Supplementary Appendix to "Instructions" (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.

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.

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 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.

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

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.

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

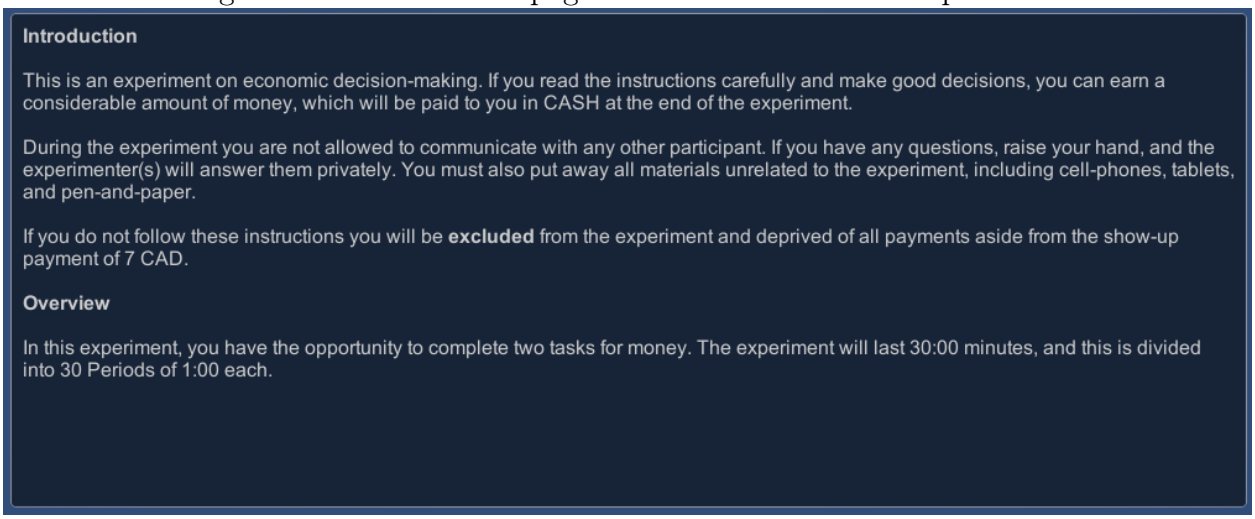


Figure B.3: Instructions page 2: description of Task 1

