labor literature on quitting (e.g. Griffeth, Hom, and Gaertner, 2000; Abowd, Kramarz, and Roux, 2006; Abraham, Spletzer, and Harper, 2007; Farber, 2010; Buchinsky, Fougere, Kramarz, and Tchernis, R, 2010). There also exists a related experimental literature on quitting in tournament settings (Eriksson, Poulson and Villeval, 2009; Fershtman and Gneezy, 2011; Rosaz, Slonim, and Villeval, 2012).} We would miss an important element of how the treatments affect worker performance without data on quitting. The high ambiguity contracts, for instance, looks far less attractive when quits are considered and the fixed payment contract is particularly disastrous because of the high level of quitting.

While not the primary purpose of our work, we also contribute to the literature studying choice and ambiguity. Ellsberg (1961) famously identified patterns in decision making which could not be reconciled with subjective expected utility maximization. Numerous theories have been developed to explain ambiguity aversion,\footnote{For examples, see Hurwicz, 1951; Segal, 1987; Gilboa and Schmeidler, 1989; Schmeidler, 1989; Ergin and Gul, 2004; Halevy and Feltkamp, 2005; Ahn, 2008; Klibanoff, Marinacci, and Mukerji, 2005 and 2009; Nau, 2006; Seo, 2009; and Cerreia-Vioglio, Maccheronia, Marinacci, 2011.} and a number of experiments have tried to sort out between the various theories. The results typically indicate a great deal of heterogeneity, with many subjects exhibiting ambiguity neutrality and no one theory explaining the behavior of all subjects who are not ambiguity neutral.\footnote{Halevy (2007) finds heterogeneity in the population, with many subjects displaying ambiguity neutrality and no one theory explaining the behavior of all ambiguity averse individuals. Ahn, Choi, Gale, and Kariv (2009) also find a high degree of heterogeneity, including many individuals who are ambiguity neutral. Although primarily concerned with the interaction between ambiguity and auctions, Chen, Katuscak, and Ozdenoren (2007) report evidence of ambiguity loving behavior. Binmore, Stewart, and Voorhoeve (2012) do not find evidence of ambiguity aversion per se, instead observing choice patterns consistent with the principle of insufficient reason. Charness, Karni, and Levin (2013) find that most subjects are ambiguity neutral, with a mixture of types for those who are sensitive to ambiguity. See also Charness and Gneezy (2010) and Abdellaoui, Baillon, Placido, and Wakker (2011).} While not intended to separate specific theories or to assign subjects to types, our results speak to this literature. Our subjects are, on average, sensitive to ambiguity. While many individuals are presumably ambiguity neutral, ambiguity aversion is an important phenomenon at an aggregate level. Subjects’ reactions to ambiguity appear to depend on the nature of the ambiguity. The effect of low ambiguity relative to the piece rate treatment is consistent with standard models of ambiguity aversion, but the positive effect of moving from low to high ambiguity does not fit well with such models. We conjecture that the type of extreme ambiguity subjects encounter in the high ambiguity treatments pushes them to adopt simple rules of thumb, such as valuing the coding task at the midpoint of the high and low possible bonuses. This would be consistent with the difference between the two high ambiguity treatments and would also explain why the coding task is surprisingly attractive with high ambiguity.\footnote{Our high ambiguity treatments differ in the best possible outcome rather than the worst. The strong effect we observe from changing the best outcomes suggests that subjects are not as focused on worst case scenarios as theories like the MMEU model of Gilboa and Schmeidler assume.} We cannot confirm this conjecture directly with our data, as the experiment was not designed to examine
the process underlying subjects’ choices, but the possibility that subjects’ decision rules depend on the degree of uncertainty is an interesting one that we hope to explore in future work.

The implications of our experiments for real world managers depend on the nature of ambiguity in field settings. We suspect the sort of modest ambiguity incorporated into the low ambiguity treatment is probably more realistic than the extreme ambiguity of the high ambiguity treatment. Workers presumably have an idea about the basis of evaluation, but often don’t know the precise mapping between performance and incentive pay. Our results imply that this type of ambiguity imposes costs on firms by making hiring more difficult (or more expensive), but doesn’t affect performance once workers are hired. Improving workers’ information about their likely performance also makes hiring more difficult, and affects performance by making quits more likely. Overly optimistic workers are a good thing from an employer’s point of view.

The usual caveats about generalizing results from experiments conducted in one specific setting apply here, and we don’t expect the effects of ambiguity to be as strong in all labor market as what we observe. We anticipate that similar effects are most likely in entry level labor markets. This conjecture is partially based on our results, as ambiguity has the most bite when subjects are deciding whether or not to take a contract, but also stems from the nature of ambiguity in employment contracts. Our subjects face uncertainty about the relationship between performance and pay or uncertainty about likely performance. As individuals gain experience, they are likely to learn along both of these dimensions, reducing any effects due to ambiguity.

Our work does not directly address why incentive contracts are designed to include ambiguity. A number of potential answers suggest themselves, including those generally advanced for incomplete contracts (i.e. complexity, flexibility to respond to unforeseen circumstances) as well as reasons related to other-regarding behavior (i.e. black box compensation is intended to limit infighting over compensation). Understanding why a contract would include ambiguity (as well as determining the optimal level of ambiguity) requires knowing how workers will respond to ambiguity in incentive contracts. The purpose of our paper is to take this first necessary step.

An unusual feature of our experiment is that all of the subjects participate through AMT. While use of AMT has become common in the social sciences, it remains rare in experimental economics. We could have conducted a standard laboratory experiment, but there are several reasons we choose to use AMT instead. First, using AMT means that our experiment falls somewhere between a lab experiment and a natural field experiment, gaining some of the advantages of each. As in a lab experiment, we have tight control over the environment. For example, we systematically vary the value of subjects’ outside

---

8 See Horton, Rand, and Zeckhauser (2011) for a lengthy discussion of using AMT for economic experiments as well as replications of well-known laboratory studies using AMT.
option (the bonus for the transcription task), making identification possible when we use the Heckman correction to control for selection. At the same time we take advantage of the naturally occurring features of AMT as an online labor market. Quitting, specifically, is not mentioned to subjects as part of the design but instead represents a normally available option within AMT. Second, use of AMT yields a more diverse population than is normally seen in the lab. Our subjects come from 31 different countries and vary widely in age, income, and educational background. Finally, running an experiment on AMT is inexpensive. We think this point is quite important for the future of experimental economics. It cost us less than $500 to gather the data for this study. If we had paid the standard rate for subjects participating in laboratory studies at Florida State University, the dataset would have cost $10 - $15 thousand. High costs of running experiments have long been a barrier to entry into the field. Most graduate students can afford to spend $500 on a study, but not $14,000. If experimental economics wants to continue to thrive, we believe it needs to embrace methods of gathering data that are less financially onerous.

There are tradeoffs involved in using AMT. Running experiments online, along with the peculiarities of AMT, raises the possibility that the researcher will lose control of the experimental environment. For examples, subjects might try to use bots, search the internet for helpful information, or collude with other subjects to complete the coding task. Running an experiment on AMT demands an enormous degree of care by the researcher to avoid loss of control. Large portions of this paper explain steps we took to maintain control and to check the data for loss of control. Potential users of AMT for economic experiments need to understand that using AMT has advantages (inexpensive, broad and interesting population, embedded within a real labor market), but is more challenging than running experiments in the laboratory if the researcher is being careful.

2. Procedures, Experimental Design, and Treatments

A. Amazon Mechanical Turk: Our experiment was conducted using the online labor market Amazon Mechanical Turk (AMT). AMT users participate as either workers or requestors (i.e. employers). Requestors post jobs, referred to as “human intelligence tasks” or “HITs”, on the AMT website along with a flat wage that is paid for completion. Workers see the HITs that are posted and, if the HIT appeals to them, complete it. Once the requestor approves their work (i.e. certifies that the work has been completed), the worker is paid the flat wage. The expectation is that workers will receive the flat wage as long as a reasonable attempt has been made to complete the HIT. The implied standard is low, as will be seen by our criterion for accepting work, and the flat wage does not serve as a performance bonus. Requesters may also pay workers a bonus beyond the flat wage. The use of bonuses is quite flexible and payment of a bonus is completely at the requester’s discretion. Requesters control how much information

---

9 To be honest, we did not conduct this experiment on AMT to save money. The senior co-author could afford to pay for similar experiments in the lab. We were motivated by interest in exploring the use of AMT to run experiments and a desire to run our experiments within a functioning labor market.
workers have about a bonus and when workers receive this information. Using bonuses as performance pay is a normal practice on AMT. Payments for both the flat wage and the bonus are made electronically through Amazon.com accounts.

We are not the first researchers to use AMT or other online labor markets (i.e. Odesk, Elance, and Guru) as a platform for conducting experiments, although this remains relatively rare within economics.\textsuperscript{10} Use of AMT comes with advantages and disadvantages. The most obvious benefit is the low cost of running experiments. HITs in AMT typically pay 5 to 50 cents for less than 20 minutes of work (Buhrmester, Kwang, and Gosling, 2011). Our entire dataset cost less than $500 to gather. Obtaining a similar dataset at the rates we typically pay subjects in the lab would have cost somewhere between $10 - $15 thousand dollars. While cost should not be the only consideration in choosing how to conduct research, AMT can play a useful role in opening up experimental economics to researchers who do not have large budgets.

A second advantage of AMT is that it allows researchers to access a more diverse set of subjects than are typically present in laboratory experiments. Table 1 summarizes the characteristics of our subjects. Our subjects’ ages range from 18 to 69. A total of 31 countries are represented with the majority coming from India and a large minority from the USA. Most are college educated and 40% report incomes in excess of $25,000 per year, a figure which jumps to 70% for American subjects.\textsuperscript{11}

Finally, as noted previously, using AMT means our experiments take place within an existing labor market. This lets us take advantage of naturally occurring features of this market, in particular the ability of subjects to quit the experiment after accepting a contract.

\textbf{[Insert Table 1]}

There are several concerns with experiments run on AMT. The scale of payoffs is much lower than experimenters typically use in laboratory studies. The average subject who completed the experiment only earned $.62. This is reasonable compensation by AMT standards, but low compared to what a college student in the US or Europe typically gets paid for participating in an experiment. Even given that this was a fast experiment (subjects averaged 23 minutes to complete the entire study), the average earnings per hour are only $1.62/hour. This raises the possibility that the incentives were not salient for the subjects. Three factors argue against this. First, virtually all of our subjects (97%) report that earning money is one of their motivations for taking HITs on AMT.\textsuperscript{12} 24% of our subjects report that AMT is

\textsuperscript{10} Economics studies using online labor markets include Chen, Ho, Kim (2010), Liu, Yang, Adamic, and Chen (2011), Paolacci, Chandler, and Ipeirotis (2010), and Horton, Rand, and Zeckhauser (2011). The latter two papers provide excellent discussions of using AMT to conduct economic experiments.

\textsuperscript{11} Indian subjects had significantly lower incomes than the other subjects (t = 12.00; p < .01). Only 24% of Indian subjects report an income in excess of $25,000 as opposed to 62% of other subjects.

\textsuperscript{12} The pre-experimental survey asked subjects why they complete tasks on AMT. Subject chose from six options which were drawn from the survey in Ipeirotis (2010). Subjects could check as many or few items as desired. We
their primary source of income. Second, as will be seen below, our subjects respond strongly to changes in incentives. This suggests that they were aware of the incentives and regarded them as an important component of their decision making. In other words, the incentives were salient. Finally, it is important to not look at the earnings solely through the eyes of a university professor working in a wealthy nation. The majority of our subjects are Indian. To point out the obvious, India is a relatively poor country. GDP per capita is about $48 thousand/year for the US, $44 thousand/year for Germany, and $46 thousand/year for Japan. These figures dwarf India’s per capita GDP of about $1500/year.\footnote{GDP figures retrieved February 11, 2013 using the World Bank’s World Development Indicators Online (WDI) database.} Average pay for a college graduate in India is only 547 rupees/day, which works out to slightly more than $10/day.\footnote{National Sample Survey Organisation. (2011).} A large fraction of our subjects are not wealthy. Even eliminating subjects under 25 years old (many of whom are probably students), 35\% of our subjects report earning less than $12,500/year. This figure jumps to 51\% for the Indian subjects. For many of our subjects, the hourly pay probably does not seem terribly low.

A second major concern is loss of control due to conducting the study online. Unlike a laboratory study, we cannot be sure that subjects are not consulting outside sources. We intentionally choose a task where consulting the internet would not be helpful. Subjects were not told that Charness and Dufwenberg (2006) was the source of the game and the messages and were not given any information that would make it easy to find this information via a Google search. Beyond looking at material directly relevant to the experiment, subjects could have been distracted by browsing the web during the experiment. We have no way of knowing whether our subjects had distractions while they participated in the study, but this would be uncorrelated with the treatments in the experiment, and casual empiricism suggests that participants in real world labor markets often spend time browsing the internet when they should be working.

AMT has a number of specific features that differ from those faced in a typical lab setting. Many elements of our design were intended to prevent a loss of control due to these features. These AMT specific modifications are described in Section 2.4. Some of AMT’s oddities we viewed as interesting features rather than bugs to be eliminated. For example, the ability to quit midway through the experiment yielded valuable data that helps us understand the effects of ambiguity in incentive contracts.

Like any population used by an experimenter, using workers on AMT has its advantages and disadvantages. The best known study comparing AMT and laboratory experiments, Horton, Rand, and Zeckhauser (2011), replicates a number of standard experiments on AMT and report results that look little...
different from the typical results of laboratory experiments. We doubt that the use of AMT had a major effect, one way or the other, on the conclusions we reached from our experiments.

Figure 1

B. Tasks: Workers had the option of participating in two tasks: coding messages (Task A) or a simple transcription task (Task B).

Task A used data from Charness and Dufwenberg (2006). In the Charness and Dufwenberg experiments, subjects played the trust game with hidden actions shown in Figure 1. Player A starts the game by choosing whether to trust Player B (by choosing In) or not (by choosing Out). If Player A chooses In, Player B can either be trustworthy (by choosing Roll) or not (by choosing Don’t). The chance element following Roll makes it impossible for Player A, who only observes his own payoff, to determine for certain that Player B has not been trustworthy if a payoff on 0 is observed.

The critical feature of the Charness and Dufwenberg data for our experiment was pre-game messages sent by Player Bs. These can be used by Player B to persuade Player A to choose “In”, often by promising to choose “Roll”. Charness and Dufwenberg report all of the messages verbatim in the published version of their paper. Task A asked subjects to code fifteen of these messages. For each message, the subject was shown a jpg image of a hand-written copy of the text (see Figure 2). We used hand-written copies of the messages as a form of human verification. This makes it more difficult to copy and paste the text into a search engine or a translation program, as well as making impossible for an internet bot to complete the task. Subjects were asked to classify messages into six possible categories,

15 The messages were hand copied by multiple RAs, so the handwriting varied across messages. Common examples of human verification are the CAPTCHAs commonly seen on websites that are vulnerable to exploitation by bots (e.g. websites selling tickets to popular events, popular on-line email systems like gmail or Yahoo! mail). These are designed to verify that information is being entered by a human user rather than being generated by an internet bot.
shown below the text of the message. These categories were drawn from a pre-experimental coding of the fifteen messages done by the co-authors. Subjects were free to check multiple categories for a message, a possibility that was stressed in the instructions. They also had the option to indicate that none of the six categories applied. All subjects choosing Task A saw the same fifteen messages in the same order. The fifteen messages were chosen to confront the subjects with a broad variety of messages; the messages vary widely in their content, length, and difficulty of coding.

