

SIMON FRASER UNIVERSITY

Department of Economics

Working Papers

17-12

"Instructions"

David J. Freeman,
Erik O. Kimbrough,
Garrett M. Petersen, and
Hanh T. Tong

July 2017



Instructions

David J. Freeman, Erik O. Kimbrough, Garrett M. Petersen, and Hanh T. Tong*

July 28, 2017

Abstract

A meta-analysis of instruction delivery and reinforcement methods in recent laboratory experiments reveals a wide and inconsistently-reported variety of practices and limited research evaluating their effectiveness. Thus we experimentally compare how methods of delivering and reinforcing experiment instructions impact subjects' understanding. We report a one-shot individual decision task in which misunderstanding can be unambiguously identified in behavior and find that misunderstanding is prevalent in our control treatment which uses relatively standard experimental instructions. We find combinations of reinforcement methods that can eliminate half of subjects' misunderstanding. Providing paper instructions is among the most effective of such methods.

JEL classification: C91

Keywords: Attention, Comprehension, Instructions

1 Introduction

Experiments start by providing instructions designed to ensure that subjects understand how their actions and the actions of others determine payoffs. Such understanding is crucial to the economic interpretation of subjects' behavior – without it, the experimenter has lost control

*Address: Department of Economics, Simon Fraser University, 8888 University Dr, Burnaby, BC V5S 1A6, Canada. We thank Jim Sylvester for programming the experiment, Cameron Young for research assistance, Kevin Laughren, seminar and conference audiences at Simon Fraser University and the 2017 Economic Science Association World Meetings in San Diego for helpful comments. Freeman thanks Simon Fraser University for research funding (SSHRC-SFU Institution Grant SMALL-2013-631004). Kimbrough thanks SSHRC for additional funding (SSHRC Insight Grant 435-2015-0798). This study has been vetted by the Simon Fraser University Office of Research Ethics (study 2014s0066).

(Smith, 1982). Clear and effective instructions are especially important in experiments where subjects' initial decisions are of interest, or when subjects make a small number of decisions or receive limited feedback. However, there is scant evidence on the effectiveness of practices for delivering instructions and reinforcing their content. This paper reports a meta-analysis of current practices and a new experiment to evaluate the effectiveness of these practices at establishing subject understanding.

Almost from the field's inception, experimental economists recognized that the effectiveness of instructions in establishing understanding may depend on how they are delivered and reinforced (Fouraker and Siegel 1963). Prominent textbooks in the field give detailed guidelines on how to deliver instructions and suggest complementary methods to increase subjects' comprehension, including reading instructions aloud, and using demonstrations, quizzes, and practice rounds (Friedman and Sunder 1994, Davis and Holt 1993, Cassar and Friedman 2004). Casual observation nevertheless suggests that there is wide variation in how practitioners deliver instructions and use reinforcement methods. Our meta-analysis reviews the methods for delivering and reinforcing instructions as reported in experimental studies recently published in six leading journals and confirms this observation. We find that almost all experimenters complement their instructions with at least one reinforcement method, though the methods used vary substantially. This suggests that the field lacks clear norms for how instructions ought to be delivered and reinforced in experimental studies. Troublingly, we were unable to classify roughly 10% of papers because they failed to provide sufficient details on their methods.

In spite of the variation in practices, there is little evidence comparing their effectiveness. Thus we conduct an experiment to evaluate the impact of instructions on understanding. We study a one-shot individual timing decision in which each subject chooses when to do a task. The task can be performed at most once, and the subject is paid the most for doing it at the correct time and least for doing it earlier. Moreover, the subject is better off not doing the task at all than doing it too early. This information is explicitly stated in the instructions.

Thus a subject’s lack of understanding or retention of the instructions is clearly revealed by doing the task too early. We hold constant the content of instructions, and our treatments vary how instructions are delivered and reinforced.

In our control treatment subjects complete self-paced computerized instructions including practice rounds. Nearly half of all subjects in this treatment do the task too early. Additional treatments study the impact of adding a pre-experiment comprehension quiz with and without providing the answers, neither of which leads to a significant reduction in misunderstanding. We thus study the additional impact of introducing monetary incentives for quiz performance, of going through the computerized instructions twice (both before and after the quiz), and of providing paper instructions alongside computerized instructions. We find that all three of these treatments lead to significant improvements relative to the control – each eliminates about half of the observed misunderstanding.

By studying an individual decision task, our experiment eliminates strategic and other-regarding motives that might confound the identification of misunderstanding. By studying a one-shot decision without feedback, we obtain a clean measure of understanding and retention of the instruction that is not confounded by learning. Two existing papers of which we are aware have studied the impact of instruction delivery and reinforcement on play in repeated public goods games (Bigoni and Dragone, 2012; Ramalingam *et al.*, 2015).¹ The more relevant of these is Bigoni and Dragone (2012), who find that subjects’ comprehension, as measured by correct quiz answers and response times, was worse when a shortened version of instructions was given on screen compared to their baseline paper instructions, shortened paper instructions, and shortened on-screen instructions with active examples requiring subject input. They do not study the effect of instructions on decisions made.

Our findings motivate advice on how to report and deliver instructions. First, experimenters need to be aware that the way that instructions are delivered and reinforced has

¹Our discussion here is restricted to instruction delivery and reinforcement. We have little to say about how variation in the content of the instructions may affect behavior, by providing or failing to provide subjects with payoff-relevant information, or alternatively by influencing the framing of the experimental task. See Alekseev *et al.* (2017) for a discussion of the use of context in instructions.

consequences for behavior. Second, in light of this, we suggest that all experimental papers should clearly report how they deliver and reinforce instructions, as this can be crucial for close replication and interpretation. Journals' efforts to require experimenters to share copies of their instructions are laudable, and these could be complemented by standardized reporting of how instructions are delivered and reinforced. Third, while we find multiple ways of improving understanding, we suggest that future experiments use paper instructions as a low-cost means of doing so.

2 Meta-analysis

We report how instructions are delivered and reinforced in 243 experimental studies published between January 2011 and June 2016 in *Experimental Economics* and five prominent general interest economics journals. We selected all papers in these journals that contained at least one lab experiment in which participants were given instructions on the experimental procedure. We checked whether instructions were delivered on paper, on screen, both, or neither. We also checked for the use of various practices intended to reinforce the content of the instructions, including reading the instructions aloud, demonstrations, practice rounds, and pre-experiment quizzes. Details of our classification procedure are given in Appendix A. The results of our meta-analysis are given in Table 1.

[Table 1 goes here]

We were unable to determine how instructions were delivered in 10% of the studies we reviewed. If behavior is sensitive to how instructions are delivered, then this oversight hampers replication. Of the remaining 218 studies, 66% deliver instructions exclusively on paper, 20% deliver instructions exclusively on screen, while another 4% use both. We find this noteworthy since the majority of these experiments are themselves computerized. The remaining 10% of these 218 studies do not deliver individualized instructions by either paper or computer. Most such studies are lab-in-the-field experiments studying non-student

populations and deliver instructions orally along with some of the reinforcement methods discussed below. We suspect that experimental economists' revealed preference for paper instructions is driven by the fact that subjects can refer back to them throughout the experiment, which may not always be the case with computer instructions. This may mitigate subjects' tendency to forget important information (something we see in our experiment as well).²

94% of all studies use at least one method of reinforcement which suggests that experimenters are almost universally concerned about subject comprehension. Instructions are read aloud in 55% of studies. This helps create common information, and we find it noteworthy that reading aloud is done more often in experiments that deliver instructions only on paper than only on screen (59% vs. 39%). The correlation with the use of paper instructions is consistent with the idea that paper instructions are often intended to help establish common information.

