Instructions

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

October 21, 2018

Abstract

A survey 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' comprehension and retention of payoff-relevant information. We report a one-shot individual decision task in which non money-maximizing behavior can be unambiguously identified and find that such behavior is prevalent in our baseline treatment which uses plain, but relatively standard experimental instructions. We find combinations of reinforcement methods that can eliminate half of non money-maximizing behavior, and we find that we can induce a similar reduction via enhancements to the content of instructions. Residual non money-maximizing behavior suggests this may be an important source of noise in experimental studies.

JEL classification: C91

Keywords: Attention, Comprehension, Instructions

[*] Address: Freeman, Petersen and Tong, Department of Economics, Simon Fraser University, 8888 University Dr, Burnaby, BC V5S 1A6, Canada. Kimbrough, Smith Institute for Political Economy and Philosophy, Chapman University, One University Drive, Orange, CA 92866, USA. We thank Jim Sylvester for programming the experiment, Cameron Young for research assistance, and Bob Slonim, two anonymous referees, Kevin Laughren, and seminar and conference audiences at Simon Fraser University, Chapman 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).

1 Introduction

Experiments start by providing instructions designed to ensure that subjects understand how their actions and others' actions determine payoffs. Such understanding is crucial to the economic interpretation of subjects' behavior – without it, the experimenter has lost control (Smith, 1982). Almost from the field's inception, experimental economists have recognized that the effectiveness of instructions in establishing understanding may depend on how they are delivered and reinforced (Fouraker and Siegel 1963). Prominent textbooks 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 suggests wide variation in how practitioners deliver instructions and use reinforcement methods. We review the methods for delivering and reinforcing instructions as reported in experimental studies recently published in six leading journals and confirm 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 experimental economics lacks clear norms for how instructions ought to be delivered and reinforced. Troublingly, we were unable to classify roughly 22% of papers because they failed to provide sufficient details on their methods.

Despite observed variation in practices, there is scant evidence comparing their effectiveness. Thus we conduct an experiment to evaluate the impact of methods of delivering instructions and reinforcing their content on behavior. We study a one-shot timing decision in which each subject is performing a default Task 1 for money and must decide when (or whether) to switch over and complete Task 2. Task 2 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 Task 2 at all than doing it too early. This information is explicitly stated in the instructions. Doing the task too early – *non money-maximizing behavior* (NMB) – could reflect idiosyncratic preferences, or result from a failure to comprehend or retain information from the instructions. Variation in NMB across treatments, which hold the distribution of preferences constant in expectation, thus reflects variation in comprehension and retention. For most treatments, we hold constant the content of instructions and vary how instructions are delivered and reinforced. We include one additional treatment with enhanced instructions as a robustness check.

In our first treatment subjects complete self-paced computerized instructions including practice rounds and then take a comprehension quiz before beginning the study (providing us an alternative measure of their comprehension upon completion of the instructions). Nearly half of subjects in this treatment do the task too early, exhibiting NMB. A second treatment provides subjects with the quiz answers, and this generates a moderate, but statistically insignificant reduction in NMB. 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 baseline – but each only eliminates about half of the observed NMB, as does our treatment with enhanced instructions.

By studying an individual decision task, our experiment eliminates strategic and other-regarding motives that might confound the identification or interpretation of NMB. By studying a one-shot decision without feedback, we obtain a clean measure of understanding and retention of the instructions that is not confounded by learning. We are aware of two existing papers that have studied the impact of instruction delivery and reinforcement on play in repeated public goods games (Bigoni and Dragone, 2012; Ramalingam *et al.*, 2018).[1] The more relevant of these is Bigoni and Dragone (2012), who find that shortened on-screen instructions led to lower quiz scores and longer response times as compared to their baseline

[1] Our 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. See also Converse and Presser (1986) for a discussion of effective survey design which offers potentially useful guidance for economists.

paper instructions, shortened paper instructions, and shortened on-screen instructions with active examples requiring subject input. However, they find no effect of instructions on observed behavior.

2 Literature Survey

We report how instructions are delivered and reinforced in 260 experimental studies published between January 2011 and December 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. For each paper, we checked whether instructions were delivered on paper, on screen, both, or neither. We also recorded 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. Since ensuring subjects' initial comprehension may be particularly important when experiments are one-shot or provide limited feedback, we further classified the nature of each experiment based on whether or not a main task was one-shot, and whether or not subjects received feedback. This allows us to assess whether experimenters adapt their instruction protocols to the nature of the task being studied. Details of our classification procedure are given in Appendix A. The results of our survey are given in Table 1.

We were unable to determine how instructions were delivered in 22% of the studies we reviewed. If behavior is sensitive to how instructions are delivered, this oversight hampers replication. Of the remaining 204 studies, 61% deliver instructions exclusively on paper, 24% deliver instructions exclusively on screen, while another 5% use both. We find this noteworthy since the majority of these experiments are themselves computerized. The remaining 10% of these 204 studies use neither paper nor computer instructions. Most such studies are lab-in-the-field experiments studying non-student populations and deliver instructions

Table 1: Instruction delivery and reinforcement in economics experiments

		Delivery method					
		Computer only	Paper only	Computer and Paper	Neither	Unclear	Total
	Total	**48**	**124**	**11**	**21**	**56**	**260**
Reinforcement	Read aloud	19	79	4	21	17	140
	Practice/Demonstration	30	63	10	15	29	147
	Demo or guided practice	21	56	8	13	19	117
	Unguided practice	16	22	4	4	16	62
	Quiz	16	54	8	6	17	101
	Feedback	10	35	5	4	10	64
	Incentive	0	3	0	0	0	3
	Require 100%	5	23	3	3	7	41
	Feedback unclear	5	18	3	2	7	35
	One-shot	15	43	4	12	10	84
	Feedback between decisions	24	73	7	6	42	152

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

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.[2]

85% of all studies use at least one method of reinforcement which suggests that experimenters are almost universally concerned about subject comprehension and retention. Instructions are read aloud in 54% of studies. We find that 57% 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, 80% use guided demonstrations or guided practice rounds, and 42% use unguided practice rounds; some studies use both.

In addition to reinforcing the content of instructions, experiments can also test subjects' comprehension thereof with pre-experiment quizzes (39% of studies). At least 63% of these reinforced understandings and corrected misunderstandings by providing answers to the quiz, and 41% required a perfect score to commence the experiment. Only three of the studies paid subjects for quiz performance. We note that 35% of studies that used a quiz did not clearly report whether or how subjects were given feedback on the quiz.

Given our prior that reinforcement may be especially important when feedback is limited, we find it surprising that one-shot experiments less frequently incorporate practice or demonstrations ($\rho = -.19$, $p < .01$, $n = 260$) and quizzes ($\rho = -.15$, $p = .02$, $n = 260$) in their instructions; see Appendix A for more detail.

Our survey reveals wide variation in how experimenters deliver and reinforce instructions. Nevertheless, there are commonalities which seem to reflect some notion of 'best practices.'

[2] Reading instructions aloud and/or publicly distributing paper instructions may also help establish common information in strategic settings (Friedman and Sunder 1994, p. 77).

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

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 a one-shot, 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 initiate the second task. Doing the second 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 misses a platform, the poodle falls to the ground and the game restarts with no penalty. Each participant receives \$0.25 per period of Task 1, so long as they jump a minimum cumulative height. This height was chosen so that it would be trivially easy to complete but not automatic – effectively guaranteeing an attentive subject this payment each period.[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 zero to 100. The task is 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

[4] Only 5 out of 308 subjects ever failed to attain the required height in a period; 4 did so once and one subject did so twice. These failures account for only 0.1% of all Poodle Jump periods.

[5] Only one subject started but failed to complete the slider task.

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.

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 precludes the subject from maximizing their monetary payoffs. We use the NMB acronym to refer to such behavior below.

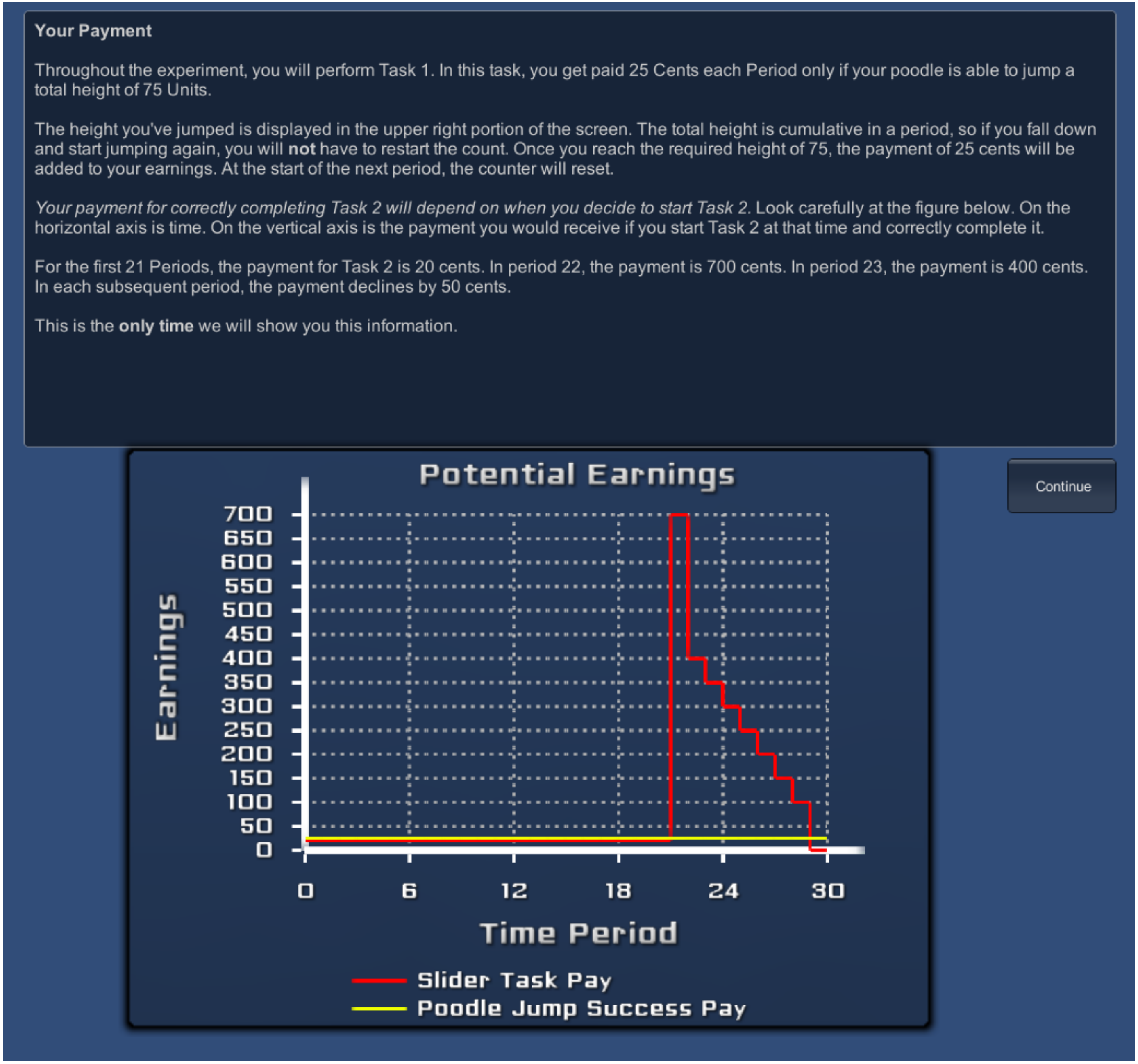
NMB can thus reveal that a subject failed to comprehend or retain a particularly key piece of payoff-relevant information from the instructions.[6] As hinted at earlier, our design restricts the set of possible preference-based explanations for NMB. Moreover, since we sample subjects from the same distribution of preferences in each treatment, variation in NMB across treatments identifies changes in comprehension and retention.

3.2 Treatment Design

We employ a between-subjects design with seven treatments. We study the effectiveness of different ways of delivering and reinforcing the experiment's instructions on NMB 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 comprehend or retain information from the

[6] We note that neither a subject who understood and retained this information but simply forgot to switch nor a subject who (for whatever reason) did not understand this information but only switched at or after period 22 would be coded as exhibiting NMB by this measure.

Figure 1: Screenshot showing how payoffs were described to subjects



Your Payment

Throughout the experiment, you will perform Task 1. In this task, you get paid 25 Cents each Period only if your poodle is able to jump a total height of 75 Units.

The height you've jumped is displayed in the upper right portion of the screen. The total height is cumulative in a period, so if you fall down and start jumping again, you will **not** have to restart the count. Once you reach the required height of 75, the payment of 25 cents will be added to your earnings. At the start of the next period, the counter will reset.

Your payment for correctly completing Task 2 will depend on when you decide to start Task 2. Look carefully at the figure below. On the horizontal axis is time. On the vertical axis is the payment you would receive if you start Task 2 at that time and correctly complete it.

For the first 21 Periods, the payment for Task 2 is 20 cents. In period 22, the payment is 700 cents. In period 23, the payment is 400 cents. In each subsequent period, the payment declines by 50 cents.

This is the **only time** we will show you this information.

Table 2: Summary of treatments

Treatment	Quiz	Answers	Additional Reinforcement	# of Subjects
NO QUIZ	No	No	No	43
QUIZ	Yes	No	No	76
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
ENHANCED	Yes	No	Only through enhanced on-screen instructions	37

instructions there is the potential for variation in delivery and additional reinforcement to reduce NMB. Our treatments test the impact of various more-or-less standard procedures employed by experimenters to improve comprehension and retention. 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.

The NO QUIZ treatment presents the instructions on screen with no additional reinforcement. The NO QUIZ treatment gives us information on NMB when subjects read instructions on their own.

The QUIZ treatment was identical to the NO QUIZ except that each subject completed a six question comprehension quiz on paper at the end of the on-screen instructions; subjects were informed that there would be a quiz prior to beginning the instructions, but no feedback was given on the quiz. The QUIZ treatment allows us to assess whether the presence of the quiz affects NMB, and the quiz itself gives a secondary measure of comprehension. When we analyze our data, we use this as our baseline treatment for comparison to the other treatments below.

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 corrected possible misunderstandings revealed in quiz answers and reinforced key pieces of information from the instructions. As noted by Cassar and Friedman (2004), a quiz is a good way to "make sure that the subjects understand the rules" (p. 71); thus we expect providing the answers to the quiz will correct failures of comprehension or retention and reduce NMB.

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. As noted by Friedman and Sunder (1994), "[when] a subject does not seem to understand the instructions [...] the experimenter may reread the relevant part of the instructions or go through an example" (p. 77). Repeating the instructions TWICE achieves both of these objectives and thus should reduce NMB.

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. Pay for performance is standard in experimental economics because economists believe it motivates subjects to think carefully and participate actively in experiments (Hertwig and Ortmann, 2001). By paying for performance on the quiz, we anticipate that subjects will exert more effort in carefully reading the instructions, thereby reducing NMB.

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] We thus expect PAPER to improve comprehension as measured by quiz scores and reduce NMB both for this reason, and through improving retention given the quiz score since written instructions are available throughout the session.

The ENHANCED treatment was identical to the QUIZ treatment but with enhanced

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

on-screen instructions.[8] Compared to the other treatments, the on-screen instructions were lengthened from five to seven screens in length. In these enhanced instructions, Figure 1 appeared four times (instead of only once), and subjects were presented with four worked-out examples that explained the payoff that would result from different possible switching times. Unlike in our other treatments, the last page of the enhanced instructions included Figure 1, and each subject waited on that page while other subjects completed the instructions and while they completed the quiz. With the benefit of hindsight, we emphasized the details we knew past subjects had failed to grasp. This treatment is also consistent with the advice of Friedman and Sunder (1994), applied between-subjects, and we expect the ENHANCED instructions to similarly reduce NMB.

For reasons explained above, we hypothesize that each additional form of reinforcement reduces NMB. Specifically, we conjectured that having a QUIZ would have a similar level of NMB as NO QUIZ, but relative to these treatments, ANSWERS would reduce NMB, each of our remaining interventions on top of that (INCENTIVE, TWICE, and PAPER) would further reduce NMB, and ENHANCED would also reduce NMB relative to QUIZ. We hypothesized that higher quiz scores will be associated with lower rates of NMB, and that in the INCENTIVE, PAPER, and ENHANCED treatments most or all reductions in NMB are reflected in higher quiz scores, while the ANSWERS and TWICE treatments reduce NMB given quiz scores.