We scored a message’s coding as correct if it matched the modal coding for subjects who coded that particular message. This scoring rule is based on the coding methodology developed by Houser and Xiao (2011). Coding is an inherently subjective exercise, but for all fifteen questions the modal coding is similar to what we would have chosen as informed experts. In other words, we think an objective expert would agree that subjects credited with a higher number of correctly coded messages had performed better at the coding task. The results section discusses how the results change with a broader measure of performance at the coding task – the effect is minimal.

We chose this particular task, coding messages from experimental data, for several reasons. We wanted a real effort task that could easily be done in an online environment, but not one where the subjects could use access to the internet or their computer to get help. Part of the experimental design involves giving subjects information (based on practice questions) about their ability to perform the task prior to choosing a task. We therefore needed a task where performance could be evaluated in a clear fashion, but a subject would not be able to determine their performance without the experimenters giving them information. Given the subjective nature of coding messages, there is no way a subject could know

---

16 For providing feedback on the practice questions, we used data from an initial set of the sessions without feedback to determine whether a coding was correct. The modal coding of the five practice questions for this subset of the data matches the modal coding for all experimental participants, so the feedback was accurate.

17 For some messages, the subjects split almost evenly between a small number of codings. In these cases there isn’t an obvious “best” coding, but the modal coding is always in the set of obvious candidates for the best coding.
their performance without being told. This inherent ambiguity matched the type of situation we had in mind in the real world, where it isn’t necessarily obvious how well somebody is doing at their job.

The Charness and Dufwenberg game was specifically chosen for several reasons. The messages are publicly available and de-identified, easing IRB approval of the study. The game is relatively simple and easy to explain. The messages are short enough to be easily read and coded, but categorization is sufficiently difficult that subjects were not faced with a trivial task.

Figure 3: Task B Example

In Task B, the transcription task, workers were shown a series of fifteen words (see Figure 3). Each word had to be typed into a text box. The words were shown as JPG images to prevent copying and pasting or use of bots. We used typed text for the JPG image rather than hand-written versions to facilitate transcription. The transcription task is intended to be a sure thing, and we had no cases where a worker incorrectly transcribed a word. Task B takes the same amount of time to complete as Task A, a point which was emphasized to the subjects, and requires some effort (albeit less than Task A) as the subject must pay attention to new words appearing and type them correctly into the text box.

C. Procedures and Experimental Design: Experimental “sessions” were posted on AMT in small batches of 15 to 25 individuals. We made no attempt to hide that workers would be participating in a study. Given the large number of studies posted on AMT and the nature of the tasks workers were asked to complete, it seemed unlikely that workers would fail to guess that they were in a study. Workers were told up front that the experiment consists of two phases, with the first phase of the experiment paying a 20 cent flat wage for completing a survey and the second phase of the experiment paying an unspecified bonus for doing a task. They were subsequently told that the first phase would last 5 minutes and that the second would take 8 – 10 minutes. (These estimates were optimistic as subjects spent more time than expected for the second phase.) It was stressed to the subjects that they would only be paid if they completed both phases of the experiment.
Figure 4 illustrates the timing of the experiment. After indicating consent, workers began the first phase of the experiment by filling out a survey. This collected demographic information (age, gender, nationality, income, and educational background) and their reasons for working on AMT. The survey also contained a questionnaire measuring their risk attitudes, and a measure of ambiguity aversion. The questions measuring of risk attitudes, taken from the German Socio-Economic Panel (GSOEP), ask about real world decisions such as how willing they are to take risks when driving or how much they would put in a risky investment if they had a financial windfall. Dohmen, Falk, Hoffman, Sunde, Schupp, and Wagner (2011) have established the behavioral validity of these questions as a measure of risk attitudes. To measure ambiguity aversion, we presented subjects with the standard pair of hypothetical questions used to generate the Ellsberg paradox:

“Suppose there is a bag containing 90 balls. You know that 30 are red and the other 60 are a mix of black and yellow in unknown proportion. One ball is to be drawn from the bag at random. You are offered a choice to (a) win $100 if the ball is red and nothing if otherwise, or (b) win $100 if it's black and nothing if otherwise. Which do you prefer?”

“The bag is refilled as before, and a second ball is to drawn from the bag at random. You are offered a choice to (c) win $100 if the ball is red or yellow, or (d) win $100 if the ball is black or yellow. Which do you prefer?”

Choices of (a) and (d) are inconsistent with subjective expected utility and can be interpreted as evidence of ambiguity aversion (e.g. Hogarth and Villeval, 2010; Dominiak, Dürsch, and Lefort, 2012).

Copies of all experimental materials, including the survey, can be seen in Appendix A.

After completing the survey, subjects were given descriptions of Task A (coding) and Task B (transcription). The description of Task A includes descriptions of Charness and Dufwenberg’s trust game with hidden actions, how to perform the coding, and the coding categories. Subjects were shown several examples of messages and coding. We intentionally did not give subjects a citation to Charness and Dufwenberg or the name of the game to make it difficult to search for additional information on the web. The description of Task B (transcription) was shorter given the trivial nature of the task, and

---

18 A full copy of the survey, along with all of the experimental materials, is contained in Appendix A.
19 We also considered using Budner’s (1962) measure of ambiguity attitudes. This has the advantage of being less abstract than the Ellsberg questions, but Budner’s questions are aimed at a much broader definition of ambiguity than what most decision theorists would have in mind.
focused on making it clear that Task B was a sure thing. Before selecting a task, workers coded five messages for practice. Workers were then asked how they thought they did at coding the practice messages in comparison to other workers. The wording of this question was deliberately vague to avoid reducing the ambiguity in the high ambiguity treatments (“How well do you think you performed the practice task in comparison to other Turkers?”). Responses were on a Leikert scale ranging from 0 (“Well below average”) to 4 (“Well above average”), with 2 representing “average” performance. The mean response was 2.43, indicating subjects on average thought they were better than average. Men were significantly more confident than women (t = 2.39; p < .05), and Indians were significantly more confident than other subjects (t = 4.73; p < .01). Regardless of treatment, subjects did not have information about their performance on the practice questions prior to answering this question.

At this point, subjects began the second phase of the experiment by choosing between Task A or Task B. Task B served as an outside option. It had a fixed and known payoff. We randomly varied the payoff for Task B between 30, 40, and 50 cents. This was resolved in advance and subjects knew the value of the outside option. Based on calibration exercises we expected that the average bonus for Task A would be about 40 cents, so the average payoff for Task B was set to roughly match the average for Task A. Varying the payoffs for Task B gives us an exogenous source of variation that can be used to identify a selection equation. The instructions stressed that Task B is a sure thing and takes the same amount of time as Task A. In other words, subjects were given no reason to base their task choice on one task taking less time to complete than the other.

Task A consisted of coding 15 messages. Subjects were forced to spend at least 30 seconds on each message before they could continue. This paralleled Task B, where subjects had to spend at least 30 seconds on each screen before the next word appeared. There were 15 words in Task B, so both tasks required 7½ minutes to complete if subjects did not go beyond the required 30 seconds. Although subjects consistently ran over the required 7½ minutes, the average time to complete the experiment was the same for Task A or Task B. Consistent with what subjects were told when choosing a task, the two tasks required the same amount of time.

The treatments varied what information subjects had about Task A prior to choosing a task. We varied information along two dimensions: what subjects knew about how the bonus was determined and what subjects knew about their own likely performance. This is a between subjects design, so each subject chose between tasks only once and participated in a single treatment.

[[Insert Table 2]]

---

20 They were not told that other values were possible for the outside option to avoid any mistaken impression that the outside option was a gamble.
21 By pure luck, the average bonus for Task A in our data was exactly as calibrated.
Table 2 summarizes how many subjects were recruited for each cell of each treatment and how many usable observations were generated for each cell. With a few exceptions described below, 35 subjects were recruited for each cell. To prevent repeat observations, we blocked AMT users who had already participated in the experiment. We dropped 19% of the observations (162/863) as unusable. Virtually all of these observations (151/162) were dropped because subjects failed the screen for English comprehension as described below. The remaining 11 subjects provided such incomplete information as to make their data unusable.22

Treatments: Information about Determination of Bonus

Fixed: Task A pays 40 cents regardless of performance. This contract is a sure thing and gives no incentives for performance. This was included to confirm that subjects were responsive to incentives. Even though both tasks take the same amount of time, we did not expect subjects with fixed pay for Task A to always choose the task with the higher payoff since subjects may be willing to sacrifice money to do an easier task or one that is more interesting. Nevertheless, the willingness to do Task A should be a decreasing function of the pay for the outside option (Task B).

Piece Rate (PR): Subjects were paid a piece rate of 9 cents for each message coded correctly. The instructions for choosing a task told subjects the number of messages to be coded, the 9 cent piece rate, and the minimum ($0.00) and maximum ($1.35) possible bonuses from Task A. They were given detailed instructions about the rule for determining whether a message was coded correctly. The goal was to eliminate any ambiguity about the incentive contract. The PR treatment gives us a baseline to determine the effects of ambiguity about the incentive contract.

Low Ambiguity (LA): Subjects were paid a piece rate of 9 cents for each message coded correctly. The instructions included detailed instruction about the rule for determining whether a message was coded correctly. Subjects were not told the piece rate, the number of questions, or even that a piece rate system was being used. Instead, they were informed the bonus would “depend on how well you do at coding messages correctly.” In other words, subjects knew that there would be a positive relationship between how many messages were coded correctly and their bonus, but had no information about that relationship might be. They were told the minimum ($0.00) and maximum ($1.35) possible bonuses from Task A, but since they did not know the number of messages it was not possible to infer the piece rate or even that a piece rate system was being used. The incentive contract in the LA treatment is more ambiguous than the incentive contract in the PR treatment. In the PR treatment, a subject who knows the probability of coding a message correctly also knows the distribution of bonuses. In the LA treatment, knowing the probability of coding a message correctly is not sufficient to know the distribution of bonuses.

---

22 These eleven individuals filled out less than half of the survey (including the risk and ambiguity measures). Subjects either filled out the entire survey or filled out less than half of it.
**High Ambiguity (HA):** Subjects were paid a piece rate of 9 cents for each message coded correctly. The instructions included the minimum ($0.00) and maximum ($1.00 or $1.35) possible bonuses from Task A. Subjects were given no additional information about how the bonus would be determined. Specifically, subjects were not told the piece rate (or even that a piece rate system was being used), the number of questions, or the rule used to determine if a message had been coded correctly. The HA treatment represents a higher level of ambiguity than the LA treatment, since subjects have no information about how the bonus is determined beyond the minimum and maximum possible bonus. In the LA treatment subjects know how performance will be measured and that there is a positive relationship between performance and the bonus.

As a sub-treatment, two different possible maximum bonuses were used in the HA treatment. Initially we only used the true maximum bonus of $1.35 (HA135). As the data accumulated, we noticed that the take-up rate for Task A was surprisingly high in HA135 given the high level of ambiguity. We suspected subjects were using simple rules of thumb that would be sensitive to the maximum possible bonus. To test this, we added a cell to the HA treatment where subjects were told that the maximum possible bonus was $1.00 (HA100). This statement was essentially true; while subjects could earn as much as $1.35 in theory, in reality no subject ever earned more than $0.99 (and we had gathered hundreds of observations before beginning the HA100 treatment). Since the purpose of this sub-treatment was to determine the effect of changing the maximum possible bonus in the HA treatment, the sub-treatment was only run with an outside option of 50 cents and we over-sampled the two HA cells with an outside option of 50 cents to generate more power for making this comparison. 

**Treatments: Information about Likely Performance**

**Feedback:** Prior to choosing a task, subjects are told the number of practice messages they coded correctly. The correlation between the number of practice messages coded correctly and the number of messages coded correctly in Task A (for subjects who chose and completed Task A) is .51, so the number of practice messages coded correctly is a good predictor of performance at Task A. We did not include cells with feedback in HA treatment since giving subjects feedback requires giving them the rule for determining whether a message is coded correctly.

Receiving feedback reduces the uncertainty faced by subjects by reducing their uncertainty about the likelihood of coding a message correctly. For example, consider a subject in the PR treatment who considers it equally likely that their probability of coding any given message correctly is either 0, ½, or 1.

---

23 Looking at Table 2, observant readers will notice that one cell of the PR treatment was also over-sampled. This was not intentional – a HIT was accidentally posted twice on AMT. Three other cells were off by a couple of observations, which weren’t mistakes per se but instead are due to the oddities of how HITs are posted on AMT and our need to pay bonuses in a timely fashion.

24 The correlation between these variables is easily significant at the 1% level.
If they code one message correctly, the best and worst case scenarios can be eliminated. The subject still faces risk, but the uncertainty has been eliminated. Feedback also improves subjects’ calibration about their ability to code messages correctly, an important issue since (as noted previously) subjects are overconfident on average.

*No Feedback:* Workers do not receive feedback the number of practice messages coded correctly.

After selecting a task, subjects finished the experiment by performing their selected task. To get paid, subjects had to complete the task. Unlike a lab experiment, subjects could quit the experiment by leaving the website or prematurely submitting the HIT. We have 106 observations (out of 701 usable observations) where subjects quit after starting one of the tasks. The instructions and materials featured prominent reminders that subjects must complete the entire task to get paid, so we think it is unlikely that many quits are due to errors. Quitting a HIT is like quitting any job, and presumably reflects the subject’s judgment that the opportunity cost of completing the HIT outweighs the benefits. We view quitting as a natural feature of this and many employment environments. Rather than treating it as a bug, our approach has been to try to understand how the experimental treatments affect the likelihood of quitting.

Bonuses were paid within a week of completing the experiment. Subjects knew that payment of the bonus would be delayed. This is not unusual for AMT and would not have surprised the subjects.

*D) AMT Specific Design Issues:* Our experiments contained a number of features addressing AMT specific issues. AMT workers come from all over the world, and differ widely in their ability to comprehend English. The experimental materials informed subjects that English fluency was required to complete the HIT, but this requirement was unenforceable. The survey therefore included two questions testing workers’ English comprehension. Both questions had only one correct answer and the correct answer was given as part of the question. For example, one of these questions reads, “Over the weekend, Bob watched two football games. On the scale below, mark the number of football games Bob watched over the weekend.” Any subject who reads this question carefully and understands English should be able to answer it correctly.

There has been a persistent problem on AMT with the quality of work, especially for workers located in lower income countries. This was the primary reason cited by AMT for blocking new workers from outside of the US (http://turkrequesters.blogspot.com/2013/01/the-reasons-why-amazon-mechanical-turk.html). The concern stems primarily from workers who try to complete tasks as fast as possible to maximize their revenues, clicking buttons rapidly without spending much time to check whether their responses make any sense. AMT actively tries to screen out workers who have a history of providing poor quality work. Our HIT was not particularly attractive for this type of worker as it was relatively intricate and had built in timers that prevented workers from clicking through as fast as possible. Nonetheless, we have little doubt that we had workers whose sole goal was to complete the job as fast as
possible. The English comprehension screen also served as a screen for this sort of behavior. The answers for the two English comprehension questions were located in different columns. This prevented workers who were mechanically checking the first or last columns from passing the comprehension test without actually reading and understanding the questions.

We dropped observations from subjects who did not correctly answer the two English comprehension questions. Indian subjects were significantly more likely to be dropped than subjects from elsewhere. We suspect that most dropped subjects comprehended English but weren’t reading the questions carefully. Given the nature of our experiment, we did not want to use data from subjects who are not actually reading the material.