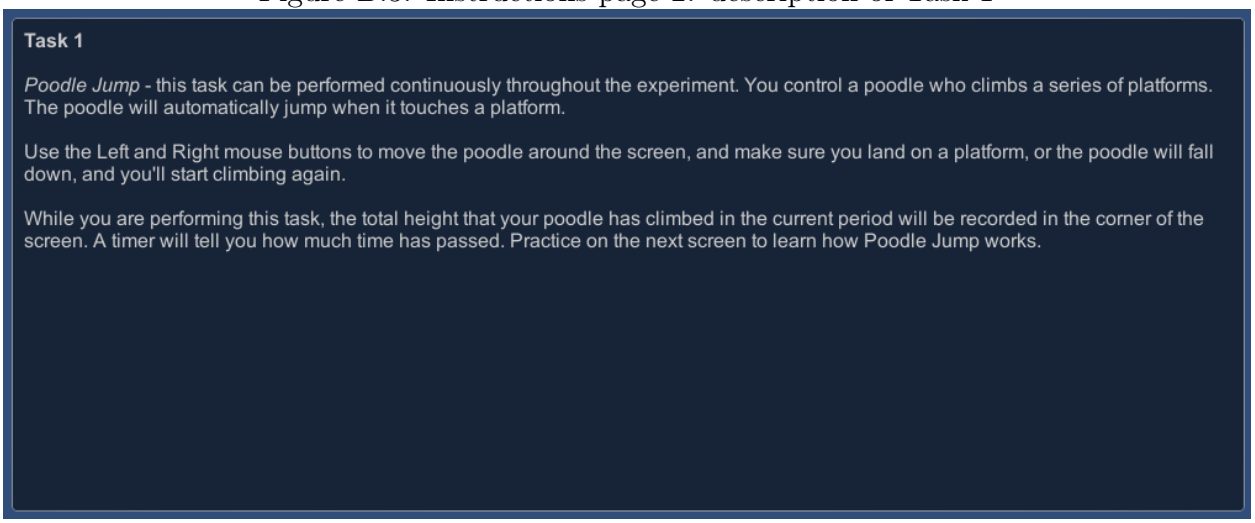


Figure B.4: Instructions page 3: Task 1 practice

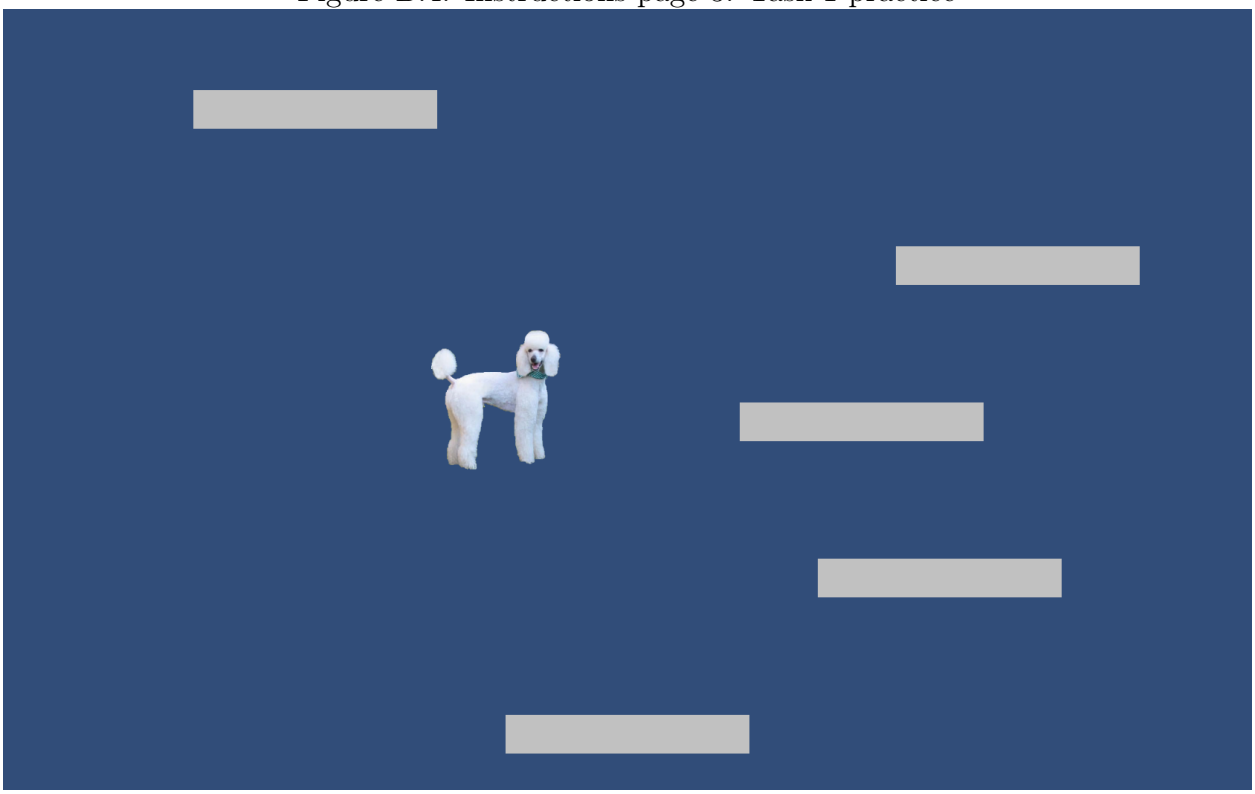


Figure B.5: Instructions page 4: description of Task 2

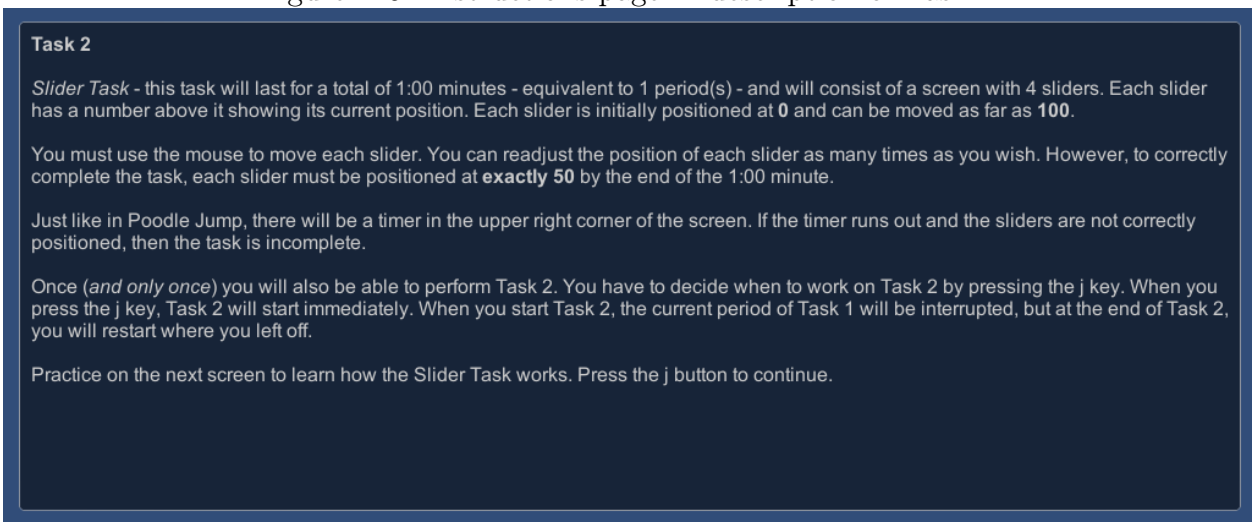
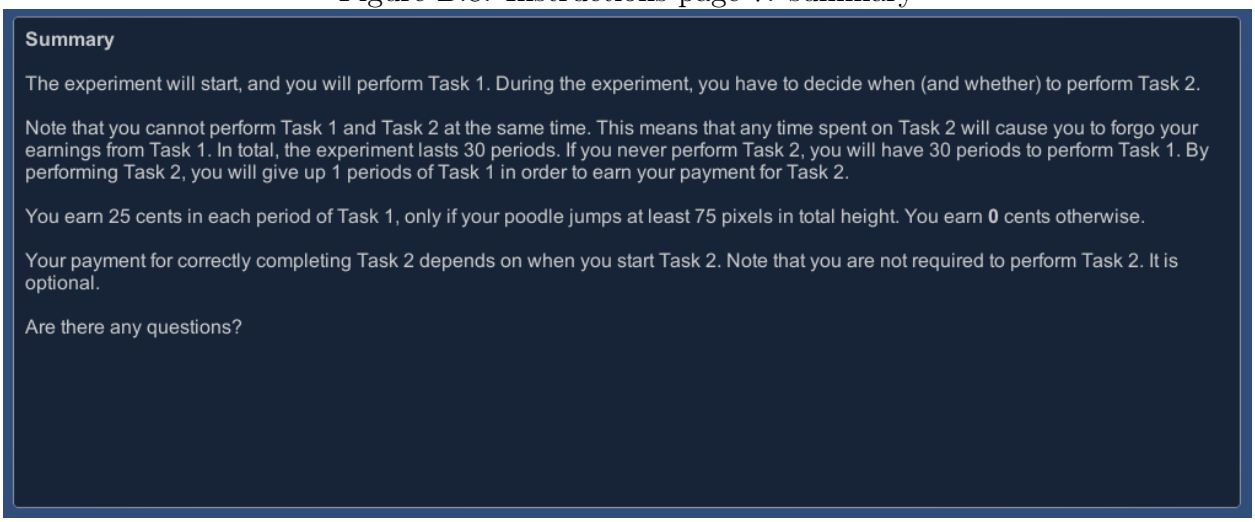


