Ambiguity in Performance Pay: An Online Experiment

PRELIMINARY DRAFT

April 12, 2013

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

David B. Johnson
Florida State University

Abstract: Many incentive plans do not include an explicit mapping between workers’ effort and compensation. Potential employees therefore face both risk and uncertainty when choosing whether to accept an employment contract. Using an online labor market, Amazon Mechanical Turk, we study the willingness of individuals to accept a contract to perform a real-effort task as well as their performance at this task. We manipulate the information potential workers have about how a performance bonus is assigned and their ability at the task. We find that ambiguity about the relationship between a worker’s performance and their bonus affects both the willingness of individuals to accept a contract and their willingness to complete the task. Ambiguity has little effect on performance subject to completing the task. Information about ability reduces willingness to accept a contract and also affects willingness to complete the task. Task performance is not affected by information about ability.
1. Introduction

Incentive contracts that tie 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 good example of this sort of incentive scheme, with common versions using a fixed and known formula to calculate a sales person’s commission as a function of their sales revenue. 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. This is the single most important incentive system that most junior faculty face. While senior faculty can usually give some 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? Is a publication in Econometrica worth two publications in Games and Economic Behavior? More or less?

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 journals ranked at least as highly as Games and Economic Behavior. Does a newly minted PhD really know how difficult it will be for them to publish six such articles in six years? It seems unlikely that new professors know 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 (a.k.a. uncertainty). Beyond not knowing what their incentive pay will be, they don’t know the process generating their incentive pay. From a technical point of view, this is equivalent to not knowing the distribution of incentive pay subject to effort. Our paper studies how ambiguity in incentive contracts affects employees. 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?

An equally important question is why an incentive contract might include ambiguity. A number of potential answers suggest themselves, including those generally advanced for incomplete contracts (i.e., complexity and flexibility to respond to unforeseen circumstances) as well as reason related to other-regarding behavior (i.e., black box compensation is intended to limit infighting over compensation). It is

---

1 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). The percentage varies from dealership to dealership, and it is common to have incentive plans beyond the commission. (http://carsalesprofessional.com/car-salesman-commission/)
obviously important to understand why a contract would include ambiguity (as well as determining the optimal level of ambiguity), but answering this question requires knowing how workers will respond to ambiguity in incentive contracts. The purpose of our paper is to take this necessary step.

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. The coding task is a challenging task where it is difficult for a worker to tell how well they are doing. As such, it is a good task for inducing ambiguity. Subjects earn a bonus based on their performance at the coding task. The transcription task is a sure thing. It takes a fixed amount of time to complete, identical to the amount of time for the coding task, and yields a certain and known payoff.

We vary subjects’ information about the coding task along two dimensions: (1) the relationship between their performance at coding 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 at coding and their 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 correctly coded messages. 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 but do not know the relationship between correct codings and the bonus (low ambiguity).
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. This represents the maximum a subject actually earned rather than the $1.35 maximum a subject could have earned.

To get a measure of subjects’ likely performance at the coding task, we had them code five practice messages prior to selecting a 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 involved telling subjects that we were classifying their codings as correct or incorrect, weakening the ambiguity in this treatment.

To generate hypotheses about treatment effects, we develop a model using a multiple priors approach related to the well-known maxmin expected utility model of Gilboa and Schmeidler (1989). The model predicts that pure reductions in ambiguity should increase willingness to select the coding task. This implies that the take-up rate for the coding task increase as we move from high ambiguity ($1.35) to low ambiguity to the piece rate treatment. The model also predicts that the take-up rate should be higher in the high ambiguity treatment with a maximum possible bonus of $1.35 versus $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 subject to choosing the coding task, as any such prediction would rely on strong assumptions about the relationship between the state of the world and the second derivative of the reward with respect to the state of the world.

Some of the results are in line with our hypotheses, but others are not. Relative to the piece rate treatment, subjects are significantly less likely to select and complete the coding task in the fixed and, more importantly, low ambiguity treatments. Holding the maximum possible bonus fixed at $1.35, subjects are significantly more likely to select and complete the coding task with high ambiguity than with low ambiguity. Willingness to select and complete the coding task under high ambiguity responds strongly to the maximum possible bonus, as predicted, with the proportion doing the coding task dropping significantly when the maximum bonus is reduced from $1.35 to $1.00. Providing subjects with feedback about their performance on the practice messages also significantly lowers willingness to select and complete 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’ overconfidence, 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. Real effort experiments often don’t allow for quitting, but this is quite natural in a field environment like AMT. Subjects working on AMT aren’t in a lab where they cannot easily leave. If they don’t like the experiment or think they can make more money doing something else, they can always switch to another webpage. 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. Relative to the piece rate treatment, the fixed bonus treatment and both high ambiguity treatments significantly increase the probability of quitting from the coding task. The low ambiguity treatment does not have a significant effect on quitting, but giving subjects feedback about their performance on the practice questions significantly increases quits (when performance is payoff relevant). The fixed treatment and the high ambiguity treatment with a maximum bonus of $1.35 both lower performance at the coding task, but these effects vanish if we control for the skill level of subjects who select the coding task or only look at subjects who complete the task. Like Dohmen and Falk (2011), we find that much of treatment effect on performance is due to selection. What is novel is the important role played by quitting as a determinant of performance.

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 the information subjects have about an incentive contract and/or likely performance is an important determinant of how well a contract works, both in terms of attracting and keeping workers. Second, our experiment examines an aspect of performance that has not received much previous attention, quitting the task. Worker quits are a major cost to firms, and form an important element of worker
performance. If we did not have data on subjects quitting the coding task, we would miss an important effect of the treatments on worker performance.

While not the primary purpose of our work, we also contribute to the literature studying choice under 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 (i.e., Hurwicz, 1951; Segal, 1987; Gilboa and Schmeidler, 1989; Schmeidler, 1989; Ahn, 2003; Ergin and Gul, 2004; Halevy and Feltkamp, 2005; Klibanoff, Marinacci, and Mukerji, 2005 and 2009; Nau, 2006; Seo, 2009; Cerreia-Vioglio, Maccheroni, Marinacci, 2011). A number of experiments have tried to sort out between the various theories. 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, Katsučák, 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.

Our experiment is not intended to separate specific theories or to assign subjects to types, and, unlike most of the preceding studies, we do not have subjects choosing between carefully constructed gambles. However, it is clear that our subjects are, on average, sensitive to ambiguity. The surprising result is that this sensitivity depends on the nature of the ambiguity. The treatment effect of low ambiguity is consistent with a fairly standard model of uncertainty aversion in the spirit of MMEU, but the reaction to high ambiguity does not fit well with such models. The lack of any difference between high ambiguity with a $1.35 maximum bonus and the known piece rate treatment could be interpreted as evidence of ambiguity neutrality, but this would not explain the results of the low ambiguity treatment. We conjecture that the type of extreme ambiguity subjects encounter in the high ambiguity treatments pushes them to adopt very 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 high ambiguity treatment is surprisingly attractive. We cannot confirm this conjecture directly with our data, as the experiment was not designed to closely 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 is in field settings. We think 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 has a cost 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

---

2 See also Abdellaoui, Baillon, Placido, and Wakker (2011). This is more concerned with developing and applying a theory of choice under uncertainty than testing between various theories.
employer’s point of view. The usual caveats about generalizing results from one specific setting apply, but our results have important practical implications for the use of incomplete contracts.

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. 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. Running an experiment on AMT is also 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 of pay for a laboratory study at Florida State, the dataset would have cost roughly $14,000. High costs of running experiments have long been a barrier to entry into the field. Most graduate students could afford to spend $500 on a study, but not $14,000. If we want experimental economics to continue to thrive, we should embrace methods of gathering data that are less financially onerous.

There are tradeoffs involved in using AMT. The data is cheap, but 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, while less expensive, 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. Payment of a bonus is completely at the requester’s discretion. Requesters control how much information 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.

---

3 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.

4 The senior co-author on this paper can obviously afford to pay more than $500 to complete a study. This study began in part as a personal experiment with using AMT as well as a test of whether coding of messages from experiments could be done cheaply and efficiently via crowdsourcing (it can!).
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. 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 has a valuable role to play in opening up experimental economics to researchers who do not have large research 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 encompass a far broader range of ages than the typical college population, ranging from 18 to 69. (Due to IRB restrictions, we did not use subjects who were younger than 18.) 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.

There are several obvious 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 perfectly 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 (the average subject only took 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. 24% of our subjects report that AMT is 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

\[\text{Insert Table 1}\]

---

5 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).

6 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.

7 The pre-experimental survey asked subjects why they complete tasks on AMT. Subject could choose from six options which were drawn from the survey in Ipeirotis (2010). Subjects could check as many or few items as desired. We classified a subject as motivated by earning money if they choose “fruitful way to spend free time and get some cash,” “for ‘primary’ income purposes (e.g. gas, bills, groceries, credit cards),” “for ‘secondary’ income purposes, pocket change (for hobbies, gadgets, going out),” or “I am currently unemployed, or have only a part time job.” For the latter category, a subject was only classified as motivated by earning money if they did not also check a category which was inconsistent with a desire to earn money (i.e. “to kill time” or “I find the tasks to be fun”).
GDP of about $1500/year.\(^8\) Average pay for a college graduate in India is only 547 rupees/day, which works out to slightly more than $10/day.\(^9\) 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 there is little reason to believe that participants in real world labor markets never 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 interesting data.

To summarize, like any population used by an experimenter, using workers on AMT has its advantages and disadvantages. We had specific reasons for using AMT. We wanted to gather a large number of observations, and would have had financial difficulty doing so if we had to pay a large amount per subject. The varied population available on AMT was also a major attraction. We doubt that the use of AMT had a major effect, one way or the other, on the validity of our results. Increasing our confidence in the validity of experiments run on AMT, Horton, Rand, and Zeckhauser (2011) replicate a number of standard experiments on AMT and report results that look little different from the typical results of laboratory experiments.\(^10\)

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

Task A used data from Charness and Dufwenberg (2006). In 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

---

\(^8\) GDP figures retrieved February 11, 2013 using the World Bank’s World Development Indicators Online (WDI) database.