Tasks A and B use time delayed buttons. The button to move to the next message/word did not appear until 30 seconds had passed. This not only ensured that the minimum time required to complete each task was the same, but also made it impossible for workers to click through the HIT as fast as possible.

Another concern with using AMT is that subjects either used multiple accounts simultaneously to be paid for the task more than once or colluded with other AMT workers by sharing information about how to code the messages. To limit collusion, the posting of HITs was scattered over time with only a limited number of slots available at any point in time. This makes it difficult for a group to collude since only a small number of slots were posted at any given time and the pool of potential workers is large. Because subjects are paid via their Amazon account (each worker has a theoretically unique AMT account which has to be linked to a unique email address and paypal account), the same person could not participate more than once unless they used multiple accounts. Nonetheless, the concern remains that there were workers who either simultaneously worked on multiple accounts or colluded to complete the task.\(^{25}\) The nature of our experiment makes this less likely as it is relatively involved and time consuming. Moreover, workers who simultaneously use multiple accounts are exactly the type of workers who are likely to fail the English screen. Nevertheless, to alleviate concerns about collusion, we monitored online forums for evidence that our HIT was being discussed. We found no evidence that information about the HIT (specifically, information about how to code the messages or how the bonus was determined) was being shared. Ex post, we checked our data for autocorrelation. If subjects worked together or

\(^{25}\) We explored collecting the IP addresses of subjects. This ran into two problems: (1) Our IRB would not allow us to collect this information without telling subjects, which would have made many subjects unwilling to participate (including many who were doing nothing untoward), and (2) IP addresses do not uniquely identify individuals. Consider individuals using computers in a lab or connecting to the internet via a public wireless network.
simultaneously completed multiple versions of the experiment, we should see correlation across time of completion. There is no significant autocorrelation either in task selection or performance.\footnote{\textit{\cite{26}}} Workers in AMT can view the HTML file being used to generate the HIT. This creates the possibility that subjects could eliminate the ambiguity or otherwise game the system by viewing the code. Several features of the experiment limited this possibility. Each treatment of the experiment used a different HTML file. Subjects could therefore could not look at the HTML and gain more information than they were supposed to have by viewing the instructions for other treatments. Bonus payments were determined manually by the research team, so the HTML could not be used to determine the incentive contract or the correct coding for any particular message. In sessions with feedback, the portion of the HTML used for determining whether or not the worker correctly coded the practice messages was intentionally convoluted.\footnote{\textit{\cite{27}}} This made it extremely difficult to use the HTML to cheat on the practice messages. While there are no incentives to cheat on the practice messages, we thought it best to eliminate this possibility.

3. Theory, Hypotheses, and Conjectures: We model ambiguity using a multiple prior approach. Unlike more conventional models where individuals’ beliefs are given by subjective probabilities over possible outcomes, individuals’ beliefs are given by subject probabilities over states of the world corresponding to different probability distributions over outcomes, specifically over bonuses as a function of effort. Intuitively, individuals don’t know how bonuses are determined, but have beliefs over possible ways bonuses might be chosen. A rational individual would correctly reduce the compound lottery, exhibiting risk aversion but not ambiguity aversion, but multiple prior models assume that individuals are overweighting some priors relative to their subjective probabilities.\footnote{\textit{\cite{28}}} Overweighting states of the world associated with a low expected utility leads to ambiguity aversion.

The best known model of this type is the maxmin expected utility model (MMEU) of Gilboa and Schmeidler (1989) which puts 100\% weight on the worst prior, but many other variants exist including the smooth ambiguity model of Klibanoff, Marinacci, and Mukerji (2005) that we employ. In this model, the expected utility from each state is transformed by a concave function and then weighted by the subjective probabilities. Just as transforming expected payoffs with a concave function yields risk aversion, transforming expected utilities leads to ambiguity aversion. Unlike MMEU, the smooth ambiguity model puts weight on all priors. This gives the model two desirable properties: (1) Unlike

\footnote{\textit{\cite{26}}} It is unlikely that subjects were colluding across HITs broadly separated in time. Given the large number of tasks being posted at any time on AMT, a subject would have needed to be exceptionally attentive and motivated to notice that similar HITs were being posted repeatedly. Former participants cannot even see the ad for the HIT, making it less likely that they would try to get a friend to participate or switch accounts to sign in a second time.

\footnote{\textit{\cite{27}}} To be more specific, a large amount of extraneous code was added to the HTML. Finding the code that determined whether a specific practice question was correct would have been equivalent to finding the proverbial needle in a haystack.

\footnote{\textit{\cite{28}}} See Halevy (2007) for experiments linking ambiguity attitudes with the ability to reduce compound lotteries.
MMEU, reducing the subjective probability of the worst state affects behavior even if this state remains possible, and (2) changing the subjective probability of states other than the worst state affects behavior.

Adapting the Klibanoff et al model to our experiment, a subject faces a choice between an option with an unknown value (Task A) and a sure thing (Task B). Let \( b \in [b_{\min}, b_{\max}] \) be the bonus for Task A. If a subject chooses Task A, they also must choose an effort \( e \in [e_{\min}, e_{\max}] \) to put into coding messages. Let the cost of effort be given by \( c(e) \), where \( c'(e) > 0 \) and \( c''(e) > 0 \). We assume that costs of effort are known rather than a source of uncertainty. A subject’s payoff from Task A is given by \( b - c(e) \). Let \( K \) be the sure payoff from Task B.

Let \( \Delta \) be the set of possible states of the world. A state of the world gives the probability of each possible bonus as a function of effort. This subsumes both the relationship between performance and bonuses (i.e. the incentive contract) and the function determining performance as a function of effort (i.e. the subject’s ability at the coding task). The objective probability of earning \( b \) in state \( s \in \Delta \) subject to choosing \( e \) is given by \( f(s,b,e) \) and the expected bonus is calculated in the usual fashion. Let \( \phi \) be a differentiable function with \( \phi' > 0 \) and \( \phi'' < 0 \) and let \( \mu \) be the subjective probability distribution over states of the world. The subject’s value for Task A is given by (1). Task A is chosen if this value is greater than \( \phi(K) \).

\[
\max_{e \in [e_{\min}, e_{\max}]} \left[ \int_{s \in \Delta} \phi \left( \frac{\int_{b_{\min}}^{b_{\max}} (b \cdot f(s,b,e) - c(e))db}{\mu} \right) \right] \mu
\]

(1)

The assumption that \( \phi \) is concave implies ambiguity aversion in the following sense: holding effort fixed, a mean preserving spread (contraction) over the states of the world makes the coding task less (more) attractive versus a sure outside option. This leads to the following proposition implying that changes in information that represent a pure reduction in uncertainty (i.e. mean preserving contractions) will lead to more choices of Task A.

**Proposition 1:** Let \( \mu \) and \( \mu' \) be subjective probability distributions over \( \Delta \). Suppose a subject chooses the coding task and effort \( e^* \) under \( \mu \). If \( \mu' \) is a mean preserving contraction of \( \mu \), the subject will also choose the coding task under \( \mu' \).

**Proof:** Let \( e^{**} \) be the optimal effort if Task A is chosen under \( \mu' \). Since \( \phi \) is a concave function, (2) must be true.

\[
\int_{s \in \Delta} \phi \left( \frac{\int_{b_{\min}}^{b_{\max}} (b \cdot f(s,b,e^*) - c(e^*))db}{\mu} \right) \right) \mu' > \int_{s \in \Delta} \phi \left( \frac{\int_{b_{\min}}^{b_{\max}} (b \cdot f(s,b,e^*) - c(e^*))db}{\mu} \right) \mu > \phi(K)
\]

(2)

By revealed preference, (3) must be true and the result follows. Q.E.D.
\[
\int_{s \in \Delta} \phi \left( \int_{b_{\min}}^{b_{\max}} (b \cdot f(s,b,e^{**}) - c(e^{**}))db \right) d\mu' > \int_{s \in \Delta} \phi \left( \int_{b_{\min}}^{b_{\max}} (b \cdot f(s,b,e^*) - c(e^*))db \right) d\mu' \quad (3)
\]

Moving from HA135 to LA to PR reduces the uncertainty facing subjects. Comparing the HA135 and LA treatments, subjects are given two additional pieces of information in the LA treatment: (1) the rule for how performance is evaluated and (2) the bonus is an increasing function of performance. The latter eliminates two payoff rules that are quite plausible under high ambiguity: the experimenter automatically gives all subjects the worst possible bonus and the experimenter automatically gives all subjects the highest possible bonus. If the primary effect of moving from HA135 to LA is to reduce the subjective probability of these extreme states, Proposition 1 implies that the coding task becomes more attractive.\(^{29}\) Likewise, under the LA treatment there are an infinite number of monotonically increasing rules for determining the bonus. Some of these rules are quite bad for subjects (i.e a negligible piece rate for coding a message correctly) while others are more desirable (i.e. a step function giving the maximum possible bonus if any message is coded correctly). If moving from LA to PR primarily affects behavior by moving weight away from extreme states, Proposition 1 again implies that the coding task becomes more attractive.\(^{30}\) This leads to our first hypothesis.

**Hypothesis 1:** Controlling for subject characteristics and the value of the outside option, the probability of choosing Task A should be increasing as we shift from the HA135 treatment to the LA treatment to the PR treatment.

So far we have focused on the effect of information on uncertainty, but giving subjects information may also improve subjects’ calibration in the following sense. Let \(s^*\) be the true state of the world and let \(EB(s,e)\) be the objective expected bonus in state \(s\) subject to choosing \(e\). A subject’s calibration error (CE) as a function of \(e\) is defined by (4). This is the difference between a subject’s expected bonus, imposing ambiguity neutrality, and the objective expected bonus. If the calibration error is positive (negative), the subject is *overconfident* (*underconfident*).

\[
CE(e) = \int_{s \in \Delta} b_{\min}^{b_{\max}} (b \cdot f(s,b,e))db d\mu - EB(s^*,e) \quad (4)
\]

Giving subjects feedback about their performance on the practice messages is likely to both reduce uncertainty and improve calibration. Coding is an unfamiliar task for our subjects. They are unlikely to know how good they are at this task (i.e. they are unlikely to know the probability of coding a message correctly as a function of effort). Giving them feedback should reduce this uncertainty. There is also

\(^{29}\) It is not necessary that subjects place no weight on these possibilities in the LA treatment. This is the advantage of using the smooth ambiguity model rather than MMEU – for behavior to change it is sufficient that the subjective probability of extreme states is reduced.

\(^{30}\) Subject in the PR treatment can still have positive weight on states that always yield the lowest possible amount since he may believe their ability at the task is sufficiently low as to prevent earning a positive bonus.
reason to believe our subjects’ beliefs about their ability to code messages were miscalibrated. Numerous experiments have found that individuals are overconfident about their ability to perform tasks (e.g. Niederle and Vesterlund, 2007), and this seems to be true in our experiment as well since the average subject thought they were above average at the coding task.\textsuperscript{31} Giving feedback about performance on the practice messages should give subjects more accurate beliefs about their abilities. Put in a more technical fashion, the reduction in uncertainty is no longer mean preserving if the calibration error is reduced. Because reducing the uncertainty and reducing the calibration error tend to work in opposite directions, the predicted effect is inherently ambiguous. We present two opposing hypotheses about the effect of feedback on task selections. These prediction apply mainly to the treatments where the bonus is dependent on performance at Task A (PR and LA). In the Fixed treatment, the bonus does not depend on performance at Task A. Subjects may still care about their performance at the task due to intrinsic motivation, but we expect them to care less.

*Hypothesis 2a:* Controlling for subject characteristics, the value of the outside option, and information about the incentive contract, if the primary effect of feedback about performance on the practice questions is a reduction in uncertainty then the probability of choosing Task A should increase when feedback is provided in treatments where the bonus depends on performance (PR and LA).

*Hypothesis 2b:* Controlling for subject characteristics, the value of the outside option, and information about the incentive contract, if the primary effect of feedback about performance on the practice questions is a reduction in overconfidence then the probability of choosing Task A should decrease when feedback is provided in treatments where the bonus depends on performance (PR and LA).

The HA135 and HA100 treatments (with an outside option of 50 cents) only differed in what subjects were told was the maximum possible bonus. In HA100 the subjects are told that a bonus greater than 100 cents cannot be achieved while in HA135 they are given no reason to believe such bonuses are impossible. We can model this change in information as either changing beliefs about the state of the world or changing beliefs subject to the state of the world. We start with the latter approach as it is simple to incorporate into the model and leads to natural predictions. For each $s \in \Delta$ let $f(s,b,e,b)$ objective probability of earning $b$ in state $s$ subject to choosing $e$ and restricting the bonus to be weakly less than $b$ where $b < b_{\text{max}}$. Equation (5) derives $f(s,b,e,b)$ from $f(s,b,e)$ via Bayes Rule:

$$f(s,b,e,b) = \frac{f(s,b,e)}{\int_{b_{\text{max}}}^{b} f(s,\beta,e)d\beta} \quad \text{if } b \leq b$$

$$= 0 \quad \text{if } b > b$$

\textsuperscript{31} There is surprisingly little correlation (.038) between their beliefs about their ability and the number of practice messages coded correctly. This also suggests that miscalibration is common.
Let $EB(s, e, b)$ be the objective expected bonus in state $s$ subject to choosing $e$ and restricting the bonus to be weakly less than $b$. The following proposition implies that moving from HA135 to HA100, holding the outside option fixed, reduces the proportion of subjects choosing Task A.

**Proposition 2:** Assume that for all states the probability distribution of bonuses as a function of effort has full support over the range $[b_{\min}, b_{\max}]$. Suppose a subject chooses the sure thing (Task B) with a maximum bonus of $b_{\max}$. If the maximum bonus is reduced to $b$ and the subjective probability distribution over states is unchanged, the subject must still choose the sure thing.

**Proof:** For every state of the world $j$, $EB(s, e, b) < EB(s, e)$. Suppose the subject chooses Task A with the restricted bonus. Since $\phi$ is an increasing function, the value of Task A must be strictly higher without the restriction. A contradiction follows Q.E.D.

The preceding implicitly assumes that all states of the world remain possible with restriction of the maximum bonus. This need not be true. A trivial example is the state of the world where the bonus equals $b_{\max}$ regardless of performance. This state cannot occur if the bonus is restricted to a maximum of $b < b_{\max}$. Having some states become impossible is easily incorporated into the theory by allowing the subjective probability of states to change in a specific fashion. First, assign zero subjective probability to all states of the world that never yield a payoff less than or equal to $b$. Update the subjective probability of the remaining states of the world using Bayes Rule. For all remaining states, update the probability of each possible bonus using Bayes Rule as above. Proposition 2 extends to this richer version of the model.

**Hypothesis 3:** Controlling for subject characteristics and holding the value of the outside option fixed, the probability of choosing Task A should be decreasing as we shift from the HA135 treatment to the HA100 treatment.

The preceding hypotheses focus on the likelihood of choosing Task A, but many of the results presented in Section 4 concern performance, both in terms of the probability of quitting and the number of messages coded correctly. Unfortunately, the theory provides little guidance on treatment effects for these measures. Predicting how the treatments affect the number of messages coded correctly is equivalent to making a prediction about how the optimal choice of effort is affected by the treatments. This in turn requires knowing the sign of the third derivative of $\phi$. We can generate predictions by assuming specific functional forms for $\phi$, but there is no basis for making such assumptions. We therefore make no formal prediction. That said, we conjectured *ex ante* that reducing ambiguity would increase the number of messages coded correctly subject to choosing Task A. Intuitively, the marginal cost of increased effort is a sure thing while the marginal return is uncertain. A decrease in uncertainty should make the marginal return more attractive. Thus, we anticipated more messages would be coded correctly in treatment with lower ambiguity (PR vs. LA and LA vs. HA135). We had no strong conjecture about the effects of feedback on the number of messages coded correctly. Once again the issue
is one of uncertainty vs. calibration. If feedback makes subjects less confident about the abilities, it should lower the marginal benefit of effort on the coding task. This runs against any positive effect from reduced uncertainty.