Figure B.6: Instructions page 5: Task 2 practice

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

Figure B.8: Instructions page 7: summary



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

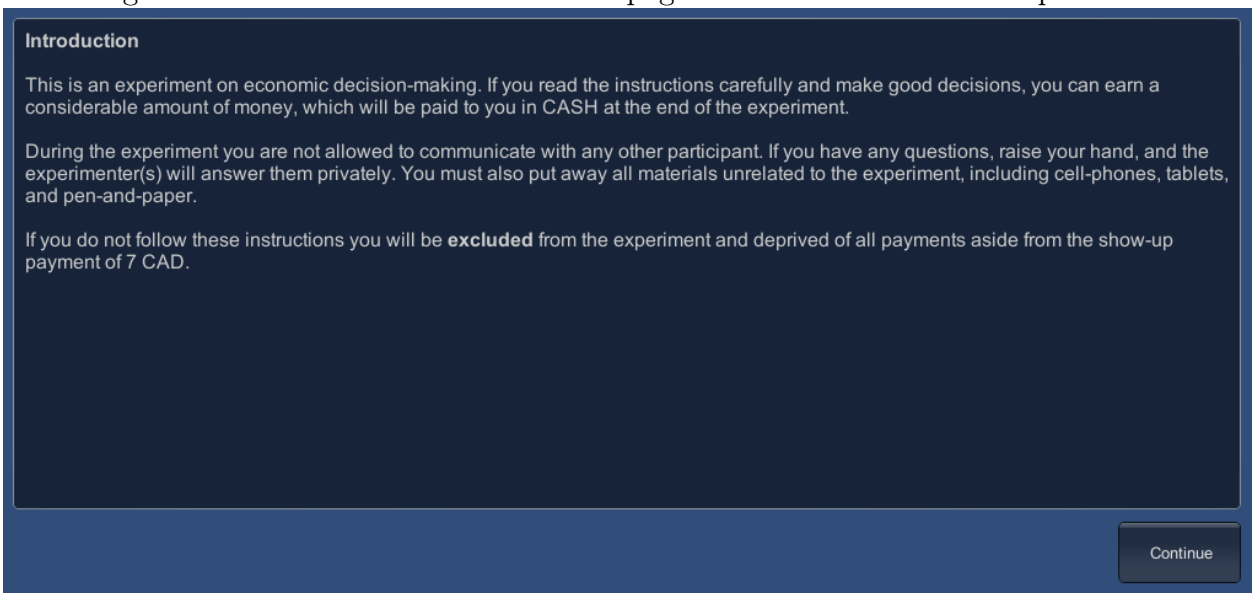


Figure B.10: ENHANCED Instructions page 2: overview and payment