Our experiment differs from existing studies on instructions in two regards. First, this is an individual decision task, so there is neither complexity from strategic behavior nor other-regarding concerns. 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 NMB and attribute variation in NMB to variation in the delivery and reinforcement of instructions. Nonetheless, we believe that our experiment provides a good analogy to other experiments, particularly those where a

[8] The ENHANCED treatment was added later on a suggestion from the editor.

decision of interest is only one of multiple decisions the subject makes. We also conjecture that more complicated experiments face at least as much risk of misunderstanding as exists in our simple experiment (even if most existing experiments are unable to diagnose it).

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; subjects in the INCENTIVE treatment were informed that they would be paid for their quiz performance above and beyond their earnings from the experiment. 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, PAPER, and ENHANCED treatments; the correct answers were revealed after all participants had completed the quiz except in the QUIZ and ENHANCED treatments. In the TWICE treatment, subjects completed the on-screen instructions a second time, including practice rounds. Then the experiment started. At the end of some sessions, we conducted a post-experiment questionnaire (Appendix D).[9]

We recruited 308 participants to 45 sessions through Simon Fraser University's CRABE recruiting system, with no subject participating in more than one session. Each session lasted under an hour. Average earnings were 18.37 CAD including a 7 CAD show-up payment. We collected no other demographic data nor other behavioral measures.

[9] We have responses from 72 subjects because this was added at the suggestion of a referee.

4 Results

We use a subject's decision to do Task 2 at any time before period 22 as NMB, which is our behavioral measure of their failure to pay attention to, comprehend, absorb, or retain information from the instructions. Table 3 shows the share of NMB by treatment. All p-values reported below are two-sided.

Finding 1: NMB is prevalent.

In our NO QUIZ and QUIZ treatments, 44% and 47% of subjects exhibited NMB by doing Task 2 before period 22. This is despite the fact that these treatments include both demonstrations and practice periods. Even in our most effective treatment, the corresponding share is 18%. These findings suggest that failures to comprehend or retain information from instructions may be an important source of noise.[10] This justifies concern about the effectiveness of instruction delivery and reinforcement methods.

Finding 2: Combining reinforcement methods reduces *NMB*.

We find that additional reinforcement reduces NMB: we reject the joint hypothesis that NMB occurs at the same rate across all treatments (Fisher's exact test, $p < .01$, $n = 308$). Compared to NO QUIZ and QUIZ, we observe somewhat less NMB in the ANSWERS treatment (33%), but we do not detect any statistically significant differences between these treatments (Fisher's exact test of equal NMB rates across these treatments, $p = .35$, n =155). In each of the INCENTIVE (24%), TWICE (18%), and PAPER (23%) treatments that provide additional reinforcement, subjects exhibited significantly less NMB than in the QUIZ treatment (Fisher's exact tests, $p < .02, .01, .01$, $n = 114, 114, 116$ respectively). While the ENHANCED treatment (22%) reduces NMB (Fisher's exact test, $p = .01$, $n = 113$), it does not eliminate it.[11] Our findings suggest that more detailed instructions and extensive

[10] In Appendix C, we show that we find similar results if we account for trembles by defining NMB based on doing Task 2 before period 21.

[11] We cannot reject the hypothesis INCENTIVE, TWICE, PAPER, and ENHANCED lead to similar improvements (Fisher's exact test of no association, $p = .96$, $n = 153$).

Table 3: *Non Money-maximizing Behavior* across treatments

	NO QUIZ	QUIZ	ANSWERS	INCENTIVE	TWICE	PAPER	ENHANCED
NMB	.442	.474	.333	.237	.184	.225	.216
Quiz Score (avg.)	n/a	4.10	4.06	4.32	4.53	5.43	4.59
QUIZ	.849						
ANSWERS	.362	.220					
INCENTIVE	.064	.016	.442				
TWICE	.017	.004	.186	.779			
PAPER	.062	.010	.316	1.00	.781		
ENHANCED	.056	.013	.302	1.00	.779	1.00	

First row reports the fraction of NMB by treatment.

Second row reports the average quiz score by treatment. Remaining entries report a p-value from a Fisher's exact test of differences in NMB between treatments.

reinforcement each improve comprehension and retention of the instructions.

Finding 3: Lower quiz scores are associated with NMB. Providing quiz answers while also making incorrect answers salient can reduce NMB among lower performers.

Quiz scores provide an alternative measure of subject comprehension immediately after the instructions. In the QUIZ treatment which provides neither feedback nor additional reinforcement, quiz score and NMB are negatively related (Goodman-Kruskal γ, $p < 0.01$, $n = 76$); indeed 13 of 76 subjects had a perfect score on the quiz, and none of them subsequently exhibited NMB in the experiment. In fact, across all of our treatments we find it striking that only one of the 73 people with a perfect quiz score exhibited NMB.[12] This indicates that full *comprehension* at the completion of the instructions appears to be a sufficient condition for avoiding NMB in our experiment and that *retention* is a second-order issue.

Our quiz score data enable us to test whether the INCENTIVE, PAPER, and ENHANCED treatments improved subjects' comprehension as demonstrated on the quiz, compared to the pooled distribution of quiz scores from the QUIZ, ANSWERS, and TWICE treatments, which followed identical procedures up to the collection of the quiz.[13] Average quiz scores by treatment are reported in Table 3. To our surprise, neither the INCENTIVE nor the ENHANCED treatment significantly improved quiz scores (rank-sum tests, $p = .59, .14$, $n = 188, 187$, respectively). The PAPER treatment, which made the answers accessible to subjects during the quiz, improved scores significantly (rank-sum test, $p < 0.01$, $n = 190$), and the linear regression in Table 4, column 3 shows that PAPER had the largest effect on quiz score of all of our treatments.[14]

Quiz score data also allow us to further assess *how* our treatments reduce NMB.

[12] One person with a perfect quiz score in the TWICE treatment switched 28 seconds too early.

[13] We find no significant differences in the distribution of quiz scores in the QUIZ, ANSWER, and TWICE treatments (Kruskal-Wallis test, $p = .15$, $n = 150$).

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

Table 4: Treatment effects on Non Money-maximizing Behavior and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB		Quiz Score	NMB	
	(1)	(2)	(3)	(4)	n
NO QUIZ	-0.128				
	(-0.889, 0.632)				
ANSWERS	-0.588	0.051	-0.050	-0.138	112
	(-1.424, 0.248)	(-2.884, 2.987)	(-0.586, 0.487)	(-0.308, 0.048)	
ANSWERS × Quiz Score		-0.206		0.00715	
		(-0.941, 0.528)		(-0.069, 0.085)	
INCENTIVE	-1.065**	-2.531*	0.211	-0.219	114
	(-1.948, -0.182)	(-5.332, 0.271)	(-0.354, 0.775)	(-0.392, -0.030)	
INCENTIVE × Quiz Score		0.361		-0.028	
		(-0.257, 0.978)		(-0.111, 0.049)	
TWICE	-1.383***	-2.181	0.421	-0.255***	114
	(-2.329, -0.436)	(-5.235, 0.873)	(-0.156, 0.998)	(-0.425, -0.065)	
TWICE × Quiz Score		0.207		-0.057	
		(-0.467, 0.880)		(-0.145, 0.021)	
PAPER	-1.131**	7.334*	1.320***	0.134	116
	(-2.009, -0.253)	(-0.810, 15.478)	(0.922, 1.718)	(-0.189, 0.343)	
PAPER × Quiz Score		-1.485*		-0.177***	
		(-2.970, 0.001)		(-0.273, -0.085)	
ENHANCED	-1.182**	-0.557	0.489*	-0.188*	113
	(-2.096, -0.269)	(-4.596, 3.482)	(-0.030, 1.008)	(-0.382, 0.013)	
ENHANCED × Quiz Score		-0.116		-0.067*	
		(-0.993, 0.760)		(-0.151, 0.004)	
Quiz Score		-0.679***			
		(-1.053, -0.306)			
Intercept	-0.105	2.683***	4.105***		
	(-0.561, 0.350)	(0.993, 4.374)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

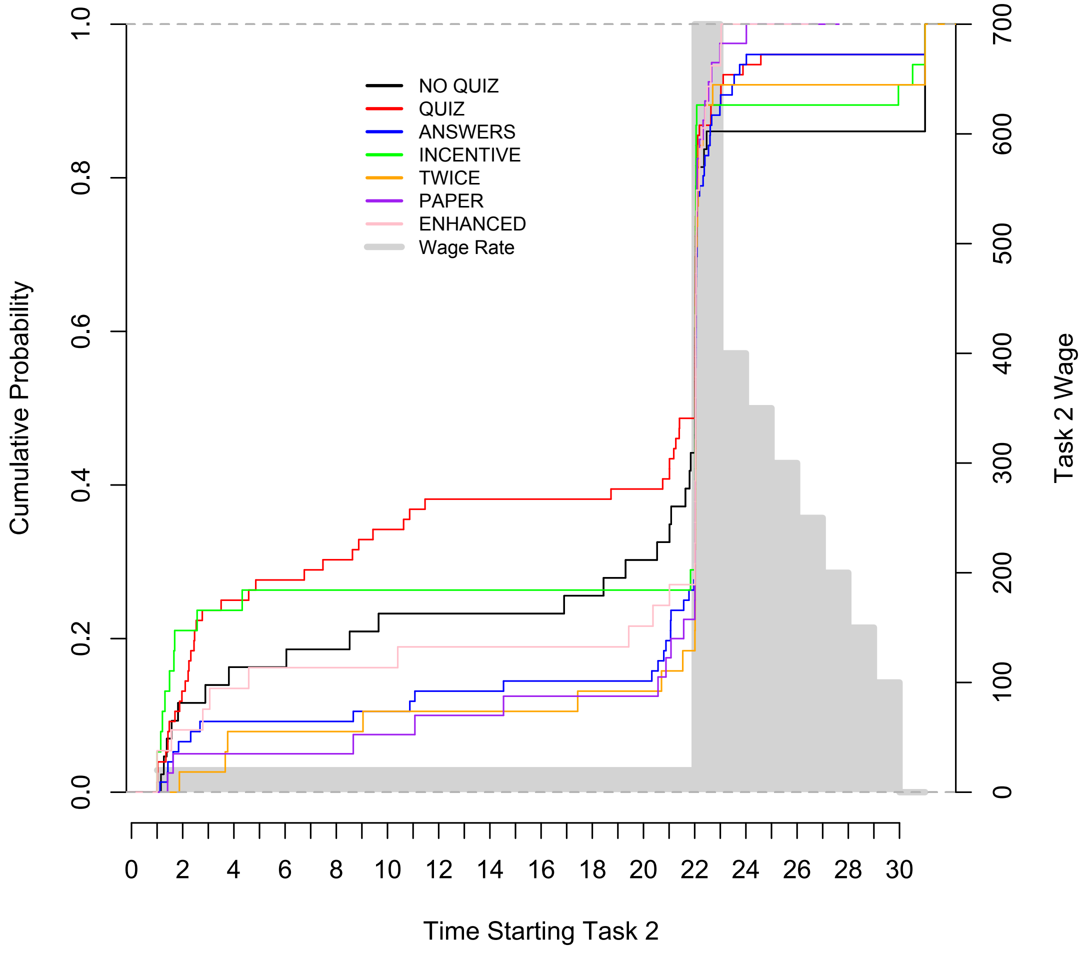
Goodman-Kruskal γ tests revealed that quiz score had a significant ($p < .05$ in each test, $n = 76, 36, 38, 40, 37$ respectively for each of QUIZ, ANSWERS, TWICE, PAPER, and ENHANCED) negative relationship with NMB in each treatment except INCENTIVE (where $p = .054$, $n = 38$) and NO QUIZ (where scores were not available). To decompose the extent to which treatment effects operate via (i.e. are *mediated* through) improved comprehension demonstrated on the quiz, we perform mediation analysis (applying the approach of Imai *et al.* 2010) in column 4 of Table 4, based on a model of NMB as a logistic-linear function of quiz score, treatment, and their interactions (column 2), and a linear regression to model treatment effects on quiz scores (column 3). The INCENTIVE and TWICE treatments have sizable and significant direct effects but insignificant and small mediated effects.[15] This indicates that these treatments primarily reduce NMB by clearing up (TWICE) and making salient (INCENTIVE) failures of comprehension demonstrated on the quiz. In contrast, the PAPER treatment has the largest mediated effect of all treatments, which is statistically significant, but only a small and insignificant direct effect beyond that. Mediated and direct effects of the ENHANCED treatment are each borderline insignificant, indicating a mix of both types of effects, but point estimates indicate a larger direct effect.

Robustness Checks Figure 2 shows empirical CDFs of completion times for Task 2, by treatment. For robustness, we show in Appendix C that we would arrive at similar qualitative conclusions to those reported in Table 4 using any of three alternative measures of NMB which vary the strictness of the criteria by which we classify behavior as NMB.

Our post-experiment questionnaire was only partially able to diagnose causes of NMB in our experiment (see Appendix D for a full analysis). While subjects' responses are correlated with behavior and quiz scores, they fail to provide any indication of the differences between the QUIZ and ENHANCED treatments in NMB revealed in the experiment.

[15] In the case of TWICE, this is reassuring since any mediated effect can only arise due to sampling variation.

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



5 Discussion

Our experiments indicate that even when using combinations of reinforcement methods including demonstrations, practice periods, and a quiz, many subjects' behavior reveals that they fail to pay attention to, understand, or retain information from the instructions. Combining these with further reinforcement methods reduced NMB, as did increasing the level of detail in the instructions. Each of these methods leads to a similar improvement but does not eliminate NMB.

In our setting, we feel confident attributing variation in the anomalous behavior that we observe to a variation in the 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 NMB. Note that though we are able to reduce NMB in our design, some residual NMB persists even in the best case. While the extent of such NMB is likely to vary with experimental context (e.g. subject pool, design, feedback), its presence is noteworthy and has implications for the power and interpretation of experimental tests.

Finally, our findings motivate advice on how to report and deliver instructions. First, experimenters should be aware that the way instructions are delivered and reinforced has consequences for behavior. Second, we suggest providing paper instructions when possible, since this requires no extra lab time, is almost free, and is about as effective in reducing NMB in our experiment as other reinforcement methods. Third, 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.[16] Journals' efforts to require experimenters to share

[16] For example, recent work by Chen *et al.* (2018) demonstrates, via new experiments following different instructions protocols, that a recent failed replication attempt arose because of differences in how instructions were delivered.

copies of their instructions are laudable, and these could be complemented by standardized reporting of how instructions are delivered and reinforced.

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.

Chen, R., Chen, Y. and Riyanto, Y. (2018). Public knowledge in coordination games: Learning from non-replication.

Converse, J. and Presser, S. (1986). *Survey Questions: Handcrafting the Standadized Questionnaire*. Sage.

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.

Hertwig, R. and Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, **24** (3), 383–403.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, **25** (1), 51–71.

Ramalingam, A., Morales, A. and Walker, J. (2018). Instruction length and content: Effects on punishment behaviour in public goods games. *Journal of Behavioral and Experimental Economics*, **73**, 66–73.

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

Supplementary Appendix to Freeman, Kimbrough, Petersen, and Tong (2018)

A Review of current practice

Inclusion/Exclusion criteria

We included experimental papers published between January 2011 and December 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%).

Coding Criteria: Some main task(s) is (are) one shot

We classified the main task or tasks for each experiment. If at least one of the main tasks is one shot (that is, subject can be viewed as making a single decision) in one or more of the treatments, we coded that paper as having a one shot main task under this column. When researchers use a choice list or the strategy method – where multiple similar decisions are made almost simultaneously, and could in-principle be viewed as one decision – we view this task as a one-shot task. In contrast, when decisions are made in a sequence, even without feedback, we would not consider those to constitute a one-shot task. Anderson *et al.*'s (2011) study provides an edge case. In their experiment, each subject plays six public goods games with different parameter values, but all six choices are presented at the same time. Since all choices are instances of the same basic task and are presented at once, we coded their experiment as one shot. If these tasks had been presented sequentially on separate screens, we would not have coded this as one shot. An interesting boundary case is a dynamic game with an evolving state variable (e.g. the money supply variable in Petersen and Winn (2014)); subjects in such games make repeated decisions in the same task, but with different incentives depending on the state. We have coded these as repeated (i.e. not one shot) because there is typically feedback between decisions and the state dependence is usually not so severe that subsequent decisions differ fundamentally from those made in initial round. The opportunities for learning from repetition thus usually dominate (though not necessarily always), and we note that we did not explicitly account for this in our coding.