Conjecture 1: Controlling for selection and subject characteristics, the number of messages coded correctly subject to choosing Task A should be increasing as we shift from the HA135 treatment to the LA treatment to the PR treatment.

The theory also provides minimal insight about treatment effects on quitting. Subjects who select Task A and quit have not received any additional information about the payoff rule beyond what they knew when choosing a task. If they have learned anything additional that is payoff relevant, it is about the difficulty of the task. It is hard to know how much information subjects gain about their ability by coding more messages since they receive no feedback about their performance while performing Task A, but we anticipate experience with Task A serves as a weak substitute for feedback about performance on the practice messages. We therefore conjectured that the likelihood of quits would be affected by feedback in the same direction as the effect of feedback on selecting Task A. In other words, if feedback makes subjects more (less) likely to choose Task A, it should make them more (less) likely to quit subject to having chosen Task A.\textsuperscript{32}

More realistically, what we think subjects learn from performing Task A is how they feel about doing the task. Unlike the practice messages, the messages in Task A are on 30 second timers. This was chosen to prevent subjects from sprinting through the coding task, but is probably more time than most individuals need. If a job is boring and the pay doesn’t look so attractive, it isn’t crazy to walk away. We therefore anticipated more quits when there is more uncertainty about the bonus – in other words, we expect higher ambiguity to lead to more quits.

Conjecture 2: Controlling for selection and subject characteristics, the probability of quitting subject to choosing Task A should be affected by feedback in the same direction as the effect of feedback on selecting Task A. We expect fewer quits as we move from the HA135 treatment to the LA treatment to the PR treatment.

4. Results: Table 3 summarizes the data by treatment. All results are based on the 701 subjects who passed the English screen and filled out the survey. We report results on both whether subjects selected Task A and whether they selected \textit{and completed} Task A, as the latter measure is arguably more important for firms recruiting workers.

[Insert Table 3]

\textsuperscript{32} To understand why these are the predicted effects, note that the effect of seeing more messages should largely occur when the subject does \textit{not} receive feedback about their performance on the practice messages. Suppose feedback tends to make subjects more pessimistic, choosing Task A less. In the absence of feedback, seeing additional messages should also make them more pessimistic and hence more likely to quit.
A. Task Selection and Quits: Ambiguity about the incentive contract affects task selection, but not exactly as expected. Subjects are less likely to select Task A in the LA treatment than the PR treatment, consistent with Hypothesis 1, but the high rate of subjects choosing (or choosing and completing) Task A in the HA135 treatment relative to the LA treatment runs contrary to this prediction.

The theory did not make a clear prediction for how feedback about performance on the practice messages would affect task selection. Table 3 shows that feedback consistently lowers the percentage of subjects who select (or select and complete) Task A. The effect of feedback is larger in the pooled data from the PR and LA treatments where performance is relevant for a subject’s earnings. These results are consistent with Hypothesis 2B, suggesting that the primary effect of feedback is a reduction in calibration errors, or, more specifically, a reduction in overconfidence.

Task selection under high ambiguity is sensitive to the maximum possible payoff, as predicted by Hypothesis 3, with choice of Task A less common in HA100 than HA135. This difference remains large if we limit the data to observations with an outside option of 50 cents (56.9% select Task A in HA135 vs. 36.7% in HA100).

Comparing outside options of 30, 40, and 50 cents, the percentages choosing Task A are 19.0%, 35.0%, and 47.4% respectively in Fixed, the one treatment where the incentives are completely unambiguous. This provides strong evidence that the incentives are salient for subjects. It is mildly surprising that many subjects choose Task A even when it unambiguously lowers earnings, but presumably these subjects derive enjoyment from doing the more interesting task. The percentage selecting (or selecting and completing) Task A is lower in Fixed than any other treatment. This makes sense; if subjects are overconfident on average, they will systematically overestimate their earnings from Task A in the other treatments.

The theory did not yield clear hypotheses about the likelihood of quits, but we conjectured quits would increase with greater ambiguity. This is true, particularly with high ambiguity. We also predicted that the effect of feedback on quits would have the same sign as the effect on selecting Task A. This prediction was incorrect, as feedback decreased the likelihood of selecting Task A and increased quits.

Definitive conclusions about task selection and quitting cannot be drawn from the raw data summarized in Table 3. By design the value of the outside option varies between subjects. This was intended to be balanced between treatments, but the random pattern of who fails the English screen as well as oversampling in the HA treatment means that the sample is not truly balanced. Likewise, even with random assignment there is variation in subjects’ demographic characteristics across treatments. To make statistically valid statements about the effects of the treatments on task selection, Table 4 reports the results of regressions examining whether subjects choose Task A, quit subject to having chosen Task A, and, combining task selection and quits, choose and complete Task A.
All regressions on Table 4 include a dummy for subjects who did not indicate that earning money was among their reasons for completing tasks on AMT. These subjects were significantly less likely to select Task A and significantly more likely to quit if they chose Task A. These estimates are not reported to save space, but full copies of the regression output are attached as an appendix.

Models 1 and 2 are probits with a dummy for whether the subject chose Task A as the dependent variable. Robust standard errors are reported in parentheses. Model 1 estimates treatment effects while controlling for the value of the outside option. The base is the PR treatment without feedback. Note that there are two separate dummies for feedback, with the effect for the FIXED treatment estimated separately from the effect for the PR and LA treatments. Since performance does not affect the bonus in FIXED, we predicted a weaker effect from feedback in this treatment.

The parameter estimate for LA is negative and significant at the 5% level in Model 1. Consistent with Hypothesis 1, low ambiguity lowers subjects’ willingness to choose Task A. HA135 has a positive effect relative to the PR treatment and the difference between the parameter estimates for the LA and HA135 treatments is significant at the 1% level ($z = 3.15; p \leq .01$). Contrary to Hypothesis 1, high ambiguity raises willingness to choose Task A relative to low ambiguity and holding the maximum possible bonus fixed. Task selection under high ambiguity is sensitive to the maximum possible bonus, consistent with Hypothesis 3, as the difference between the estimates for HA135 and HA 100 is significant at the 5% level ($z = 2.12; p \leq .05$). Both parameter estimates for the effect of feedback are negative, and the effect is significant at the 10% level for the PR and LA treatments. This is consistent with Hypothesis 2b. The effect of having a fixed payment for Task A is negative and significant at the 1% level. If the dataset is limited to observations from FIXED, the dummy for having an outside option of 50 cents is negative and significant at the 1% level ($z = 3.16; p < .01$). This confirms that subjects respond strongly to the incentives in the FIXED treatment.

Model 2 adds controls for subject demographics: age, gender, nationality, income, and education. For nationality we include dummies for India and the USA, with all other nationalities serving as the base. Given that Indians and Americans make up 88% of the sample, this captures most of the variation by nationality. The survey questions for income and education only allowed for categorical answers, so the independent variables are dummies for the various categories. Model 2 also includes controls for the subject’s risk and ambiguity attitudes as well as a control for their confidence in their ability to perform

---

33 As a simpler way of measuring the effect of changing the upper limit on payoffs in the high ambiguity treatment, we have run a multinomial logit using only observations from the HA135 and HA100 treatments with an outside option of 50 cents. The only independent variables in this regression are dummies for HA100 and not being motivated by earning money. The estimated effect of HA100 on the probability of selecting and completing Task A is negative and significant at the 5% level (parameter = -.860; std err = .395).

34 The base is an outside option of 30 cents.
Task A. The survey on risk attitude includes two general questions on risk attitudes as well six questions asking about risk attitudes in specific real world settings. The answers over these six questions are highly correlated, so we combined their answers into a single measure using factor analysis. We classified subjects as ambiguity averse, ambiguity neutral, or ambiguity loving based on their answers to the two Ellsberg questions. Subjects who were ambiguity averse were coded as a 1 while those who were ambiguity loving were coded as a -1. After completing the practice questions subjects were asked about their performance relative to other Turkers. We use this as our measure of confidence. The number of practice questions answered correctly is also included as an independent variable. This allows us to separate the effect of confidence from the effects of skill at Task A. To save space, Table 2 does not report most of the added variables in Models 2, 4, and 6 since they are neither statistically significant nor of great economic interest, but full regression output has been included in the appendix.

The main results are little changed between Models 1 and 2. The parameter estimates for Fixed and LA are still negative and significant, and the difference between HA135 and LA remains significant at the 1% level. The difference between HA135 and HA100 is slightly weaker but remains significant at the 10% level. Of the added variables, the only one that has a significant effect is gender as men are weakly more likely to choose Task A than women. We were surprised that neither the risk nor ambiguity measures were statistically significant.35

Models 3 and 4 look at the decision to quit subject to having selected Task A. These are probits with a Heckman correction since there is selection into the population of subjects that choose Task A. For the Heckman model to be identified, we need an exogenous source of variation that is correlated with decisions to select Task A but not with actions subject to choosing Task A. By design, the value of the outside option serves this purpose. The base is the PR treatment. The independent variables in Models 3 and 4 are the same as those in Models 1 and 2.36 Model 3 is a basic regression that only controls for the treatment while Model 4 adds controls for demographics and the behavioral measures.

Looking at Model 3, the HA100, and HA135 treatments both significantly increase the frequency of quitting. The LA treatment also increases quits, but this effect is not significant, so the data only provides partial support for Conjecture 2 vis-à-vis the effect of ambiguity on quits. Feedback also increases to frequency of quitting, albeit weakly, in the treatments where task performance is payoff relevant (PR and LA). This runs contrary to Conjecture 2. The treatment with the strongest effect on quits is FIXED, with the estimate easily significant at the 1% level.

The results are somewhat weakened by the inclusion of additional variables in Model 4, but the the HA100 and HA135 treatments remain statistically significant at the 10% level, as does feedback in the PR

---

35 The risk measures also fail to achieve joint significant ($\chi^2 = 3.84$, 3 d.f., p > .10).
36 The maximum likelihood estimation for Model 4 would not converge with separate dummies all six categories for education, so we combined categories to create three categories.
and LA treatments. Looking at the added variables, subjects who are less risk averse or less ambiguity averse are significantly more likely to quit. The reasons for these effects will become clearer when we break down the data by treatment. Subjects who are more confident about their ability to code messages quit significantly less.\(^ {37}\)

The dependent variable for Models 5 and 6 is a dummy for whether the subject selected and completed Task A. The specifications are otherwise identical to Models 1 and 2 respectively. The results are similar to those reported in Models 1 and 2, as should be expected since relatively few subjects quit. One notable exception is the negative effect of feedback in the PR and LA treatments, which is now significant at the 5% level. Feedback in these two treatments weakly decreases the proportion of subjects choosing Task A and weakly increases the percentage who quit. The combination of these two weak effects yields a strong effect on whether subjects select and complete Task A.

To further explore the effect of feedback, we reran Models 5 and 6 with interactions between feedback and the number of practice messages coded correctly (i.e. the feedback received by subjects). In the PR and LA treatments, this interaction term was positive and significant at the 10% level. The estimate for feedback remains negative and is now significant at the 1% level.\(^ {38}\) The result is quite intuitive. Consistent with feedback reducing calibration errors, subjects receiving bad feedback become less likely to select and complete Task A while those receiving positive feedback become more likely to do so.

Our main results on task selection and quitting are summarized as follows:

**Result 1:** Relative to the PR treatment, the LA treatment lowers the probability of choosing Task A. This is consistent with Hypothesis 1. Relative to the LA treatment, the HA135 treatment significantly raises the probability of choosing Task A. This is inconsistent with Hypothesis 1.

**Result 2:** For the PR and LA treatments (pooled), feedback about performance on the practice questions lowers the probability of choosing Task A. This is consistent with Hypothesis 2b. This effect is sensitive to the content of the feedback, providing further evidence that feedback reduces calibration errors.

**Result 3:** The HA100 treatment lowers the probability of choosing Task A relative to HA135. This is consistent with Hypothesis 3.

**Result 4:** Quits are more frequent in both HA treatments relative to the PR treatment, consistent with Conjecture 2. Feedback increases quits, contrary to Conjecture 2.

\(^ {37}\) The results of Model 4 could reflect overfitting. There are a large number of explanatory variables and only 35 observations where subjects drop out. As a check of our results, we modified Model 4 so the only variables added to Model 3 were the hypothetic investment question, the ambiguity aversion measure, and the confidence measure. All three variables remain significant at the 10% level or better and the estimated treatment effects are little changed.

\(^ {38}\) For the modified version of Model 5, the parameter estimate for the interaction term is \(.156\) with a standard error of \(.089\) while the parameter estimate for “Feedback * (Piece Rate + Low Ambiguity)” becomes \(-.606\) with a standard error of \(.231\). Similar results obtain for the modified version of Model 6.
B. Performance: Our measure of performance is how many messages were coded correctly in Phase 2. A message was counted as correct if the subject’s coding matches the modal coding for the population. This is the criterion that was used to determine whether or not a subject was paid the piece rate for coding a message correctly.

Returning to Table 3, the number of messages coded correctly is lower in HA135 and HA100 than in PR. This is especially true if all subjects who chose Task A are considered rather than limiting the dataset to those who selected and completed Task A. The FIXED treatment has an even more powerful effect on performance relative to PR than the HA135 and HA100 treatments.

We cannot reach any firm conclusions about treatment effects on performance without controlling for variation in subject characteristics across treatments. This is especially important because there is strong selection present in which subjects select Task A, particularly on the ability of subjects at the coding task. This can be seen in Figure 5 which graphs the average number of practice questions coded correctly, broken down by treatment (pooling data with and without feedback) and whether the subject choose Task A or Task B. For PR, Fixed, and HA100 there are clear gaps between subjects who choose Task A and

---

39 To be more specific, the population is subjects who are included in the dataset and coded the message in question.
those who choose Task B. For each treatment we ran a Wilcoxon rank-sum test to determine whether the distribution of performance on the practice questions (i.e. number of practice questions coded correctly) was different for subjects who choose Task A versus Task B. Three (***) and two (**), and one (*) stars shown above the bars indicate significant differences at the 1%, 5%, and 10% levels respectively. There are significant differences in the PR, FIXED, and HA100 treatments. Performance on the practice questions is a good measure of skill at Task A, so the results shown on Figure 5 strongly indicate that selection will play a role in determining treatment effects on performance.

Table 5 reports the results of regressions with the number of questions coded correctly in Task A as the dependent variable. To correct the standard errors for the effects of selection, these are Heckman selection models. As before, the value of the outside option provides an exogenous source of variation that is relevant for the selection decision but not post-selection.40 The left panel of Table 5 looks at the number of messages coded correctly by subjects who choose Task A, regardless of whether or not they completed the task. Many subjects who quit did so after coding some messages – the average number of messages coded correctly was 0.83 for subjects who quit (versus 4.36 for those who select and complete Task A). The right panel only considers the number of messages coded correctly by subjects who choose and completed Task A.

[Insert Table 5 here]

Model 1 only includes the treatment effects as independent variables. Both HA100 and HA135 have significant negative effects on performance, as does Fixed. Model 2 adds controls for demographics (age, income, education, and gender), risk and ambiguity attitudes, confidence, and practice questions coded correctly. Not surprisingly, the parameter estimate for performance on the practice messages is large, positive, and easily significant at the 1% level. With the additional controls, the treatment effects are greatly reduced and lose statistical significance. The absence of statistical significance is driven by controlling for how many practice questions were coded correctly as adding this variable alone is enough to eliminate any significant treatment effects. The effect of the treatments on coding performance is largely due to selection effects, echoing the results of Dohmen and Falk (2011).