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

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

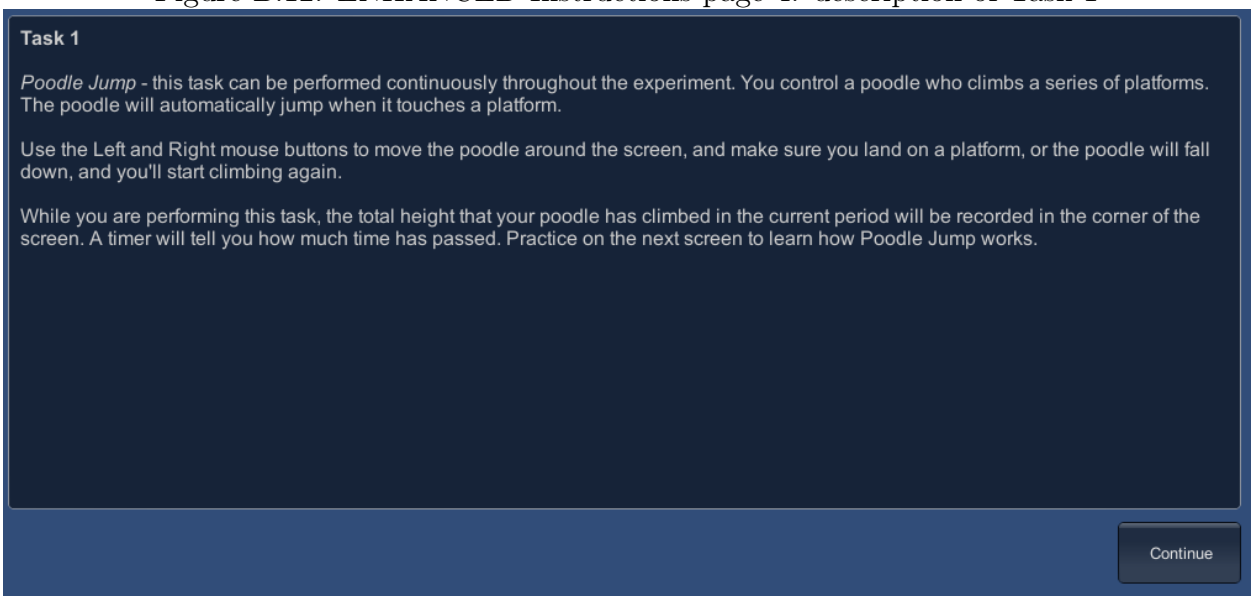


Figure B.13: ENHANCED Instructions page 5: Task 1 practice

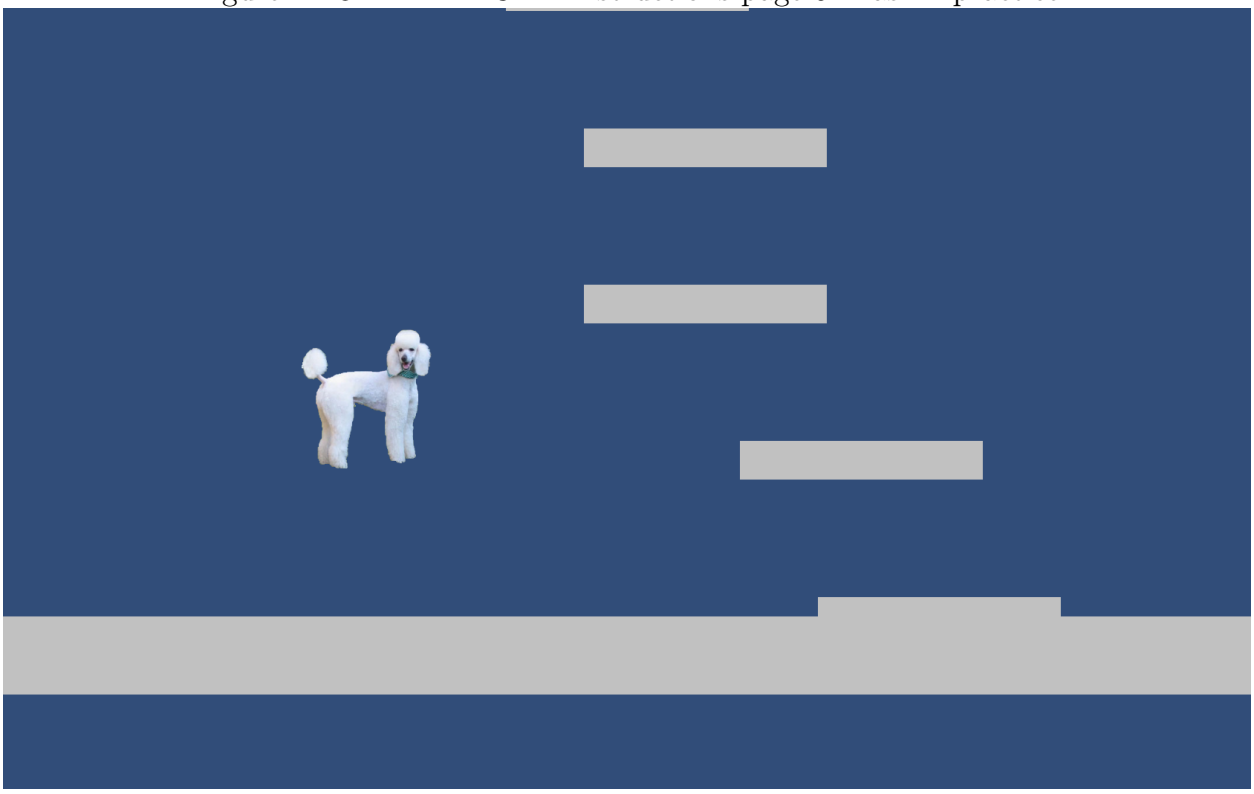


Figure B.14: ENHANCED Instructions page 6: description of Task 2