Coding Criteria: Some main task(s) has (have) feedback between decisions

If at least one of the main tasks was repeated with feedback between rounds in one or more of the treatments, we coded that paper as having a repeated main task with feedback under this column (e.g. a repeated public goods game in which subjects learned their payoff after each round (e.g. Bayer *et al.* (2013)). We considered it sufficient for a subsequent round to involve choices in the same basic task as the preceding one for which feedback was given. For example, in Noussair and Stoop (2015), subjects in one treatment completed two dictator games in a row, with different reward media (money and time) with feedback between them – we viewed these as repetitions of the same task with feedback.

Coding Criteria: More than one task

We coded whether an experiment has more than one incentivized task. In some cases, an experiment required subjects to input multiple separate decisions associated with the same broader task – in these instances, we coded this a single task (as discussed above). Sometimes a single task has multiple decisions (e.g. a centipede game as in Cox and James (2012) or a public goods game with punishment as in Harris *et al.* (2015)). Similarly, in an experiment that required subjects to vote on a sanctioning scheme that would then be implemented in a public goods game (Kamei *et al.*, 2015), we viewed the vote and the subsequent game as one task. Many experiments coded as having more than one task would follow up a main task with a secondary preference elicitation.

Cross-Check

Each paper was independently coded by two coders, who read each of the 260 papers in the review along with any instructions available in their online supplementary materials. For each of the 11 categories coded, both coders marked them as true (=1), false (=0), or

Table A.1: Correlation between experiment type and delivery and reinforcement

	One-shot	*p*-value	Feedback between decisions	*p*-value
Paper only	.048	.437	.008	.899
Computer only	-.011	.863	-.082	.189
Both	.018	.770	.022	.722
Neither	.157	.011	-.180	.004
Read aloud	.112	.072	-.092	.141
Practice/Demonstration	-.191	.002	.190	.002
Quiz	-.146	.019	.159	.010

Table reports pairwise correlations between delivery/reinforcement category (rows) and experiment type (columns) and their *p*-values.

uncertain (=0.5). Both coders agreed most of the time, only disagreeing (including cases where one coder was uncertain) in 363 out of 11 × 260 judgments, and only disagreeing fundamentally (i.e. one coder marking a "0" and the other a "1" on a given paper-category judgment) in 200 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.

Correlations amongst practices

One-shot experiments account for about one third of the experiments using computerized instructions (31%) or paper instructions (35%). 57% of experiments that use neither paper nor on-screen instructions are one-shot games; most of these studies are field experiments in which experimenters read instructions aloud or go through the instruction one-on-one with subjects.

Table A.2: Instruction practices by feedback

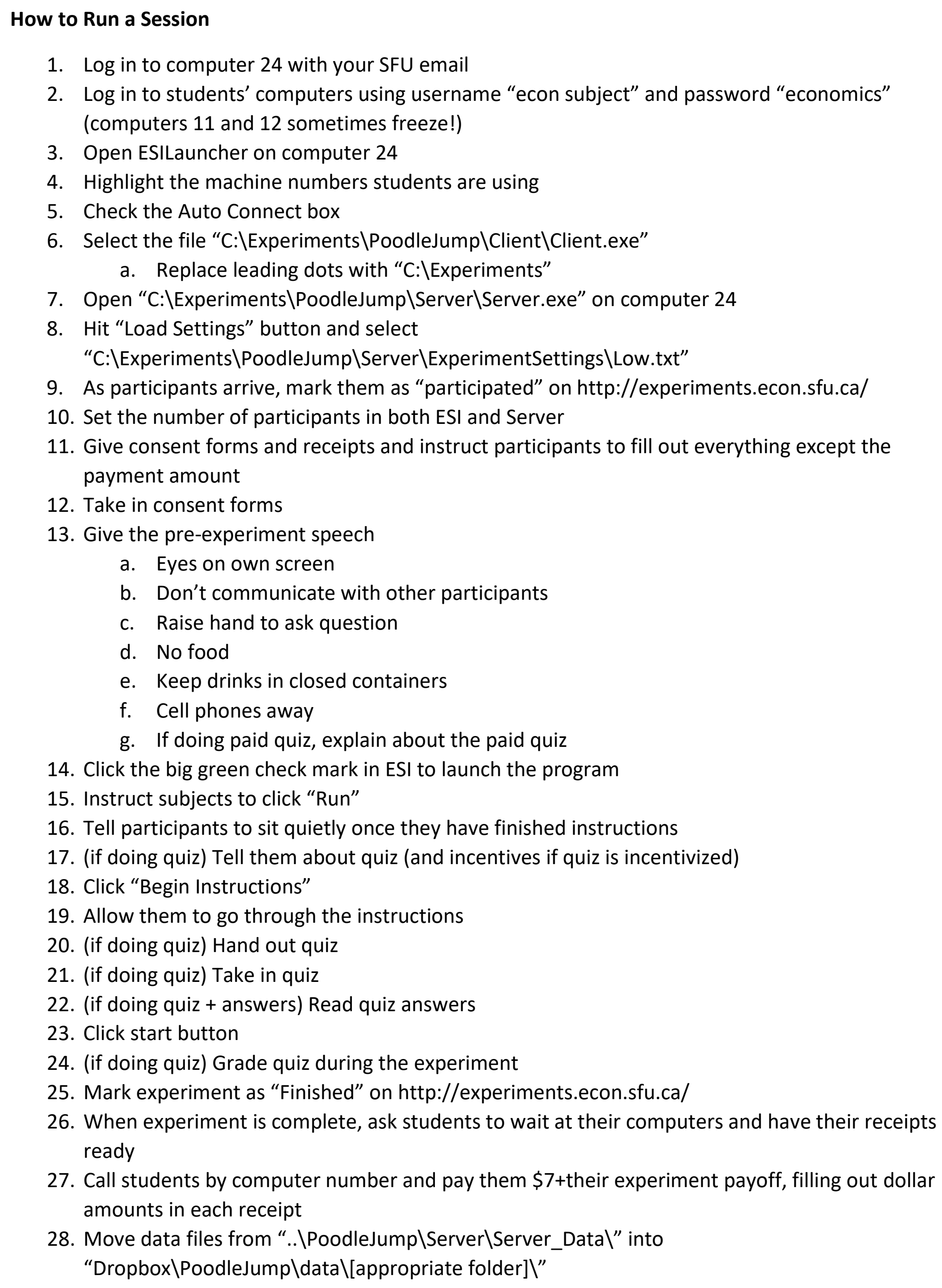
	One-shot	Feedback between decisions
Total	84	152
Read aloud	52	76
Practice/Demonstration	36	98
Quiz	24	69

We also find that one-shot experiments tend to be less likely to use each of the reinforcement methods (except for reading aloud) – even though such experiments give no feedback, making each subject's initial understanding of the instructions crucial. We suspect that this is because one-shot experiments tend to be simpler and therefore easier to explain. Instructions are read aloud more often in one-shot game experiments (62%) than in experiments with feedback between decisions (50%). Other reinforcement methods are used less often in one-shot experiments than in experiments with feedback between decisions (respectively, 43% versus 65% use some form of practice or demonstration, while 29% versus 45% use a quiz). These differences result in a significant negative association between one-shot experiments and use of practice/demonstration ($\rho = -.191$, $p = .002$) and quizzes ($\rho = -.146$, $p = .019$) in the instructions.

B Experimental Instructions

The experimental sessions all followed the script in Figure B.1.

Figure B.1: Experimenter's script for running a session

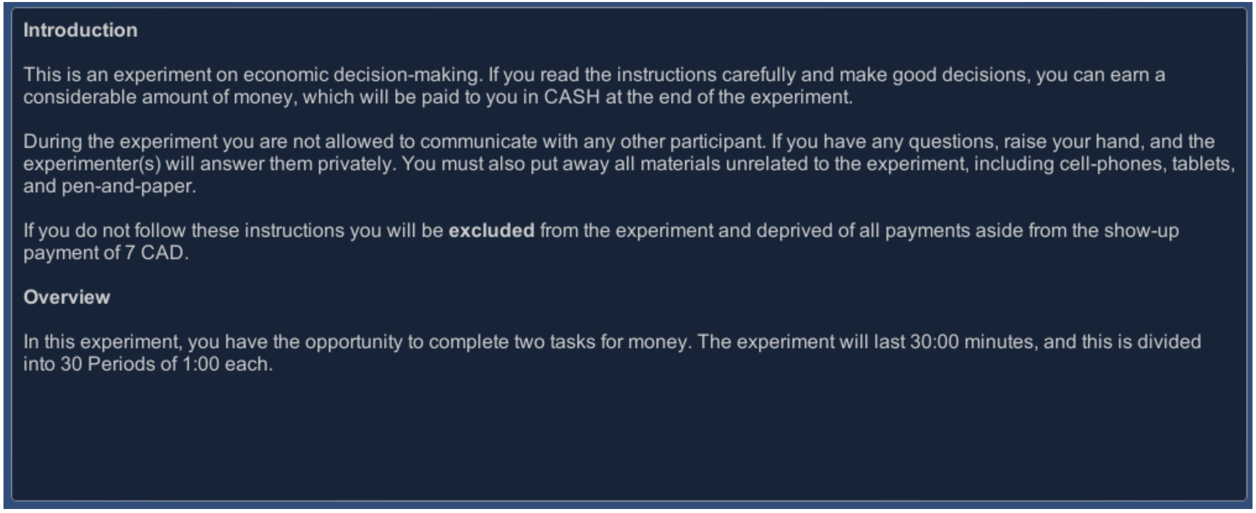


How to Run a Session

1. Log in to computer 24 with your SFU email
2. Log in to students' computers using username "econ subject" and password "economics" (computers 11 and 12 sometimes freeze!)
3. Open ESILauncher on computer 24
4. Highlight the machine numbers students are using
5. Check the Auto Connect box
6. Select the file "C:\Experiments\PoodleJump\Client\Client.exe"
    a. Replace leading dots with "C:\Experiments"
7. Open "C:\Experiments\PoodleJump\Server\Server.exe" on computer 24
8. Hit "Load Settings" button and select "C:\Experiments\PoodleJump\Server\ExperimentSettings\Low.txt"
9. As participants arrive, mark them as "participated" on http://experiments.econ.sfu.ca/
10. Set the number of participants in both ESI and Server
11. Give consent forms and receipts and instruct participants to fill out everything except the payment amount
12. Take in consent forms
13. Give the pre-experiment speech
    a. Eyes on own screen
    b. Don't communicate with other participants
    c. Raise hand to ask question
    d. No food
    e. Keep drinks in closed containers
    f. Cell phones away
    g. If doing paid quiz, explain about the paid quiz
14. Click the big green check mark in ESI to launch the program
15. Instruct subjects to click "Run"
16. Tell participants to sit quietly once they have finished instructions
17. (if doing quiz) Tell them about quiz (and incentives if quiz is incentivized)
18. Click "Begin Instructions"
19. Allow them to go through the instructions
20. (if doing quiz) Hand out quiz
21. (if doing quiz) Take in quiz
22. (if doing quiz + answers) Read quiz answers
23. Click start button
24. (if doing quiz) Grade quiz during the experiment
25. Mark experiment as "Finished" on http://experiments.econ.sfu.ca/
26. When experiment is complete, ask students to wait at their computers and have their receipts ready
27. Call students by computer number and pay them $7+their experiment payoff, filling out dollar amounts in each receipt
28. Move data files from "..\PoodleJump\Server\Server_Data\" into "Dropbox\PoodleJump\data\[appropriate folder]\"

We include copies of all instructions pages as seen by each subject in all treatments. First, we show the screenshots that apply for all except for the ENHANCED treatment. 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.2: Instructions page 1: introduction to the experiment



Introduction

This is an experiment on economic decision-making. If you read the instructions carefully and make good decisions, you can earn a considerable amount of money, which will be paid to you in CASH at the end of the experiment.

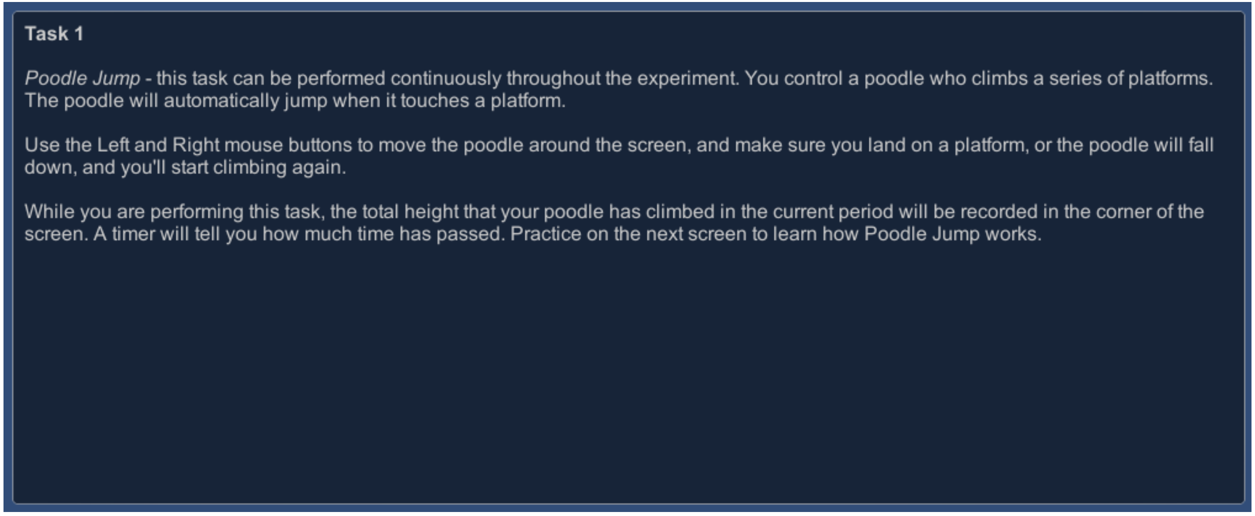
During the experiment you are not allowed to communicate with any other participant. If you have any questions, raise your hand, and the experimenter(s) will answer them privately. You must also put away all materials unrelated to the experiment, including cell-phones, tablets, and pen-and-paper.

If you do not follow these instructions you will be **excluded** from the experiment and deprived of all payments aside from the show-up payment of 7 CAD.

Overview

In this experiment, you have the opportunity to complete two tasks for money. The experiment will last 30:00 minutes, and this is divided into 30 Periods of 1:00 each.

Figure B.3: Instructions page 2: description of Task 1



Task 1

Poodle Jump - this task can be performed continuously throughout the experiment. You control a poodle who climbs a series of platforms. The poodle will automatically jump when it touches a platform.

Use the Left and Right mouse buttons to move the poodle around the screen, and make sure you land on a platform, or the poodle will fall down, and you'll start climbing again.

While you are performing this task, the total height that your poodle has climbed in the current period will be recorded in the corner of the screen. A timer will tell you how much time has passed. Practice on the next screen to learn how Poodle Jump works.