Model 3 replicates Model 1, but only includes performance by subjects who choose and completed Task A. The effects of the HA100, HA135, and FIXED treatments are reduced in magnitude and only FIXED still achieves statistical significance. Model 4 parallels Model 2 – controlling for performance on the practice questions reduces the magnitude of all the estimated treatment effects, none of which even vaguely approach statistical significance. It is somewhat surprising that FIXED does not have an effect. Even though subjects in FIXED have no financial incentive to make an effort at the coding task, their

40 Other than the value of the outside option, the independent variables are same for the first and second stage of all the models reported on Table 5.
performance is not reduced from the PR treatment once we account for selection. The primary effect of
the incentive contract is on who does the job, not on how well they do the job.

Result 5: Relative to the PR treatment, subjects who choose Task A code fewer messages correctly in the
HA100, HA135, and FIXED treatments. These treatment effects are due to selection and vanish after
controlling for performance on the practice questions. The estimated treatment effects are also reduced if
the sample is restricted to subjects who select and complete Task A. The results are only weakly
consistent with Conjecture 1.

C. Additional Regression Results: As noted in Section 2.D, a potential concern with using AMT is that
subjects either simultaneously used multiple accounts or colluded on tasks. We think this is unlikely, for
reasons explained previously, but observations are not statistically independent if subjects game the
system in either fashion. This lack of independence should be local in nature – observations gathered at
about the same time are likely to be correlated, but not those gathered at widely spaced times. Our data
was gathered in small batches via fifty HITs spread across four months. We never had two HITs posted
at the same time and the HITs are separated in time. Each HIT can therefore be treated as a cluster. As a
robustness check, we reran the regression analysis from Tables 4 and 5 with the standard errors corrected
for clustering at the HIT level. This weakens the assumption that observations gathered at the same time
are independent, allowing for the possibility that some of the subjects in a HIT might have been colluding
(or might even have been the same person with different IDs). Correcting the standard errors for
clustering at the HIT level slightly strengthened the statistical significance of the main results, but had no
effect on our qualitative conclusions.

To examine whether the effect of feedback and subject characteristics vary across treatments, we
reran Model 6 from Table 4 using data from subsets of the data (FIXED, PR, LA, and HA100 pooled with
HA135). This exercise yielded several useful insights. First, we drew conclusions about the effects of
feedback based on pooling effects across the PR and LA treatments. Running Model 6 separately for the
PR and LA treatments, the feedback effect is significant in both cases. This implies that the effect is not
driven by just one of the two treatments. Second, ambiguity aversion has a positive and significant effect
on the probability of selecting and completing Task A in the FIXED treatment, but not in any other
treatment. Digging deeper, ambiguity aversion decreases the chance of quits significantly in the FIXED
treatment but not in other treatments. Quitting and searching for another task on AMT is a gamble.
When Task A is an unambiguous sure thing, ambiguity averse subjects are relatively unwilling to bear the
uncertainty of searching for a new task. Finally, the positive effect of subject confidence on the
likelihood of selecting and completing Task A comes primarily from the high ambiguity treatments.

The results in Table 5 examine how many messages were coded correctly in Task A, but for some of
the messages there was little difference between the frequency of the modal coding and the second most
frequent coding. As a check on our results, we generated an alternate performance measure that counted a coding as correct if the subject chose the modal coding or any coding whose frequency was within 5% of the frequency of the modal coding. Running the regressions from Table 5 with this alternative measure of performance leads to similar but slightly weaker results. Specifically, the estimate for HA100 is no longer significant in Model 1 and the estimate for FIXED is no longer significant in Model 3. The overall pattern of the results is unchanged, as the treatment effects largely vanish once a control for subject ability is added.

5. Discussion and Conclusions: Our paper has several important takeaway messages. First, ambiguity matters (but not always in the expected fashion). There are advantages to having a flexible incentive contract, but there can also be costs. The primary effects of ambiguity occur when workers are deciding whether to accept and complete a job. Assuming the modest ambiguity present in the LA treatment is representative of the ambiguity workers face in field settings, ambiguity is harmful from a firm’s point of view, making it harder to hire workers. The more extreme ambiguity of the HA treatment reverses this conclusion. Subjects seem to be using a simple rule of thumb that is easily manipulated. With an overly rosy view of the best possible outcome, workers accept jobs that they might otherwise avoid. Second, our results about feedback suggest that (worker) ignorance is bliss from a firm’s point of view. If workers tend to be overconfident about their abilities, this high confidence helps firms by making workers more willing to attempt difficult tasks. Giving workers information about their true ability only serves to dispels this beneficial overconfidence. Third, a number of the effects of our various treatments involve quitting. This illustrates the importance of allowing quitting as part of the experimental design.

Experimenters normally try to avoid having subjects leave an experiment, but without this feature we would have misunderstood the treatment effects of feedback and ambiguity.

Finally, we hope our experiments illustrate the value and difficulties of using AMT. We think AMT is potentially an extremely useful tool, but one that must be handled with great care. It does not reduce the effort required from an experimenter. We devoted a huge amount of effort (and a huge number of pages in this paper) to be as careful as possible to limit the potential problems that arise with use of AMT, and to check our data that the results are not due to the peculiarities of AMT. For researchers who are willing to invest the necessary effort, AMT opens up experiments to researchers who cannot easily afford to spend many thousands of dollars on an experiment. It also provides access to a naturally occurring labor market with more diverse participants than are found in the usual experimental subject pool. We did not choose to run our experiments on AMT due to a lack of other options. We can afford to run similar experiments in a standard laboratory setting. We choose to use AMT because we think it offers an interesting environment for running experiments that we hope other researchers will adopt.
Bibliography


Rosaz, J., Slonim, R., & Villeval, M. C. (2012). Quitting and peer effects at work.


Table 1: Subject Characteristics and Task Performance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>701</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td>Age</td>
<td>701</td>
<td>31.62</td>
<td>9.95</td>
</tr>
<tr>
<td>Indian</td>
<td>701</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td>American</td>
<td>701</td>
<td>0.29</td>
<td>0.46</td>
</tr>
<tr>
<td>Canadian and European</td>
<td>701</td>
<td>0.08</td>
<td>0.28</td>
</tr>
<tr>
<td>College Graduate</td>
<td>701</td>
<td>0.75</td>
<td>0.43</td>
</tr>
<tr>
<td>Income &lt; $12.5K</td>
<td>701</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>Income ≥ $25K</td>
<td>701</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>Primary Income</td>
<td>701</td>
<td>0.24</td>
<td>0.43</td>
</tr>
<tr>
<td>Motivated by Earning Money</td>
<td>701</td>
<td>0.97</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 2: Count of Subjects by Treatment

<table>
<thead>
<tr>
<th>Outside Option</th>
<th>Recruited</th>
<th>Usable Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$0.30</td>
<td>$0.40</td>
</tr>
<tr>
<td>Piece Rate, No Feedback</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Piece Rate, Feedback</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Fixed Pay, No Feedback</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>Fixed Pay, Feedback</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Low Ambiguity, No Feedback</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>Low Ambiguity, Feedback</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.35)</td>
<td>35</td>
<td>34</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3: Summary of Results by Treatment

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Select Task A</th>
<th>Dropped Out if Select Task A</th>
<th>Select and Complete Task A</th>
<th># Correct if Select Task A</th>
<th># Correct if Complete Task A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piece Rate, No Feedback</td>
<td>55.3%</td>
<td>0.0%</td>
<td>55.3%</td>
<td>4.74</td>
<td>4.74</td>
</tr>
<tr>
<td>Piece Rate, Feedback</td>
<td>47.3%</td>
<td>9.3%</td>
<td>42.9%</td>
<td>4.51</td>
<td>4.97</td>
</tr>
<tr>
<td>Fixed Pay, No Feedback</td>
<td>37.8%</td>
<td>22.6%</td>
<td>29.3%</td>
<td>3.06</td>
<td>3.62</td>
</tr>
<tr>
<td>Fixed Pay, Feedback</td>
<td>30.1%</td>
<td>28.6%</td>
<td>21.5%</td>
<td>3.39</td>
<td>4.30</td>
</tr>
<tr>
<td>Low Ambiguity, No Feedback</td>
<td>46.5%</td>
<td>7.5%</td>
<td>43.0%</td>
<td>4.10</td>
<td>4.41</td>
</tr>
<tr>
<td>Low Ambiguity, Feedback</td>
<td>35.6%</td>
<td>9.7%</td>
<td>32.2%</td>
<td>4.26</td>
<td>4.68</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.35)</td>
<td>63.3%</td>
<td>9.5%</td>
<td>57.3%</td>
<td>3.78</td>
<td>4.06</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>36.7%</td>
<td>13.6%</td>
<td>31.7%</td>
<td>3.23</td>
<td>3.63</td>
</tr>
</tbody>
</table>
Table 4: Regression Analysis of Treatment Effects on Selection and Completion of Task A

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Select Task A</th>
<th>Dropout from Task A</th>
<th>Select and Complete Task A</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed Pay</td>
<td>-0.498***</td>
<td>-0.619***</td>
<td>1.265***</td>
</tr>
<tr>
<td></td>
<td>(0.186)</td>
<td>(0.197)</td>
<td>(0.378)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>-0.194 (0.216)</td>
<td>-0.216 (0.220)</td>
<td>0.925** (0.459)</td>
</tr>
<tr>
<td></td>
<td>0.254 (0.171)</td>
<td>0.186 (0.176)</td>
<td>0.748** (0.323)</td>
</tr>
<tr>
<td>Low Ambiguity</td>
<td>-0.288** (0.138)</td>
<td>-0.314** (0.143)</td>
<td>0.394 (0.323)</td>
</tr>
<tr>
<td>Feedback * Fixed</td>
<td>-0.211 (0.198)</td>
<td>-0.126 (0.207)</td>
<td>0.152 (0.373)</td>
</tr>
<tr>
<td>Feedback *</td>
<td>-0.226* (0.137)</td>
<td>-0.241* (0.143)</td>
<td>0.537* (0.315)</td>
</tr>
<tr>
<td>Outside 40</td>
<td>-0.343*** (0.130)</td>
<td>-0.395*** (0.134)</td>
<td>-0.373*** (0.131)</td>
</tr>
<tr>
<td>Outside 50</td>
<td>-0.658*** (0.125)</td>
<td>-0.734*** (0.131)</td>
<td>-0.614*** (0.126)</td>
</tr>
<tr>
<td>Male</td>
<td>0.177* (0.105)</td>
<td>0.065 (0.226)</td>
<td>0.177* (0.105)</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>0.003 (0.002)</td>
<td>0.007* (0.004)</td>
<td>0.001 (0.002)</td>
</tr>
<tr>
<td>Investment Question</td>
<td>0.038 (0.082)</td>
<td>-0.370** (0.188)</td>
<td>0.143* (0.082)</td>
</tr>
<tr>
<td>Ambiguity Aversion</td>
<td>0.059 (0.062)</td>
<td>-0.313** (0.143)</td>
<td>0.114* (0.064)</td>
</tr>
<tr>
<td>Confidence</td>
<td>-450.922</td>
<td>-435.995</td>
<td>-551.379</td>
</tr>
<tr>
<td>Log Pseudo Likelihood</td>
<td>-521.723</td>
<td>-437.278</td>
<td>-421.453</td>
</tr>
</tbody>
</table>

Notes: All regressions based on 701 observations. Numbers in parentheses are robust standard errors. Three (***)**, two (**), and one (*) stars indicate statistical significance at the 1%, 5%, and 10% levels.
<table>
<thead>
<tr>
<th>Selection Type</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Pay</td>
<td>-1.512**</td>
<td>-0.425</td>
<td>-1.050*</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.567)</td>
<td>(.569)</td>
<td>(.612)</td>
<td>(.629)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>-1.296**</td>
<td>-0.133</td>
<td>-0.971</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>(0.656)</td>
<td>(.588)</td>
<td>(.700)</td>
<td>(.612)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.35)</td>
<td>-0.920**</td>
<td>-0.341</td>
<td>-0.679</td>
<td>-0.157</td>
</tr>
<tr>
<td></td>
<td>(0.447)</td>
<td>(.380)</td>
<td>(.442)</td>
<td>(.374)</td>
</tr>
<tr>
<td>Low Ambiguity</td>
<td>-0.371</td>
<td>-0.067</td>
<td>-0.205</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>(0.460)</td>
<td>(.386)</td>
<td>(.461)</td>
<td>(.392)</td>
</tr>
<tr>
<td>Feedback * Fixed</td>
<td>0.333</td>
<td>0.093</td>
<td>0.693</td>
<td>0.075</td>
</tr>
<tr>
<td>(Piece Rate + Low Ambiguity)</td>
<td>(0.628)</td>
<td>(.550)</td>
<td>(.677)</td>
<td>(.632)</td>
</tr>
<tr>
<td>Feedback * Correct Practice</td>
<td>-0.046</td>
<td>0.088</td>
<td>0.226</td>
<td>0.316</td>
</tr>
<tr>
<td>(Piece Rate + Low Ambiguity)</td>
<td>(0.462)</td>
<td>(.392)</td>
<td>(.468)</td>
<td>(.413)</td>
</tr>
<tr>
<td>Correct Practice</td>
<td></td>
<td>0.952***</td>
<td></td>
<td>0.947***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.115)</td>
<td></td>
<td>(.123)</td>
</tr>
<tr>
<td>Log Pseudo Likelihood</td>
<td>-1197.894</td>
<td>-1132.348</td>
<td>-1087.037</td>
<td>-1027.152</td>
</tr>
</tbody>
</table>

**Notes:** All regressions based on 701 observations. Numbers in parentheses are robust standard errors. Three (***) and two (**), and one (*) stars indicate statistical significance at the 1%, 5%, and 10% levels.
Appendix A: Full Test of Hit

Consent

Informed Consent: I chose to voluntarily to participate in this research project. The purpose of this work is to study the use of crowd sourcing to perform basic data analysis tasks. I have been recruited for this study through Mechanical Turk. Only persons 18 years of age or older may participate, and I affirm that I am 18 years of age or older. This study involves reading hand written messages in English. Only individuals who read and write English may participate. I affirm that I can read and write in English.

The initial phase of this study will take about 5 minutes to complete and will involve me answering a number of survey questions. If I complete the first phase of the study, I will be eligible for a second phase where I can earn a bonus by either completing a basic data analysis task, coding verbal responses, or performing a transcription task. The second phase of the study will last 8 - 10 minutes.

I will earn 20 cents (USA) for successfully completing the first phase of the study. I will then be given details on the second phase of the study, including an explanation of what bonuses I will be eligible to earn.

I am free to withdraw from the study at any time and without incurring the ill will of the researchers. I will only be paid if I complete both phases of the task.

There are no known risks or benefits from this study beyond those from any typical activity you might do in an online environment. This study will benefit society by helping researchers to better understand the use of crowd sourcing to analyze data.

The confidentiality of any personal information will be protected to the extent allowed by law. To the extent allowed by law, only the researcher and any research assistants conducting this experiment will have access to the data from this study. My name will not be reported with any results related to this research.

I can obtain further information from David Johnson (xsfsturk@gmail.com). If I have questions concerning my rights as a research subject, I can call the FSU Human Subjects Committee office at 1-850-644-8836 or email them at humansubjects@magnet.fsu.edu.