We find that 56% of studies use demonstrations or practice rounds to reinforce subject understanding of the experiment. Examples of such practices include physical demonstrations of how risk will be resolved,³ guided examples of possible actions and their consequent outcomes, and unpaid practice rounds. Of the studies that use at least one of these forms of reinforcement, 84% use guided demonstrations or guided practice rounds, and 30% use unguided practice rounds; some studies use both.

While reinforcement methods discussed earlier enhance inflows of information to subjects, experiments can also test the retention of this information. Thus pre-experiment quizzes were used by 40% of all studies. At least 73% of these reinforced understandings and corrected misunderstandings by giving answers to the quiz. In addition, 40% required a perfect score to commence the experiment. Only three of the studies in our sample paid subjects for quiz performance. We note that 22% of studies that used a quiz did not clearly report whether or

²Publicly distributed paper instructions may also help establish common information in strategic settings (Friedman and Sunder 1994, p. 77).

³Davis and Holt (1993, p. 23) and Friedman and Sunder (1994 p. 67) suggest that the use of physical randomization devices may enhance credibility.

how subjects were given feedback on the quiz. As noted earlier, this information is necessary for close replication and we see its frequent omission as a limitation of current practices.

Our meta-analysis shows that there is wide variation in how experimenters deliver and reinforce instructions. Nevertheless, there are commonalities which seem to reflect some notion of ‘best practices’. Few studies have tested whether current practices are effective – our experiment is designed to fill this gap.

3 Experimental Design

3.1 Overview of Experiment

We design an individual choice experiment in which each subject performs two tasks, a base task which provides a low flow of payoffs throughout the experiment, and a second task which can only be completed once and results in a potentially large lump sum payoff. The amount of the lump sum depends on the time at which they complete the second task. Doing the task too early results in a lower payoff than doing it at the right time (or not doing it at all).

Task 1 is the Poodle Jump game (based on a popular mobile game Doodle Jump), where players guide a bouncing poodle up a series of platforms by pressing two buttons. When a subject missed a platform, the poodle fell to the ground and the game restarted with no penalty. Each participant receives \$0.25 per minute of Task 1, so long as they have jumped a minimum cumulative height. This height was chosen so that it would be trivially easy to complete but not automatic, at a level that effectively guaranteed a subject this payment each period so long as they were attentive to the task.⁴

Task 2 is a simplified version of the slider task (Gill and Prowse, 2012). Players can switch from Task 1 to Task 2 at any time by pressing the ‘j’ key, but they can only do this once. In the slider task, players are presented with four sliders which can be moved from

⁴Only 29 subjects ever failed to attain the required height in a period; 28 did so once and one subject did so twice. These failures account for only 0.5% of all Poodle Jump periods.

zero to 100. The task has been successfully completed when all four sliders are dragged to 50 and the player clicks “Continue.”⁵ Task 2’s payoff depends on when the subject presses ‘j’. For the first 21 periods, each period being one minute long, it pays \$0.20. However, in period 22 it jumps to \$7, falling to \$4 in period 23, then dropping by \$0.50 in every period thereafter until period 30 when the experiment ends. These payoffs are demonstrated in Figure 1. Doing Task 2 in period 22 maximizes a subject’s payoff; whereas, doing it before period 22 minimizes a subject’s payoff. If a subject fails to do Task 2 in period 22, they would always earn higher payoffs by doing it as soon as possible thereafter.

[Figure 1 goes here]

The challenge for subjects is to recognize and remember the correct time to press the ‘j’ key to complete Task 2, given the attention required to successfully complete Task 1 in each period. However, subjects have strong incentive to complete Task 2 at the right time: doing Task 2 at the right time raises payoffs by \$6.75 relative to not doing it at all, and by a minimum of \$3 compared to completing it at any other time. Moreover, doing it before period 22 leads the subject to forgo the opportunity to do it at the optimal period or thereafter, and also results in a lower payoff than never doing Task 2. Thus, doing Task 2 before period 22 provides a behavioral measure of a subject’s failure to understand the experiment.⁶

3.2 Treatment Design

We employ a between-subjects design with six treatments. We study the effectiveness of different ways of delivering and reinforcing the experiment’s instructions on subjects’ understanding, using our aforementioned measure. Many experimenters implicitly assume that subjects fully understand their instructions. If this is true, we should not observe any difference between treatments. However, if subjects do not always understand instructions there

⁵Only one subject who started the slider task failed to complete it.

⁶In our design, forgetfulness is a natural source of noise distinct from a lack of understanding. This is why we focus on doing Task 2 before period 22 as a measure of misunderstanding. Doing Task 2 too late or not at all could reflect forgetfulness rather than misunderstanding.

is the potential for variation in delivery and additional reinforcement to improve understanding. All treatments are summarized in Table 2. All treatments started with a common set of self-paced on-screen instructions, which included a graphical explanation of payoffs as well as reinforcement from practice rounds for both tasks and practice switching between tasks.

[Table 2]

The CONTROL treatment presents the instructions on screen with no additional reinforcement. The CONTROL treatment gives us baseline information on subject comprehension when they read instructions on their own.

The QUIZ treatment was identical to the CONTROL except that each subject completed a six question comprehension quiz on paper at the end of the on-screen instructions; no feedback was given on the quiz, but subjects were informed that there would be a quiz prior to beginning the instructions. The QUIZ treatment allows us to assess whether the presence of the quiz affects subjects' understanding; the quiz itself gives a secondary measure of understanding.

The ANSWERS treatment was identical to the QUIZ treatment, except that subjects were presented the answers to the quiz orally after all had completed it. This provided additional reinforcement of the key pieces of information from the instructions that were tested on the quiz and corrected possible misunderstandings revealed in quiz answers.

The TWICE treatment was identical to the ANSWERS treatment except that after completing the quiz and answers, the experimenter unexpectedly restarted the instructions for the participants to work through a second time. This allowed subjects to further review any content they missed on the first go and provided additional reinforcement.

The INCENTIVE treatment was identical to the ANSWERS treatment except that subjects were paid \$0.50 for each correct quiz answer, and were informed of this before starting the instructions. We hypothesized that this would lead subjects to pay more attention to the material in the instructions, and make any mistakes from the quiz more salient, thereby improving understanding.

Finally, the PAPER treatment was identical to the ANSWERS treatment except that the experimenter also distributed paper printouts of the instructions (in addition to the on-screen instructions), which participants could keep and reference at any time, even while completing the quiz.⁷ This is a common practice even for computerized experiments.

Our experiment differs from existing studies on instructions in two regards. First, this is an individual decision task, so there is no complexity from strategic behavior or altruistic preferences. Second, it is a one-shot task – each subject can only press ‘j’ once – so participants who fail to understand the instructions cannot learn through trial and error. These features allow us to cleanly identify mistakes and attribute variation in mistakes to variation in the delivery and reinforcement of instructions.

3.3 Procedures

Upon entering the lab, the experimenter assigned participants to visually isolated computer terminals. Participants were told not to interact with one another for the duration of the experiment. In all treatments, participants were informed that they would be given a set of instructions followed by an experiment in which they could potentially earn a significant amount of money; in the treatments with a quiz, they were also informed that there would be a quiz at the end of the instructions. The experimenter then started the self-paced on-screen instructions which included a written description of the tasks and the payoff structure, practice rounds of both tasks, practice switching between tasks, and a graphical illustration of the payoffs to both tasks in each period (a full copy of the instructions are presented in Appendix B). Once all participants completed the instructions, the experimenter distributed the quiz in the QUIZ, ANSWERS, INCENTIVE, TWICE, and PAPER treatments; the correct answers were revealed after all participants had completed the quiz except in the QUIZ treatment. In the TWICE treatment, subjects completed the on-screen instructions a second time, including practice rounds. Then the experiment started.