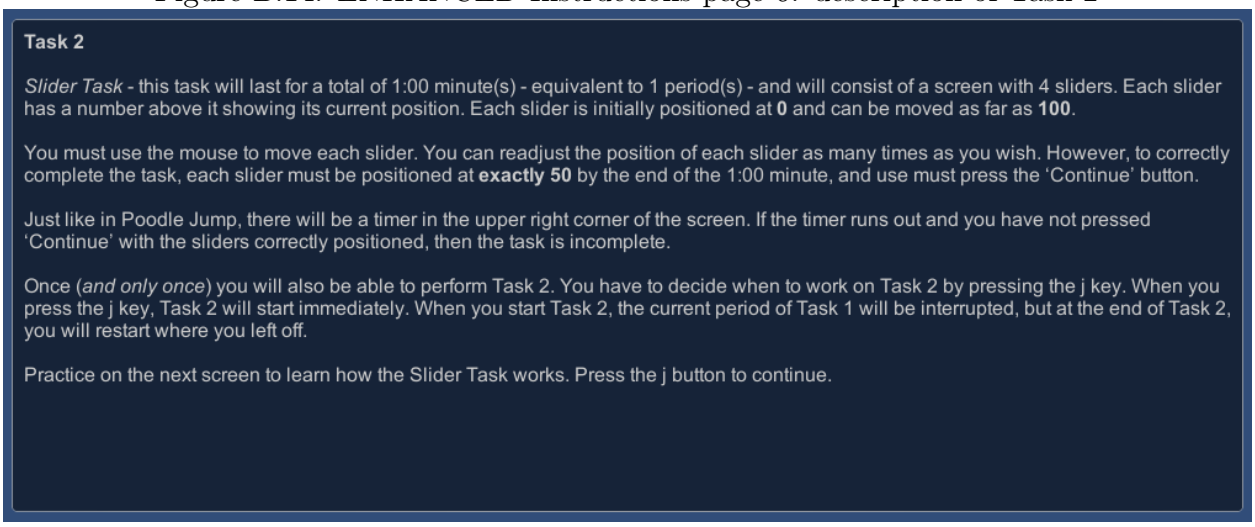


Figure B.15: ENHANCED Instructions page 7: Task 2 practice

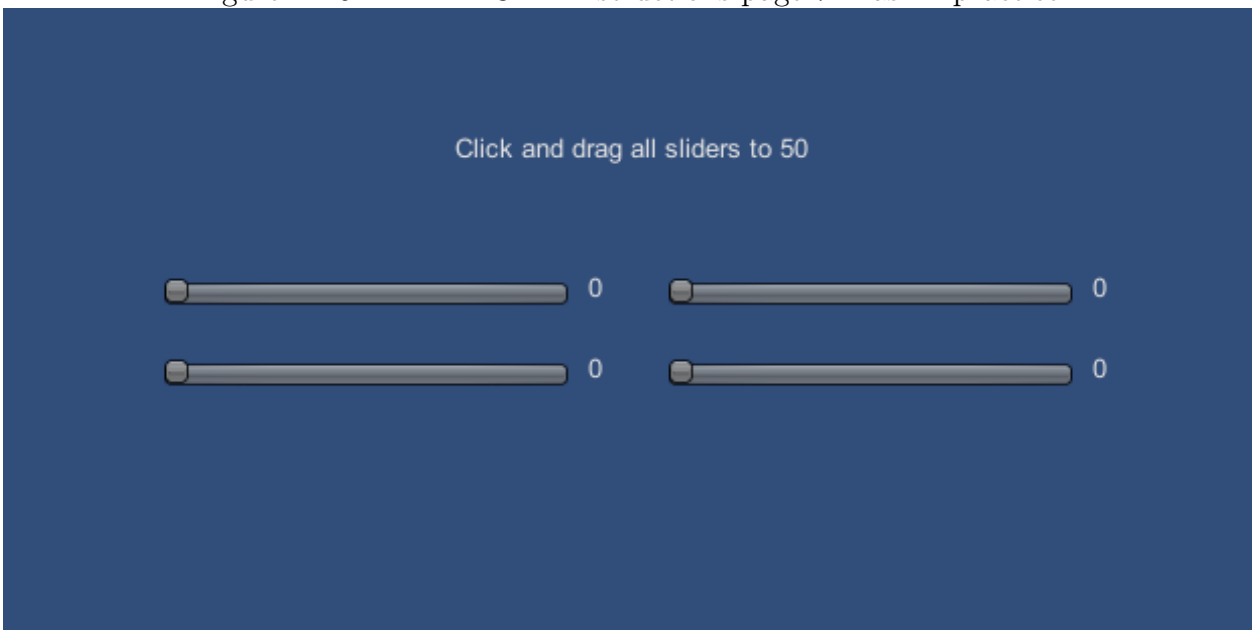


Figure B.16: ENHANCED Instructions page 8: payment recap

Figure B.17: ENHANCED Instructions page 7: summary

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.