I may ask questions at any time via email (xsfsturk@gmail.com). Please feel free to contact us at this email address if you have any questions. Should new information become available during the course of this study about risks or benefits that might affect my willingness to continue in this research project, it will be given to me as soon as possible.

By clicking on the checkbox below, you indicate consent to participate in this study.

Phase 1

1. What is your gender?

2. What is your age?
3. What country do you currently live in?

4. What is your Nationality?

5. Which of the following best describes your highest achieved education level? Dropdown box. Categories are: (1) High School; (2) High School Graduate; (3) Some college, no degree; (4) Associates degree; (5) Bachelors degree; (6) Graduate degree (Masters, Doctorate, etc.).

6. Over the weekend, Bob watched 2 football games. On the scale below, mark the number of football games Bob watched over the weekend. Scale goes from 0 to 4.

7. What is the total income of your household? Dropdown box. Categories are: (1) 12,500; (2) 12,500 - 24,999; (3) 25,000 - 37,499; (4) 37,500 - 49,999; (5) 50,000 - 62,499; (6) 62,500 - 74,999; (7) 75,000 - 87,499; (8) 87,500 - 99,999; (9) 100,000 or more.

8. Why do you complete tasks in Mechanical Turk? Please check any of the following that applies: Categories are: (1) Fruitful way to spend free time and get some cash; (2) for primary income purposes (e.g. gas, bills, groceries, credit cards); (3) for secondary income purposes, pocket change (for hobbies, gadgets, going out); (4) to kill time; (5) I find the task to be fun; and (6) I am currently unemployed, or have only a part time job.

9. Next year, Jack and Jill are planning on visiting Disneyland. Jill has been to Disneyland many times while Jack has never been (0 times). On the scale below, mark how often Jack has been to Disneyland. Scale goes from 0 to 4.

10. How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick a box on the scale, where the value 0 means: "risk averse" and the value 10 means: "fully prepared to take risks". You can use the values in between to make your estimate.

11. People can behave differently in different situations. How would you rate your willingness to take risks in the following areas? Please tick a box on the scale, where the value 0 means: "risk averse" and the value 10 means: "fully prepared to take risks". You can use the values in between to make your estimate. The six categories are: (1) while driving; (2) in financial matters; (3) during leisure and sport; (4) in your occupation; (5) with your health; (6) your faith in other people.

12. Please consider what you would do in the following situation: Imagine that you had won 100,000 dollars in the lottery. Almost immediately after you collect the winnings, you receive the following financial offer from a reputable bank, the conditions of which are as follows: There is the chance to double the money within two years. It is equally possible that you could lose half of the amount invested. You have the opportunity to invest the full amount, part of the amount or reject the offer. What share of your lottery winnings would you be prepared to invest in this financially risky, yet lucrative investment? Please tick a box in each line of the scale.

13. Suppose there is a bag containing 90 balls. You know that 30 are red and the other 60 are a mix of black and yellow in unknown proportion. One ball is to be drawn from the bag at random. You are
offered a choice to (a) win $100 if the ball is red and nothing if otherwise, or (b) win $100 if it's black and nothing if otherwise. Which do you prefer?

14. The bag is refilled as before, and a second ball is to drawn from the bag at random. You are offered a choice to (c) win $100 if the ball is red or yellow, or (d) win $100 if the ball is black or yellow. Which do you prefer?

**Phase 2**

Thank you for completing the survey. In the next part of the HIT you will have the option of completing one of two possible tasks. Completing either of the tasks will earn you a bonus beyond the 20 cents you earned for completing the survey. You will only be paid if you complete both phases of the HIT.

In the first Task (Task A) you will be asked to classify a series of short messages sent by players participating in a social science experiment involving monetary incentives, trust, and reciprocity. You will get more instructions concerning this on the next page.

In the second Task (Task B), you will be asked to type a word that appears in an image. We have calibrated the tasks so that each task takes about the same amount of time. The words are fairly simple. If you are paying attention, you should have no problem typing all of the words correctly.

Before you make your selection, we have some messages that we would like you to classify for practice. Please use these practice messages to become familiar with Task A. After you have completed the practice messages you will be asked to select between Task A and B. Your bonus will be paid in addition to the amount that you made completing the survey. **When you are ready, please click the "Begin Practice" button.**

As previously discussed in this HIT, you will be asked to classify short messages sent by players participating in a social science experiment. In this experiment, subjects were randomly grouped into pairs to play a single game. The game had two roles, an A and B player.

The game is shown in the tree below. At the start of the game the A player had the option to select IN or OUT. If the A player selected OUT, both A and B earned 5 dollars. If A selected IN, the B player was then given the option to choose ROLL or DON'T ROLL. If the B player selected DON'T ROLL, the B player would earn 14 dollars and the A player would earn 0.
Alternatively if the B player selected ROLL, the B player would earn 10 dollars while a 6 sided dice would be rolled to determine the A player’s earning. If the dice came up 1, the A player would earn 0; if the dice came up 2 through 6, the A player would earn 12 dollars. The game ended when payoffs were assigned. Subjects were paid their earnings, in cash, and had no further interaction with the other player (or any other experimental subject).

The messages you are coding were sent by B players to A players. These were sent at the beginning of the game, before the A player decided whether or not they would select IN or OUT. The players did not know each other’s identities (i.e. play was anonymous) and had no opportunity to communicate other than the pre-game message by the B player.

Before you begin coding, please look carefully at the description of the game to make certain you understand this material. If you would like clarification, you can email xsfsturk@gmail.com with questions.

**MAIN INSTRUCTIONS:**

Please classify each of these handwritten messages using the key located below the handwritten text. You can classify a message by clicking the checkbox adjacent to the classification. Note, each message may have more than one possible classification. You should check all that you think apply.

A description of the available classifications for the messages follows. Please read through this material carefully. Before you begin it is important for you to know how to code the messages. If you would like clarification, you may email xsfsturk@gmail.com with questions.

**Promise to roll (explicit or implied):** Check this box if the message being sent either explicitly states that the sender will choose Roll or implies that the sender will choose Roll.
Appeal for Trust: Check this box if the message sent is asking the recipient, either explicitly or implicitly, to trust the sender of the message.

Appeal for Individual Monetary Incentive: Check this box if the message being sent asks the subject to make a selection that would be in the recipient's own monetary interest.

Appeal to joint payoffs: Check this box if the message being sent by the sender suggests that the recipient make a decision that will result in both individuals making more money.

Appeal to fairness: Check this box if the message being sent by the sender suggests that the recipient make a fair decision such as a decision that results in both players earning a more equal amount.

Attempt to build rapport (e.g. jokes, friendly banter): Check this box if the message being sent is an attempt to establish a personal connection with the recipient through the use of jokes or friendly banter.

None of the Above: Check this box if none of the above classifications are applicable to the message.

Before you begin, please consider the examples below:

Example 1

I am going to choose roll. If you choose IN, WE BOTH MAKE MORE $$ $$.

This example would be coded for two categories, “Promise to roll” and “Appeal to joint payoffs.”

Example 2

Hopefully, I’ll make a lucky roll.

This example would be coded as one category, “Promise to roll.”

Please begin when you are ready. When you are finished, please hit the ”NEXT” button. If you have trouble reading the text, you can zoom in by holding "Ctrl" and pressing "+".

Practice Questions

1) I’m going to roll
2) Take a risk
3) I have laundry to do tonight and I really don’t want to do it! But I don’t have clean underwear left and I don’t want to go commando tomorrow. We’ll see what I decide tonight. This man acts funny doesn’t he? But he seems cool, he’s quite the character. All this mystery is kinda cool.
4) The fairest thing to do is if you opt “In”. Then I will proceed to choose “roll”. That way you and I have 5/6 chance to make money for the both of us. That is much better than just making $ 5.00 each increase both our chances. Thanks.

5) Choose in, I will roll dice, your are 5/6 likely to get 2,3,4,5, or 6 → $ 12. This way both of us will win something.

How well do you think you performed the practice task in comparison to other Turkers? Subject marked their answer using a Likert scale ranging from well below average to well above average.

If in feedback they see the following pop-up box:

---

You got 0 of 5 correct

---

FIXED

You will now select between the two tasks. Task A consists of coding messages like what you did during in the practice task. However, Task A will have more messages and will not include any messages you saw during the practice task. In Task B you will type a word in the textbox below the word. Remember, each task takes about the same amount of time to complete.

If you select Task A, you will earn a bonus of 40 cents for certain for completing the task (i.e. coding all of the messages).

If you select Task B, you will be paid a bonus of 50 cents if you successfully type all of the words. Task B is designed to be sufficiently easy that you will earn this bonus for certain as long as you pay attention. Your bonus will be paid in addition to the amount that you made
completing the survey. **You will only be paid if you complete either Task A or Task B in the second phase of the HIT.**

Please select the task that you wish to complete.

---------------------------------------------------------------------

**HA**

You will now select between the two tasks. Task A consists of coding messages like what you did during in the practice task. However, Task A will have more messages and will not include any messages you saw during the practice task. In Task B you will type a word in the textbox below the word. Remember, each task takes about the same amount of time to complete.

If you select Task A, you will earn a bonus somewhere between 0 and 135 cents for completing the task (i.e. coding all of the messages).

If you select Task B, you will be paid a bonus of 30 cents if you successfully type all of the words. Task B is designed to be sufficiently easy that you will earn this bonus for certain as long as you pay attention. Your bonus will be paid in addition to the amount that you made completing the survey. **You will only be paid if you complete either Task A or Task B in the second phase of the HIT.**

Please select the task that you wish to complete.

---------------------------------------------------------------------

**LA**

You will now select between the two tasks. Task A consists of coding messages like what you did during in the practice task. However, Task A will have more messages and will not include any messages you saw during the practice task. In Task B you will type a word in the textbox below the word. Remember, each task takes about the same amount of time to complete.

If you select Task A, you will earn a bonus somewhere between 0 and 135 cents for completing the task (i.e. coding all of the messages). The precise bonus you earn will depend on how well you do at coding messages correctly.

A message is counted as "correct" if your coding matches the coding chosen by the most people participating in this study. Consider the example shown below:
Example 1

Suppose that 50% of participants code "Promise to Roll", and "Appeal to Joint Payoffs", 20% code only "Promise to Roll", 15% code only "Appeal to Joint Payoffs", and 15% code "None of the Above". To get credit for a correct coding you must code both "Promise to Roll" and "Appeal to Joint Payoffs".

Example 2

Suppose that 35% of participants code "Promise to Roll", 25% "Appeal to Joint Payoffs" and "Promise to Roll", 30% code only "Appeal to Joint Payoffs", and 10% code "None of the Above". To get credit for a correct coding you must code only "Promise to Roll."

If you select Task B, you will be paid a bonus of 30 cents if you successfully type all of the words. Task B is designed to be sufficiently easy that you will earn this bonus for certain as long as you pay attention. Your bonus will be paid in addition to the amount that you made completing the survey. You will only be paid if you complete either Task A or Task B in the second phase of the HIT.

Please select the task that you wish to complete.

PR

You will now select between the two tasks. Task A consists of coding messages like what you did during in the practice task. However, Task A will have more messages and will not include any messages you saw during the practice task. In Task B you will type a word in the textbox below the word. Remember, each task takes about the same amount of time to complete.

If you select Task A, you will earn a bonus of 9 cents for each message you code correctly. There are 15 messages, so if you select Task A, you will earn a bonus somewhere between 0 and 135 cents for completing the task (i.e. coding all of the messages).
A message is counted as "correct" if your coding matches the coding chosen by the most people participating in this study. Consider the example shown below:

**Example 1**

![Image of a message saying: I am going to choose Roll. If you choose IN, WE BOTH MAKE MORE $$$.

Suppose that 50% of participants code "Promise to Roll", and "Appeal to Joint Payoffs", 20% code only "Promise to Roll", 15% code only "Appeal to Joint Payoffs", and 15% code "None of the Above". To get credit for a correct coding you must code both "Promise to Roll" and "Appeal to Joint Payoffs".

**Example 2**

![Image of a message saying: Hopefully, I'll make a lucky roll.

Suppose that 35% of participants code "Promise to Roll", 25% "Appeal to Joint Payoffs" and "Promise to Roll", 30% code only "Appeal to Joint Payoffs", and 10% code "None of the Above". To get credit for a correct coding you must code only "Promise to Roll."

If you select Task B, you will be paid a bonus of **30 cents** if you successfully type all of the words. Task B is designed to be sufficiently easy that you will earn this bonus for certain as long as you pay attention. Your bonus will be paid in addition to the amount that you made completing the survey. **You will only be paid if you complete either Task A or Task B in the second phase of the HIT**

Messages if Worker Selected Task A

1) You can have the 2 extra dollars. I’ll be nice and choose to roll.
2) Stay IN, I really need the money.
3) Tee hee, this is kinda Twilight Zone – ism; Why not “go for it”, eh? I hope you have a lovely evening as well.
4) Please choose In so we can get paid more.
5) If you choose in I’ll roll. P In R Why? If you choose out, we walk out with $10 each. If you choose IN & I choose IN then both of us coin. So it’s a compromise. By agreeing to this I guarantee myself more $ than risking you choose out. So if you choose out I get $10 ($5 diff.) if you choose in I get $15 vs. $19 ($4 diff.). that’s why
6) Hello fair stranger, anonymous partner _ _ _ Choose whatever you want. Far be it from me to influence your decision, but I think you should choose “in” and I should choose “roll” and we should take the chance at both earning as much as we can. 5 chances out of 6 say it’ll work, and
I’m totally broke, looking to rake in stray cash however I can. I feel the luck in the air. E In R I
don’t really have much else to say. Hope you’re doing well, whoever you are. Yes. That’s all.
Random note from random human
7) Have a Happy day
8) If you stay in, the chances of the die coming up other than 1 are 5 in 6 – pretty good. Otherwise,
we’d both be stuck at $5. (If you opt out)
9) If you will choose “In”, I will choose to roll. This way, we both have an opportunity to make
more than $5!
10) You’ll still be gaining more than if I had chosen Don’t roll.
11) Good luck I do not know what I’m going to do, so I have no hints on how to advise you on
choosing “in” or “out.” Though it would be beneficial for me to pick don’t roll and hope you
pick “in”, I also like to give you a chance to gain some cash. Who knows?
12) Choose “In” so we can both make some $$ What are the chances me rolling a 1? I’ll try my best.
13) If I roll a 2–6 (you’ll know when you receive the $, you will give $5.00 to a stranger. P In R
[[then there is a line, under which is written “Sign here if you are so kind]]] Thanks.
14) Hi, well I’m going to Roll so you have at least a shot for more money. I hope it works out.
15) CHOOSE IN, SO WE CAN ROLL AND GET $12 AND $10.