⁷The PAPER treatment potentially reduces forgetfulness since all relevant information is accessible throughout the experiment.

We recruited 228 participants to 23 sessions through Simon Fraser University’s CRABE recruiting system, with no subject participating in more than one session. Each session lasted under an hour. All payments were given in Canadian dollars, with average earnings of \$18.35 including a \$7 show-up fee. We collected no other demographic data nor other behavioral measures.

4 Results

We use a subject’s decision to do Task 2 at any time before period 22 as a behavioral measure of their failure to pay attention to, understand, absorb, or retain information from the instructions. We refer to such behaviors as *misunderstanding*. Figure 2 shows the share of subjects who misunderstood the instructions by treatment.

[Figure 2 goes here]

Finding 1: Subject misunderstanding is prevalent.

In our CONTROL treatment, 44% of subjects did Task 2 before period 22, demonstrating misunderstanding. This is despite the fact that even our CONTROL treatment includes both demonstrations and practice periods. Even in our most effective treatment, the corresponding share is 18%. These findings suggest that misunderstanding may be an important source of noise that is distinct from trembles.⁸ This justifies concern about the effectiveness of instructions.

Finding 2: Combining reinforcement methods reduces subject misunderstanding.

Column 1 of Table 3 reports the marginal effects from a logistic regression analysis of treatment effects displayed in Figure 2. The dependent variable takes a value of 1 when a subject misunderstood the instructions and 0 otherwise, and the specification includes an intercept

⁸In the Appendix, we show that we find similar results if we account for trembles by defining misunderstanding based on doing Task 2 before period 21.

term and treatment dummies.

We find a similar extent of misunderstanding in the QUIZ (48%) and ANSWERS (33%) treatments as in the CONTROL (44%) treatment, and these differences are statistically insignificant. In each of the INCENTIVE (24%), TWICE (18%), and PAPER (23%) treatments that provide additional reinforcement, subject misunderstanding was significantly lower than in the CONTROL treatment at the 5% level.

Our findings suggest that extensive reinforcement does improve subject understanding. In particular, providing paper instructions seems to be a low-cost way of improving understanding that does not require extra time in the lab.

[Table 3 goes here]

Finding 3: Higher quiz scores are associated with less misunderstanding, but both pre- and post-quiz interventions can reduce misunderstanding.

Quiz scores provide an alternative measure of subject understanding immediately after the instructions. In the QUIZ treatment which provides no feedback nor additional reinforcement, quiz score and misunderstanding are negatively correlated ($r = -0.22$); indeed 4 of 33 subjects had a perfect score on the quiz, and none of them subsequently revealed misunderstanding in the experiment, showing a clear link between these two measures of understanding.

Figure 3 shows the distribution of quiz scores by treatment; we lump together the QUIZ, ANSWERS, and TWICE treatments in this table since they followed identical procedures up to the collection of the quiz. We find only marginally significant differences in the distribution of quiz scores in the QUIZ, ANSWER, and TWICE treatments ($p=.10$, Kruskal-Wallis test). We expected that the INCENTIVE treatment would improve quiz performance (since we're economists), but surprisingly, it did not ($p=0.41$, rank-sum test). We also predicted improved quiz scores in the PAPER treatment, since it made the answers easily accessible to subjects when they wrote the quiz, and the data strongly support this prediction ($p<0.01$, rank-

sum test).⁹ On average, subjects in the PAPER treatment correctly answered 5.4 out of 6 questions, compared to only 4.2 in the other non-incentivized quiz treatments and 4.3 in the INCENTIVE treatment.

Table 3, Column 2 reports the marginal effects from a logistic regression analysis of treatment effects, controlling for quiz performance; the CONTROL treatment is excluded from this estimation since it had no quiz. As expected, there is a negative and statistically significant relationship between quiz performance and subject misunderstanding. After controlling for quiz score, the additional effect of the PAPER treatment is insignificant – suggesting that its treatment effect in Column 1 is mostly reflected in higher quiz scores. The TWICE and INCENTIVE treatments significantly reduce misunderstanding when controlling for quiz performance. The second round of instructions in the TWICE treatment was a surprise restart that occurred after the quiz, so any reduction in misunderstanding it effected must have taken place after the quiz. Surprisingly, the INCENTIVE treatment led to improved performance that was not reflected in quiz scores. We conjecture that this treatment operated by making subjects’ mistakes on the quiz more salient.

[Figure 3 goes here]

Robustness Checks Figure 4 shows empirical CDFs of completion times for Task 2, by treatment. For robustness, we show in Appendix C that we would arrive at similar results to those reported in Figure 2 and Table 3 using any of three alternative measures of misunderstanding. First, we allow for trembles by only classifying a subject as having misunderstood if they did Task 2 before period 21, making allowance for a subject who started Task 2 no more than 60 seconds before it was optimal. Second, we look at a stricter measure of misunderstanding, classifying a subject as misunderstanding unless they did Task 2 exactly in period 22 – which counts both those who completed Task 2 too early, too late, or never completed it as misunderstanding. Third, we reclassify only those who did Task

⁹The positive effect of paper instructions on quiz performance is consistent with the evidence reported in Bigoni and Dragone (2012).

2 before period 22 or never at all as misunderstanding. We arrive at the same qualitative conclusion using any of these three measures.

[Figure 4 here]

5 Discussion

Our meta-analysis shows that in current practice, there is wide variation in how instructions are delivered and reinforced. Our experiment shows that there is also wide variation in the effectiveness of commonly used practices. Since we find that subject understanding is sensitive to how instructions are delivered and reinforced, it is important that future studies clearly report the details of their procedures in order to allow for close replication. Our meta-analysis showed that this standard is not met by many published papers.

Our experiments indicate that even when using combinations of reinforcement methods including demonstrations, practice periods, and a quiz, many subjects' behavior reveals that they failed to pay attention to, understand, or retain information from the instructions. Combining these with further reinforcement methods reduced subjects' misunderstanding. In particular, we found that giving paper instructions, giving the instructions twice, or incentivizing the quiz each led to significant improvements. Each of these methods led to a similar improvement.

In our setting, we feel confident attributing the anomalous behavior that we observe to a failure to understand or absorb the instructions. In other experiments designed to test for anomalous behavior, the distinction between truly anomalous behavior of interest and a failure to understand the instructions may not be so clearcut. This justifies a concern with how instructions are given and the use of behavioral checks of understanding.

Our findings broadly suggest that experimenters' attempts to reinforce the instructions or make them more salient can be effective at reducing subject misunderstanding. That being said, different reinforcement methods used here and elsewhere differ in their costs and

benefits. Giving paper instructions requires no extra lab time and is almost free, whereas the other two methods that we found to be similarly effective are costly in terms of lab time and money. Based on this evidence, we suggest that future lab experiments use paper instructions rather than computer-only instructions where possible.

Finally, though we are able to improve understanding in our design, we still observe some residual misunderstanding, even in the best case. While the extent of such misunderstanding is likely to vary with experimental context (e.g. subject pool, design), its presence is noteworthy and has implications for the power and interpretation of experimental tests.

References

- ALEKSEEV, A., CHARNESS, G. and GNEEZY, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, **134**, 48–59.
- BIGONI, M. and DRAGONE, D. (2012). Effective and efficient experimental instructions. *Economics Letters*, **117** (2), 460–463.
- CASSAR, A. and FRIEDMAN, D. (2004). *Economics lab: an intensive course in experimental economics*. Routledge.
- DAVIS, D. D. and HOLT, C. A. (1993). *Experimental Economics*. Princeton University Press.
- FOURAKER, L. E. and SIEGEL, S. (1963). *Bargaining Behavior*. McGraw-Hill New York.
- FRIEDMAN, D. and SUNDER, S. (1994). *Experimental methods: A primer for economists*. Cambridge University Press.
- GILL, D. and PROWSE, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, **102** (1), 469–503.