Figure B.4: Instructions page 3: Task 1 practice

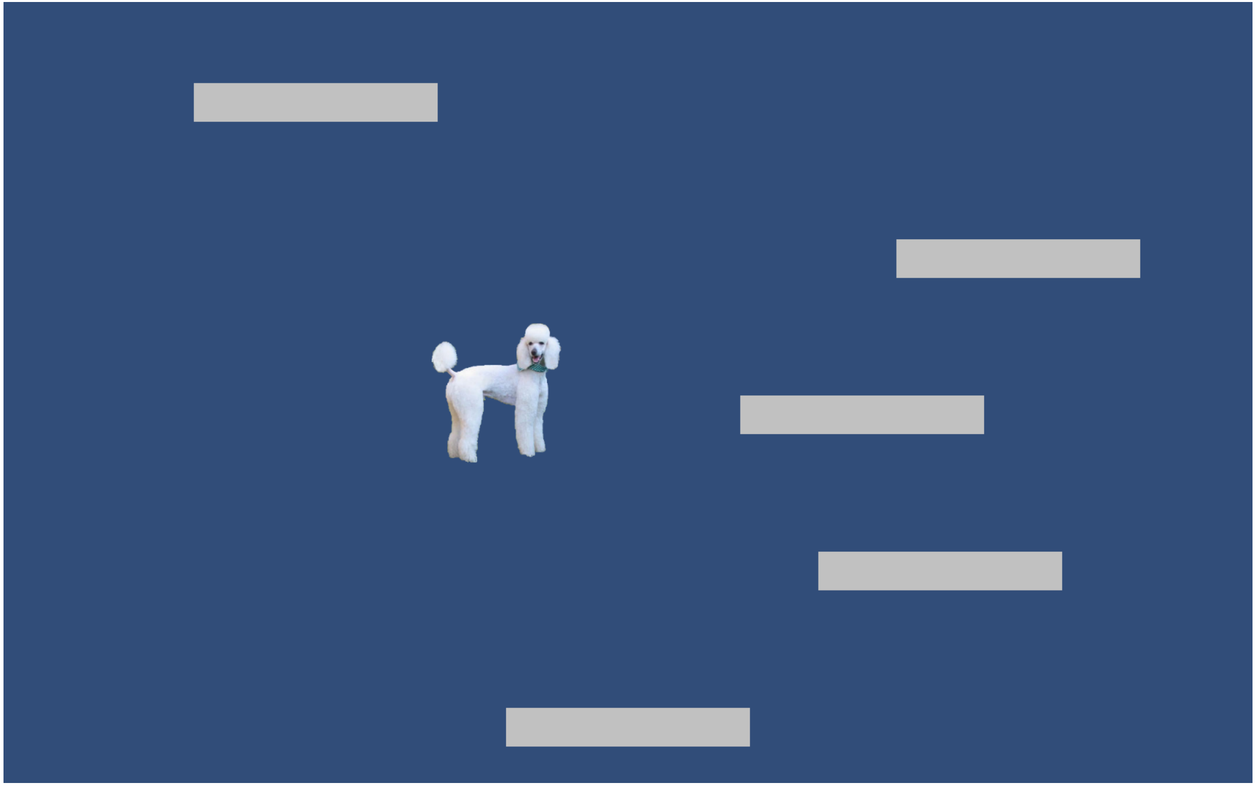
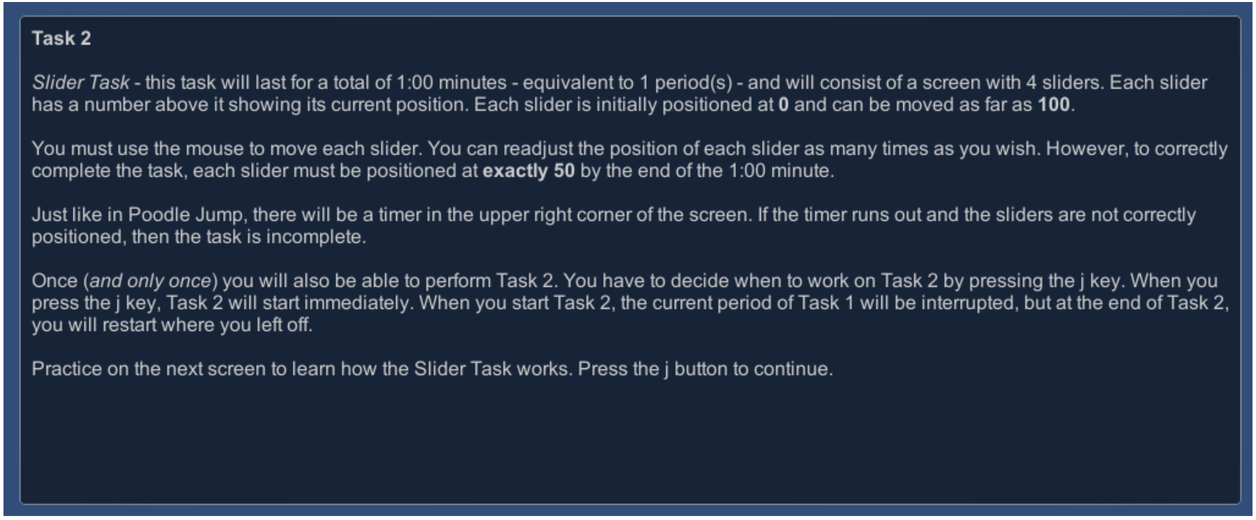


Figure B.5: Instructions page 4: description of Task 2



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**.

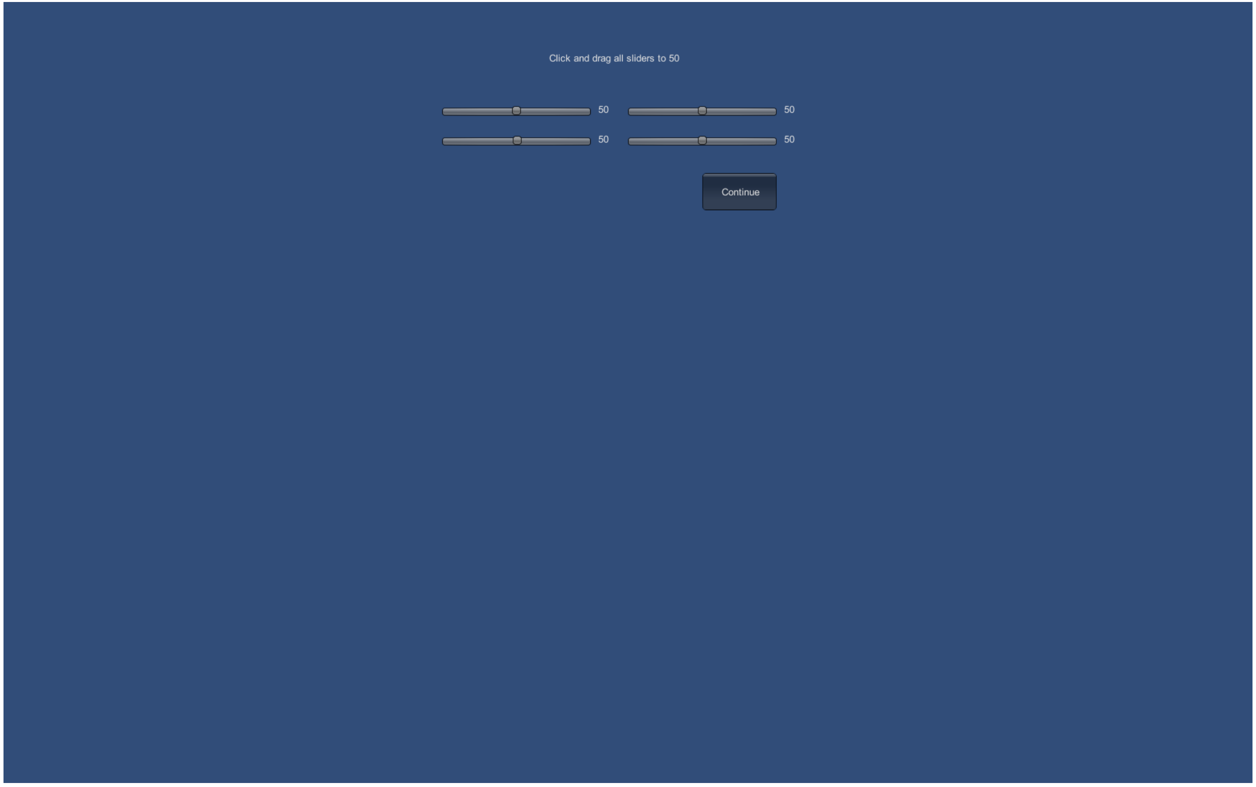
You must use the mouse to move each slider. You can readjust the position of each slider as many times as you wish. However, to correctly complete the task, each slider must be positioned at **exactly 50** by the end of the 1:00 minute.

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.

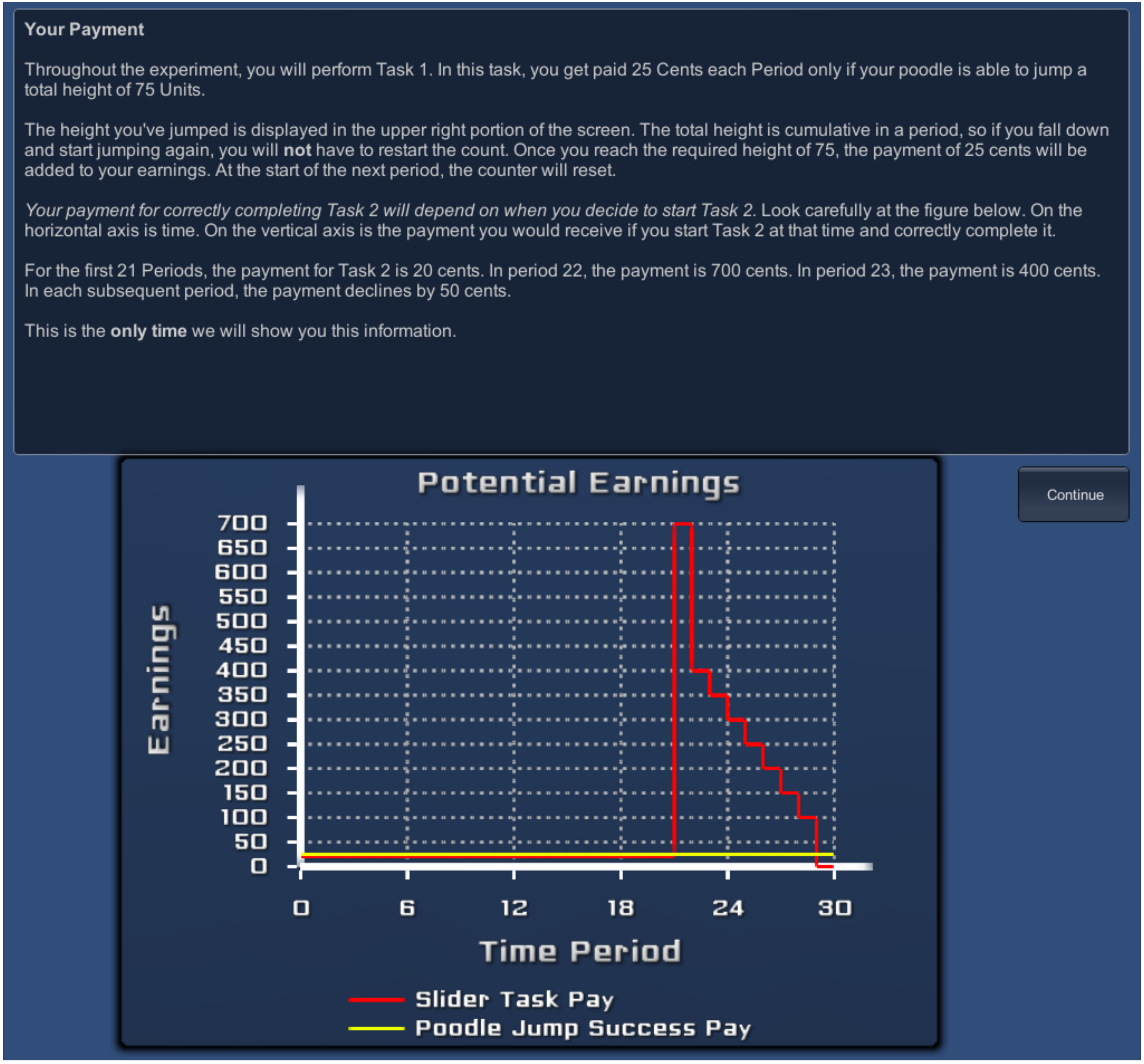
Practice on the next screen to learn how the Slider Task works. Press the j button to continue.

Figure B.6: Instructions page 5: Task 2 practice



Click and drag all sliders to 50

Figure B.7: Instructions page 6: payment schedule description



Your Payment

Throughout the experiment, you will perform Task 1. In this task, you get paid 25 Cents each Period only if your poodle is able to jump a total height of 75 Units.

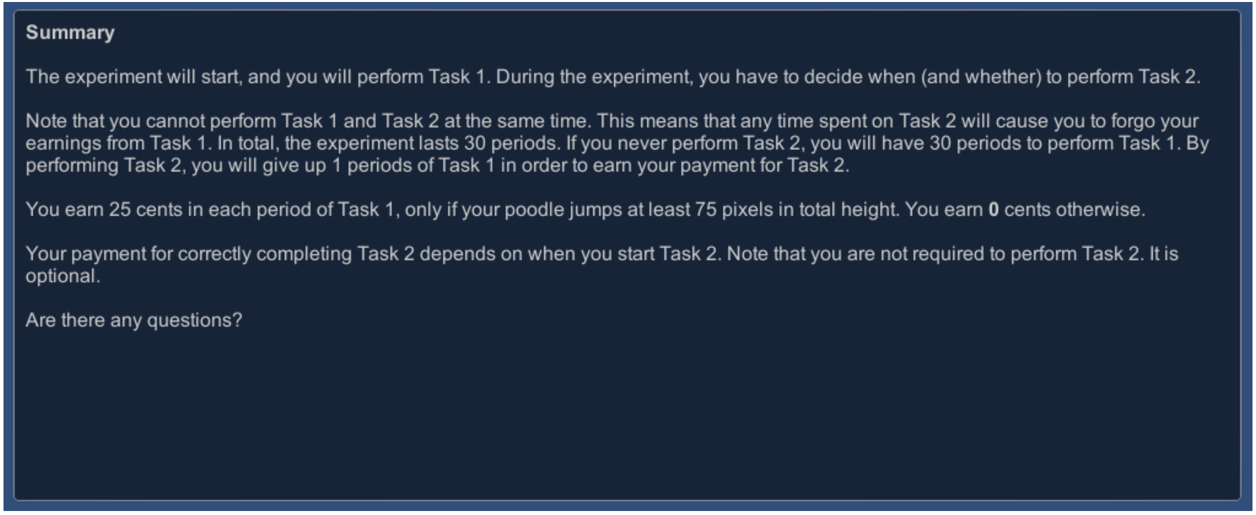
The height you've jumped is displayed in the upper right portion of the screen. The total height is cumulative in a period, so if you fall down and start jumping again, you will **not** have to restart the count. Once you reach the required height of 75, the payment of 25 cents will be added to your earnings. At the start of the next period, the counter will reset.

Your payment for correctly completing Task 2 will depend on when you decide to start Task 2. Look carefully at the figure below. On the horizontal axis is time. On the vertical axis is the payment you would receive if you start Task 2 at that time and correctly complete it.

For the first 21 Periods, the payment for Task 2 is 20 cents. In period 22, the payment is 700 cents. In period 23, the payment is 400 cents. In each subsequent period, the payment declines by 50 cents.

This is the **only time** we will show you this information.

Figure B.8: 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.

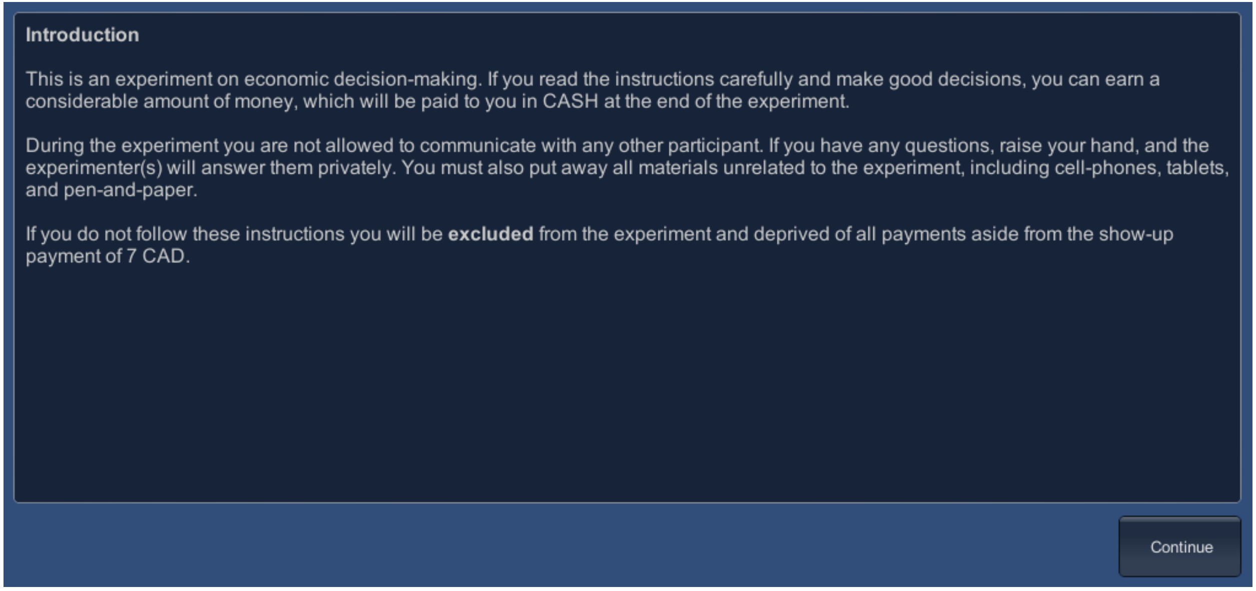
You earn 25 cents in each period of Task 1, only if your poodle jumps at least 75 pixels in total height. You earn **0** cents otherwise.