Words if worker selected Task B

1) Automobile
2) Calamari
3) Dinosaur
4) Hootenanny
5) Juxtapose
6) Luminary
7) Mallard
8) Molasses
9) Plateau
10) Rupture
11) Subliminal
12) Turtle
13) Unleaded
14) Zombie
15) Attention
Appendix B: Full Regression Output from Stata

Variable Key:

Fixed: Dummy For Fixed ("Fixed Pay")
Ha_Lcap: Dummy For HA100 ("High Ambiguity (Max = $1.00)")
Ha_Hcap: Dummy For HA135 ("High Ambiguity (Max = $1.35)")
La: Dummy For LA ("Low Ambiguity")
Ioutside_40: Dummy For Task B Bonus Payment Of 40 Cents
Ioutside_50: Dummy For Task B Bonus Payment Of 50 Cents
Notincome: Dummy For A Subject That Is Not Income Motivated
Feedback_N~D: Dummy For Sessions With Feedback That Are Not Fixed Bonus
_income_2: Dummy For Income ("12,500 - 24,999")
_income_3: Dummy For Income ("25,000 - 37,499")
_income_4: Dummy For Income ("37,500-49,999")
_income_5: Dummy For Income ("50,000 - 62,499")
_income_6: Dummy For Income ("62,500 - 74,999")
_income_7: Dummy For Income ("75,000 - 87,499")
_income_8: Dummy For Income ("87,500-99,999")
_income_9: Dummy For Income ("100,000 or more")
_educatio~2: Dummy For Education ("High School Graduate")
_educatio~3: Dummy For Education ("Some College, No Degree")
_educatio~4: Dummy For Education ("Associates Degree")
_educatio~5: Dummy For Education ("Bachelor’s Degree")
_educatio~6: Dummy For Education ("Graduate Degree (Masters, Doctorate, Etc.")
Age: Worker’s Age
Male: Dummy For Male ("Male")
India: Dummy For Indian ("India")
American: Dummy For American ("America")
Answerq10r~K: General Risk Question ("Risk Averse" (0) – “Fully Prepared To Take Risks” (10))
Answerq17r~K: Hypothetical Gamble From 100,000 Windfall
Riskfactor: Factor Analysis Variable Created Based Upon Subjects Willingness To Take Risks (In The Following Facets Of Their Life: (1) While Driving; (2) In Financial Matters; (3) During Leisure And Sport; (4) In Your Occupation; (5) With Your Health; (6) Your Faith In Other People.)
Ammearse: Dummy variable for Ambiguity Aversion (“Ambiguity Averse”)
Answerqc: Worker’s Confidence In Comparison To Other Workers
Correctp: Number Of Practice Messages Correctly Coded
### Regressions for Table 4

**Probit regression**  
Number of obs = 701  
Wald chi2(9) = 60.96  
Prob > chi2 = 0.0000  
Log pseudolikelihood = -450.9215  
Pseudo R2 = 0.0654

| selecta | Coef. | Std. Err. | z   | P>|z|    | [95% Conf. Interval] |
|---------|-------|-----------|-----|-------|----------------------|
| fixed   | -0.4975761 | 0.1868353 | -2.67 | 0.008 | -0.8628845 to -0.1322677 |
| ha_lcap | -0.1939955 | 0.2156401 | -0.90 | 0.368 | -0.6166423 to 0.2286513 |
| ha_hcap | 0.2538041 | 0.1707389 | 1.49 | 0.137 | -0.0809263 to 0.5885344 |
| la      | -0.2877694 | 0.1376718 | -2.09 | 0.037 | -0.5576231 to -0.179596 |
| feedback_f_d | -0.2106486 | 0.1976642 | -1.07 | 0.287 | -0.5980632 to 0.1767661 |
| feedback_n_d | -0.2257972 | 0.1374722 | -1.64 | 0.100 | -0.4952377 to 0.0436434 |
| outside_40 | -0.3431604 | 0.1297826 | -2.64 | 0.008 | -0.5975295 to -0.087913 |
| outside_50 | -0.6578861 | 0.1250987 | -5.26 | 0.000 | -0.903075 to -0.4126972 |
| notincome | 0.2053486 | 0.2985392 | 0.69 | 0.492 | -0.3797775 to 0.7904748 |
| _cons   | 0.5014388 | 0.1429670 | 3.51 | 0.000 | 0.2212286 to 0.7816491 |

Prob regression  
Number of obs = 701  
Wald chi2(32) = 87.54  
Prob > chi2 = 0.0000  
Log pseudolikelihood = -435.9945  
Pseudo R2 = 0.0964

| selecta | Coef. | Std. Err. | z   | P>|z|    | [95% Conf. Interval] |
|---------|-------|-----------|-----|-------|----------------------|
| fixed   | -0.7335369 | 0.1311365 | -5.59 | 0.000 | -0.992959 to -0.4731135 |
| ha_lcap | -0.3948228 | 0.1340166 | -2.95 | 0.003 | -0.6574904 to -0.1321551 |
| ha_hcap | 0.7335369 | 0.1311365 | -5.59 | 0.000 | -0.992959 to -0.4731135 |
| la      | -0.1459471 | 0.1393670 | 1.07 | 0.286 | -0.430526 to 0.1395298 |
| _cons   | 0.5104388 | 0.1429670 | 3.51 | 0.000 | 0.2212286 to 0.7816491 |

Prob > chi2 = 0.0000  
Wald chi2(9) = 60.96  
Pseudo R2 = 0.0654  
Number of obs = 701  
Log pseudolikelihood = -435.9945  
Pseudo R2 = 0.0964  
Number of obs = 701  
Log pseudolikelihood = -435.9945  
Pseudo R2 = 0.0964
Probit model with sample selection

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>Censored obs</th>
<th>Uncensored obs</th>
<th>Wald chi2(7)</th>
<th>Log pseudolikelihood</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>701</td>
<td>385</td>
<td>316</td>
<td>16.61</td>
<td>-551.3785</td>
<td>0.0201</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dropout</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fixed</td>
<td>1.264815</td>
<td>.3775103</td>
<td>3.35</td>
<td>0.001</td>
<td>.5249081</td>
<td>2.004721</td>
</tr>
<tr>
<td>ha_lcap</td>
<td>.924715</td>
<td>.4588665</td>
<td>2.02</td>
<td>0.044</td>
<td>.0253531</td>
<td>1.824077</td>
</tr>
<tr>
<td>ha_hcap</td>
<td>.7480652</td>
<td>.3233241</td>
<td>2.31</td>
<td>0.021</td>
<td>.1143617</td>
<td>1.381769</td>
</tr>
<tr>
<td>la</td>
<td>.3937943</td>
<td>.3230085</td>
<td>1.22</td>
<td>0.223</td>
<td>-.392908</td>
<td>1.026879</td>
</tr>
<tr>
<td>feedback_f-d</td>
<td>.1524535</td>
<td>.3729623</td>
<td>0.41</td>
<td>0.683</td>
<td>-.578539</td>
<td>.8834462</td>
</tr>
<tr>
<td>feedback_n-d</td>
<td>.536569</td>
<td>.3149030</td>
<td>1.70</td>
<td>0.088</td>
<td>-.806296</td>
<td>1.153768</td>
</tr>
<tr>
<td>notincome</td>
<td>.3330488</td>
<td>.4849318</td>
<td>0.69</td>
<td>0.492</td>
<td>-.617400</td>
<td>1.283498</td>
</tr>
<tr>
<td>_cons</td>
<td>-2.151843</td>
<td>.5009865</td>
<td>-4.30</td>
<td>0.000</td>
<td>-3.133759</td>
<td>-1.169927</td>
</tr>
</tbody>
</table>

|                  |               |               |               |             |                      |             |
| selecta          |               |               |               |             |                      |             |
| fixed            | -.4979827     | .1861214      | -2.68         | 0.007       | -.8627738            | -.1331915   |
| ha_lcap          | -.1956535     | .2159284      | -0.91         | 0.365       | -.6188653            | .2275584    |
| ha_hcap          | .2532124      | .1706646      | 1.48          | 0.138       | -.0812828            | .5877077    |
| la               | -.2876088     | .1377445      | -2.09         | 0.037       | -.557583             | -.0176346   |
| feedback_f-d     | -.2101586     | .197562       | -1.06         | 0.287       | -.5973731            | .1770558    |
| feedback_n-d     | -.225280      | .157623       | -1.64         | 0.102       | -.4950161            | .0444562    |
| notincome        | .2041518      | .2971523      | 0.69          | 0.492       | -.3782561            | 1.7865596   |
| _outside_40      | -.3483157     | .1364376      | -2.55         | 0.011       | -.6157285            | -.080903    |
| _outside_50      | -.6577107     | .1250372      | -5.26         | 0.000       | -.902779             | -4.126424   |
| _cons            | .5029804      | .1430408      | 3.52          | 0.000       | .2226256             | .7833352    |

| /athrho           | .1334873      | .7110049      | 0.19          | 0.851       | -1.260127            | 1.527102    |

| rhol              | .1327001      | .69852        | -0.851        | 0.909923    |                      |             |

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.04  Prob > chi2 = 0.851

Probit model with sample selection

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>Censored obs</th>
<th>Uncensored obs</th>
<th>Wald chi2(27)</th>
<th>Log pseudolikelihood</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>701</td>
<td>385</td>
<td>316</td>
<td>1096.79</td>
<td>-521.723</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

|                  | Robust        |               |               |             |                      |             |
| dropout          |               |               |               |             |                      |             |
| fixed            | 1.150873      | .4279792      | 2.69          | 0.007       | .3120487             | 1.989696    |
| ha_lcap          | .9231801      | .4898112      | 1.88          | 0.059       | -.0368322            | 1.883192    |
| ha_hcap          | .5924453      | .3530649      | 1.68          | 0.093       | -.0995492            | 1.28444     |
| la               | .1630376      | .3206428      | 0.50          | 0.616       | -.4674107            | .789486     |
| feedback_f-d     | .3125763      | .3939355      | 0.78          | 0.434       | -.4702245            | 1.095377    |
| feedback_n-d     | .5626039      | .3244246      | 1.73          | 0.083       | -.0732566            | 1.198464    |
| notincome        | .291758       | .5095999      | 0.57          | 0.567       | -.7070395            | 1.290556    |
| income_2         | .2133706      | .2881892      | 0.74          | 0.459       | -.3514659            | .7782111    |
| income_3         | -.1267121     | .4023781      | -0.31         | 0.753       | -.9153587            | .6619345    |
| income_4         | -.236918      | .4613267      | -0.52         | 0.603       | -.1438766            | .6644919    |
| income_5         | -.4.797855    | .4017056      | -11.94        | 0.000       | -.5.58518             | 4.010527    |
| income_6         | -.2534835     | .6194784      | -0.41         | 0.682       | -.1.467639            | .960672     |
| income_7         | .6414725      | .5163733      | 1.24          | 0.214       | -.3706007            | 1.653546    |
| income_8         | .8960416      | .7537037      | 1.19          | 0.234       | -.5811906            | 2.373274    |
| income_9         | -.4.490137    | .4463626      | -10.06        | 0.000       | -.5.364991            | 3.615282    |
| _educatio-1      | .181373       | .4338735      | 0.42          | 0.676       | -.6690035            | 1.031749    |
| _educatio-2      | .2893125      | .3490908      | 0.83          | 0.407       | -.3945069            | .973932     |
| age              | .0114901      | .0115211      | 1.00          | 0.319       | -.011090             | .0340711    |
| male             | .0649168      | .2256364      | 0.29          | 0.774       | -.3737225            | .507156     |
| Variable  | Coef.   | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|-----------|---------|-----------|-------|------|----------------------|
| fixed     | -0.619896 | 0.1963328 | -3.16 | 0.002 | -1.004701 -0.235098 |
| ha_lcap   | -0.2180092 | 0.2210635 | -0.99 | 0.324 | -0.6512856 0.2152673 |
| ha_hcap   | 0.1951816 | 0.1758826 | 1.05  | 0.292 | -0.195419 0.5299051 |
| la        | -0.3341682 | 0.1432509 | -2.35 | 0.019 | -0.617998 -0.0503382 |
| feedback_f_d | -0.1252181 | 0.2068297 | -0.61 | 0.545 | -0.5305986 0.2801606 |
| feedback_n_d | -0.2408309 | 0.1430842 | -1.68 | 0.092 | -0.5212708 0.396089 |
| notincome | 0.1451669 | 0.3102027 | 0.47  | 0.640 | -0.4628191 0.751529 |
| outside_40 | -0.3983969 | 0.1508282 | -2.64 | 0.008 | -0.6940147 -0.102779 |
| outside_50 | -0.7330704 | 0.1314524 | -5.58 | 0.000 | -0.9907123 -0.4754285 |
| income_2   | -0.1812234 | 0.1400204 | -1.29 | 0.196 | -0.4556854 0.0932116 |
| income_3   | -0.1378454 | 0.1665582 | 0.83  | 0.409 | -0.46188035 0.4609493 |
| income_4   | 0.1596011 | 0.1842893 | 0.87  | 0.386 | -0.2015993 0.5208015 |
| income_5   | -0.0750733 | 0.2548205 | 0.29  | 0.768 | -0.4423658 0.5745123 |
| income_6   | -0.0089207 | 0.2939605 | -0.03 | 0.976 | -0.5956566 0.5778152 |
| income_7   | 0.0186797 | 0.2833137 | 0.07  | 0.947 | -0.5366049 0.5739644 |
| income_8   | -0.4102878 | 0.4218738 | -0.97 | 0.331 | -1.237066 0.4164847 |
| income_9   | 0.0114381 | 0.20948 | 0.04  | 0.971 | -0.5951315 0.6100708 |
| educati-a-2 | 0.4934212 | 0.3522025 | 1.42  | 0.157 | -0.2365503 0.1126818 |
| educatio-3 | 0.1996737 | 0.3595922 | 0.55  | 0.580 | -0.5063261 0.9046607 |
| educatio-4 | 0.3993882 | 0.4294894 | 0.93  | 0.352 | -0.4423956 1.241172 |
| educatio-5 | 0.4127121 | 0.3398293 | 1.21  | 0.225 | -0.2533413 1.078765 |
| educatio-6 | 0.4818867 | 0.3499613 | 1.38  | 0.169 | -0.2040428 1.16778 |
| age       | -0.0054917 | 0.0051477 | -1.07 | 0.286 | -0.0155811 0.0045977 |
| male      | -0.1773414 | 0.1051812 | 1.69  | 0.092 | -0.362881 0.3834928 |
| india     | 0.0492968 | 0.1624856 | -0.27 | 0.785 | -0.3511379 0.2805442 |
| american  | -0.2480114 | 0.1785587 | -1.39 | 0.165 | -0.5979794 0.1019565 |
| answertq0r-k | -0.0199665 | 0.0236606 | -0.80 | 0.423 | -0.0653405 0.0274075 |
| answertq1r-k | 0.0028349 | 0.0019472 | 1.46  | 0.145 | -0.0009816 0.0066514 |
| riskfactor | 0.0985275 | 0.0772649 | 1.28  | 0.202 | -0.052909 0.249964 |
| ammeasure2 | 0.0387045 | 0.0824034 | 0.47  | 0.639 | -0.1228052 0.2020212 |
| answeqg | 0.0582253 | 0.1625371 | 0.35  | 0.724 | -0.0643451 0.1807958 |
| correctp  | -0.0040121 | 0.0503746 | -0.08 | 0.937 | -0.1027445 0.0947203 |
| cons      | 0.2745087 | 0.4692304 | 0.59  | 0.519 | -0.6451659 1.194183 |

Wald test of indep. eqns. (rho = 0): ch2(9) = 67.79 Prob > chi2 = 0.0000

Probit regression
Number of obs = 701
Wald chi2(9) = 67.79
Prob > chi2 = 0.0000
Log pseudolikelihood = -437.27801 Pseudo R2 = 0.0736