RAMALINGAM, A., MORALES, A. and WALKER, J. (2015). *Instruction Length and Content: Effects on Punishment Behaviour in Public Goods Games*. Tech. rep., School of Economics, University of East Anglia, Norwich, UK.

SMITH, V. L. (1982). Microeconomic systems as an experimental science. *The American Economic Review*, **72** (5), 923–955.

Table 1: Meta-analysis of instruction delivery and reinforcement
Delivery method

	Computer only	Paper only	Computer and Paper	Neither	Unclear	Total
Total	44	143	9	22	25	243
Read aloud	17	85	3	22	7	134
Practice/Demonstration	25	75	8	15	13	136
Demo or guided practice	18	68	6	12	10	114
Unguided practice	10	18	4	3	6	41
Quiz	14	64	6	5	7	96
Feedback	9	47	4	4	6	70
Incentive	0	3	0	0	0	3
Require 100%	4	25	3	2	4	38
Feedback unclear	4	14	2	0	1	21

Note: Each entry is the number of papers classified in that respective category. Indented categories are subsets of the last non-indented category above.

Reinforcement

Table 2: Summary of treatments

Treatment	Quiz	Answers	Additional Reinforcement	Number of Subjects
CONTROL	No	No	No	43
QUIZ	Yes	No	No	33
ANSWERS	Yes	Yes	No	36
INCENTIVE	Yes	Yes	Pay 0.50 CAD per correct quiz answer	38
TWICE	Yes	Yes	Instructions restarted unexpectedly	38
PAPER	Yes	Yes	Instructions duplicated in paper printout	40

Table 3: Marginal effects from logistic regression analysis of subject misunderstanding

	Dependent variable	
	Misunderstanding	
QUIZ	0.038 (0.103)	
ANSWERS	-0.091 (0.086)	-0.111 (0.082)
INCENTIVE	-0.172** (0.075)	-0.170* (0.074)
TWICE	-0.218*** (0.069)	-0.201*** (0.071)
PAPER	-0.183** (0.074)	-0.066 (0.097)
Score		-0.111*** (0.026)
Misunderstanding in CONTROL	0.44	
Observations	228	185
Log Likelihood	-136	-96.1
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 1: Screenshot showing how payoffs were described to subjects