\(^10\) This paper also provides an excellent discussion of using AMT to conduct experiments. For another in depth discussion of the advantages and disadvantages of AMT, see Paolacci, Chandler, and Ipeirotis (2010). Numerous studies have examined the demographics of workers on AMT (e.g. Paolacci *et al* (2010) and Huang, Zhang, Parkes, Gajos, and Chen (2010)) as well as the quality of work produced (e.g. Marge, Banerjee, and Rudnicky (2010)).
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 B. 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 their paper. Task A asked subjects to classify 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 messages as a form of human verification. This makes it more difficult to copy and paste the text into either a search engine or a translation program, as well as making impossible for an internet bot to complete the task. Multiple people were used to write out the messages, so the hand writing varied from message to message, but we made certain that all of the messages were legible. Subjects were asked to classify messages into six possible categories, 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. We scored a message classification as correct if it matched the modal classification from the population of experimental participants. This scoring rule

---

11 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.

12 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.
draws on the coding methodology developed in Xiao and Houser (2011). 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. Not surprisingly, the messages that generated the least agreement about the correct coding were those that were the longest and required coding multiple categories. Coding is an inherently subjective exercise, but for all fifteen questions the modal coding is close to what we would have chosen as informed experts. Our data analysis includes robustness checks for different measures of whether a coding is correct.

Figure 2: Task A Example

4)

- Promise to Roll (explicit or implied)
- Appeal for Trust
- Appeal for Individual Monetary Incentive
- Appeal to Joint Payoffs
- Appeal to Fairness
- Attempt to Build Rapport (e.g., jokes, friendly banter)
- None of the Above

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 have known 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 categorized, but categorization is sufficiently difficult that subjects were not faced with a trivial task.

A secondary reason for conducting this research on AMT and using message classification as the task was that one of authors (Cooper) conducts large amounts of research that involves coding verbal protocols. This can be extremely expensive and time consuming, so he was interested in exploring the usefulness of AMT as a way of getting material coded. As it turns out, AMT is great way to get fast, cheap, and reliable coding done!

In Task B, the transcription task, workers were shown a series of fifteen words (see Figure 3). Each word had to be typed correctly into a text box. The words are shown as JPG images to prevent copying and pasting or use of robots. To make the task as unambiguous as possible, we chose words that are relatively
Figure 3: Task B Example

Figure 4: Experiment Design

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 seems unlikely that workers would have failed 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 much 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.13 This collected demographic information (age, gender,

13 A full copy of the survey, along with all of the experimental materials, is contained in Appendix A.
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, A., Dürsch, P., & Lefort, J. P., 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 (message classification) 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 focused on making it clear that Task B was a sure thing. Before selecting a task, workers coded five messages for practice. After completing this task, workers were asked how they thought they did on the practice task in comparison to other workers. The wording of this question was deliberately vague to avoid destroying 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 average response was 2.42, indicating subjects on average thought they were better than average. Men were significantly more confident than women (t = 2.38; p < .05), and Indians were significantly more confident than other subjects (t =4.75; 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 serves as an outside option. It has a fixed and known payoff. We randomly varied the payoff for Task B between 30, 40, and 50 cents. This was resolved prior to the experiments and subjects knew what

---

14 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.
value had been drawn. Based on calibration exercises we expected that the average bonus for Task A would be about 40 cents (this is quite close to the realized average of 42.8 cents for subjects who select and complete Task A), so payoffs for Task B were set to roughly match those 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 stress that Task B is a sure thing and takes the same amount of time as Task A.

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. In practice, 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.

The treatments for experiment vary what information subjects had about Task A prior to choosing a task. We vary information along two dimensions: what subjects know about the incentive contract and what subjects know about their own likely performance. This is a between subjects design, so each subject choose between tasks only once and participated in a single treatment.

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 (162/864) 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.

*Treatments: Information about the Incentive Contract*

*Fixed:* Task A pays 40 cents regardless of performance. This contract is a sure thing and gives no incentives for performance. There were two reasons to include this treatment. The primary purpose was 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. We can imagine subjects who are willing to sacrifice some 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). The Fixed treatment also allows us to measure the effect of having an incentive contract on performance in the absence of ambiguity. This is not the most novel feature of our experimental design, but still an issue of interest to economists.

*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

---

15 They were not even told that there had been a random draw to avoid any mistaken impression that the outcome for the outside option was random.
16 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.
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 some 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. Even if a subject knows the probability that a single message is
coded correctly, they cannot know the distribution over the total number of messages coded correctly or
the distribution over bonuses from Task A.

**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 at least know how performance will be measured.

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 that
subjects were overly sensitive to the maximum possible bonus. To test this, we added cells of the HA
treatment were subjects were told that the maximum possible bonus was $1.00 (HA100). This was not
deceptive since the maximum bonus earned by any subject up to that point in time was only $0.99 (and
this was after we had gathered hundreds of observations). Since the primary 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. 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 questions coded correctly and the number of
questions coded correctly in Task A (for subjects who completed Task A) is .40, so knowing the number

---

17 Subsequently there was a subject who earned $1.08.
18 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.
of practice questions coded correctly is a good measure of likely performance at Task A.\(^\text{19}\) 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 does not eliminate the uncertainty about how likely a subject is to code a message correctly, since coding practice questions correctly is not perfectly correlated with performance at Task A, but we think it is fairly obvious that the uncertainty is reduced. Feedback also improves subjects’ calibration about their ability at the task, 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 can always quit the experiment by leaving the website or prematurely submitting the HIT. We have 106 observations (out of 702 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 of the 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 this outcome. As will be seen, the information subjects have about their incentive contract and likely performance affects the likelihood of quitting.

Bonuses were paid shortly after the experiment was completed. Payment always took place within a week of completing an experiment. Subjects knew that payment of the bonus would be delayed. This is not unusual for AMT and would not have been surprising for the subjects.

**D) AMT Specific Design Issues:** Our experiments contained a number of features that address AMT specific issues. AMT workers come from all over the world, and differ widely in their ability to comprehend English. The experimental materials inform subjects that English fluency is required to complete the HIT, but this requirement is unenforceable. The survey therefore included two questions testing workers’ English comprehension. Both questions have only one correct answer and the correct answer is 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 source of concern is primarily 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. Our HIT is not particularly attractive for this type of worker, as it is relatively intricate and has built in timers (see below) that prevent workers from clicking through as fast as possible, and AMT actively tries to screen out workers who have a history of providing poor quality work.

\(^{19}\) The correlation between these variables is easily significant at the 1% level.
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 serves 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. Unsurprisingly, Indian subjects were significantly more likely to be dropped than subjects from other nations (95% of the other subjects are from the US, Canada, or Western Europe). We suspect that most of the dropped subjects comprehended English but weren’t reading the questions carefully. Given the nature of our experiment, we did not want to use observations 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. With time delayed buttons we can ensure that the minimum time required to complete each task is the same. Because each task takes the same amount of time to complete, workers’ preferences for Task A relative to Task B should depend on the value of the outside option (Task B payoff), their beliefs about the bonus from Task A, and the inherent attractiveness of performing Task A versus Task B. It should not depend on the unobservable opportunity cost of their time. Using time delayed buttons also makes it impossible for workers to click through the HIT as fast as possible. Presumably workers who view the opportunity cost of their time as being relatively high are more likely to use this approach and also were more likely to quit in our experiment.

Another concern with using AMT is that subjects either use multiple accounts simultaneously to be paid for the task more than once or are colluding 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), we were able to prevent the same person from participating more than once unless they used multiple accounts. In spite of the preceding, the concern remains that there are workers who are either simultaneously working on multiple accounts or colluding to complete the task. The nature of our task makes this less likely, as it is relatively involved and time consuming. The sorts of workers who try to work on multiple accounts are also the types 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 the message coding or the bonus) was being shared. Ex post, we checked our data for autocorrelation. If subjects were working together or

---
20 Subjects could spend as much time as desired beyond the 30 seconds. The realized average time to complete the experiment was identical for the two tasks (as noted previously).
21 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 completing multiple versions of the experiment, we should see correlation across time of completion. There is no significant autocorrelation either in task selection or performance.\textsuperscript{22}

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. The use of timers meant that the HTML could not be used to determine how many messages would be coded in Task A (the HTML can only be viewed for the current screen). In sessions with feedback, the portion of the HTML used for determining whether or not the worker correctly coded the practice messages is made intentionally convoluted.\textsuperscript{23} This was done to make it extremely difficult to use the HTML to cheat on practice questions. While there are no incentives to cheat on the practice questions, we thought it best to eliminate this possibility.

3. Theory and Hypotheses

We develop a relatively simple theory of choosing a task to give some structure to our hypotheses. We take a multiple priors approach to modeling ambiguity, but combine the maxmin approach due to Gilboa and Schmeidler (1989) with subjective expected utility. The maxmin model of Gilboa and Schmeidler is a well-known model of ambiguity aversion that is both tractable and provides an explanation for the Ellsberg Paradox. \[\text{Add some material about how this relates to the theory lit. Maybe Gilboa 1988 or Jaffray 1988.}\]

A subject in our experiment faces a choice between an option with an unknown value (Task A) and a sure thing (Task B). Let R be the reward for Task B. Let \{r_0, r_1, r_2, \ldots, r_n\} be the potential rewards from Task A. In terms of our experiment, potential rewards for Task A range from 0 to 135 cents. If a subject chooses Task A, they then choose an effort \(e \in [e_{\text{min}}, e_{\text{max}}]\) to put into coding. Let the cost of effort be given by \(c(e)\), where \(c'(e) > 0\) and \(c''(e) > 0\). We assume that costs of effort represent are known, not a source of uncertainty. The objective probability of earning \(r_i\) is given by \(f(e, r_i)\) and the expect reward is calculated in the usual fashion. Let \(\text{ER}(e)\) stand for the objective expected reward.