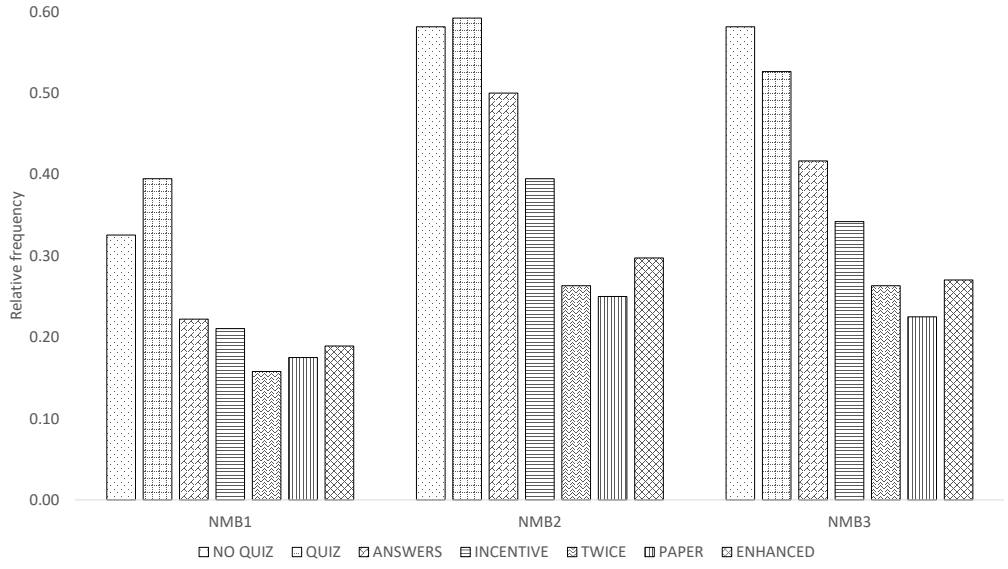


Table C.1: Treatment effects on NMB1 and Quiz Scores

	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*	
	(-1.746, 0.096)	(-2.648, 3.061)	(-0.586, 0.487)	(-0.329, 0.008)	112
ANSWERS × Quiz Score		-0.324		0.005	
		(-1.062, 0.413)		(-0.056, 0.070)	
INCENTIVE	-0.894*	-1.380	0.211	-0.164*	
	(-1.810, 0.022)	(-4.202, 1.422)	(-0.354, 0.775)	(-0.331, 0.021)	114
INCENTIVE × Quiz Score		0.127		-0.022	
		(-0.508, 0.762)		(-0.091, 0.039)	
TWICE	-1.247**	-0.677	0.421	-0.199**	
	(-2.244, -0.249)	(-3.940, 2.586)	(-0.156, 0.998)	(-0.367, -0.010)	114
TWICE × Quiz Score		-0.135		-0.044	
		(-0.847, 0.578)		(-0.119, 0.016)	
PAPER	-1.123**	7.787**	1.320***	0.163	
	(-2.070, -0.176)	(1.053, 14.521)	(0.922, 1.718)	(-0.118, 0.375)	116
PAPER × Quiz Score		-1.632**		-0.133***	
		(-2.901, -0.363)		(-0.223, -0.046)	
ENHANCED	-1.028**	0.249	0.489*	-0.139*	
	(-1.981, -0.074)	(-3.675, 4.174)	(-0.030, 1.008)	(-0.325, 0.060)	113
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	
	(-1.179, 0.434)	(-5.163, 2.209)	(-0.586, 0.487)	(-0.268, 0.070)	112
ANSWERS × Quiz Score		0.192		0.009	
		(-0.624, 1.009)		(-0.098, 0.116)	
INCENTIVE	-0.800*	-2.840	0.211	-0.164*	
	(-1.605, 0.004)	(-6.591, 0.911)	(-0.354, 0.775)	(-0.344, 0.013)	114
INCENTIVE × Quiz Score		0.443		-0.042	
		(-0.353, 1.239)		(-0.157, 0.069)	
TWICE	-1.402***	-2.201	0.421	-0.254***	