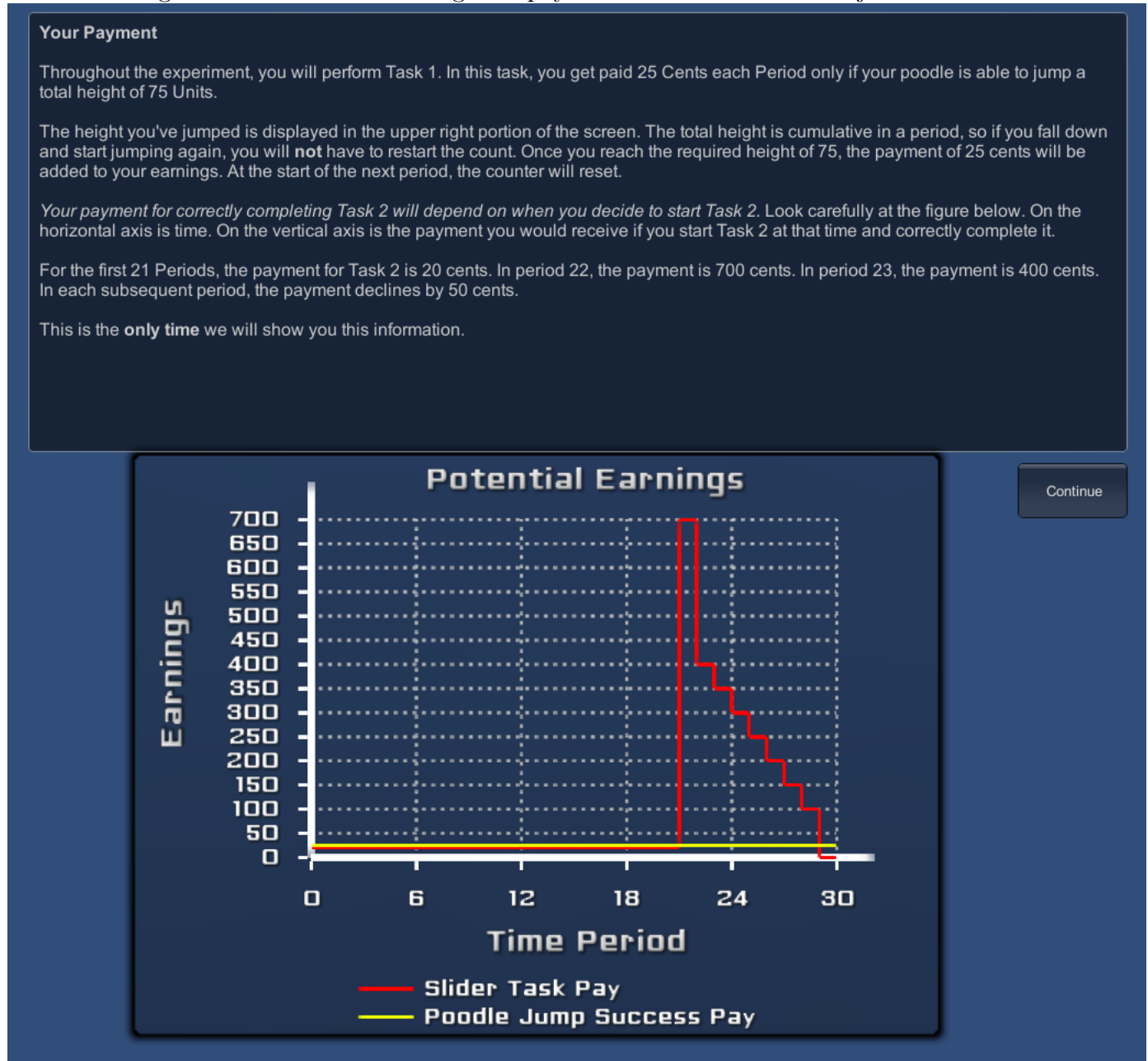


Figure 2: Percentage of subjects revealing *misunderstanding*, by treatment

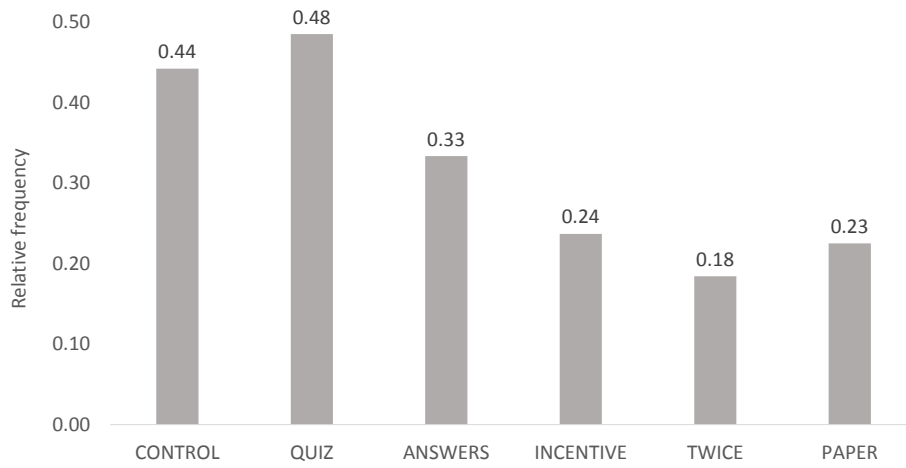


Figure 3: Histogram of quiz score distribution, by treatment

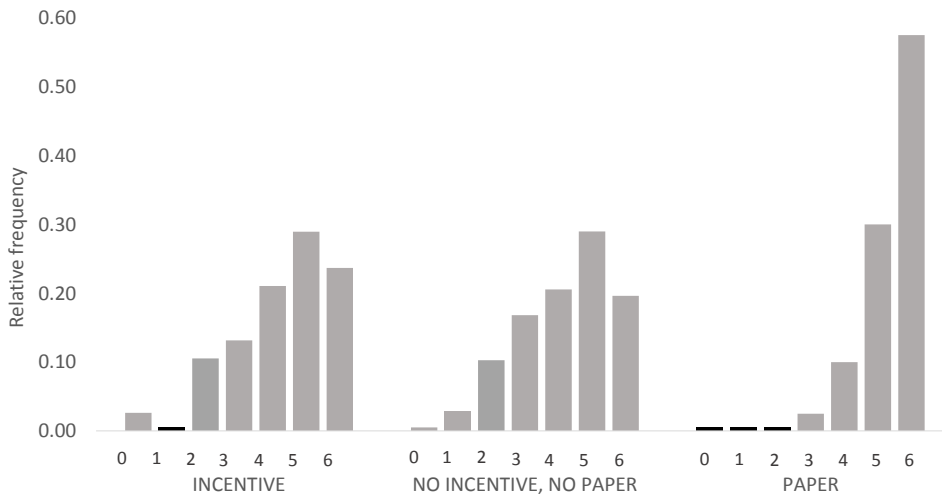
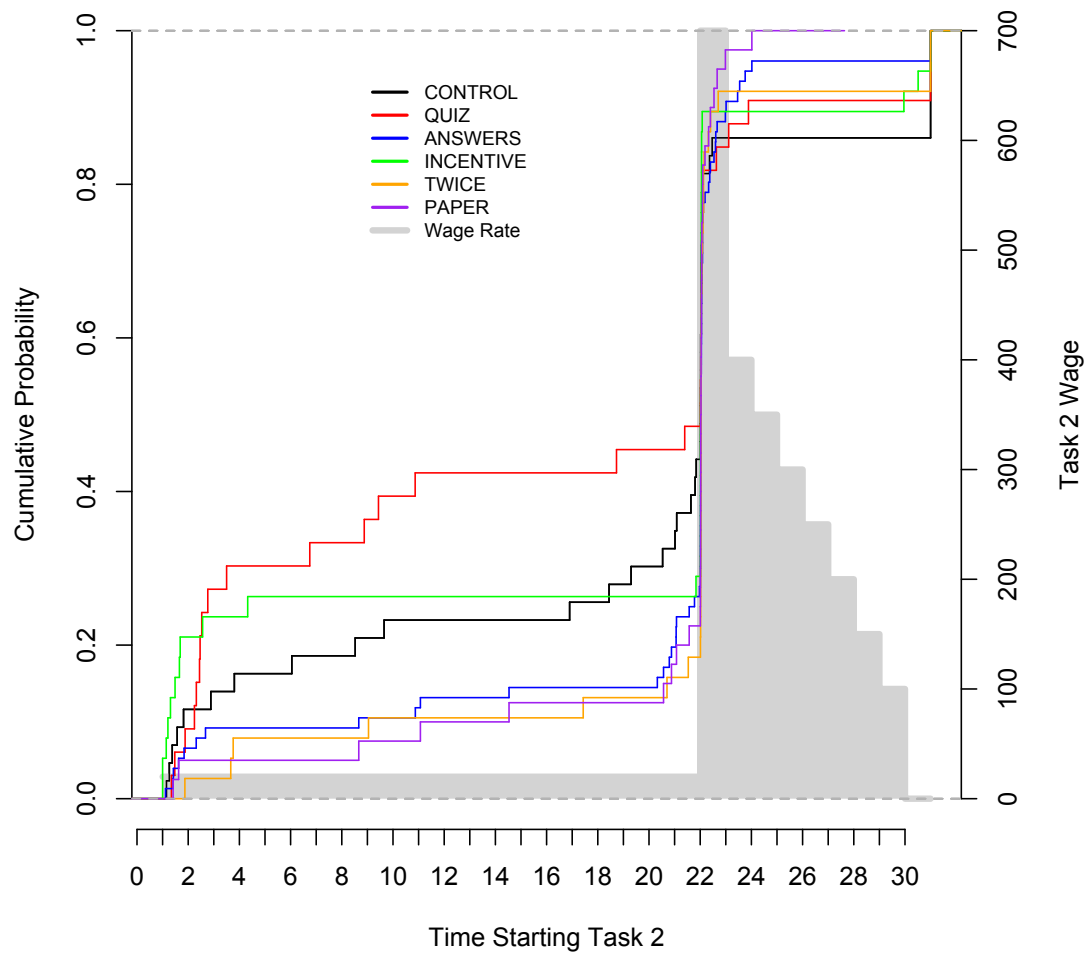


Figure 4: Empirical CDFs of Task 2 completion times, by treatment.



Supplementary Appendix to Freeman, Kimbrough, Petersen and Tong (2017)

A Review of current practice

Inclusion/Exclusion criteria

We included experimental papers published between January 2011 and June 2016 in six journals: the American Economic Review, Econometrica, the Quarterly Journal of Economics, the Journal of Political Economy, the Review of Economic Studies, and Experimental Economics. Articles from the AER: Papers and Proceedings were excluded. In order to be included, a paper had to include at least one lab experiment. We excluded field experiments and online experiments that were not conducted in a controlled environment, but we include “lab-in-the-field” experiments that were conducted in a controlled environment.

To classify each included experiment, we reviewed both the text of each paper and supplementary materials available online through the journal’s website, with the exception of uncompiled code (e.g. z-Tree code).

Coding Criteria: Delivery

Delivery methods could include paper instructions or computer instructions. Values in the supplementary table are 1 for yes, 0 for no, 0.5 for uncertain. In some cases, an alternative delivery method was used; for example, Etang *et al.* (2011) studied subjects in rural Cameroon and used purely verbal instructions because many subjects were illiterate.

We code the study as having paper instructions if it is directly stated or clearly implied that a set of paper instructions were used. Some papers were explicit about their use of printed instructions, while others required us to infer the existence of paper instructions

from other details. For instance, Mittone and Ploner (2011, p. 207) write that "after the choices are collected, instructions for the beliefs elicitation phase are distributed." Distribution implies a written set of instructions, though this is not explicitly stated. Sometimes we inferred the form of instructions from the instructions themselves, for instance in Altmann *et al.* (2014), the instructions included screenshots, from which we inferred that they must have been printed on paper.

We code the study as having computer instructions if it is directly stated or clearly implied that computerized instructions were used. Sometimes this was explicit, while other times it had to be inferred. For instance, in papers that included copies of their instructions online, some instructions told participants to click on something to proceed to the next screen. This implies that the instructions are computerized, even if it is not explicitly stated in the text of that paper. Cox and James (2012, Supplement p. 2) end their instructions by telling their subjects, "When you have finished reading and have asked any questions you might have, please click Done."

Many papers are unclear on whether the instructions are given on paper or on computers. If there was no explicit statement of the form of instructions in the paper itself, and no clear indication from the instructions where these were available online, the paper was coded as uncertain.

Coding Criteria: Reinforcement

We coded four different forms of reinforcement.

1. Read aloud. We code an experiment as having read aloud its instructions if it is stated or clearly implied that the instructions were presented orally. Most often this meant that the experimenter read the instructions for the participants to hear. Some studies, such as Aycinena *et al.* (2014, p. 110), included voice recordings of the instructions, which we coded as read aloud as indicated by the following quote "They were provided with instructions and

were also shown a video which read these instructions aloud.”