Suppose a subject believes there are \(m\) possible states of the world, where each state of the world corresponds to a different distribution over rewards as a function of effort. Let \(\gamma_j\) be the subjective probability the subject puts on the \(j\)-th state of the world, where \(\sum_{j=1}^{m} \gamma_j = 1\) and let \(g_{ij}(e, r_i)\) be the subjective probability of earning \(r_i\) in the \(j\)-th state of the world subject to choosing effort \(e\), where \(\sum_{i=1}^{n} g_{ij}(e, r_i) = 1\). If

\textsuperscript{22} 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.

\textsuperscript{23} 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.
the subject correctly reduced the resulting compound lottery, the subjective expected reward (SER) is given by the following:

$$\text{SER}(e) = \sum_{j=1}^{m} \gamma_j \left[ \sum_{i=1}^{n} (g_{ij}(e,r) \tau_j) \right]$$  \hspace{1cm}(1)$$

Ignoring the issue of ambiguity aversion, there is no reason to believe that a subject’s subjective beliefs are calibrated correctly. Let \((\text{SER}(e) - \text{ER}(e))\) be defined as a subject’s calibration error. If the calibration error is positive (negative), the subject is overconfident (underconfident).

To capture the idea that a subject is ambiguity averse, we model them as putting extra weight on the worst possible state of the world. Define \(r(e)\) and \(j(e)\) as follows:

$$r(e) = \min_{j=1,2,...,m} \left( \sum_{i=1}^{n} (g_{ij}(e,r) \tau_i) \right)$$  \hspace{1cm}(2)$$

$$j(e) = \arg \min_{j=1,2,...,m} \left( \sum_{i=1}^{n} (g_{ij}(e,r) \tau_i) \right)$$  \hspace{1cm}(3)$$

The subject’s alpha maxmin expected reward (AMMER) is given by the following equation:

$$\text{AMMER}(e) = \frac{\alpha \gamma_j(e) \tau_j(e) + (1 - \alpha) \text{SER}(e)}{1 - \alpha \gamma_j(e)}$$  \hspace{1cm}(4)$$

One critical feature of this specification should be noted; the additional weight on the worst possible state of the world is a function of its subjective probability. The worst state of the world is always overweighted, but less so if it is considered unlikely. This captures the idea that some situations contain more uncertainty than others, and ambiguity aversion should be lower at settings with less uncertainty.

Within this framework, we model a change in a subject’s information as changes in \(\gamma\) and \(g\) (the weights on different states of the world and subjective probabilities subject to the state of the world). Abusing notation, define \(\text{SER}(\gamma,g,e)\) to be the subjective expected reward as a function of \(\gamma\), \(g\), and \(e\). \(\text{AMMER}(\gamma,g,e)\) is defined in an analogous fashion. Let \(\text{SER}(\gamma,g,e \mid j)\) be the subjective expected reward subject to the \(j\)-th state of the world being realized. A change in a subject’s information from \((\gamma, g)\) to \((\hat{\gamma}, \hat{g})\) is a pure reduction in uncertainty if it fulfills the following conditions.

1) \( g = \hat{g} \)

2) \( \text{SER}(\gamma,g,e) = \text{SER}(\hat{\gamma},g,e) \forall e \in [e_{\min}, e_{\max}] \)

3) Consider state \(j\) and \(k\) such that \(\text{SER}(\gamma,g,e \mid j) < \text{SER}(\gamma,g,e \mid k) < \text{SER}(\gamma,g,e)\). The following condition must hold:

$$\frac{\gamma_j}{\gamma_k} > \frac{\hat{\gamma}_j}{\hat{\gamma}_k}$$

4) Consider state \(j\) and \(k\) such that \(\text{SER}(\gamma,g,e \mid j) > \text{SER}(\gamma,g,e \mid k) > \text{SER}(\gamma,g,e)\). The following condition must hold:
The first condition implies that the change in information only affects a subject’s beliefs about the likelihood of different states of nature, not his beliefs about what happens subject to a specific state of nature. The second condition implies that the change in information does not affect the subject’s calibration error. The final two conditions imply that weight is moved away from more extreme states of the world. It follows that $\gamma_{\hat{\gamma}}$ is reduced.

It is simple to prove the following proposition. This proposition implies that changes in information that represent a pure reduction in uncertainty will lead to weakly more choices of Task A.

**Proposition 1:** Suppose a subject chooses the option with an unknown value subject to information $(\gamma, g)$. If $(\hat{\gamma}, \hat{g})$ is a pure reduction in uncertainty from $(\gamma, g)$, the subject will also choose the option with an unknown value subject to information $(\hat{\gamma}, \hat{g})$.

**Proof:** Let $e^*$ be the effort chosen subject to information $(\gamma, g)$. By the definition of a pure reduction in uncertainty, $r(e^*)$ is unaffected by the change in information and $\gamma_{(\hat{\gamma}, \hat{g})}$ is reduced. The derivative of (4) with respect to $\gamma_{(\hat{\gamma}, \hat{g})}$ is negative since $r(e^*)$ is less than $\text{SER} (\gamma, g, e^*)$. It follow that

$$\text{AMMER} (\hat{\gamma}, g, e^*) - c(e^*) > \text{AMMER} (\gamma, g, e^*) - c(e^*) > R.$$ 

Let $e^{**}$ be the effort chosen subject to information $(\hat{\gamma}, g)$. By revealed preference,

$$\text{AMMER} (\hat{\gamma}, g, e^{**}) - c(e^{**}) > \text{AMMER} (\hat{\gamma}, g, e^*) - c(e^*) > R.$$ Q.E.D.

The structure of the model does not explicitly allow for the possibility of quitting. However, given that quitting takes place after choosing Task A, this can be treated as part of the reward structure. In other words, the subjects can be assumed to know the value of quitting (i.e. their time value) and can assign a subjective probability to this being the outcome. It follows that the possibility of quitting does not affect our conclusions.

Moving from the HA135 treatment to the LA treatment to the PR treatment, uncertainty about the contract is being reduced while the range of possible bonuses is unaffected. If we treat these changes as pure reductions in uncertainty, we get the following 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.

Obviously, assume that these are pure reductions in uncertainty is a strong assumption. If giving subjects additional information alters the calibration error, behavior need not shift in the directions predicted by Hypothesis 1. This is particularly germane in the case of giving subject’s feedback about their performance on the practice questions. Because coding is an unfamiliar task for our subjects, it seems to assume they face true uncertainty about their likely performance at this task. Giving them feedback should act as a reduction in uncertainty. However, there are also numerous results suggesting that
individuals are overconfident about their ability to perform tasks (e.g. Niederle and Vesterlund, 2007). This suggests that giving subjects feedback about their performance on the practice question will reduce their calibration error, specifically reducing their overconfidence. Because the reduction in uncertainty and the reduction in overconfidence work in opposite directions, we present two opposing hypotheses about the effect of feedback on task selections. These effects should primarily apply in 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. While subjects might still care about their performance at the task, we expect them to care less. We therefore don’t expect feedback to have much of an effect in the Fixed treatment.

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. Put differently, in HA100 the subjects know for certain that a reward greater than 100 will not be achieved. We could model this as affecting a subject’s beliefs about the state of the world or as changing their beliefs subject to the state of the world. We take the latter approach as it is simple to incorporate into the model and leads to natural predictions. For each state, let $\tilde{g}_{ij}(e,r_i)$ be the subjective probability of earning $r_i$ in the j-th state of the world subject to choosing effort $e$ and the reward being restricted to be less than or equal to a dollar. We derive $\tilde{g}_{ij}(e,r_i)$ from $g_{ij}(e,r_i)$ via Bayesian updating.

$$\tilde{g}_{ij}(e,r_i) = \begin{cases} g_{100}(e,r_i) & \text{if } r_i \leq 100 \\ \sum_{r=0}^{100} g_{ir}(e,r) & \text{if } r_i > 100 \end{cases}$$

It is now straightforward to prove the following proposition. This proposition implies that moving from HA135 to HA100, holding the outside option fixed at 50 cents, will reduce the proportion of subjects choosing Task A.

**Proposition 2:** Suppose a subject chooses the sure thing subject to information $(\gamma, g)$ with a maximum reward of 135. If the maximum reward is reduced to 100 and the information is otherwise unchanged, the subject must still choose the sure thing as long as $j(e)$, the state of the world that minimizes the subjective expected reward as a function of effort, is unaffected by this change.

**Proof:** For every state of the world $j$, $\text{SER}(\gamma, \tilde{g}, e | j) < \text{SER}(\gamma, g, e | j)$. Thus the second term of the numerator of (4) must be decreased if $e$ is held fixed. Since $\gamma_j(e)$ must be increased and $r(e)$ is decreased by limiting the maximum reward, it follows that $\text{AMMER} (\gamma, \tilde{g}, e) < \text{AMMER} (\gamma, g, e)$. Let $e^*$ be the
optimal effort with a maximum reward of 135 and $e^{**}$ be the optimal effort with a maximum reward of 100. By revealed preference, $\text{AMMER}(\gamma, g, e^{**}) < \text{AMMER}(\gamma, g, e^{*}) < \text{AMMER}(\gamma, g, e^{*}) < R$. Q.E.D.

Note that we must assume that restricting the maximum reward does not affect the identity of the worst state of the world holding effort fixed. This assumption is necessary because the weight on the worst state of the world is a function of its subjective probability. Even though $g(e)$ must be reduced by restricting the possible rewards, $\gamma_{j(e)}$ could also be reduced if the identity of the worst state changes. This could lead to an increase in AMMER.

Proposition 2 implies the following hypothesis:

**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.

Many of the results presented in Section 4 concern performance, both in terms of messages coded correctly and the probability of quitting, subject to information about the incentive contract and information about performance on the practice questions. We did not have any specific hypotheses about these performance measures ex ante. To make a prediction would require knowing how the derivative of the subjective expected reward with respect to effort varies with respect to changes in the state of the world. Assuming that this derivative is increasing in the state of the world, already a heroic assumption, is not sufficient to generate predictions since pure reductions in uncertainty reduce the weight on both the best and worst states of the world. To make predictions ex ante requires an assumption about the second derivative with respect to the state of the world as well. Since there was no natural assumption to make about this, it seemed wiser to let the data speak for itself.