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?

Next, we include screenshots from the instructions from the ENHANCED treatment. Note that, unlike in the other treatments, the final summary screen remained displayed in the ENHANCED while subjects wrote the quiz.

Figure B.9: ENHANCED Instructions page 1: introduction to the experiment



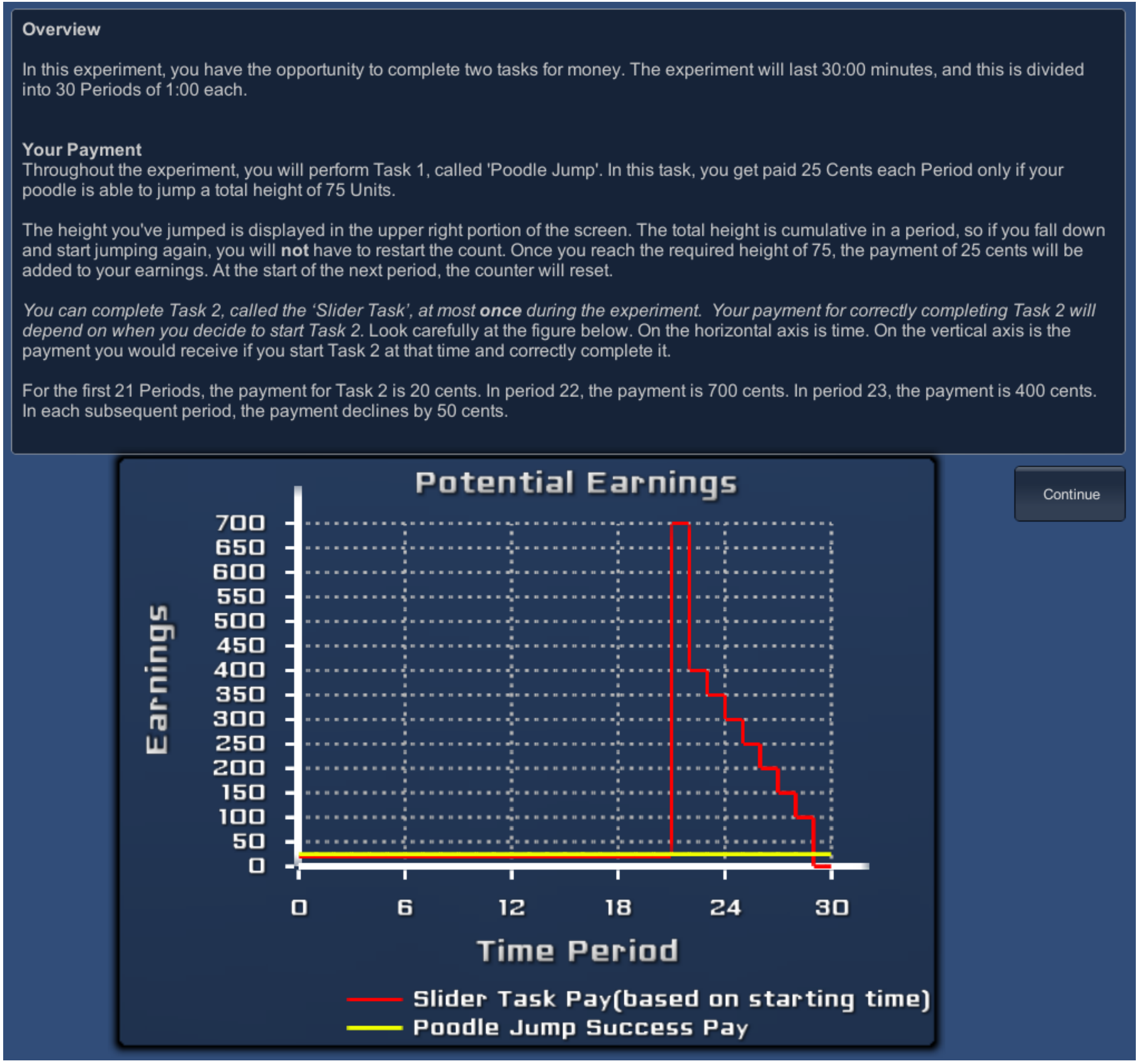
Introduction

This is an experiment on economic decision-making. If you read the instructions carefully and make good decisions, you can earn a considerable amount of money, which will be paid to you in CASH at the end of the experiment.

During the experiment you are not allowed to communicate with any other participant. If you have any questions, raise your hand, and the experimenter(s) will answer them privately. You must also put away all materials unrelated to the experiment, including cell-phones, tablets, and pen-and-paper.

If you do not follow these instructions you will be **excluded** from the experiment and deprived of all payments aside from the show-up payment of 7 CAD.

Figure B.10: ENHANCED Instructions page 2: overview and payment



Overview

In this experiment, you have the opportunity to complete two tasks for money. The experiment will last 30:00 minutes, and this is divided into 30 Periods of 1:00 each.

Your Payment

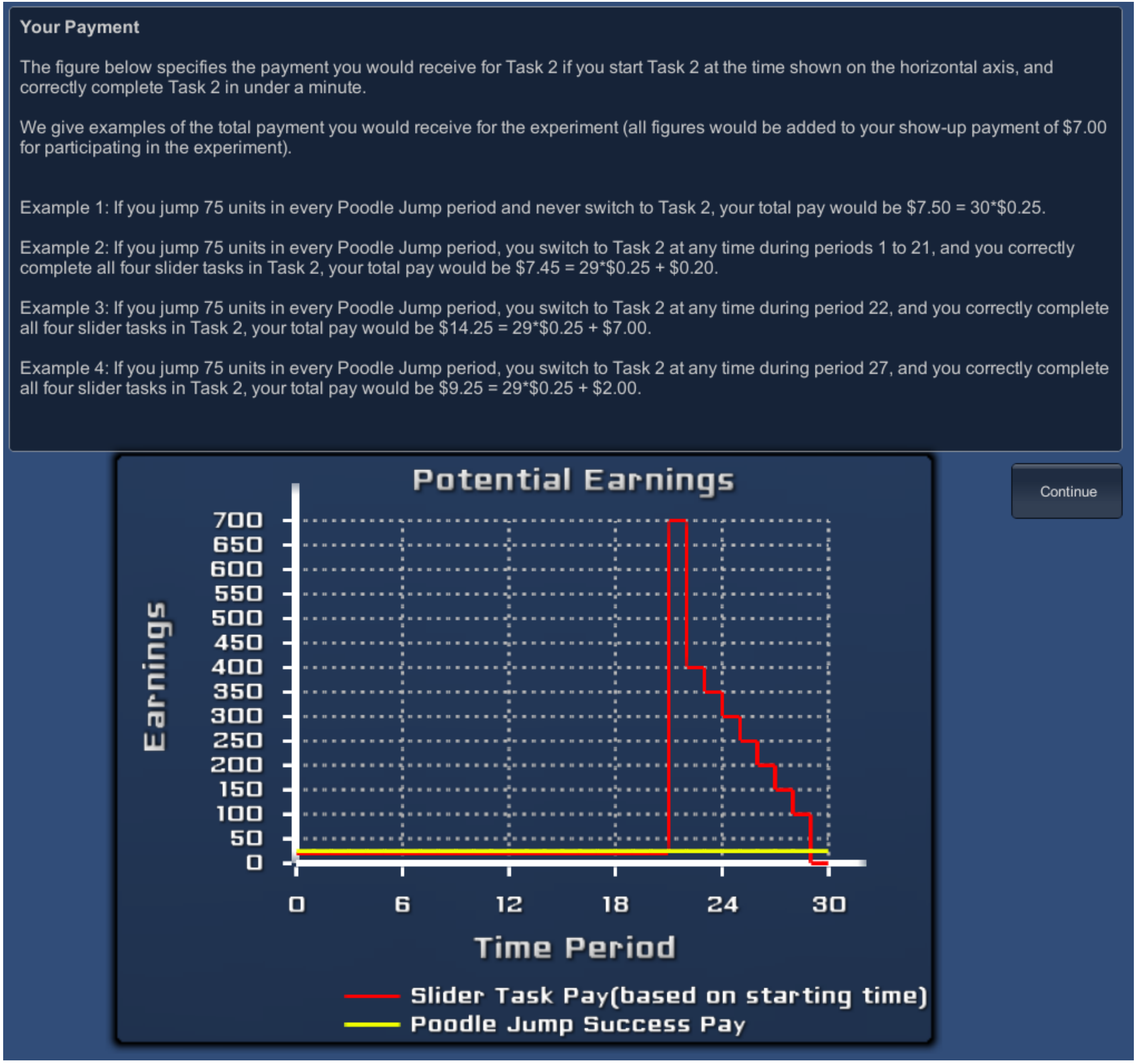
Throughout the experiment, you will perform Task 1, called 'Poodle Jump'. In this task, you get paid 25 Cents each Period only if your poodle is able to jump a total height of 75 Units.

The height you've jumped is displayed in the upper right portion of the screen. The total height is cumulative in a period, so if you fall down and start jumping again, you will **not** have to restart the count. Once you reach the required height of 75, the payment of 25 cents will be added to your earnings. At the start of the next period, the counter will reset.

You can complete Task 2, called the 'Slider Task', at most ***once*** *during the experiment. Your payment for correctly completing Task 2 will depend on when you decide to start Task 2.* Look carefully at the figure below. On the horizontal axis is time. On the vertical axis is the payment you would receive if you start Task 2 at that time and correctly complete it.

For the first 21 Periods, the payment for Task 2 is 20 cents. In period 22, the payment is 700 cents. In period 23, the payment is 400 cents. In each subsequent period, the payment declines by 50 cents.

Figure B.11: ENHANCED Instructions page 3: payment examples



Your Payment

The figure below specifies the payment you would receive for Task 2 if you start Task 2 at the time shown on the horizontal axis, and correctly complete Task 2 in under a minute.

We give examples of the total payment you would receive for the experiment (all figures would be added to your show-up payment of $7.00 for participating in the experiment).

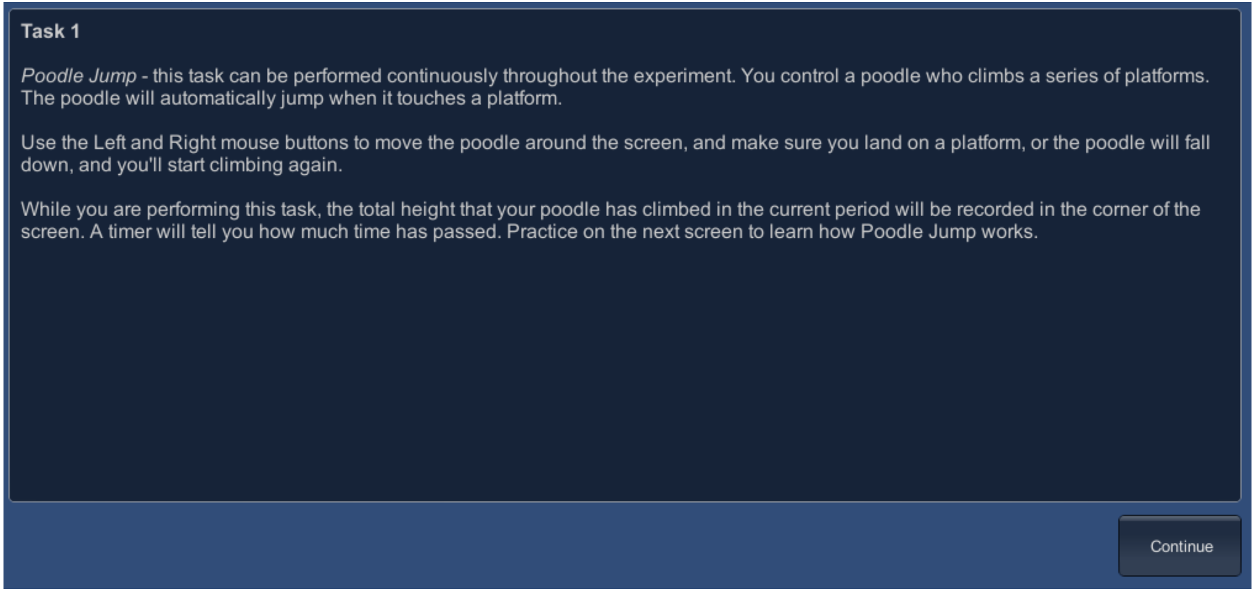
Example 1: If you jump 75 units in every Poodle Jump period and never switch to Task 2, your total pay would be $7.50 = 30*$0.25.

Example 2: If you jump 75 units in every Poodle Jump period, you switch to Task 2 at any time during periods 1 to 21, and you correctly complete all four slider tasks in Task 2, your total pay would be $7.45 = 29*$0.25 + $0.20.

Example 3: If you jump 75 units in every Poodle Jump period, you switch to Task 2 at any time during period 22, and you correctly complete all four slider tasks in Task 2, your total pay would be $14.25 = 29*$0.25 + $7.00.

Example 4: If you jump 75 units in every Poodle Jump period, you switch to Task 2 at any time during period 27, and you correctly complete all four slider tasks in Task 2, your total pay would be $9.25 = 29*$0.25 + $2.00.

Figure B.12: ENHANCED Instructions page 4: description of Task 1



Task 1

Poodle Jump - this task can be performed continuously throughout the experiment. You control a poodle who climbs a series of platforms. The poodle will automatically jump when it touches a platform.

Use the Left and Right mouse buttons to move the poodle around the screen, and make sure you land on a platform, or the poodle will fall down, and you'll start climbing again.

While you are performing this task, the total height that your poodle has climbed in the current period will be recorded in the corner of the screen. A timer will tell you how much time has passed. Practice on the next screen to learn how Poodle Jump works.