2. Demonstration or guided practice. We code a paper as including demonstration or guided practice if we can infer that it used walk-throughs of the experimental interface, examples, or demonstrations of aspects of the experiment during the instructions phase. Walk-throughs involve actively-guided practice by the subject. Examples include hypothetical descriptions of potential actions and consequent outcomes. For instance, Brookins and Ryvkin (2014) give subjects an example of the likelihood of success, conditional on the group members’ investment. Demonstrations actively highlight one or more aspects of the experiment, for example, throwing a die to show subjects how uncertainty will be resolved as in Ericson and Fuster (2011). The mere use of graphical or tabular methods to communicate information, or providing screenshots in paper instructions, was considered neither demonstration nor guided practice.

3. Unguided practice. If the experiment included one or more unpaid practice rounds without guidance, we coded this as unguided practice. Sometimes this was explicit in the body of the paper, while other times it was only indicated in the instructions themselves.

4. Quiz. Quizzes or questionnaires were only included if they occurred after the instructions and before the experiment. Many experiments include questionnaires to check participants’ understanding ex post, but these are not counted as they do not reinforce participants’ understanding of the instructions before the experiment.

When a quiz was given, we checked whether feedback was given after the quiz and before the experiment. If it was clearly stated that subjects were given the correct answers to the quiz, “Feedback” was coded as a 1. If subjects must get 100% to proceed with the experiment, we infer that feedback was given. Many papers give quizzes to “ensure comprehension of instructions” but do not explicitly indicate whether answers were given. For example Cabrera *et al.* (2013, p. 432) indicate that “subjects completed a quiz to make sure they had fully

understood the logic of the game.” It is ambiguous whether this implies that feedback was given to promote subject understanding ex-ante or instead quiz performance was used by the experimenters to assess subject comprehension ex-post. Such papers are coded as uncertain with respect to quiz feedback. We also separately code whether subjects were paid for correct quiz answers (Incentivized) and whether participants were required to get all questions correct before continuing (Require 100%).

Cross-Check

The meta-analysis was independently coded by two coders, who read each of the 243 papers in the review along with any instructions available in their online supplementary materials. For each of the 9 categories coded, both coders marked them as true ($=1$), false ($=0$), or uncertain ($=0.5$). Both coders agreed most of the time, only disagreeing (including cases where one coder was uncertain) in 296 out of 9×243 judgments, and only disagreeing fundamentally (i.e. one coder marking a “0” and the other a “1” on a given paper-category judgment) in 134 such judgments. The area with the most disagreement was the presence of demonstration, examples, or guided practice. These are particularly difficult to identify, as they are often buried in lengthy instructions and the difference between explanation and demonstration is somewhat subjective. We note that false negatives are more likely than false positives – it is easy to miss an example or demonstration in instructions but hard to see one where it doesn’t actually exist. After each person coded independently, both coders reconciled disagreements to put together the data for Table 1. Typically, when only one coder was uncertain, disagreement was resolved in favor of the certain coder. In the case of genuine disagreement coders discussed and settled on the most likely classification.

B Experimental Instructions

We include copies of all instructions pages as seen by each subject. Note that the printed instructions for the paper treatment did not include the screenshots shown in Figure B.4 and Figure B.6, since they completed practice periods for Tasks 1 and 2 as part of the on-screen instructions, like all other subjects.