4. Results

As noted previously, unless otherwise stated, all results are based on the 702 subjects who passed the English screen and filled out the survey. Table 3 summarizes our data by treatment. The first three columns examine task selection (whether subjects select and complete Task A) while the final two look at performance (number of messages coded correctly). Task selection can be viewed as a multi-step process. To complete Task A, a subject must both select this task and not quit before completing it. We therefore distinguish between subjects who select Task A and those who select and complete Task A.

[Insert Table 3]

A. Task Selection: 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 Task A in the HA135 treatment relative to the PR treatment runs contrary to this prediction. The frequency of quitting Task A is higher in the HA135 treatment than in the PR treatment, but the proportion of subjects who select and complete Task A still remains slightly higher in the HA135 treatment than in PR treatment.

We did not predict ex ante how feedback about performance on the practice questions affects task selection. Table 3 shows that feedback consistently lowers the percentage of subjects who select and complete Task A. These effects are 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 the percentage of subjects choosing Task A smaller in HA100 than HA135. This difference remains large if we either limit the data to observations with an outside option of 50 cents (56.9% select Task A in HA135 vs. 36.7% in HA100) or consider whether subjects select and complete Task A.

Subjects respond strongly to incentives in Fixed, the one treatment where the incentives are unambiguous. Comparing outside options of 30, 40, and 50 cents, the percentages choosing Task A are 19%, 35%, and 47% respectively. This provides clear evidence that the incentives are salient for subjects. It is only mildly surprising that the majority of subjects choose Task A even when it unambiguously lowers earnings, since Task A is somewhat interesting while Task B is boring. The percentage selecting and completing Task A is lower in Fixed than any other treatment. This makes sense given that subjects are overconfident on average, suggesting that they are systematically overestimating their earnings from Task A in the other treatments.

Strong conclusions about task selection cannot be drawn from the raw data summarized in Table 3. By design the value of the outside option varies between subjects. While this was intended to be balanced between treatments, the random pattern of who fails the English screen means that the sample is not truly balanced. Likewise, even with random assignment to treatments 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, completing the process, choose and complete Task A.

[Insert Table 4]

All of the 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 available from the authors upon request.

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 only includes controls for the treatment and 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 anticipated a weaker effect from feedback in this treatment.

The parameter estimate for LA is negative and significant at the 5% level. 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 ($z = 3.07; p \leq .01$). Contrary to Hypothesis 1, high ambiguity raises willingness to choose Task A relative to low ambiguity 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 ($z = 2.07; p \leq .05$). Both parameter estimates for the effect of feedback are negative, but neither is statistically significant at conventional levels. 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 ($z = 3.15; p < .01$). This provides further evidence that subjects respond to the incentives in the one treatment where the bonus payment is completely unambiguous.

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 actually 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 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. As described above, the ambiguity measure is based on whether the subject’s answers were consistent with the Ellsberg Paradox (indicating ambiguity aversion). 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 as they are neither statistically significant nor of great economic interest. As noted previously, 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 somewhat surprised that neither the risk nor ambiguity measures were statistically significant. These measures do not predict whether subjects will select Task A, but, as will be seen below, they have a significant impact on subjects’ decisions to quit subject to choosing Task A.

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 select Task A. For the Heckman model to be identified, we need an exogenous source of variation that is correlated with

---

24 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).

25 As an alternative we have broken down nationalities into five categories (North American, Indian, EU, Southeast Asia, and other). This does not significantly improve the fit. The main results are unaffected, although the difference between HA135 and HA100 returns to significance at the 5% level ($z = 1.98$).

26 The risk measures also fail to achieve joint significant ($\chi^2 = 3.84, 3$ d.f., $p > .10$).
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.\textsuperscript{27} 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, Fixed, HA100, and HA135 all significantly increase the frequency of quitting. Feedback also increases to frequency of quitting, albeit weakly, in the cases where performance at the task is payoff relevant (PR and LA). These results are somewhat weakened by the inclusion of additional variables in Model 4, although all four parameter remain statistically significant at least at the 10% level. Looking at the added variables, subjects who are less risk averse or less ambiguity averse are significantly more likely to quit. Subjects who are relatively confident about their abilities quit significantly less.\textsuperscript{28}

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 effect of feedback in the PR and LA treatments, which now has a significant (at the 5% level) negative effect. Feedback in these two treatments weakly decreases the proportion of subjects choosing Task A and increases the percentage who quit. The combination of these two effects has a strong effect on whether subject select and complete Task A. This result is consistent with Hypothesis 2B. Due to similar logic, confidence also has a weakly positive effect on whether a subject selects and completes 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.\textsuperscript{29} The result is quite intuitive. Subjects who receive bad feedback become less likely to select and complete Task A while those receiving positive feedback become more likely to do so. This is consistent with feedback affecting task selection by reducing calibration errors.

To summarize, the regression analysis largely confirms our observations based on the raw data in Table 3. LA and Fixed both significantly decrease the likelihood (relative to PR) that a subject selects and completes Task A. HA135 significantly increases the probability of selecting and completing Task A relative to LA, but this probability falls significantly between the HA135 and HA100 treatments. When the effects on selecting Task A and quitting from Task A are combined, feedback significantly reduces the probability of selecting and completing Task A in PR and LA.

\textsuperscript{27} 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.

\textsuperscript{28} The results of Model 4 have to be treated cautiously since we are including a large number of explanatory variables and there are only 35 observations where subjects drop out. As a check on 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. The estimates and significance for the risk and ambiguity measures are little affected relative to Model 4, but the estimate for confidence shrinks and no longer achieves statistical significance. The estimate for feedback in the PR and LA treatments dips just below significance at the 10% level.

\textsuperscript{29} 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.
Result 1: Relative to the PR treatment, the LA treatment lowers the probability of choosing and completing Task A. This is consistent with Hypothesis 1.

Result 2: Relative to the PR treatment, the HA_{135} treatment slightly raises the probability of choosing and completing Task A. This is inconsistent with Hypothesis 1. The HA_{100} lowers the probability of choosing and completing Task A relative to HA_{135}. This is consistent with Hypothesis 3. Relative to the PR treatment, the HA_{135} treatment increases the probability of choosing Task A and quitting.

Result 3: For the PR and LA treatments (pooled), feedback about performance on the practice questions lowers the probability of choosing and completing Task A. This effect is sensitive to the content of the feedback, providing further evidence that feedback reduces calibration errors. This is consistent with Hypothesis 2b.

Result 4: Relative to the PR treatment, the FIXED treatment lowers the probability of choosing and completing Task A and increases the probability of choosing Task A and quitting.

B. Performance: The hypotheses developed in Section 4 focus on the effects of the various treatments on whether subjects choose Task A, but their performance at Task A is also of interest. Performance has two dimensions, whether subjects complete the task and, if so, how well do they do at coding the messages. The former element of performance has already been addressed through the probit regressions reported in Table 4, so we now turn to the latter element.

Our measure of performance is how many messages were coded correctly in Phase 2. A message is counted as correct if the subject’s coding matches the modal coding. This is the criterion that was used to determine whether or not a subject was paid the piece rate for coding a message correctly.

In analyzing subjects’ performance at Task A, a central issue is selection into Task A. Figure 5 illustrates this point. For each treatment (pooling data with and without feedback), Figure 5 graphs the average number of practice questions coded correctly broken down by whether the subject choose Task A or Task B. For PR, Fixed, and HA_{100} there are clear gaps between two groups. For each treatment we ran Wilcoxon rank-sum test to determine whether the distribution of correct codings on the practice questions 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. Performance on the practice questions is a good measure of skill at Task A, and the number of practice questions coded correctly strongly correlates with performance on Task A. The results shown on Figure 5 therefore suggest that selection could play an important role in determining 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.\(^\text{30}\) 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 of the subjects who quit did so after coding some messages, and the average number of messages coded correctly was 0.86 for subjects who quit (versus 4.35 for those who select and complete Task A). The right panel only considers the number of messages coded correctly by subject who choose and completed Task A.

[Insert Table 5 here]

Model 1 is a basic model that only includes the treatment effects as independent variables. Fixed has a strong negative effect on performance and HA100 has a weak negative effect. The estimate for HA10 is slightly larger than the estimate for HA135, but this narrowly misses significance at the 10% level. 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 questions 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

\(^{30}\) 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.
of statistical significance is driven by the addition of the control for how many practice questions were coded correctly. Adding this variable alone is enough to eliminate any treatment effects. The effect of the treatments on coding performance seems to be entirely due to selection effects, echoing the results of Dohmen and Falk (2011).

The negative treatment effects reported in Model 1 rely in part on the subjects who select Task A and quit. Model 3 replicates Model 1, but only looks at messages coded correctly by subjects who completed Task A. The effects of the Fixed and HA135 treatments are reduced in magnitude and no longer achieve statistical significance. Model 4 parallels Model 2. Once again, controlling for performance on the practice questions reduces the magnitude of all the estimated treatment effects, none of which even vaguely approach statistical significance.

Result 5: Relative to the PR treatment, subjects who choose Task A code fewer messages correctly in the Fixed and HA135 treatments. Similar to the results of Dohmen and Falk (2011), these treatment effects are due to selection and vanish after controlling for performance on the practice questions.

As noted in Section 2.D, a potential concern with using AMT is that subjects are either simultaneously completing tasks on multiple accounts or colluding on tasks. We think this is unlikely, for reasons explained previously, but observations will not be 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 our regression analysis 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.

Our dataset divides roughly equally into two groups, Indians and everyone else, with 95% of the non-Indian subjects coming from the USA, Canada, and Western Europe. Beyond the obvious observation that the non-Indians come from countries that are wealthier and better developed than India, the non-Indian subjects have significantly different personal characteristics from the Indians (significantly wealthier, less educated, and more likely to be female). A natural question is how our main results look if we limit the population to non-Indians. As it turns out, the effect is minimal.