Figure B.13: ENHANCED Instructions page 5: Task 1 practice

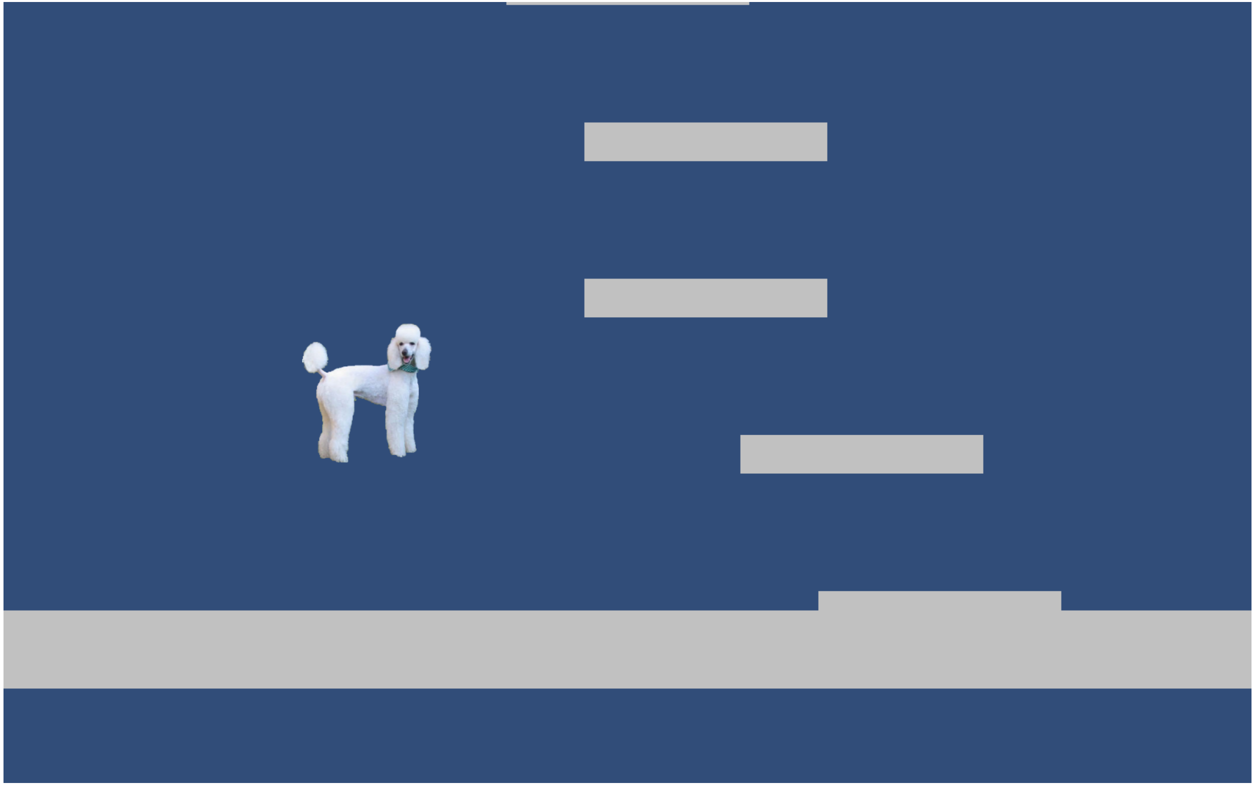
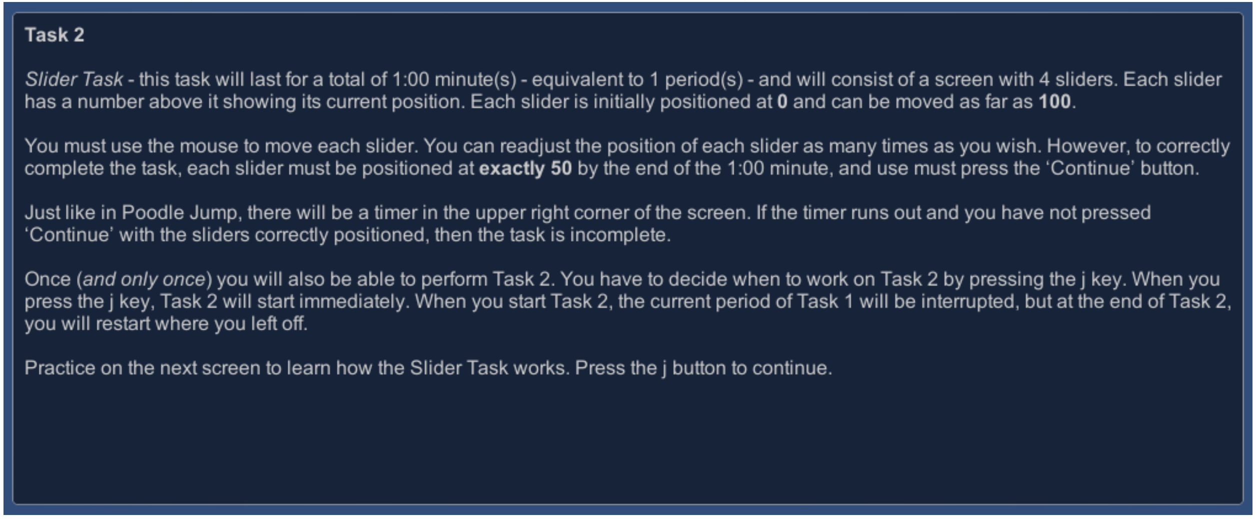


Figure B.14: ENHANCED Instructions page 6: description of Task 2



Task 2

Slider Task - this task will last for a total of 1:00 minute(s) - 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**.

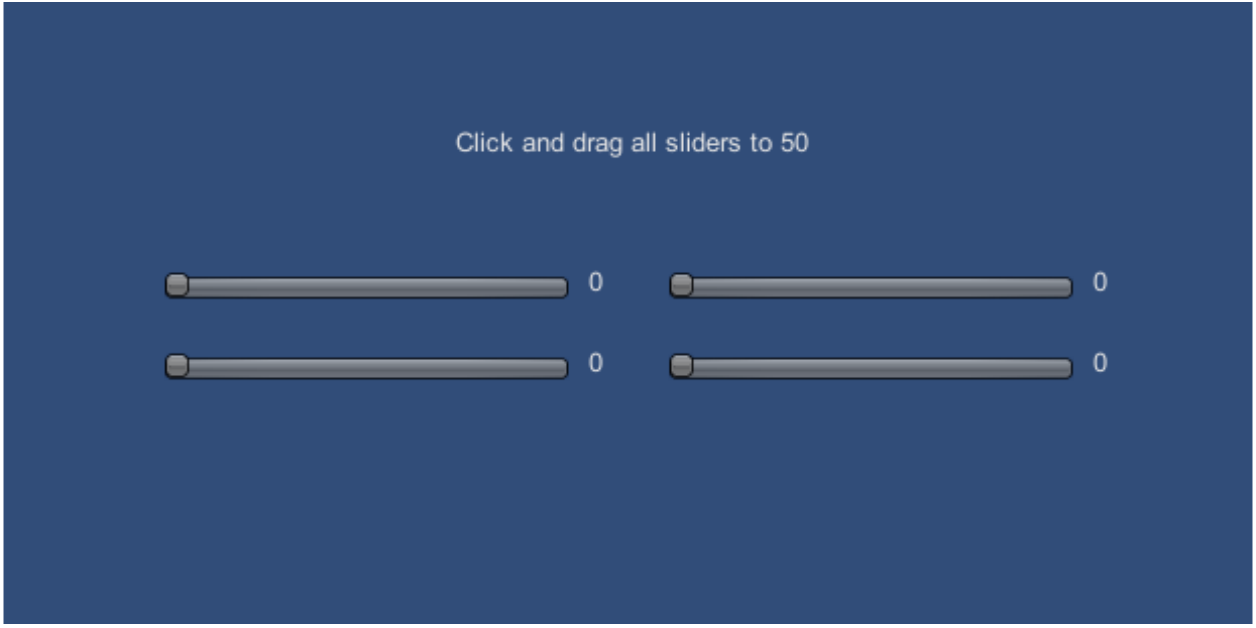
You must use the mouse to move each slider. You can readjust the position of each slider as many times as you wish. However, to correctly complete the task, each slider must be positioned at **exactly 50** by the end of the 1:00 minute, and use must press the ‘Continue’ button.

Just like in Poodle Jump, there will be a timer in the upper right corner of the screen. If the timer runs out and you have not pressed ‘Continue’ with the sliders 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.

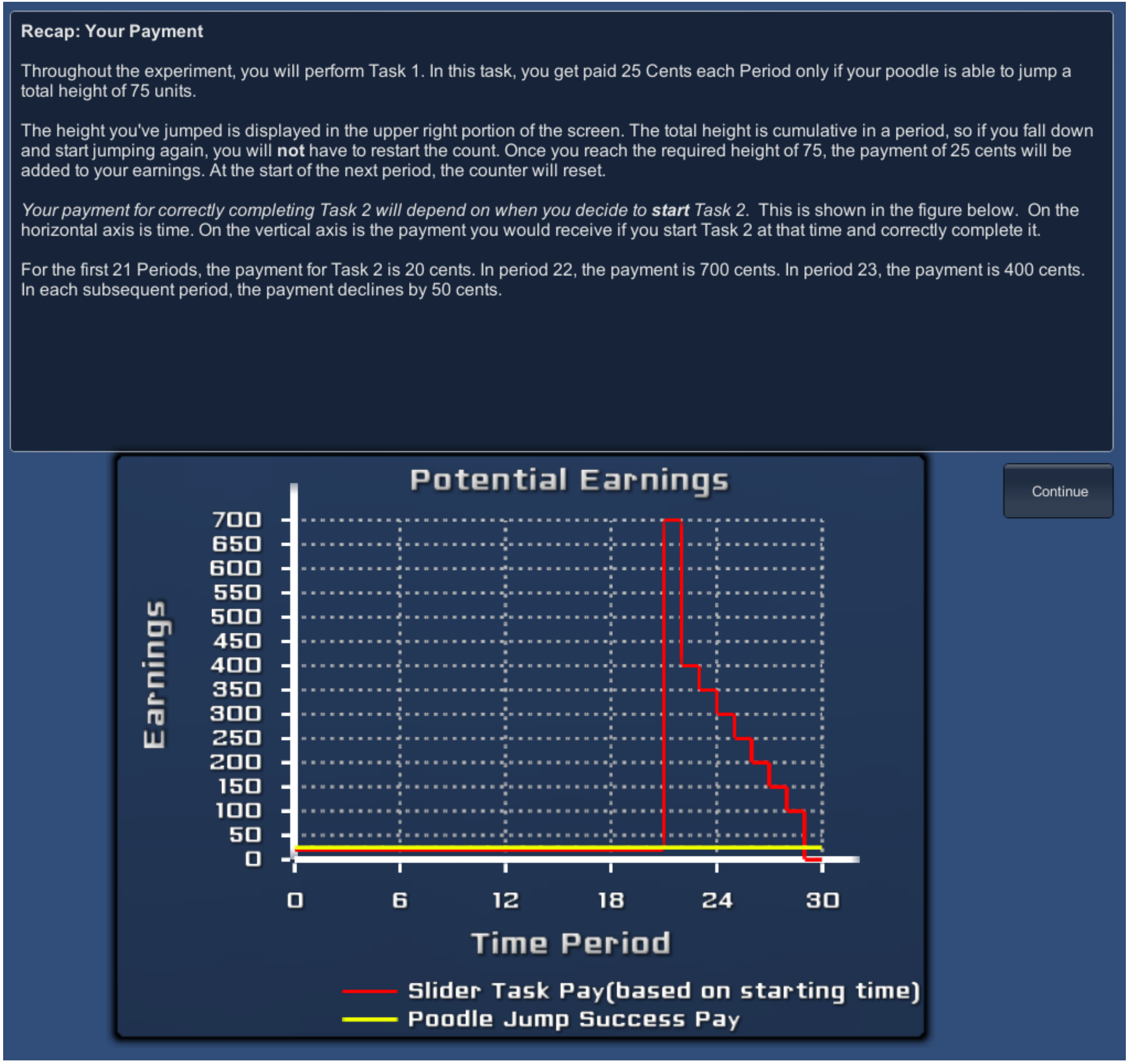
Practice on the next screen to learn how the Slider Task works. Press the j button to continue.

Figure B.15: ENHANCED Instructions page 7: Task 2 practice



Click and drag all sliders to 50

Figure B.16: ENHANCED Instructions page 8: payment recap



Recap: Your Payment

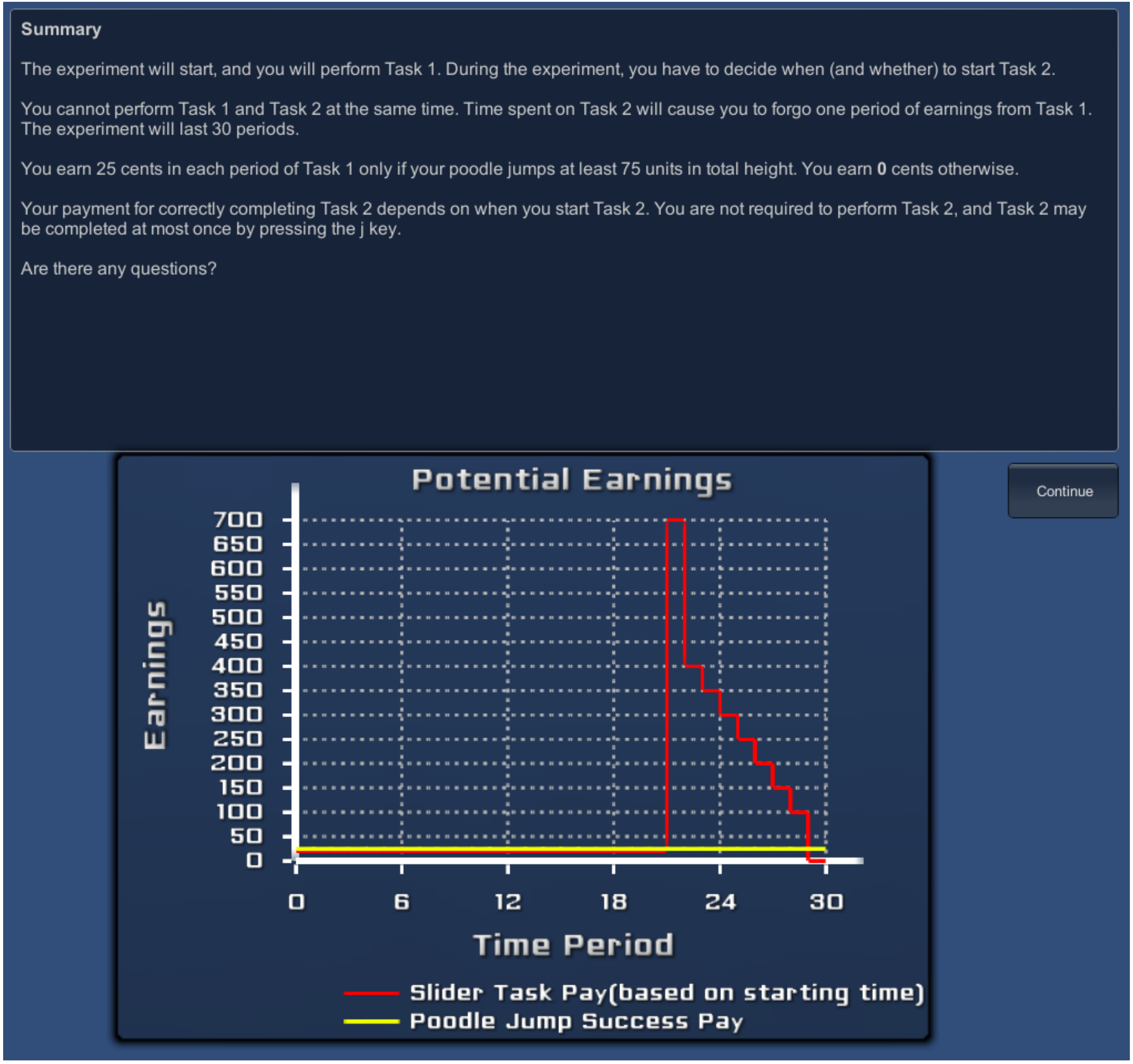
Throughout the experiment, you will perform Task 1. In this task, you get paid 25 Cents each Period only if your poodle is able to jump a total height of 75 units.

The height you've jumped is displayed in the upper right portion of the screen. The total height is cumulative in a period, so if you fall down and start jumping again, you will **not** have to restart the count. Once you reach the required height of 75, the payment of 25 cents will be added to your earnings. At the start of the next period, the counter will reset.

Your payment for correctly completing Task 2 will depend on when you decide to ***start*** *Task 2.* This is shown in the figure below. On the horizontal axis is time. On the vertical axis is the payment you would receive if you start Task 2 at that time and correctly complete it.

For the first 21 Periods, the payment for Task 2 is 20 cents. In period 22, the payment is 700 cents. In period 23, the payment is 400 cents. In each subsequent period, the payment declines by 50 cents.

Figure B.17: ENHANCED 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 start Task 2.

You cannot perform Task 1 and Task 2 at the same time. Time spent on Task 2 will cause you to forgo one period of earnings from Task 1. The experiment will last 30 periods.

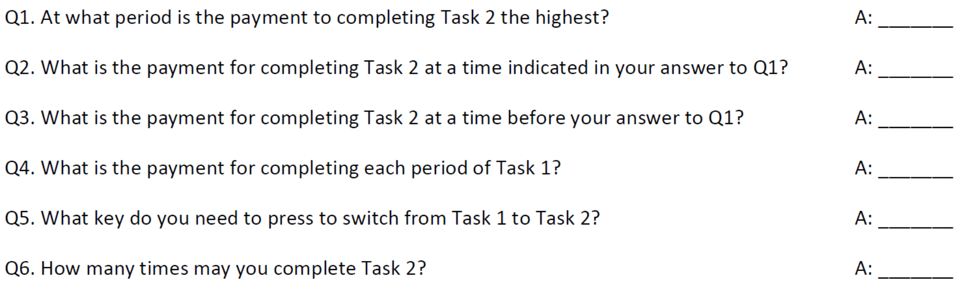
You earn 25 cents in each period of Task 1 only if your poodle jumps at least 75 units in total height. You earn **0** cents otherwise.

Your payment for correctly completing Task 2 depends on when you start Task 2. You are not required to perform Task 2, and Task 2 may be completed at most once by pressing the j key.

Are there any questions?