Figure B.1: Instructions page 1: introduction to the experiment

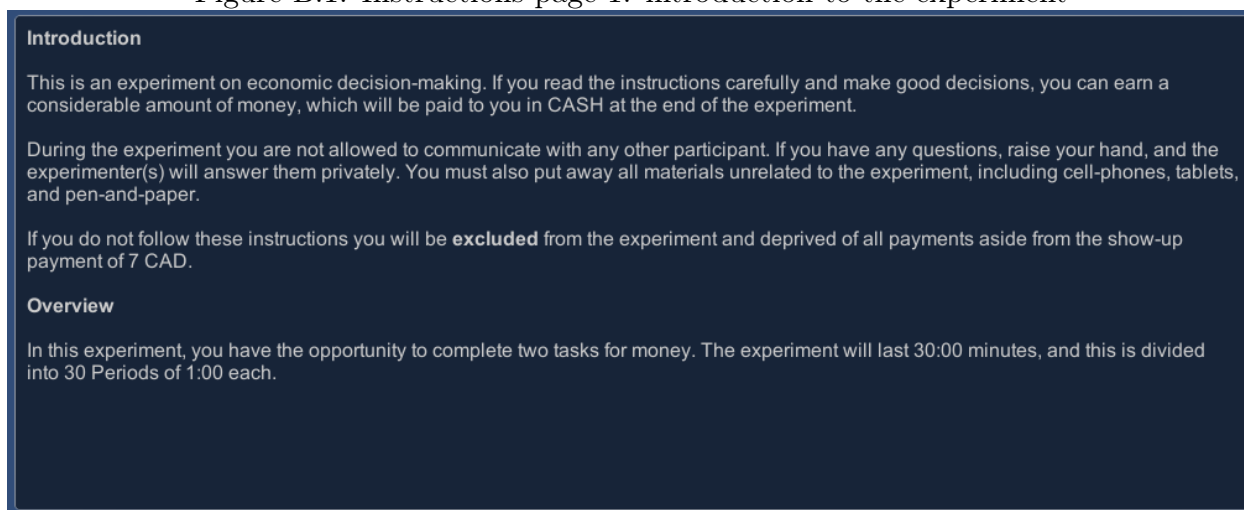


Figure B.2: Instructions page 2: description of Task 1

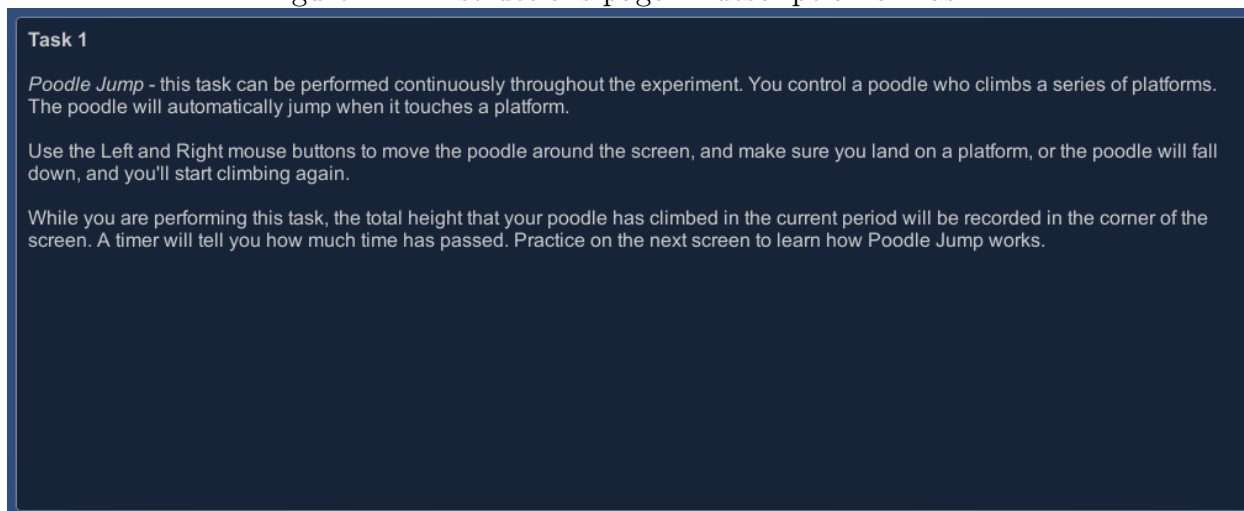


Figure B.3: Instructions page 3: Task 1 practice

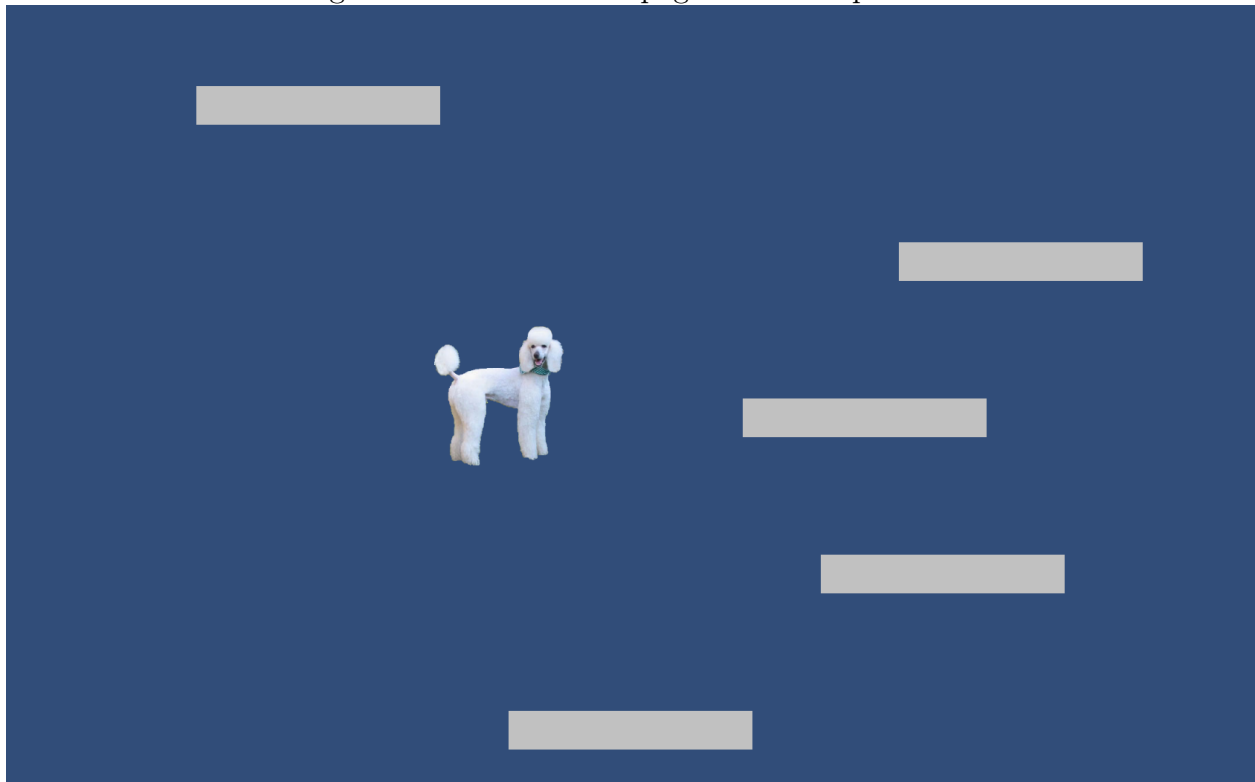


Figure B.4: Instructions page 4: description of Task 2

Task 2

Slider Task - this task will last for a total of 1:00 minutes - equivalent to 1 period(s) - and will consist of a screen with 4 sliders. Each slider has a number above it showing its current position. Each slider is initially positioned at **0** and can be moved as far as **100**.

You must use the mouse to move each slider. You can readjust the position of each slider as many times as you wish. However, to correctly complete the task, each slider must be positioned at **exactly 50** by the end of the 1:00 minute.

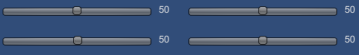
Just like in Poodle Jump, there will be a timer in the upper right corner of the screen. If the timer runs out and the sliders are not correctly positioned, then the task is incomplete.

Once (*and only once*) you will also be able to perform Task 2. You have to decide when to work on Task 2 by pressing the j key. When you press the j key, Task 2 will start immediately. When you start Task 2, the current period of Task 1 will be interrupted, but at the end of Task 2, you will restart where you left off.

Practice on the next screen to learn how the Slider Task works. Press the j button to continue.

Figure B.5: Instructions page 5: Task 2 practice

Click and drag all sliders to 50



Continue

The image shows a dark blue rectangular area containing a practice task. At the top, the text 'Click and drag all sliders to 50' is displayed. Below this text are four horizontal sliders arranged in two rows and two columns. Each slider is a light gray bar with a small dark gray knob in the center. To the right of each slider is the number '50'. Below the sliders is a dark gray rectangular button with the word 'Continue' in white text.

Figure B.6: Instructions page 6: payment schedule description

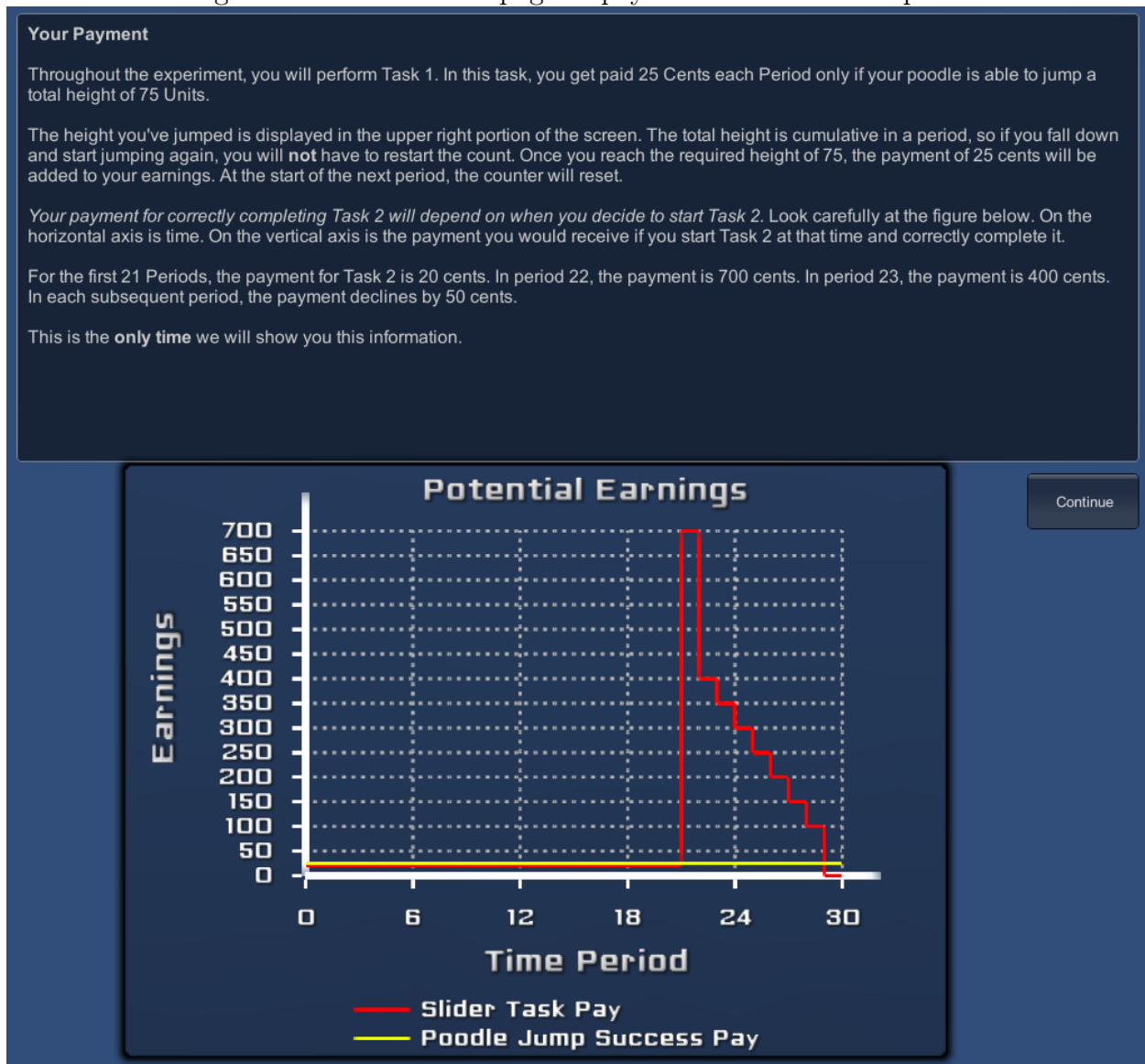


Figure B.7: Instructions page 7: summary

Summary

The experiment will start, and you will perform Task 1. During the experiment, you have to decide when (and whether) to perform Task 2.