	(-2.267, -0.538)	(-6.525, 2.122)	(-0.156, 0.998)	(-0.424, -0.078)	114
TWICE × Quiz Score		0.130		-0.084	
		(-0.879, 1.138)		(-0.203, 0.031)	
PAPER	-1.471***	8.269	1.320***	0.056	
	(-2.331, -0.612)	(-4.351, 20.889)	(0.922, 1.718)	(-0.288, 0.236)	116
PAPER × Quiz Score		-1.652		-0.284***	
		(-4.001, 0.698)		(-0.389, -0.182)	
ENHANCED	-1.233***	-2.724	0.489*	-0.216**	
	(-2.083, -0.383)	(-6.883, 1.434)	(-0.030, 1.008)	(-0.402, -0.033)	113
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	
	(-1.252, 0.369)	(-3.559, 2.839)	(-0.586, 0.487)	(-0.278, 0.070)	112
ANSWERS × Quiz Score		-0.075		0.009	
		(-0.851, 0.702)		(-0.089, 0.106)	
INCENTIVE	-0.759*	-2.207	0.211	-0.157*	
	(-1.810, 0.022)	(-5.334, 0.921)	(-0.354, 0.775)	(-0.336, 0.028)	114
INCENTIVE × Quiz Score		0.127		-0.037	
		(-0.508, 0.762)		(-0.141, 0.063)	
TWICE	-1.135***	-0.365	0.421	-0.183**	
	(-1.996, -0.274)	(-4.401, 3.670)	(-0.156, 0.998)	(-0.356, -0.004)	114
TWICE × Quiz Score		-0.193		-0.074	
		(-1.163, 0.776)		(-0.181, 0.026)	
PAPER	-1.342***	5.617	1.320***	0.074	
	(-2.220, -0.464)	(-2.618, 13.852)	(0.922, 1.718)	(-0.252, 0.278)	116
PAPER × Quiz Score		-1.143		-0.240***	
		(-2.649, -0.363)		(-0.340, -0.143)	
ENHANCED	-1.099**	0.321	0.489*	-0.158*	
	(-1.962, -0.235)	(-3.790, 4.431)	(-0.030, 1.008)	(-0.349, 0.028)	113
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 an-

swers, 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.