For task selection we reran Model 6 from Table 4 with only non-Indian subjects. Relative to PR, LA continues to have a significant negative effect on the probability of choosing and completing Task A. HA135 has virtually no effect relative to PR, but HA100 now has a significantly negative effect and the difference between the two high ambiguity treatments remains significant. Feedback continues to have a strong negative effect in the PR and LA treatments. For task performance we reran Models 1 and 4 from Table 5 with only non-Indian subjects. In the redone version of Model 1, the negative effect of Fixed is only significant at the 10% level and negative effect of HA135 no longer achieves statistical significance. Turning to Model 4, none of the treatment effects are statistically significant (as in the original). The non-Indian population is not a carbon copy of the Indian population, but our main conclusions would be little affected if we were limited to the non-Indian population.
5. Discussion and Conclusions

The purpose of this paper was to examine how ambiguity about the relationship between performance and incentive pay as well as ambiguity about likely performance affects the willingness of individuals to accept and complete a job. We also study the effects on performance subject to completing the job.

Ambiguity has a strong effect on willingness to accept and complete the coding task. Some of these effects are as expected: low ambiguity, decreasing the maximum payoff under high ambiguity, and adding feedback about likely performance all reduce willingness to select and complete the coding task. Less expected is the strong positive effect of moving from low ambiguity to high ambiguity.

The effects on performance are primarily through the probability of quitting. Fixed pay, high ambiguity, and feedback about likely performance all increase quitting relative to the piece rate treatment. The effects on performance subjects to having completed the coding task are minimal. This points to the importance of allowing quitting – without this, we would have missed the primary effect of the treatments on task performance.

A somewhat unusual feature of our experimental design is the use of AMT. Throughout this paper we have been very careful to limit some of 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. 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 by the experiments, but it does open up experiments to individuals who cannot easily afford to spend many thousands of dollars on an experiment. We naturally hope that our results will be replicated, using both lab and field populations.

The purpose of this study was not to identify the optimal role of ambiguity in an incentive contract, since answering such questions requires knowing how workers react to ambiguity. That said, a natural step for follow-up work is to study when it pays for a firm to use incomplete contracts that induce ambiguity. There are natural reasons to want ambiguity in a contract, but our study suggests that there are also likely to be costs. It is an empirical question how the costs and benefits will balance out.
Bibliography


### 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>702</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td>Age</td>
<td>702</td>
<td>31.62</td>
<td>9.94</td>
</tr>
<tr>
<td>Indian</td>
<td>702</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td>American</td>
<td>702</td>
<td>0.29</td>
<td>0.46</td>
</tr>
<tr>
<td>College Graduate</td>
<td>702</td>
<td>0.75</td>
<td>0.44</td>
</tr>
<tr>
<td>Income &lt; $12.5K</td>
<td>702</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>Income ≥ $25K</td>
<td>702</td>
<td>0.40</td>
<td>0.49</td>
</tr>
<tr>
<td>Primary Income</td>
<td>702</td>
<td>0.24</td>
<td>0.43</td>
</tr>
<tr>
<td>Motivated by Earning Money</td>
<td>702</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></th>
<th>Usable Observations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$0.30</td>
<td>$0.40</td>
<td>$0.50</td>
<td>$0.30</td>
</tr>
<tr>
<td>Piece Rate, No Feedback</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>Piece Rate, Feedback</td>
<td>35</td>
<td>35</td>
<td>50</td>
<td>23</td>
</tr>
<tr>
<td>Fixed Pay, No Feedback</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>Fixed Pay, Feedback</td>
<td>35</td>
<td>37</td>
<td>35</td>
<td>29</td>
</tr>
<tr>
<td>Low Ambiguity, No Feedback</td>
<td>35</td>
<td>34</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>Low Ambiguity, Feedback</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.35)</td>
<td>35</td>
<td>35</td>
<td>73</td>
<td>31</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>0</td>
<td>0</td>
<td>75</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>4.62</td>
<td>4.62</td>
</tr>
<tr>
<td>Piece Rate, Feedback</td>
<td>47.3%</td>
<td>9.3%</td>
<td>42.9%</td>
<td>4.40</td>
<td>4.85</td>
</tr>
<tr>
<td>Fixed Pay, No Feedback</td>
<td>37.8%</td>
<td>22.6%</td>
<td>29.3%</td>
<td>3.13</td>
<td>3.71</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>62.7%</td>
<td>9.5%</td>
<td>56.8%</td>
<td>3.81</td>
<td>4.09</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>36.7%</td>
<td>13.6%</td>
<td>31.7%</td>
<td>3.41</td>
<td>3.79</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>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
<th>Model 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Select Task A</td>
<td>Dropout from Task A</td>
<td>Select and Complete Task A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed Pay</td>
<td>-0.497*** (0.186)</td>
<td>-0.619*** (0.197)</td>
<td>1.259*** (0.381)</td>
<td>1.153*** (0.447)</td>
<td>-0.714*** (0.19)</td>
<td>-0.806*** (0.199)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>-0.198 (0.216)</td>
<td>-0.219 (0.22)</td>
<td>0.918** (0.461)</td>
<td>0.937* (0.517)</td>
<td>-0.341 (0.218)</td>
<td>-0.345 (0.223)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.35)</td>
<td>0.238 (0.17)</td>
<td>0.174 (0.175)</td>
<td>0.747** (0.32)</td>
<td>0.630* (0.341)</td>
<td>0.098 (0.168)</td>
<td>0.070 (0.173)</td>
</tr>
<tr>
<td>Low Ambiguity</td>
<td>-0.288** (0.138)</td>
<td>-0.315** (0.143)</td>
<td>0.391 (0.324)</td>
<td>0.183 (0.326)</td>
<td>-0.325** (0.138)</td>
<td>-0.337** (0.144)</td>
</tr>
<tr>
<td>Feedback * Fixed</td>
<td>-0.210 (0.198)</td>
<td>-0.125 (0.207)</td>
<td>0.150 (0.373)</td>
<td>0.331 (0.405)</td>
<td>-0.240 (0.208)</td>
<td>-0.197 (0.215)</td>
</tr>
<tr>
<td>Feedback * (Piece Rate + Low Ambiguity)</td>
<td>-0.225 (0.137)</td>
<td>-0.239* (0.143)</td>
<td>0.534* (0.316)</td>
<td>0.561* (0.325)</td>
<td>-0.286** (0.138)</td>
<td>-0.319** (0.144)</td>
</tr>
<tr>
<td>Outside 40</td>
<td>-0.352*** (0.13)</td>
<td>-0.402*** (0.134)</td>
<td></td>
<td></td>
<td>-0.381*** (0.131)</td>
<td>-0.426*** (0.135)</td>
</tr>
<tr>
<td>Outside 50</td>
<td>-0.656*** (0.125)</td>
<td>-0.732*** (0.131)</td>
<td></td>
<td></td>
<td>-0.612*** (0.126)</td>
<td>-0.677*** (0.133)</td>
</tr>
<tr>
<td>Male</td>
<td>0.183* (0.105)</td>
<td>0.040 (0.225)</td>
<td>0.184* (0.105)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothetical Investment Question</td>
<td>0.291 (0.194)</td>
<td>0.726** (0.365)</td>
<td>0.109 (0.201)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguity Aversion</td>
<td>0.045 (0.105)</td>
<td>-0.441* (0.258)</td>
<td>0.149 (0.106)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidence</td>
<td>0.059 (0.062)</td>
<td>-0.317** (0.148)</td>
<td>0.114* (0.064)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: All regressions based on 702 observations. Numbers in parentheses are robust standard errors. Three (***) , two (**) , and one (*) stars indicate statistical significance at the 1%, 5%, and 10% levels.
Table 5: Heckman Analysis of Performance

<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></td>
<td>Select A</td>
<td>Select A</td>
<td>Select and Completed A</td>
<td>Select and Completed A</td>
</tr>
<tr>
<td>Fixed Pay</td>
<td>-1.325**</td>
<td>-0.392</td>
<td>-0.847</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>(0.56)</td>
<td>(0.558)</td>
<td>(0.607)</td>
<td>(0.601)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.00)</td>
<td>-0.998</td>
<td>0.054</td>
<td>-0.701</td>
<td>0.498</td>
</tr>
<tr>
<td></td>
<td>(0.631)</td>
<td>(0.575)</td>
<td>(0.672)</td>
<td>(0.590)</td>
</tr>
<tr>
<td>High Ambiguity (Max = $1.35)</td>
<td>-0.763*</td>
<td>-0.224</td>
<td>-0.520</td>
<td>-0.034</td>
</tr>
<tr>
<td></td>
<td>(0.425)</td>
<td>(0.367)</td>
<td>(0.417)</td>
<td>(0.359)</td>
</tr>
<tr>
<td>Low Ambiguity</td>
<td>-0.258</td>
<td>0.014</td>
<td>-0.091</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>(0.442)</td>
<td>(0.374)</td>
<td>(0.443)</td>
<td>(0.359)</td>
</tr>
<tr>
<td>Feedback * Fixed</td>
<td>0.274</td>
<td>0.114</td>
<td>0.607</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>(0.614)</td>
<td>(0.559)</td>
<td>(0.65)</td>
<td>(0.632)</td>
</tr>
<tr>
<td>Feedback *</td>
<td>-0.036</td>
<td>0.096</td>
<td>0.229</td>
<td>0.309</td>
</tr>
<tr>
<td>(Piece Rate + Low Ambiguity)</td>
<td>(0.442)</td>
<td>(0.379)</td>
<td>(0.448)</td>
<td>(0.397)</td>
</tr>
<tr>
<td>Correct Practice</td>
<td></td>
<td>0.896***</td>
<td></td>
<td>0.880***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.113)</td>
<td></td>
<td>(0.119)</td>
</tr>
<tr>
<td>Log Pseudo Likelihood</td>
<td>-1184.460</td>
<td>-1126.673</td>
<td>-1071.910</td>
<td>-1020.215</td>
</tr>
</tbody>
</table>

Notes: All regressions based on 702 observations. Numbers in parentheses are robust standard errors. Three (***), 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:

![Message from webpage](image)

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

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
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

I am going to choose roll. If you choose IN, we both make more money.