Our quiz, which was included after the instructions and before the main experiment in all treatments except for NO QUIZ, featured the following six questions:

Figure B.18: 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: _______

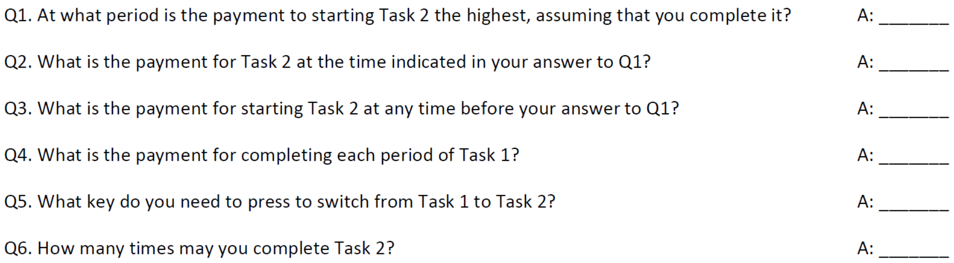
Q4. What is the payment for completing each period of Task 1? A: _______

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: _______

In our follow-up experimental sessions, we slightly re-worded some of the quiz questions to make them more clear. This new quiz was administered to all subjects in the ENHANCED treatment and some of the subjects in the QUIZ treatment.

Figure B.19: Revised post-instructions quiz



Q1. At what period is the payment to starting Task 2 the highest, assuming that you complete it? A: _______

Q2. What is the payment for Task 2 at the time indicated in your answer to Q1? A: _______

Q3. What is the payment for starting Task 2 at any 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: _______

While scores in the QUIZ treatment did increase slightly under the new quiz, from an average of 3.9 to 4.4, this difference is not statistically significant ($p = .11$, rank-sum test), and thus we pool data from all QUIZ sessions. We also did not observe any significant differences in NMB ($p = .50$, Fisher's exact test).

C Robustness checks

We redo our analysis with three alternative measures of NMB to check the robustness our results. The specifications reported in Table C.1-3 are all analogous to the specifications in Table 4, but with alternative definitions of NMB. The dependent variable "NMB1" is equal to one if the subject did Task 2 before period 21 and equal to zero otherwise; this measure of NMB allows for trembles. The "NMB2" variable defines any behavioral deviation from optimality as NMB. That is, it classifies a subject as exhibiting NMB unless they did Task 2 exactly in period 22. Finally, the "NMB3" variable classifies those who did Task 2 before period 22 or never at all as NMB. The results of these alternative specifications are broadly consistent with those reported in Table 4. Figure C.1 plots the share of subjects with NMB in each treatment, by each of these alternative measures. To check the robustness of our logit regressions, Table C.4 reports estimated linear probability models with (OLS analogues to columns 1 and 2 of Table 4); for comparison purposes note that we do not report marginal effects in Table 4 since the mediation analysis in column 4 provides the economically meaningful estimates of interest.

Figure C.1: Percentage of subjects revealing NMB, under three alternative definitions of NMB, by treatment.

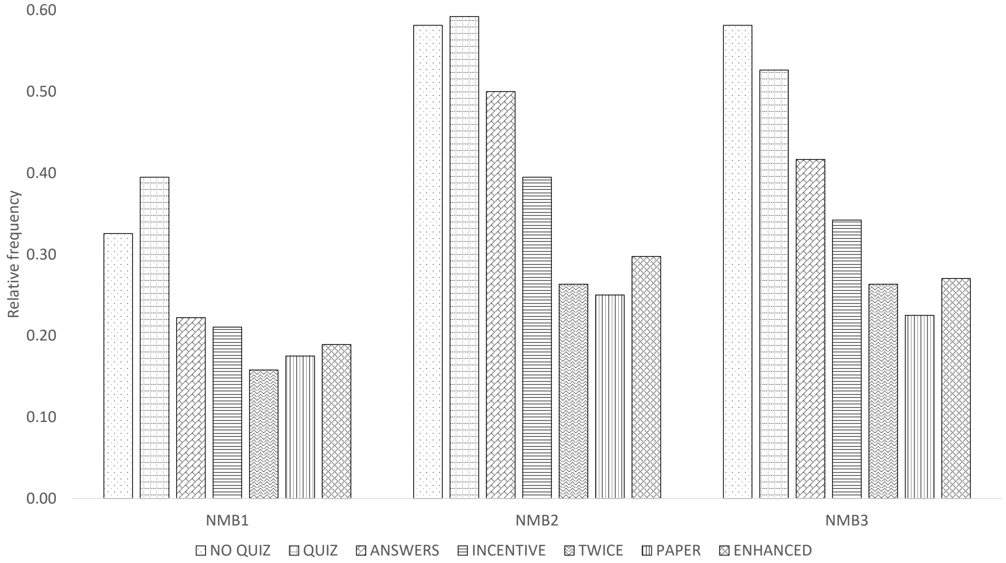


Table C.1: Treatment effects on NMB1 and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB1		Quiz Score	NMB1	
	(1)	(2)	(3)	(4)	
NO QUIZ	-0.301				n
	(-1.096, 0.495)				
ANSWERS	-0.825*	0.207	-0.050	-0.169*	112
	(-1.746, 0.096)	(-2.648, 3.061)	(-0.586, 0.487)	(-0.329, 0.008)	
ANSWERS × Quiz Score		-0.324		0.005	
		(-1.062, 0.413)		(-0.056, 0.070)	
INCENTIVE	-0.894*	-1.380	0.211	-0.164*	114
	(-1.810, 0.022)	(-4.202, 1.422)	(-0.354, 0.775)	(-0.331, 0.021)	
INCENTIVE × Quiz Score		0.127		-0.022	
		(-0.508, 0.762)		(-0.091, 0.039)	
TWICE	-1.247**	-0.677	0.421	-0.199**	114
	(-2.244, -0.249)	(-3.940, 2.586)	(-0.156, 0.998)	(-0.367, -0.010)	
TWICE × Quiz Score		-0.135		-0.044	
		(-0.847, 0.578)		(-0.119, 0.016)	
PAPER	-1.123**	7.787**	1.320***	0.163	116
	(-2.070, -0.176)	(1.053, 14.521)	(0.922, 1.718)	(-0.118, 0.375)	
PAPER × Quiz Score		-1.632**		-0.133***	
		(-2.901, -0.363)		(-0.223, -0.046)	
ENHANCED	-1.028**	0.249	0.489*	-0.139*	113
	(-1.981, -0.074)	(-3.675, 4.174)	(-0.030, 1.008)	(-0.325, 0.060)	
ENHANCED × Quiz Score		-0.273		-0.051*	
		(-1.144, 0.598)		(-0.123, 0.003)	
Quiz Score		-0.519***			
		(-0.875, -0.164)			
Intercept	-0.427*	1.662**	4.105***		
	(-0.893, 0.038)	(0.121, 3.202)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

Table C.2: Treatment effects on NMB2 and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB2		Quiz Score	NMB2	
	(1)	(2)	(3)	(4)	
NO QUIZ	-0.044				n
	(-0.812, 0.724)				
ANSWERS	-0.373	-1.477	-0.050	-0.103	112
	(-1.179, 0.434)	(-5.163, 2.209)	(-0.586, 0.487)	(-0.268, 0.070)	
ANSWERS × Quiz Score		0.192		0.009	
		(-0.624, 1.009)		(-0.098, 0.116)	
INCENTIVE	-0.800*	-2.840	0.211	-0.164*	114
	(-1.605, 0.004)	(-6.591, 0.911)	(-0.354, 0.775)	(-0.344, 0.013)	
INCENTIVE × Quiz Score		0.443		-0.042	
		(-0.353, 1.239)		(-0.157, 0.069)	
TWICE	-1.402***	-2.201	0.421	-0.254***	114
	(-2.267, -0.538)	(-6.525, 2.122)	(-0.156, 0.998)	(-0.424, -0.078)	
TWICE × Quiz Score		0.130		-0.084	
		(-0.879, 1.138)		(-0.203, 0.031)	
PAPER	-1.471***	8.269	1.320***	0.056	116
	(-2.331, -0.612)	(-4.351, 20.889)	(0.922, 1.718)	(-0.288, 0.236)	
PAPER × Quiz Score		-1.652		-0.284***	
		(-4.001, 0.698)		(-0.389, -0.182)	
ENHANCED	-1.233***	-2.724	0.489*	-0.216**	113
	(-2.083, -0.383)	(-6.883, 1.434)	(-0.030, 1.008)	(-0.402, -0.033)	
ENHANCED × Quiz Score		0.345		-0.101*	
		(-0.560, 1.249)		(-0.212, 0.007)	
Quiz Score		-1.344***			
		(-1.872, -0.816)			
Intercept	0.373	6.236***	4.105***		
	(-0.090, 0.835)	(3.637, 8.836)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

Table C.3: Treatment effects on NMB3 and Quiz Scores

	Dependent variable			Mediation analysis	
	NMB3		Quiz Score	NMB3	
	(1)	(2)	(3)	(4)	
NO QUIZ	0.223				n
	(-0.540, 0.987)				
ANSWERS	-0.442	-0.360	-0.050	-0.112	112
	(-1.252, 0.369)	(-3.559, 2.839)	(-0.586, 0.487)	(-0.278, 0.070)	
ANSWERS × Quiz Score		-0.075		0.009	
		(-0.851, 0.702)		(-0.089, 0.106)	
INCENTIVE	-0.759*	-2.207	0.211	-0.157*	114
	(-1.810, 0.022)	(-5.334, 0.921)	(-0.354, 0.775)	(-0.336, 0.028)	
INCENTIVE × Quiz Score		0.127		-0.037	
		(-0.508, 0.762)		(-0.141, 0.063)	
TWICE	-1.135***	-0.365	0.421	-0.183**	114
	(-1.996, -0.274)	(-4.401, 3.670)	(-0.156, 0.998)	(-0.356, -0.004)	
TWICE × Quiz Score		-0.193		-0.074	
		(-1.163, 0.776)		(-0.181, 0.026)	
PAPER	-1.342***	5.617	1.320***	0.074	116
	(-2.220, -0.464)	(-2.618, 13.852)	(0.922, 1.718)	(-0.252, 0.278)	
PAPER × Quiz Score		-1.143		-0.240***	
		(-2.649, -0.363)		(-0.340, -0.143)	
ENHANCED	-1.099**	0.321	0.489*	-0.158*	113
	(-1.962, -0.235)	(-3.790, 4.431)	(-0.030, 1.008)	(-0.349, 0.028)	
ENHANCED × Quiz Score		-0.314		-0.088*	
		(-1.201, 0.574)		(-0.189, 0.006)	
Quiz Score		-1.021***			
		(-1.471, -0.571)			
Intercept	0.105	4.400***	4.105***		
	(-0.350, 0.561)	(2.314, 6.486)	(3.789, 4.422)		
Observations	308	265	265		

QUIZ is the omitted category. *, **, and *** respectively denote $p < .1$, $p < .05$, $p < .01$. Robust (HC1) 95% confidence intervals are in parentheses in Columns (1)-(4). Mediation column reports estimated "direct effects" in the row of a treatment dummy, and mediated effects in the row of the interaction term between Quiz Score and that treatment dummy, both evaluated relative to the QUIZ baseline. That is, the direct effect of a treatment corresponds to $\mathbb{E}[\text{NMB}|\text{Treatment}, \text{Quiz Score} = 4.1] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$, while the mediated effect corresponds to $\mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = \mathbb{E}[\text{Quiz Score}|\text{Treatment}]] - \mathbb{E}[\text{NMB}|\text{QUIZ}, \text{Quiz Score} = 4.1]$.

Table C.4: Treatment effects on NMB – linear probability model robustness checks

	Dependent variable					
	NMB1		NMB2		NMB3	
NO QUIZ	-0.069		-0.011		0.055	
	(0.092)		(0.095)		(0.096)	
ANSWERS	-0.173*	-0.098	-0.092	-0.098	-0.110	-0.056
	(0.090)	(0.289)	(0.102)	(0.156)	(0.101)	(0.183)
INCENTIVE	-0.184**	-0.366	-0.197**	-0.316	-0.184*	-0.374
	(0.088)	(0.294)	(0.098)	(0.203)	(0.097)	(0.234)
TWICE	-0.237***	-0.283	-0.329 ***	-0.378*	-0.263***	-0.242
	(0.082)	(0.303)	(0.092)	(0.194)	(0.093)	(0.205)
PAPER	-0.220 ***	0.893*	-0.342***	0.911**	-0.301***	0.697
	(0.083)	(0.454)	(0.090)	(0.409)	(0.088)	(0.460)
ENHANCED	-0.206**	-0.126	-0.295***	-0.330	-0.256***	-0.111
	(0.086)	(0.347)	(0.095)	(0.254)	(0.094)	(0.250)
Quiz Score		-0.117***		-0.209***		-0.192***
		(0.035)		(0.022)		(0.026)
ANSWERS × Quiz Score		-0.020		-0.001		-0.016
		(0.060)		(0.040)		(0.043)
INCENTIVE × Quiz Score		0.048		0.038		0.053
		(0.059)		(0.042)		(0.048)
TWICE × Quiz Score		0.021		0.030		0.013
		(0.058)		(0.039)		(0.041)
PAPER × Quiz Score		-0.177**		-0.180**		-0.137*
		(0.079)		(0.069)		(0.078)
ENHANCED × Quiz Score		-0.005		0.030		-0.011
		(0.067)		(0.050)		(0.046)
Intercept	0.395***	0.873***	0.592***	1.450***	0.526***	1.313***
	(0.057)	(0.166)	(0.057)	(0.090)	(0.058)	(0.113)
Observations	308	265	308	265	308	265
R^2	0.044	0.194	0.082	0.387	0.072	0.340

QUIZ is the omitted category. *, **, and *** respectively denote $p < 0.1$, $p < .05$, $p < .01$.

Robust (HC1) standard errors are in parentheses.

We note that our statistical tests find significant differences between our main QUIZ treatment and each of our INCENTIVE, TWICE, PAPER, and ENHANCED treatments, but do not detect significant differences among the latter four treatments, and also detects no significant difference between the ANSWERS treatment and other treatments (see Table 2 in the main text). This raises the question of statistical power. We note that the comparisons between the QUIZ treatment and each of the INCENTIVE, TWICE, PAPER, and ENHANCED treatments appear to be appropriately powered. Across the latter four treatments, 21.6% of subject misunderstand (a fraction which ranges between 18.4-23.7% across these treatments),[1] while 47.4% of subjects in the QUIZ treatment misunderstand. A simple ex-post power calculation indicates that if we recruited $n_1 = 76$ and $n_2 = 38$ subjects to two treatments in which each subject misunderstands with probability $p_1 = .474$ and $p_2 = .216$ (respectively), then we have a 79.4% chance of detecting a statistically significant difference between treatments (at the 5% significance level). This suggests a reasonable level of power in our comparisons between the four aformentioned treatments and QUIZ. However, 33.3% of subject misunderstand in the ANSWERS treatment – an intermediate case between QUIZ and these other four treatments. If we recruited $n_1 = 76$ and $n_2 = 36$ subjects to two treatments in which each subject misunderstands with probability $p_1 = .474$ and $p_2 = .333$ (respectively), then we have only a 33.2% chance of detecting a statistically significant difference between treatments. If instead we recruited $n_1 = 38$ and $n_2 = 36$ subjects to two treatments in which each subject misunderstands with probability $p_1 = .216$ and $p_2 = .333$ (respectively), then we have only a 18.2% chance of detecting a statistically significant difference between treatments. These calculations indicate that our sample sizes are too small to reliably detect a statistically significant difference between our ANSWERS treatment and the QUIZ treatment, or between the ANSWERS treatment and any of the INCENTIVE, TWICE, PAPER, and ENHANCED treatments. If we instead view the NO QUIZ and QUIZ, pooled, as baseline instructions treatments without reinforcement, and

[1] These numbers are relatively close to each other, so we use the 21.6% for our illustrative calculations below.

the remaining treatments as enhanced instructions or reinforcement treatments, then our samples have $n_1 = 119$, $n_2 = 189$, $p_1 = .462$, and $p_2 = .238$; under these samples sizes and NMB probabilities, we had a 98.3% chance of detecting a significant difference in NMB.