Note that you cannot perform Task 1 and Task 2 at the same time. This means that any time spent on Task 2 will cause you to forgo your earnings from Task 1. In total, the experiment lasts 30 periods. If you never perform Task 2, you will have 30 periods to perform Task 1. By performing Task 2, you will give up 1 periods of Task 1 in order to earn your payment for Task 2.

You earn 25 cents in each period of Task 1, only if your poodle jumps at least 75 pixels in total height. You earn 0 cents otherwise.

Your payment for correctly completing Task 2 depends on when you start Task 2. Note that you are not required to perform Task 2. It is optional.

Are there any questions?

Figure B.8: Post-instructions quiz

- Q1. At what period is the payment to completing Task 2 the highest? A: _____
- Q2. What is the payment for completing Task 2 at a time indicated in your answer to Q1? A: _____
- Q3. What is the payment for completing Task 2 at a time before your answer to Q1? A: _____
- Q4. What is the payment for completing each period of Task 1? A: _____
- Q5. What key do you need to press to switch from Task 1 to Task 2? A: _____
- Q6. How many times may you complete Task 2? A: _____

C Robustness checks

In this section, we redo our analysis with three alternative measures of misunderstanding to check the robustness of our results. The specifications reported in Table C.1 are all analogous to the logit specifications in Table 3. The dependent variable “misunderstanding1” is equal to one if the subject did Task 2 before period 21 and equal to zero otherwise; this measure of misunderstanding allows for trembles. The “misunderstanding2” variable defines any behavioural deviation from optimality as misunderstanding. That is, it classifies a subject as misunderstanding unless they did Task 2 exactly in period 22. Finally, the “misunderstanding3” variable classifies those who did Task 2 before period 22 or never at all as misunderstanding. The results of these alternative specifications are broadly consistent with those reported in Table 3. Figure C.1 plots the share of subjects revealing misunderstanding in each treatment, by each of these alternative measures.

Figure C.1: Percentage of subjects revealing *misunderstanding*, under 3 alternative definitions of *misunderstanding*, by treatment.

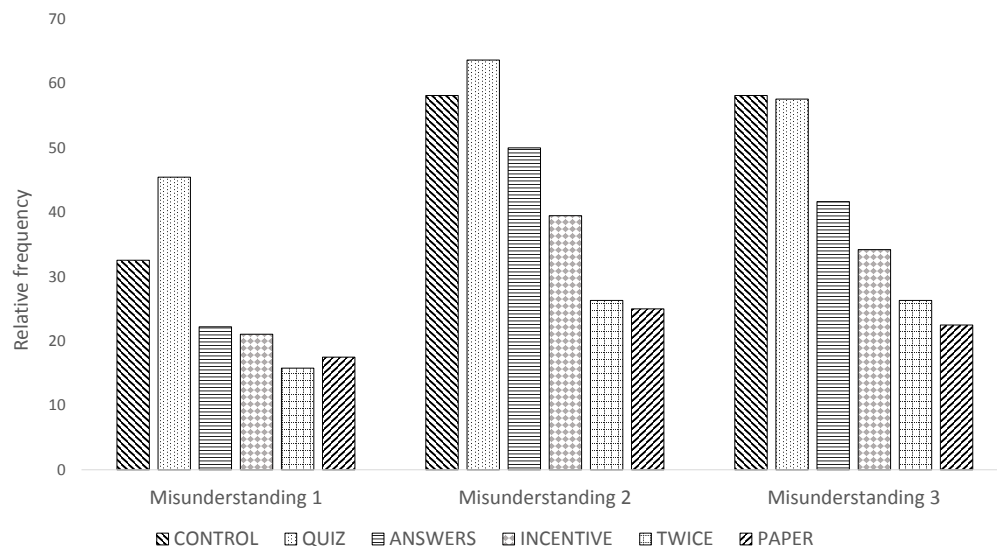


Table C.1: Marginal effects from logistic regression analysis of subject misunderstanding, robustness checks

	Dependent variable					
	Misunderstanding 1		Misunderstanding 2		Misunderstanding 3	
QUIZ	0.110 (0.104)		0.057 (0.118)		-0.005 (0.111)	
ANSWERS	-0.088 (0.077)	-0.148** (0.059)	-0.079 (0.106)	-0.149 (0.122)	-0.148 (0.093)	-0.149 (0.103)
INCENTIVE	-0.098 (0.075)	-0.143** (0.062)	-0.174* (0.096)	-0.208* (0.114)	-0.210** (0.085)	-0.178* (0.099)
TWICE	-0.146** (0.069)	-0.172*** (0.059)	-0.289*** (0.083)	-0.322*** (0.096)	-0.274*** (0.077)	-0.228** (0.092)
PAPER	-0.131* (0.070)	-0.068 (0.082)	-0.302*** (0.081)	-0.090 (0.134)	-0.308*** (0.073)	-0.072 (0.122)
Score		-0.104*** (0.023)		-0.282*** (0.047)		-0.221** (0.038)
Misunderstanding in CONTROL	0.325		0.581		0.581	
Observations	228	185	228	185	228	185
Log Likelihood	-124	-85.4	-146	-85.5	-144	-89.3
Note:	*p<0.1; **p<0.05; ***p<0.01					

References

- ALTMANN, S., FALK, A., GRUNEWALD, A. and HUFFMAN, D. (2014). Contractual incompleteness, unemployment, and labour market segmentation. *The Review of Economic Studies*, **81** (1), 30–56.
- AYCINENA, D., BALTADUONIS, R. and RENTSCHLER, L. (2014). Valuation structure in first-price and least-revenue auctions: an experimental investigation. *Experimental Economics*, **17** (1), 100–128.
- BROOKINS, P. and RYVKIN, D. (2014). An experimental study of bidding in contests of incomplete information. *Experimental Economics*, **17** (2), 245–261.
- CABRERA, S., FATÁS, E., LACOMBA, J. A. and NEUGEBAUER, T. (2013). Splitting leagues:

- promotion and demotion in contribution-based regrouping experiments. *Experimental Economics*, **16** (3), 426–441.
- COX, J. C. and JAMES, D. (2012). Clocks and trees: Isomorphic dutch auctions and centipede games. *Econometrica*, **80** (2), 883–903.
- ERICSON, K. M. M. and FUSTER, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *Quarterly Journal of Economics*, **126** (4), 1879–1907.
- ETANG, A., FIELDING, D. and KNOWLES, S. (2011). Does trust extend beyond the village? experimental trust and social distance in cameroon. *Experimental Economics*, **14** (1), 15–35.
- MITTONE, L. and PLONER, M. (2011). Peer pressure, social spillovers, and reciprocity: an experimental analysis. *Experimental Economics*, **14** (2), 203–222.