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

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.**

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**

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**

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  
_lincome_2: Dummy For Income (“12,500 - 24,999”)  
_lincome_3: Dummy For Income (“25,000 - 37,499”)  
_lincome_4: Dummy For Income (“37,500-49,999”)  
_lincome_5: Dummy For Income (“50,000 - 62,499”)  
_lincome_6: Dummy For Income (“62,500 - 74,999”)  
_lincome_7: Dummy For Income (“75,000 - 87,499”)  
_lincome_8: Dummy For Income (“87,500-99,999”)  
_lincome_9: Dummy For Income (“100,000 or more”)  
_Ieducatio~2: Dummy For Education (“High School Graduate”)  
_Ieducatio~3: Dummy For Education (“Some College, No Degree”)  
_Ieducatio~4: Dummy For Education (“Associates Degree”)  
_Ieducatio~5: Dummy For Education (“Bachelor’s Degree”)  
_Ieducatio~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.)  
Ammeasure: 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

|                    | Coef. | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|--------------------|-------|-----------|-------|------|-----------------------|
| fixed              | -.4970342 | .1863854 | -2.67 | 0.008 | -.8623429 -.1317255 |
| ha_lcap            | -.1979352 | .2155945 | -0.92 | 0.359 | -.6204927 .2246224  |
| ha_hcap            | .2377822  | .170183  | 1.40  | 0.162 | -.0957703 .5713347  |
| la                 | -.28783  | .1376516 | -2.09 | 0.037 | -.5576222 -.0180378 |
| feedback_f~d       | -.2104438 | .1976684 | -1.06 | 0.287 | -.5978669 .1769792  |
| feedback_n~d       | -.2253349 | .1374497 | -1.64 | 0.101 | -.4947314 .0440616  |
| _Ioutside_40       | -.3518474 | .1296242 | -2.71 | 0.007 | -.605906 -.0977887  |
| _Ioutside_50       | -.6558621 | .1249117 | -5.25 | 0.000 | -.9006846 -.4110396 |
| notincome          | .2079537  | .2983637 | 0.70  | 0.486 | -.3768285 .7927359  |
| _cons              | .5032259  | .1429516 | 3.52  | 0.000 | .2230458 .7834059   |
Probit regression

Number of obs = 702

Wald chi2(32) = 87.33
Prob > chi2 = 0.0000

Log pseudolikelihood = -436.81468 Pseudo R2 = 0.0958

------------------------------------------------------------------
|               Robust
selecta |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------
fixed |  -.6190595   .1966936    -3.15   0.002    -1.004572   -.2335471
ha_lcap |  -.2187146   .2197353    -1.00   0.320    -.6493879    .2119587
ha_hcap |   .1737857    .175048     0.99   0.321     -.169302    .5168735
la |  -.3152781   .1432543    -2.20   0.028    -.5960515   -.0345048
feedback_f-d |  -.1253376   .2074108    -0.60   0.546    -.5318553    .2811801
feedback_n-d |  -.2386534   .1429941    -1.67   0.095    -.5189166    .0416099
_Iincome_2 |  -.1895641   .1392741    -1.36   0.173    -.4625364    .0834081
_Iincome_3 |   .1399571   .1664156     0.84   0.400    -.1862116    .4661257
_Iincome_4 |   .1578422   .1840701     0.86   0.391    -.2029286     .518613
_Iincome_5 |   .0731506   .2547752     0.29   0.774    -.4625162    .5728509
_Iincome_6 |  -.0060969   .2961265    -0.02   0.984    -.5805299    .5783361
_Iincome_7 |   .0280014   .3519903     0.58   0.559    -.6849421    .5449445
_Iincome_8 |   .4039033   .4292039     0.95   0.341    -.4327153    1.240153
_Iincome_9 |  -.1496006   .0236857    -6.40   0.000    -.2958734    -.003323
_Ieducatio-2 |   .4912572   .3760874     1.31   0.191    -.2458605    1.228375
_Ieducatio-3 |   .2056482   .3519903     0.58   0.559    -.4842401    .8953666
_Ieducatio-4 |   .408509    .4292039     0.95   0.341    -.4327153    1.249733
_Ieducatio-5 |   .4191521   .3350828     1.25   0.211    -.2375981    1.075902
_Ieducatio-6 |   .4771508   .348714     1.37   0.171    -.206316    1.160618
age |  -.0054093   .0051319    -1.05   0.292    -.0154676     .004649
male |   .1826782   .1047153     1.74   0.081    -.0225601    .3879164
india |  -.026908   .1652853    -0.16   0.871    -.3508612    .2970452
american |  -.2316626   .1780047    -1.30   0.193    -.5805453    .1172201
answerq10r-k |  -.0196006   .0236857    -0.83   0.408    -.0660237    .0368225
answerq17r-k |   .0029165   .0019393     1.50   0.133    -.0008843    .0067174
riskfactor |   .0998858   .0773607     1.30   0.197    -.0517654    .251483
ammeasure |   .0449869   .1053102     0.43   0.669    -.1614173    .2513911
answerqc |  -.0015568   .0015437    -1.00   0.314     -.004615    .0015014
correctp |  -.0015568   .0015437    -1.00   0.314     -.004615    .0015014
_Ioutside_40 |   .4015538   .1339161     3.00   0.003     .1395487    .6640245
_Ioutside_50 |  -.7315941   .1053102    -6.90   0.000    -.9361599    -.527028
notincome |  -.1467708   .3019903    -0.49   0.623    -.7408734    .4473324
_cons |   .2397183   .4598389     0.52   0.602    -.6615494    1.140986
------------------------------------------------------------------
Probit model with sample selection

| Number of obs |  702 |
| Censored obs  |  386 |
| Uncensored obs|  316 |

Wald chi2(7) = 16.16
Prob > chi2 = 0.0237

Log pseudolikelihood = -552.4399

| Coef. | Std. Err. | z    | P>|z| | 95% Conf. Interval |
|-------|-----------|-----|-----|-----------------|

**dropout**

| fixed | 1.258869  | .3809935 | 3.30 | 0.001 | .5121357    | 2.005603 |
| ha_lcap | .9184178  | .4611126 | 1.99 | 0.046 | .0146537    | 1.822182 |
| ha_hcap | .7472265  | .320127  | 2.33 | 0.020 | .1197891    | 1.374664 |
| la     | .3908725  | .3236559 | 1.21 | 0.227 | -.2434814   | 1.025226 |

**selecta**

| fixed | -.497481  | .1861092 | -2.67 | 0.008 | -.8622483   | -.132737 |
| ha_lcap | -.1998098 | .215903  | -0.93 | 0.355 | -.6229718   | .2233523 |
| ha_hcap | .2371607  | .1700461 | 1.39 | 0.163 | -.0961235   | .570445 |
| la     | -.2876104 | .1377339 | -2.09 | 0.037 | -.5575639   | -.0176569 |

| feedback_f-d | -.2098959 | .1975584 | -1.06 | 0.288 | -.5971033   | .1773115 |
| feedback_n-d | -.2247582 | .176052  | -1.36 | 0.177 | -.6494594   | .197453 |
| notincome   | -.2065829 | .1358622 | -1.53 | 0.125 | -.474948    | .0627937 |
| _Ioutside_40 | -.3574844 | .1358622 | -2.63 | 0.009 | -.6237695   | -.0911993 |
| _Ioutside_50 | -.6555602 | .1249611 | -5.25 | 0.000 | -.9003421   | -.4107782 |

| _cons | .5048661  | .1429621 | 3.53 | 0.000 | .2246655    | .7850667 |

| athrho | .1489292  | .7187252 | 0.21 | 0.836 | -1.259746   | 1.557605 |

| rho    | .1478378  | .7030168 | -.51 | 0.609 | -.1459942   | .4326694 |

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.04 Prob > chi2 = 0.8358
Probit model with sample selection  
Number of obs      =       702  
Censored obs       =       386  
Uncensored obs     =       316  
Wald chi2(27)      =    767.20  
Log pseudolikelihood = -523.0497  
Prob > chi2        =    0.0000