Robust

| good_selecta | Coef.   | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|--------------|---------|-----------|-------|------|----------------------|
| fixed        | -0.7141639 | 0.1897777 | -3.76 | 0.000 | -1.086121 -0.3422064 |
| ha_lcap      | -0.3375562 | 0.2180074 | -1.55 | 0.122 | -0.7648429 0.0897304 |
| ha_hcap      | 0.1112215 | 0.1688854 | 0.66  | 0.510 | -0.2197879 0.4422309 |
| la           | -0.3248596 | 0.1383798 | -2.35 | 0.019 | -0.596079 -0.0536402 |
| feedback_f_d | -0.240036 | 0.2077606 | -1.16 | 0.248 | -0.6472394 0.1671673 |
| feedback_n_d | -0.2862096 | 0.1381543 | -2.07 | 0.038 | -0.5569864 0.015432 |
| outside_40   | -0.3728919 | 0.1321255 | 2.84  | 0.004 | -0.6501469 -0.1153639 |
| outside_50   | -0.6136 | 0.1257225 | -4.88 | 0.000 | -0.8600117 -0.3671883 |
| notincome    | 0.0365988 | 0.296397 | 0.12  | 0.902 | -0.5443286 0.6175263 |

53
Probit regression

Number of obs = 701
Wald chi2(32) = 95.35
Prob > chi2 = 0.0000

Log pseudolikelihood = -421.45292  Pseudo R2 = 0.1071

------------------------------------------------------------------------------
|            Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------|------------------|---------|---------|-----------------------------|
 fixed        | -0.8083375   .1991001 -4.06  0.000  -1.198567   -0.418085
 ha_lcap      | -0.3430342   .2226565 -1.54  0.123  -0.7794329   .0933466
 ha_hcap      | -0.0781567   .1734709  0.45   0.652  -0.26184     .418354
 la           | -0.3335817   .1440065 -2.32  0.021  -0.6158294   -0.051341
 feedback_f_d | -0.320366    .144682 -2.23  0.026   .3066081   -0.394652
 feedback_n_d | -0.2378077   .143831  1.65  0.098  -0.5179113   0.440998
 _income_2    | 0.1688715   .1634031  0.91  0.362  -0.1739272   .4619357
 _income_3    | 0.1686926   .1864028  1.00  0.317  -0.1878692   .5519864
 _income_5    | 0.162445    .2560191  0.63  0.526  -0.3393433   0.6642332
 _income_6    | -0.261868   .3105877  0.09  0.927  -0.5810122   0.67359
 _income_7    | -0.157028   .2824606 -0.56  0.578  -0.7106406   .3965846
 _income_8    | 0.361596    .4233415 -1.33  0.185  -0.139133    .268138
 _income_9    | 0.0292188   .3100964  0.09  0.925   -0.5785589   0.639966
 _educatio_2  | 0.3114664   .3762853  0.83  0.408   .4260391   1.049972
 _educatio_3  | 0.3500213   .3596698  0.12  0.908   .5648224   1.150512
 _educatio_4  | 0.3313634   .4264681  0.74  0.462   -0.5222693   1.149455
 _educatio_5  | 0.2561414   .3346856  0.77  0.444   -0.3995578   0.912385
 _educatio_6  | 0.3622466   .3487121  1.04  0.299   -0.321265   1.04571
 age          | -0.006887    .0052174 -1.32  0.187  -0.0171131   0.0033388
 male         | 0.1768417   .1053219  1.68  0.093   -0.0295856   0.383268
 india        | -0.1302094   .1677756 -0.78  0.438   -0.4590435   0.1986247
 american     | -0.2835427   .1791065 -1.58  0.113   -0.6345849   0.067496
 answeq10<k   | -0.0163312   .0239176 -0.68  0.495   -0.0632088   0.0305464
 answeq17<k   | 0.0009944   .0020204  0.49  0.623   -0.0029659   0.004954
 riskfactor   | 0.0258111   .0777414  0.33  0.740   -0.1265593   0.1781815
 ammeasure2   | 0.1434775   .0822283  1.74  0.081   -0.0176873   0.3046419
 answeq<1     | 0.114156    .0644135  1.77  0.076   -0.0120921   0.2404042
 correctp     | 0.0251905   .0511152  0.49  0.622   -0.0749334   0.1253745
 notoutside_40| 0.4225586   .6354972  3.14  0.002   -0.6839311   -0.15878
 notoutside_50| -0.6808131  .133008   5.12  0.000   -0.941504   -0.4201223
 notincome    | 0.0230372   .2979366 -0.08  0.938   -0.6098219   0.5690777
_cons         | 0.3031777   .4644181  0.65  0.514   -0.6071051   1.21338
------------------------------------------------------------------------------

Regressions for Table 5

Log pseudolikelihood = -1197.894  Prob > chi2 = 0.0310

|            Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------|------------------|---------|---------|-----------------------------|
 fixed        | -1.5124     .5669133 -2.67  0.008   -2.62353   -0.401279
 ha_lcap      | -1.29626    .6559649 -1.98  0.048   -2.581927   -0.0105921
 ha_hcap      | -0.9203533  .4473668 -2.06  0.040   -1.796856   -0.0432105
 la           | -0.813      .4192218 -1.94  0.053   -2.0271958   0.4099149
 feedback_f_d | 0.332165    .6278926  0.53  0.596   -0.8974304   1.563863
 feedback_n_d | -0.046422   .4616179 -0.10  0.920   -0.9511765   0.8583325
 notincome    | 0.9546672   .8981883  1.06  0.288   -0.8057496   2.715084
_cons         | 4.844269    .5057202  9.58  0.000   3.853075   5.835462

54
Ioutside | feedback_f~d | feedback_n~d |   .0876405   .3919178     0.22   0.823
|   .0925383   .5501506     0.17   0.866
regression model with sample selection)        Censored obs       =       385
------------------------------------------------------------------------------
selecta      | ammeasure2 |   .2099265   .2012015     1.04   0.297
riskfactor |
education |   .0471825    .041756     1.13   0.258
notincome |   .2038485   .2975589     0.69   0.493
correctp |   .9518903   .1152671     8.26   0.000      .725971     1.17781
回答 |    .277206     .14769     1.88   0.061
\textit{/lnsigma} |   .9495507   .0355655    26.70   0.000     .8798436    1.019258
ha_hcap |   .1824353   .1752479     1.04   0.298
ha_lcap |   .6847021    .525874     1.29   0.204
american |   .1722602    .525874    1.29   0.204
answer\textsuperscript{lr}-k |   .0954492    .0625549     1.53   0.127
answer\textsuperscript{qr}-k |   .0085394    .0055473     1.54   0.124
riskfactor |   .0337278    .0202216     1.67   0.098
ammeasure2  |   .2099265    .0212015     1.04   0.297
answercp |   .277206     .14769     1.88   0.061
correct |   .9518903   .1152671     8.26   0.000      .725971    .17781
\textit{/cons} |   .5029762   .1430657     3.52   0.000     .2225727    .7833798
----------------------------------------------------------------
 Wald test of indep. eqns. (rho = 0): chi2(1) =     0.30   Prob > chi2 = 0.5852

Heckman selection model Number of obs =       701
(regression model with sample selection) Censored obs =       385
Uncensored obs =       316
Wald chi2(19) = 189.38
Log pseudolikelihood = -1132.348     Prob > chi2 = 0.0000

| Coef.  Std. Err.  z    P>|z|     [95% Conf. Interval] |
|--------|----------------|-----|--------|----------------|
| Robust |
| totalcorrect | fixed | -.4249227 | .5691133 | -0.75 | 0.455 | -1.540364 | .6905199 |
| | ha_lcap | -.1328479 | .5975475 | -0.23 | 0.821 | -1.28442 | 1.018724 |
| | ha_hcap | -.3411654 | .3796263 | -0.90 | 0.369 | -1.085219 | .4028848 |
| | la | -.067092 | .38619 | -0.17 | 0.862 | -.2420242 | .689844 |
| | feedback_f~d | .0925383 | .5501506 | 0.17 | 0.866 | -.9857371 | 1.170814 |
| | feedback_n~d | .0876405 | .3919178 | 0.22 | 0.823 | -.6805043 | .8557853 |
| | income | .8820804 | .6230567 | 1.42 | 0.157 | -.3390884 | 2.103249 |
| | education | .0269278 | .0684949 | 0.39 | 0.694 | -.1073199 | .161754 |
| | age | -.0063242 | .0135454 | -0.47 | 0.641 | -.0328727 | .0202243 |
| | male | .0244728 | .2632421 | 0.09 | 0.926 | -.4914722 | .5404178 |
| | india | -.1.151812 | .4033252 | -2.86 | 0.004 | -.9423299 | .3619248 |
| | american | -.1.722602 | .525874 | -3.33 | 0.143 | -.212954 | .858434 |
| | answer\textsuperscript{lr}-k | -.0954492 | .0625549 | -1.53 | 0.127 | -.2180546 | .0217516 |
| | answer\textsuperscript{qr}-k | -.0085394 | .0055473 | -1.54 | 0.124 | -.0204079 | .0022372 |
| | riskfactor | -.0337278 | .202216 | -0.17 | 0.868 | -.4300748 | .3626192 |
| | ammeasure2 | .2099265 | .0212015 | 1.04 | 0.297 | -.1844211 | .6042472 |
| | answerqc | .277206 | .14769 | 1.88 | 0.061 | -.012261 | .566673 |
| | correct | .9518903 | .1152671 | 8.26 | 0.000 | .725971 | 1.17781 |
| | \textit{/cons} | .3.120207 | 1.159469 | 2.69 | 0.007 | .8476895 | 5.392725 |

selecta | fixed | -.6009964 | .1923013 | -3.13 | 0.002 | -.97979 | .2249028 |
| ha_lcap | -.2546194 | .21943 | -1.16 | 0.246 | -.6847021 | .1754633 |
| ha_hcap | .1824353 | .1752749 | 1.04 | 0.298 | -.1610917 | .5259678 |
| la | -.2747598 | .1412187 | -1.95 | 0.052 | -.5515434 | .0020238 |
| feedback_f~d | -.134283 | .202659 | -0.66 | 0.508 | -.5314874 | .2629214 |
| feedback_n~d | -.267034 | .140356 | -1.90 | 0.057 | -.5420279 | .0079599 |
| notincome | .2601503 | .3037792 | 0.68 | 0.497 | -.3892441 | .8015446 |
| _outside_40 | -.388026 | .1372799 | -2.83 | 0.005 | -.6570897 | .118962 |
| _outside_50 | -.7019447 | .1282752 | -5.47 | 0.000 | -.9533594 | .4505299 |
| income | .001665 | .0267726 | 0.06 | 0.950 | -.0508084 | .0541383 |
| education | .0471825 | .041756 | 1.13 | 0.258 | -.0346578 | .1290227 |
| age | -.0061175 | .0050516 | -1.21 | 0.226 | -.0160384 | .0037833 |
| india | -.0458081 | .1653314 | -0.28 | 0.782 | -.3695817 | 0.2782354 |
| american | -.2420643 | .1769027 | -1.37 | 0.171 | -.5887871 | .1046586 |
\[
\begin{align*}
\text{Log pseudolikelihood} & = \text{Censored obs} = 420 \\
\text{Heckman select} \\
\text{Wald test of indep. eqns. (rho = 0): chi2(1) = 0.08} & \quad \text{Prob > chi2 = 0.7799} \\
\begin{array}{lcccc}
\_Ioutside_50 & \_Ioutside_40 & \text{feedback_n~d} & \text{feedback_f~d} & \\
0.2261928 & 0.4681941 & 0.48 & 0.629 & \\
0.6926279 & 0.6770963 & 1.02 & 0.306 & \\
\end{array}
\text{Log pseudolikelihood} & = \text{Uncensored obs} = 281 \\
\text{Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02} & \quad \text{Prob > chi2 = 0.9025} \\
\begin{array}{lcccc}
\text{good selecta} & \\
\text{fixed} & -1.05106 & 0.6123327 & -1.72 & 0.086 \\
\text{ha_lcap} & -0.9714069 & 0.6996894 & -1.39 & 0.165 \\
\text{ha_hcap} & -0.6791329 & 0.4423777 & -1.54 & 0.125 \\
\text{feedback_f~d} & -0.2045895 & 0.4608711 & -0.44 & 0.657 \\
\text{feedback_n~d} & 0.6926279 & 0.6770963 & 1.02 & 0.306 \\
\text{notincome} & 1.733603 & 0.6483828 & 2.69 & 0.007 \\
\_cons & 4.808733 & 0.5476101 & 8.78 & 0.000 \\
\end{array}
\begin{array}{lcccc}
\text{good selecta} & \\
\text{fixed} & -0.7154729 & 0.1894706 & -3.78 & 0.000 \\
\text{ha_lcap} & -0.3372985 & 0.2179704 & -1.55 & 0.122 \\
\text{ha_hcap} & -0.111331 & 0.1690048 & 0.66 & 0.511 \\
\text{feedback_f~d} & -0.3258049 & 0.1382591 & -2.36 & 0.018 \\
\text{feedback_n~d} & -0.2394701 & 0.2075343 & -1.15 & 0.249 \\
\text{notincome} & 0.6926279 & 0.6770963 & 1.02 & 0.306 \\
\_cons & 4.808733 & 0.5476101 & 8.78 & 0.000 \\
\end{array}
\begin{array}{lcccc}
\_athrho & -0.0385317 & 0.314922 & -0.12 & 0.902 \\
\_lnsigma & 0.7712727 & 0.0382672 & 20.15 & 0.000 \\
\_cons & 4.808733 & 0.5476101 & 8.78 & 0.000 \\
\end{array}
\begin{array}{lcccc}
\_athrho & -0.0385317 & 0.314922 & -0.12 & 0.902 \\
\_lnsigma & 0.7712727 & 0.0382672 & 20.15 & 0.000 \\
\_cons & 4.808733 & 0.5476101 & 8.78 & 0.000 \\
\end{array}
\end{align*}
\]

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02 Prob > chi2 = 0.9025

Heckman selection model
(regression model with sample selection)

\[
\begin{align*}
\text{Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02} & \quad \text{Prob > chi2 = 0.9025} \\
\begin{array}{lcccc}
\text{good selecta} & \\
\text{fixed} & -1.05106 & 0.6123327 & -1.72 & 0.086 \\
\text{ha_lcap} & -0.9714069 & 0.6996894 & -1.39 & 0.165 \\
\text{ha_hcap} & -0.6791329 & 0.4423777 & -1.54 & 0.125 \\
\text{feedback_f~d} & -0.2045895 & 0.4608711 & -0.44 & 0.657 \\
\text{feedback_n~d} & 0.6926279 & 0.6770963 & 1.02 & 0.306 \\
\text{notincome} & 1.733603 & 0.6483828 & 2.69 & 0.007 \\
\_cons & 4.808733 & 0.5476101 & 8.78 & 0.000 \\
\end{array}
\end{align*}
\]

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02 Prob > chi2 = 0.9025

Heckman selection model
(regression model with sample selection)
| r~k | answerq17 | .0010656 | .0019859 | .54 | 0.592 |
| _Ioutside_50 | | | | | |
| _Ioutside_40 | | | | | |
| feedback_n | | | | | |
| feedback_f | | | | | |
| income | | | | | |
| education | | | | | |
| age | | | | | |
| male | | | | | |
| good_selecta | | | | | |
| ansrgrq10r | | | | | |
| ansrgrq17r | | | | | |
| riskfactor | | | | | |
| ammeasure2 | | | | | |
| totalcorrect | | | | | |

| r~k | answerq10 | .3156311 | .4131187 | .76 | 0.445 |
| feedback_n | | | | | |
| feedback_f | | | | | |
| income | | | | | |
| education | | | | | |
| notincome | | | | | |
| education | | | | | |
| notincome | | | | | |
| _cons | | | | | |

| r~k | /lnsigma | .7159771 | .0395072 | 18.12 | 0.000 |
| | correctp | | | | |
| | answerqc | | | | |
| american | | | | | |
| /athrho | | | | | |

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.01 Prob > chi2 = 0.9230

57