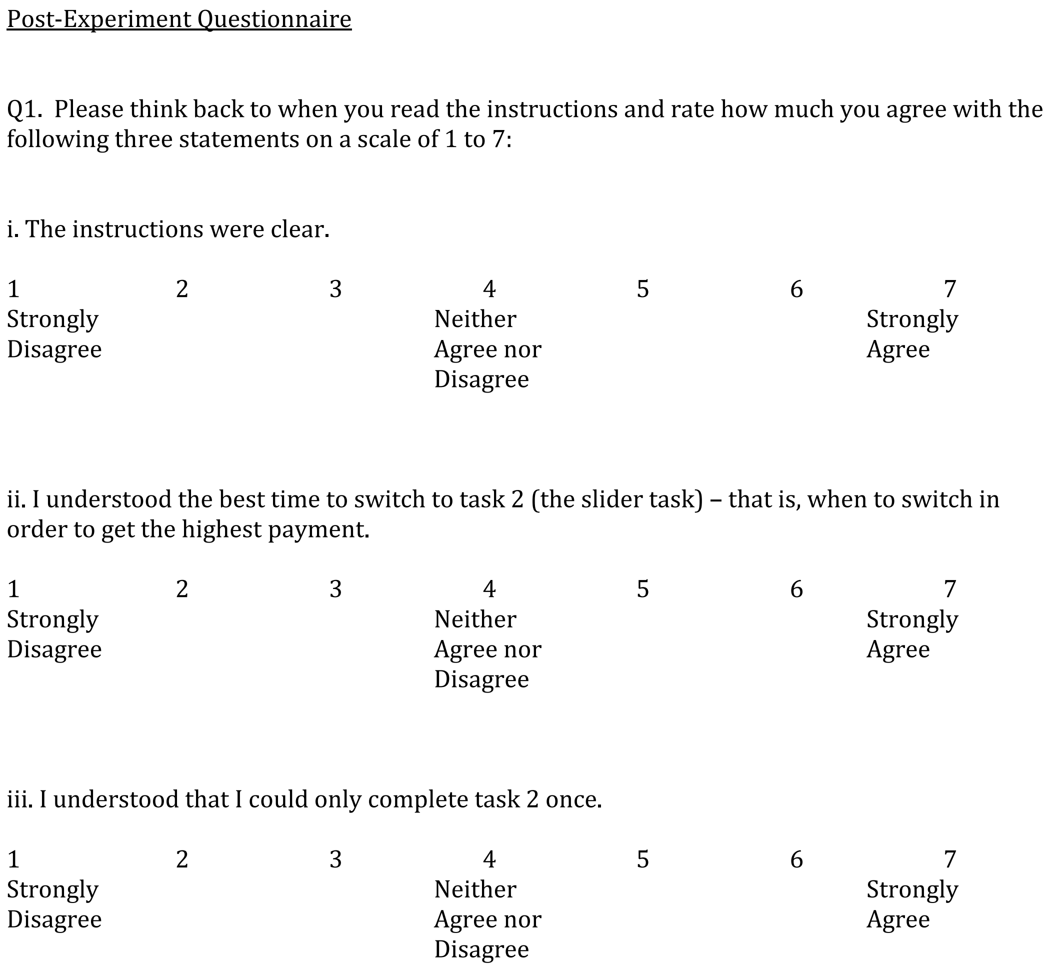
Our statistical analysis was conducted in R (R Core Team, 2017). The regressions in Table 3 (and above) used the 'lm' and 'glm' command in the base 'stats' package, with robust standard errors calculated using the 'sandwich' package (Zeileis 2004; 2006). Mediation analysis used the 'mediation' package (Tingley *et al.*, 2014). Goodman-Kruskal gamma tests use the 'DescTools' package (Signorell, 2018). We used the 'pwr' package (Champely, 2018) for the power analysis reported above. Figures made in 'ggplot2' (Wickham, 2009).

D Post-experiment questionnaire

At suggestion of a referee and the editor, we added a post-experiment questionnaire to our ENHANCED treatment, and ran additional sessions of the QUIZ treatment followed by this questionnaire to paint a more complete picture of subjects' decisionmaking processes as they went though the experiment. We asked nine questions in total.

Our first observation is that there is no statistical difference between QUIZ and ENHANCED on any of the first six quantitative questions.

Figure C.2: Post-experiment questionnaire (Page 1)



Post-Experiment Questionnaire

Q1. Please think back to when you read the instructions and rate how much you agree with the following three statements on a scale of 1 to 7:

i. The instructions were clear.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

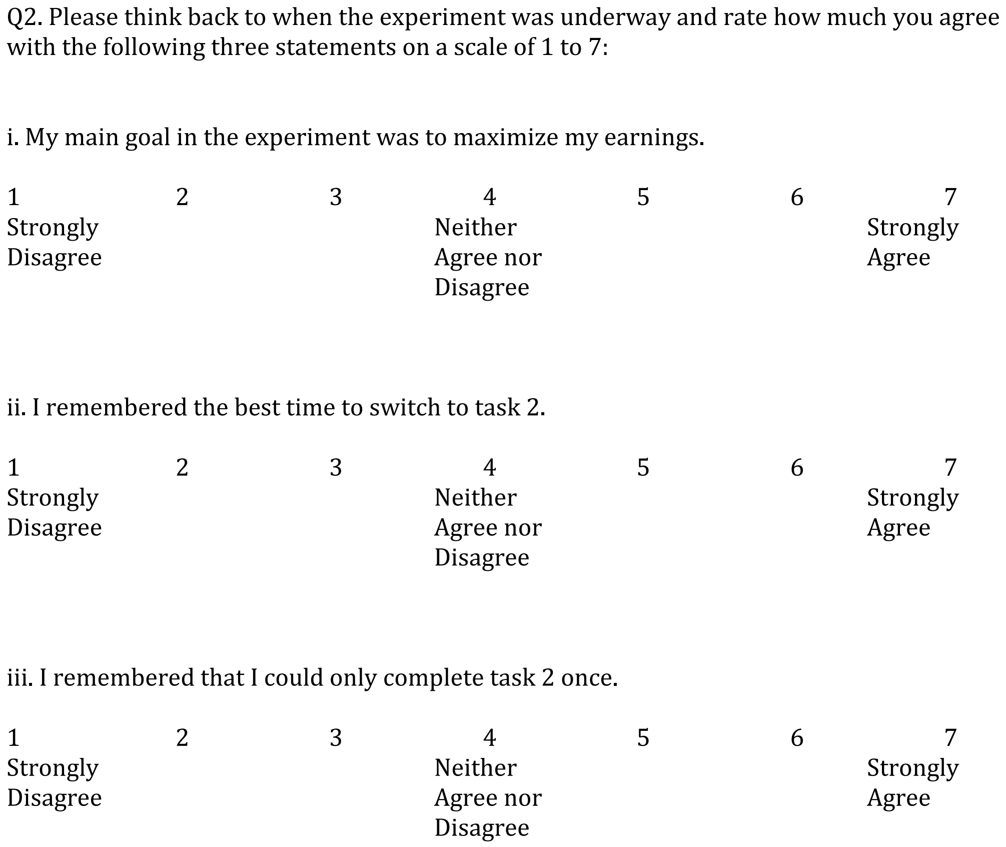
ii. I understood the best time to switch to task 2 (the slider task) – that is, when to switch in order to get the highest payment.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

iii. I understood that I could only complete task 2 once.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

Figure C.3: Post-experiment questionnaire (Page 2)



Q2. Please think back to when the experiment was underway and rate how much you agree with the following three statements on a scale of 1 to 7:

i. My main goal in the experiment was to maximize my earnings.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

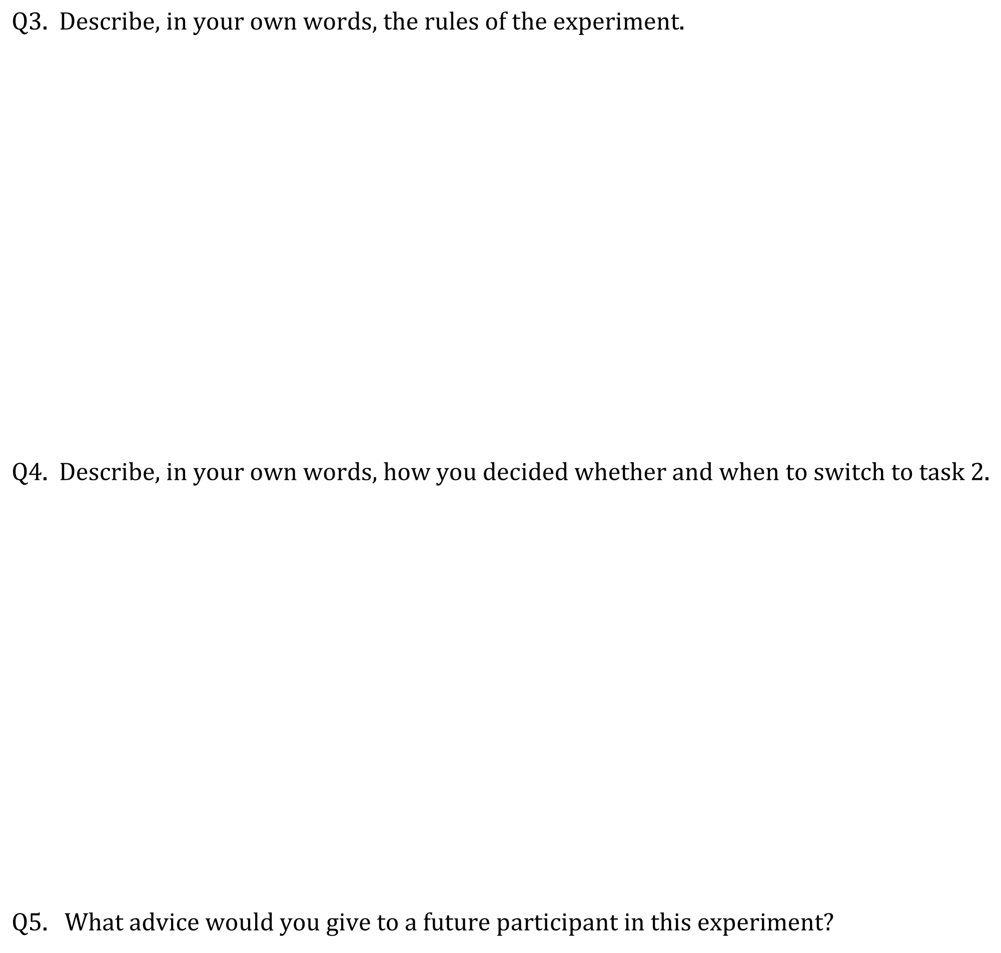
ii. I remembered the best time to switch to task 2.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

iii. I remembered that I could only complete task 2 once.

1	2	3	4	5	6	7
Strongly Disagree			Neither Agree nor Disagree			Strongly Agree

Figure C.4: Post-experiment questionnaire (Page 3)



Q3. Describe, in your own words, the rules of the experiment.

Q4. Describe, in your own words, how you decided whether and when to switch to task 2.

Q5. What advice would you give to a future participant in this experiment?

	QUIZ	ENHANCED	p-value
Comprehension			
Q1i (Clarity)	5.7 (6)	5.4 (6)	0.31
Q1ii (Understood Optimum)	5.7 (7)	5.6 (7)	0.41
Q1iii (Understood Once)	5.4 (7)	5.9 (7)	0.55
Retention			
Q2i (Maximized Earnings)	6.4 (7)	6.3 (7)	0.43
Q2ii (Remembered Optimum)	5.8 (7)	5.6 (6)	0.57
Q2iii (Remembered Once)	5.6 (7)	6.0 (7)	0.38

Mean (median) reported; p-values for rank-sum tests of equality of distributions.

Table C.5: Correlation between subjects' evaluation and misunderstanding and quiz score

	misunderstanding	p.value_misunderstanding	quiz score	p.value_score
Q1i	-0.168	0.159	0.281	0.017
Q1ii	-0.267	0.024	0.202	0.089
Q1iii	-0.406	0.0004	0.202	0.088
Q2i	0.039	0.744	0.046	0.700
Q2ii	-0.371	0.001	0.383	0.001
Q2iii	-0.356	0.002	0.196	0.100

Table D.1 shows that our post-experimental questionnaire results indicate that subjects largely felt that they both understood and retained the key pieces of information from the instructions – with the median subject indicating that they agreed or strongly agreed that they understood and remembered when they should switch (Q1ii, Q2ii), and how many times they could switch (Q1iii, Q2iii). In addition, most subjects agreed with the statement "The instructions were clear", with the median subject rating the statement a 6 out of 7. We find no significant differences between the distribution of answers to any of these questions between the QUIZ and ENHANCED treatments ($p > .3$ in all pairwise comparisons, rank-sum tests). Since we do observe a difference in NMB revealed in the experiment, our post-experimental questionnaire inadvertently reveals its limits at diagnosing reasons for NMB and the potential for improvements. That being said, Table C.3 indicates that subjects' post-experiment answers strongly correlate with both NMB in the experiment and quiz scores. Post-experiment reports of understanding (Q1ii,iii) and retention (Q2ii,iii) were each negatively correlated with NMB ($p < .03$ in all cases). In addition, the subject's post-experimental agreement with the statement "The instructions were clear" was positively correlated with their post-instructions quiz score ($\rho = .281$, $p = .017$).

22 of the 72 subjects who wrote the questionnaire mentioned the instructions in their written answers. Nearly all of these were in Q5: "What advice would you give to a future participant in this experiment?" For instance, the first three subjects to mention the instructions answered Q5 as follows: "Pay attention to the instructions." "Do the experiment with patience and read instructions very carefully." "Read the instructions and follow them for more $." These are typical answers; many subjects recognized, ex post, that paying close attention to the instructions was important for achieving the maximum payoff.

21 of the 72 subjects who wrote the questionnaire showed some kind of mistaken understanding of the experiment, even after having completed it. Many of these misunderstandings were orthogonal to our variable of interest (the time to do task 2). For instance, although our instructions clearly stated that one could get a $0.25 payoff for each period of task 1 if a certain threshold was reached, many seemed to believe that one could earn more than $0.25 by doubling or tripling the threshold. For instance, one subject wrote, "You have a poodle that jumps on to platforms, each 75 units, you get paid 25c." Another one wrote, "Roughly, I would only get 50c at most doing poodle jump for the whole period." The payoff is fixed at 25 cents, so 50 would be impossible. Many subjects appear to believe that they could earn for both tasks 1 and 2 if they completed the minimum height before switching. This is a minor misunderstanding, though it is stated in the instructions that one must forego earnings from one period of task 1 in order to perform task 2.

However, the majority of subjects do not show explicit misunderstandings in their answers, and some even demonstrate learning. One subject who did not perform task 2 at the correct time wrote, “I wasn’t aware I can only switch to task 2 only once. So I switched to task 2 in the first period.” Another wrote, “I thought it didn’t mention number of times we could do the bonus so I did it very early on.” These subjects clearly realized their mistakes after they had made them, which suggests that repeated decisions (with feedback of some form) can be a substitute for reinforcing understanding. On the other hand, some subjects failed to understand our instructions and still didn’t understand them afterwards. One such subject wrote, “If you taking task 1, you can change game into task 2, but you cannot turn back to task 1.”

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.

Anderson, L. R., DiTraglia, F. J. and Gerlach, J. R. (2011). Measuring altruism in a public goods experiment: a comparison of us and czech subjects. *Experimental Economics*, **14** (3), 426–437.

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.

Bayer, R.-C., Renner, E. and Sausgruber, R. (2013). Confusion and learning in the voluntary contributions game. *Experimental Economics*, **16** (4), 478–496.

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.

Champely, S. (2018). *pwr: Basic Functions for Power Analysis*. R package version 1.2-2.

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.

Harris, D., Herrmann, B., Kontoleon, A. and Newton, J. (2015). Is it a norm to favour your own group? *Experimental Economics*, **18** (3), 491–521.

Kamei, K., Putterman, L. and Tyran, J.-R. (2015). State or nature? endogenous formal versus informal sanctions in the voluntary provision of public goods. *Experimental Economics*, **18** (1), 38–65.

Mittone, L. and Ploner, M. (2011). Peer pressure, social spillovers, and reciprocity: an experimental analysis. *Experimental Economics*, **14** (2), 203–222.

Noussair, C. N. and Stoop, J. (2015). Time as a medium of reward in three social preference experiments. *Experimental Economics*, **18** (3), 442–456.

Petersen, L. and Winn, A. (2014). Does money illusion matter? comment. *American Economic Review*, **104** (3), 1047–62.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Signorell, A. (2018). *DescTools: Tools for Descriptive Statistics*. R package version 0.99.25.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, **59** (5), 1–38.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, **11** (10), 1–17.

— (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, **16** (9), 1–16.