|               Robust |
|---------------|-------------------|
|                Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval] |
|-------------|-------------------|----------------|-------|------------------|
| dropout     | fixed             | 1.152682   .4471652     2.58   0.010     .2762543     2.02911 |
|             | ha_lcap           | .9370076   .516994       1.81   0.070     -.076282     1.950297 |
|             | ha_hcap           | .6299156   .3406403       1.85   0.064     -.0377271    1.297558 |
|             | la                | .184108   .3260157       0.56   0.574    -.4555658    .8223897 |
| feedback_f_d | fixed             | .3311852   .4045661       0.82   0.413    -.4617498     1.12412 |
| feedback_n_d | fixed             | .5612394   .3253681       1.72   0.085     -.0764704    1.198949 |
| notincome   | fixed             | .2781471   .5111105       0.54   0.586    -.7236111    1.279905 |
|             | ha_lcap           | .191239   .2876578       0.67   0.506    -.3725058    .7550919 |
|             | ha_hcap           | -.1560803  .4059071     -0.38   0.701    -.8951646    .6830043 |
|             | la                | .6299156   .3406403       1.85   0.064     -.0377271    1.297558 |
|             | feedback_f_d      | -.6287804  .6902152     -0.44   0.659    -.9516436    .9252595 |
|             | feedback_n_d      | .6551875   .5003088       1.31   0.190    -.3259398    1.635775 |
|             | notincome         | 1.008843   .8011412       1.26   0.208     -.2539925    2.579051 |
|             | _Ioutside_40      | -.4084255  .1515226     -2.70   0.007    -.6944042    -.122446 |
|             | _Ioutside_50      | -.7304557  .1322124     -5.52   0.000     -.995872     -.464838 |
|             | _Iincome_2        | -.1882351  .1997206     -1.00   0.317    -.5750757    .2085194 |
|             | _Iincome_3        | -.1396205  .1664121       0.84   0.401    -.4654134    .3975729 |
|             | _Iincome_4        | .1571074   .1842096       0.85   0.394    -.2039368    .5181515 |
|             | _Iincome_5        | .0722689   .2550597       0.28   0.777    -.427639     .5721768 |
|             | _Iincome_6        | -.0017383  .2997671     -0.01   0.995     -.5893211    .5879448 |
|             | _Iincome_7        | .0251507   .2823583       0.09   0.929    -.5282615    .5785629 |
|             | _Iincome_8        | .0407046   .418211       -0.97   0.330    -.822672    .4263689 |
|             | _Iincome_9        | .0139567   .3093642       0.05   0.956    -.5923859    .6209933 |
|             | _Ieducatio-a2      | .4838537  .384532       1.26   0.208    -.2698152    1.237523 |
|             | _Ieducatio-a3      | .177252   .3615088       0.55   0.584    -.5108189    .9062948 |
|             | _Ieducatio-a4      | .4085532  .4299          0.95   0.342    -.4340353    1.251142 |
|             | _Ieducatio-a5      | .412903   .3412086       1.21   0.226    -.2558353    1.081659 |
|             | _Ieducatio-a6      | .4749508  .3497611       1.36   0.174    -.210584     1.160486 |
|             | age                | -.0054539  .0051567     -1.06   0.290    -.0155609    .0046532 |
|             | male               | .1838952  .1052873       1.75   0.081    -.0224641    .3902544 |
|             | india              | .0275071  .1658021       0.17   0.866    -.3524712    .2974572 |
|             | american           | .226826   .1780660       1.30   0.193    -.1780660    .6317157 |
|             | answerq10r-k       | .0196405  .0236653       0.83   0.407    -.0660237    .0267427 |
|             | answerq17r-k       | .0029238  .0019437       1.50   0.133    -.0008858    .0067333 |
|             | riskfactor         | .1004923  .0773092       1.30   0.194    -.0510309    .2520155 |
ammeasure |  .0461748   .1063609   0.43   0.664   -.1622887    .2546382
answerqc |  .0578786   .0625615   0.93   0.355   -.0647397    .1804968
correctp |  -.0012706   .0503181  -0.03   0.980   -.0998923     .097351
_cons |  .2495031   .4710409   0.53   0.596   -.6737202    1.172726
-------------+----------------------------------------------------------------
/athrho |  .1330886   .9535834   0.14   0.889   -1.7359    2.002078
-------------+----------------------------------------------------------------
rho |  .1323083   .9368904                      -.9397494    .9641741
------------------------------------------------------------------------------
Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02  Prob > chi2 = 0.8890
------------------------------------------------------------------------------
Probit regression

Number of obs   =        702
Wald chi2(9)    =      67.17
Prob > chi2     =     0.0000
Log pseudolikelihood = -438.14461                 Pseudo R2       =     0.0728

------------------------------------------------------------------------------
|               Robust
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
good_selecta |
  fixed       |  -.7136931    .189797    -3.76   0.000    -1.085688   -.3416978
  ha_lcap     |    -.34104   .2179688    -1.56   0.118    -.7682511    .0861711
  ha_hcap     |    .0977379   .1683789     0.58   0.562    -.2322788    .4277545
  la          |  -.3249029   .1383773    -2.35   0.019    -.5961175   -.0536883
  feedback_f-d|  -.2398576   .2077861    -1.15   0.248    -.6471109    .1673956
  feedback_n-d|   .0392876   .2962644     0.13   0.895      -.54138    .6199552
  _Ioutside_40|  -.3807198   .1311097    -2.90   0.004    -.6376901   -.1237495
  _Ioutside_50|  -.6118573    .125574    -4.87   0.000    -.8579779   -.3657367
  notincome   |   .0392876   .2962644     0.13   0.895      -.54138    .6199552
  _cons       |   .4740967   .1426711     3.32   0.001     .1944664    .7537269
------------------------------------------------------------------------------
Probit regression  
Number of obs = 702  
Wald chi2(32) = 94.61  
Prob > chi2 = 0.0000  
Log pseudolikelihood = -422.68008  
Pseudo R2 = 0.1055  

<table>
<thead>
<tr>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>good_selecta</td>
</tr>
<tr>
<td>fixed</td>
</tr>
<tr>
<td>ha_lcap</td>
</tr>
<tr>
<td>ha_hcap</td>
</tr>
<tr>
<td>la</td>
</tr>
<tr>
<td>feedback_f-d</td>
</tr>
<tr>
<td>feedback_n-d</td>
</tr>
<tr>
<td>_Iincome_2</td>
</tr>
<tr>
<td>_Iincome_3</td>
</tr>
<tr>
<td>_Iincome_4</td>
</tr>
<tr>
<td>_Iincome_5</td>
</tr>
<tr>
<td>_Iincome_6</td>
</tr>
<tr>
<td>_Iincome_7</td>
</tr>
<tr>
<td>_Iincome_8</td>
</tr>
<tr>
<td>_Iincome_9</td>
</tr>
<tr>
<td>_Ieducao-2</td>
</tr>
<tr>
<td>_Ieducao-3</td>
</tr>
<tr>
<td>_Ieducao-4</td>
</tr>
<tr>
<td>_Ieducao-5</td>
</tr>
<tr>
<td>_Ieducao-6</td>
</tr>
<tr>
<td>age</td>
</tr>
<tr>
<td>male</td>
</tr>
<tr>
<td>indi</td>
</tr>
<tr>
<td>american</td>
</tr>
<tr>
<td>answerq10r-k</td>
</tr>
<tr>
<td>answerq17r-k</td>
</tr>
<tr>
<td>riskfactor</td>
</tr>
<tr>
<td>ammeasure</td>
</tr>
<tr>
<td>answerqc</td>
</tr>
<tr>
<td>correctp</td>
</tr>
<tr>
<td>_Ioutside_40</td>
</tr>
<tr>
<td>_Ioutside_50</td>
</tr>
<tr>
<td>notincome</td>
</tr>
<tr>
<td>_cons</td>
</tr>
</tbody>
</table>
Regressions for Table 5

Heckman selection model                         Number of obs      =       702
(regression model with sample selection)        Censored obs       =       386
Uncensored obs     =       316
Wald chi2(7)       =    12.50
Log pseudolikelihood = -1184.461                Prob > chi2        =    0.0853

------------------------------------------------------------------------------
|               Robust
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
totalcorrect |
    fixed    |  -1.324739   .5600874    -2.37   0.018     -2.42249   -.2269874
    ha_lcap   |  -.9976087   .6314563    -1.58   0.114     -2.23524   .2400229
    ha_hcap   |  -.7628656   .4246538    -1.80   0.072    -1.595172   .0694405
       la     |  -.2576153   .4423187    -0.58   0.560    -1.124544   .6093135
feedback_f-d   |  .2743352   .6144938     0.45   0.655    -.9300506    1.478721
feedback_n-d   |  .036365   .4416467    -0.08   0.934    -.829246   .8292466
notincome     |   .7796492   .8639362     0.90   0.367    -.9136347    2.472933
     _cons    |   4.718909   .5122513     9.21   0.000     3.714915   5.722903
-------------+----------------------------------------------------------------
selecta      |
    fixed    |  -.499959   .1858877    -2.69   0.007    -.8642922   -.1356259
    ha_lcap   |  -.1986197   .2155373    -0.92   0.357    -.6210649    .2238256
    ha_hcap   |   .237049   .170351     1.39   0.164    -.0968328    .5709308
       la     |  -.2893089   .1376253    -2.10   0.036    -.5590496  -.0195683
feedback_f-d   |  -.2090558   .1973326    -1.06   0.289    -.5958207    .177709
feedback_n-d   |  -.2253039   .1375334    -1.64   0.101    -.4948644    .042567
notincome     |   .2065545   .2973355     0.69   0.487    -.3762123   .7893214
_Outside_40   |  -.3520662   .1294844    -2.72   0.007    -.6058509   -.0982815
_Outside_50   |  -.6569737   .1430418    -4.64   0.000    -.912755   -.4012619
     _cons    |   .5049486   .1430418     3.53   0.000    -.2245918    .7350353
------------------------------------------------------------------------------
/athrho       |  -.1229717   .2437462    -0.50   0.614    -.6007054    .354762
/lnsigma      |   .9038282   .0354223    25.52   0.000     .8344017    .9732546
------------------------------------------------------------------------------
rho           |  -.1223556   .2400971    -0.50   0.614    -.340919    .0965514
sigma         |   2.469037   .0874598     28.02   0.000     2.298837    2.645244
lambda        |  -.3021005   .5961648    -0.50   0.614    -.4004224   .1962214
------------------------------------------------------------------------------
Wald test of indep. eqns. (rho = 0): chi2(1) =   0.25   Prob > chi2 = 0.6139
Heckman selection model

(Heckman selection model
(regression model with sample selection)

Number of obs = 702
Censored obs = 386
Uncensored obs = 316

Log pseudolikelihood = -1126.673
Prob > chi2 = 0.0000

Wald chi2(19) = 152.29

<table>
<thead>
<tr>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coef.   Std. Err.   z   P&gt;</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>totalcorrect</td>
</tr>
<tr>
<td>fixed</td>
</tr>
<tr>
<td>ha_lcap</td>
</tr>
<tr>
<td>ha_hcap</td>
</tr>
<tr>
<td>la</td>
</tr>
<tr>
<td>feedback_f-d</td>
</tr>
<tr>
<td>feedback_n-d</td>
</tr>
<tr>
<td>notincome</td>
</tr>
<tr>
<td>income</td>
</tr>
<tr>
<td>education</td>
</tr>
<tr>
<td>age</td>
</tr>
<tr>
<td>male</td>
</tr>
<tr>
<td>india</td>
</tr>
<tr>
<td>american</td>
</tr>
<tr>
<td>answerq10r-k</td>
</tr>
<tr>
<td>answerq17r-k</td>
</tr>
<tr>
<td>riskfactor</td>
</tr>
<tr>
<td>ammeasure</td>
</tr>
<tr>
<td>correctp</td>
</tr>
<tr>
<td>_cons</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>selecta</td>
</tr>
<tr>
<td>fixed</td>
</tr>
<tr>
<td>ha_lcap</td>
</tr>
<tr>
<td>ha_hcap</td>
</tr>
<tr>
<td>la</td>
</tr>
<tr>
<td>feedback_f-d</td>
</tr>
<tr>
<td>feedback_n-d</td>
</tr>
<tr>
<td>notincome</td>
</tr>
<tr>
<td>_outside_40</td>
</tr>
<tr>
<td>_outside_50</td>
</tr>
<tr>
<td>income</td>
</tr>
<tr>
<td>education</td>
</tr>
<tr>
<td>age</td>
</tr>
<tr>
<td>male</td>
</tr>
<tr>
<td>india</td>
</tr>
<tr>
<td>american</td>
</tr>
<tr>
<td>answerq10r-k</td>
</tr>
<tr>
<td>answerq17r-k</td>
</tr>
<tr>
<td>riskfactor</td>
</tr>
<tr>
<td>ammeasure</td>
</tr>
<tr>
<td>correctp</td>
</tr>
<tr>
<td>_cons</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>/athrho</td>
</tr>
<tr>
<td>/lnsigma</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>rho</td>
</tr>
<tr>
<td>sigma</td>
</tr>
<tr>
<td>lambda</td>
</tr>
</tbody>
</table>

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02  Prob > chi2 = 0.8801
Heckman selection model                         Number of obs      =       702
(regression model with sample selection)        Censored obs       =       421
Uncensored obs     =       281

Wald chi2(7)       =      16.26
Log pseudolikelihood =   -1071.906                Prob > chi2        =    0.0229

------------------------------------------------------------------------------
|                      Robust                      |                      |
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  |
------------------------------------------------------------------------------
totalcorrect |                      |                      |
 fixed |  -.8465917   .6071093    -1.39   0.163    -2.036504    .3433206
 ha_lcap |  -.7013883   .6719863    -1.04   0.297    -2.018457    .6156807
 ha_hcap |  -.5200974   .4171968    -1.25   0.213    -1.337788    .2975932
 la |  -.0905678   .4431125    -0.20   0.838    -.9590524    .7779167
 feedback_f~d |  .607411   .6502879     0.93   0.350    -.6671298    1.881952
 feedback_n~d |  .2289199   .4478457     0.51   0.609    -.6488416    1.106681
 notincome |  1.520204   .6025843     2.52   0.012     .3391607    2.701248
 _cons |  4.673323   .5594997     8.35   0.000     3.576724    5.769923
------------------------------------------------------------------------------
good_selecta |                      |                      |
 fixed |  -.7148793   .1895033    -3.77   0.000    -.1086299    -.3434597
 ha_lcap |  -.34099   .2179366    -1.56   0.118    -.768138    0.0861579
 ha_hcap |  .0975789   .1682695     0.59   0.552     -.077883    .2730395
 la |  -.3257206   .1382697    -2.36   0.018    -.5967242    -.0547171
 feedback_f~d |  -.2394012   .2075971    -1.15   0.249    -.6462841    .1748178
 feedback_n~d |  -.2860306   .1381857    -2.07   0.038    -.5568695    -.0151916
 notincome |  -.0396855   .2964908     0.13   0.894    -.5414258    .6207968
 _outside_40 |  -.3794344   .1318708    -2.88   0.004    -.6378963    -.1209724
 _outside_50 |  -.6123673   .1254704    -4.88   0.000    -.8582848    -.3664497
 _cons |  .4744998   .1427507     3.32   0.001     .1947135     .754286
------------------------------------------------------------------------------
/athrho |                      |                      |
 fixed |  -.0702335   .2982189    -0.24   0.814    -.5457319    .5142649
 ha_lcap |  -.0701183   .2967527    -0.24   0.814    -.5457319    .5142649
 ha_hcap |  .2311879   .0871766     2.68   0.007     .0647135     .4142649
 la |  -.162105   .4867797    -0.33   0.743    -.4408695    .2965595
 feedback_f~d |  .2147177   .4892142     0.44   0.656     -.9142649    1.240783
 feedback_n~d |  .2147177   .4892142     0.44   0.656     -.9142649    1.240783
 notincome |  .2147177   .4892142     0.44   0.656     -.9142649    1.240783
 lambda |  .2147177   .4892142     0.44   0.656     -.9142649    1.240783
------------------------------------------------------------------------------
Wald test of indep. eqns. (rho = 0): chi2(1) =      0.06    Prob > chi2 =    0.8138
Heckman selection model
(regression model with sample selection)

Number of obs = 702
Censored obs = 421
Uncensored obs = 281

Wald chi2(19) = 153.74
Prob > chi2 = 0.0000

Log pseudolikelihood = -1020.215

|               Robust |
| Coef. Std. Err. z P>|z| [95% Conf. Interval] |
|---------------------|
| totalcorrect        |
| fixed               | 0.0183317 0.601604 0.03 0.976 -1.160791 1.197454 |
| ha_lcap             | 0.4979373 0.5896842 0.84 0.398 -0.6578226 1.653697 |
| la                  | -0.033977 0.359397 -0.09 0.925 -0.7383821 0.670428 |
| feedback_f_d        | 0.1669485 0.6326342 0.26 0.792 -0.987293 1.121183 |
| feedback_n_d        | 0.109436 0.4894561 0.22 0.825 -0.6588507 0.826723 |
| notincome           | 0.0251187 0.0662353 0.38 0.705 -0.1047002 0.1549375 |
| income              | 0.0110141 0.1049746 0.10 0.916 -0.1947324 0.2167606 |
| age                 | 0.0729726 0.2604866 0.28 0.779 -0.4375718 0.5835169 |
| male                | 0.038636 0.5896842 0.07 0.944 -1.104199 1.182011 |
| india               | -0.7942858 0.3872252 0.21 0.832 -1.544379 0.0557075 |
| american            | 0.0328365 0.3972252 0.08 0.936 -0.7366059 0.7922759 |
| feedback_q10r_k     | -0.0870511 0.0618822 0.26 0.792 -0.208338 0.0342358 |
| feedback_q17n_k     | -0.0023207 0.0054226 0.04 0.967 -0.0129488 0.0083073 |
| riskfactor          | 0.0247806 0.1936002 0.11 0.914 -0.3546688 0.4032294 |
| ammeasure           | 0.0509556 0.2682209 0.19 0.849 -0.476859 0.5787607 |
| answerq10r_k        | -0.0984125 0.0618822 0.16 0.872 -0.208338 0.0014125 |
| answerq17n_k        | 0.0119477 0.0119477 0.98 0.329 -0.0026066 0.0260661 |
| _cons               | 0.8801576 0.1188328 7.41 0.000 0.6472496 1.113066 |
| good_selecta        |
| fixed               | -0.7757899 0.193847 -4.00 0.000 -1.155723 -0.395869 |
| ha_lcap             | -0.3923657 0.2218361 -1.77 0.077 -0.8217565 0.042425 |
| ha_hcap             | -0.0555486 0.171888 0.32 0.747 -0.2813457 0.3924294 |
| la                  | -0.2966375 0.1417372 -2.09 0.036 -0.574437 0.003773 |
| feedback_f_d        | -0.2022842 0.2112412 -0.96 0.338 -0.613093 0.2117409 |
| feedback_n_d        | -0.3409458 0.1411517 -2.42 0.016 -0.6175981 0.0492935 |
| notincome           | 0.0502337 0.2976067 0.17 0.866 0.053043 0.5375967 |
| _outside_40         | -0.4218582 0.1352236 -3.12 0.002 -0.6868916 -0.1568248 |
| _outside_50         | -0.6441082 0.1297337 -4.96 0.000 -0.8983816 -0.3898348 |
| income              | 0.0023164 0.0267828 0.09 0.931 -0.0501769 0.0548097 |
| education           | -0.041311 0.0242194 0.98 0.329 -0.0417094 0.1245716 |
| age                 | -0.007444 0.005128 0.08 0.935 -0.0174947 0.0026066 |
| male                | 0.1585299 0.1043636 1.42 0.155 0.0596555 0.3520245 |
| india               | -0.114695 0.1654455 0.69 0.488 0.4389623 2.0957234 |
| american            | -0.2631435 0.1763716 -1.49 0.136 -0.6088254 0.0823585 |
| answerq10r_k        | 0.0124863 0.0235767 0.53 0.596 0.0586958 0.0337232 |
| answerq17n_k        | 0.001555 0.0019789 0.58 0.559 0.0027231 0.0003411 |
| riskfactor          | 0.0122558 0.0766007 0.16 0.873 -0.297488 0.3222571 |
| ammeasure           | 0.1355166 0.1057688 1.28 0.200 -0.0717864 0.3428197 |
| answerq10r_k        | 0.135396 0.0628306 2.09 0.036 0.088394 0.2568652 |
| correctp            | 0.0369073 0.0495757 0.74 0.460 0.0610073 0.1382182 |
| _cons               | 0.2282694 0.3699718 0.62 0.537 -0.496862 0.9534008 |
| /athrho              | -0.0263972 0.3872848 -0.07 0.946 -0.7865269 0.7373239 |
| /lnsigma            | 0.682253 0.0338888 17.88 0.000 0.6109846 0.761466 |
| rho                  | 0.0263911 0.3875583 0.74 0.460 0.0610073 0.1382182 |
| sigma               | 1.9686204 0.076248 1.42 0.156 1.842244 2.141413 |
| lambda              | 0.0524181 0.769748 0.07 0.941 -1.561097 1.456261 |

Wald test of indep. eqns. (rho = 0): chi2(1) = 0.00  Prob > chi2 = 0.9457